TRAIL
TRUSTED AI LABS
BY DIGITALWALLONIA4.AI / SPW-RECHERCHE

Wallonie recherche SPW

UCLouvain

FAPESP

RL-Net: Interpretable Rule Learning with Neural Networks

Lucile Dierckx,
Rosana Veroneze,
Siegfried Nijssen

Paper accepted at PAKDD2023

Motivation

Need for interpretable classifiers

- In various domains
- With good performance

→ Very present topic in the literature

- Important class: Rule-based models (IF-THEN)
- Based on pattern mining → Exponential algorithm
- Based on heuristics → Task-specific

Motivation

Recent focus on NeuSy AI

- Rely on gradient-based learning
- Keep interpretability
- Neural network literature

⟶ New studies on neural rule learning

With focus on:

- Binary classification
- Unordered rule sets

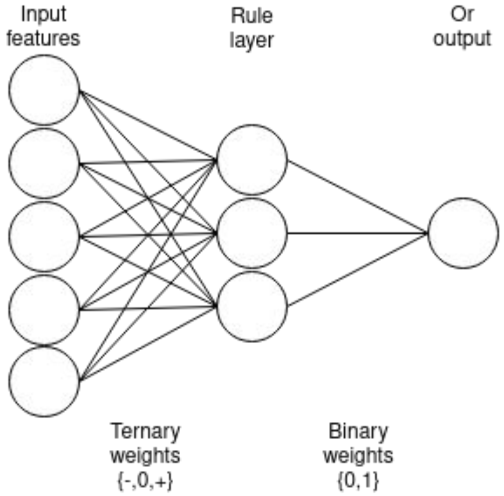
⟶ Need of multi-class, ordered rule list model

Overview

- **Objective**
- Model
- Experiments
- Results
- Conclusion

Objective

DR-Net - Binary rule set[1]



- Two-steps training
- Unordered rule sets
- Binary rules
- Unique rule consequent

⟶ Adapted for:

- Ordered rule lists
- Binary and multi-class rules + Easy adaptation multi-label

1 - Qiao, L., Wang, W., Lin, B.: Learning accurate and interpretable decision rule sets from neural networks. AAAI (2021)

Objective

Rule set

IF … THEN *Positive*
IF … THEN *Positive*
IF … THEN *Positive*
…
ELSE *Negative*

- No ordering
- Voting mechanism for multi-class

Rule list

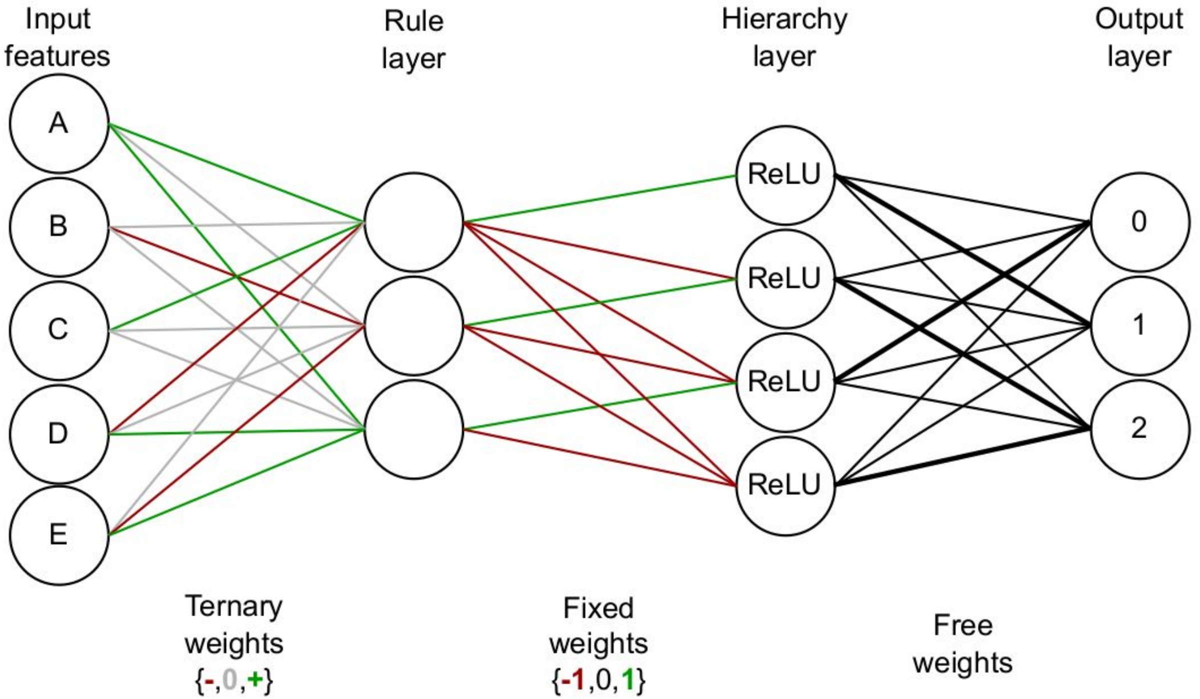
IF … THEN *Class w*
ELSE IF … THEN *Class x*
ELSE IF … THEN *Class y*
…
ELSE *Class z*

- Ordered by hierarchy
- Fully interpretable for multi-class

Overview

- Objective
- **Model**
- Experiments
- Results
- Conclusion

Model: RL-Net



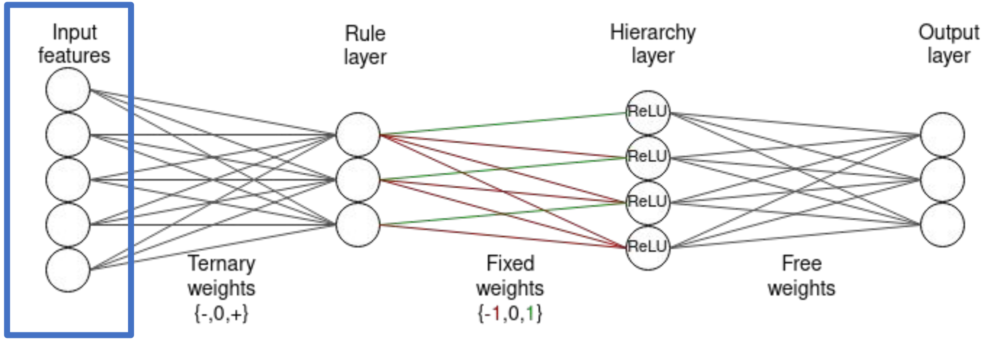
IF A and C and not D THEN 1

ELSE IF not B and not E THEN 2

ELSE IF A and D and E THEN 0

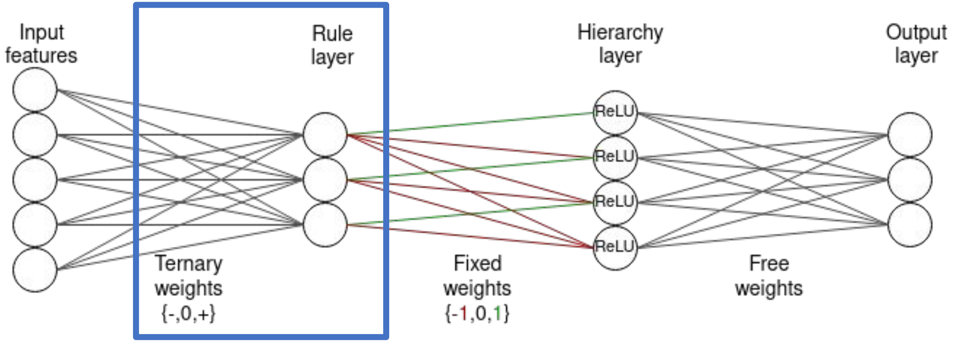
ELSE 2

Model: Input Features



- Used with tabular datasets
- Features are binarized
- No need to express negation

Model: Rule Layer



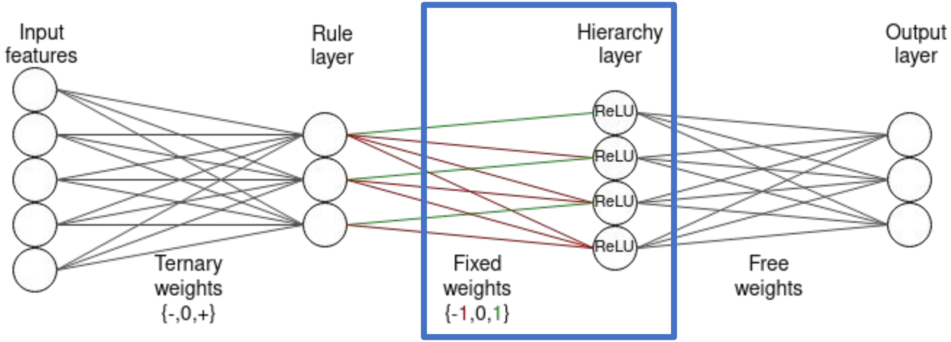
- Ternary weights with free value
- Weights express contribution of features
- Adaptative bias

⟶ Output >0 when all conditions are met

⟶ Binarized output

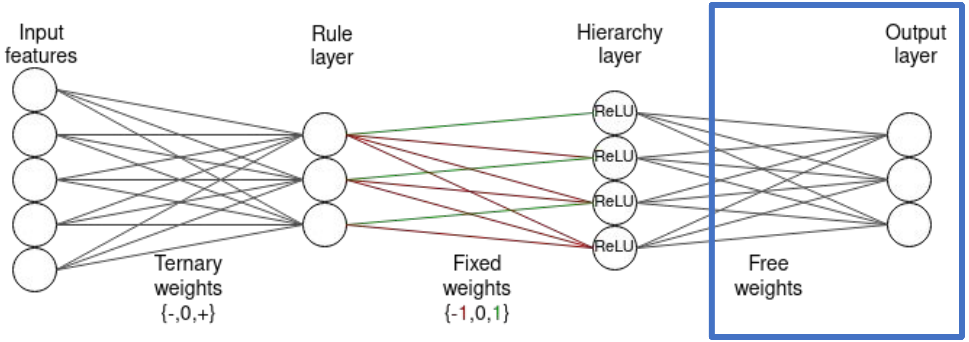
$$y = \sum w_i \cdot x_i - \sum_{w_i > 0} w_i + 1$$

Model: Hierarchy Layer



- Rule n is activated only if rules 1,...,n-1 are not
- Ternary weights set at initialization
- Non-trainable weights
- Lowest node = default ELSE rule

Model: Output Layer



- Associates one label per rule
- Free weights
- Highest weight of each rule is the label

⟶ Remains fully interpretable

Model: Gradient-Based Training

Standard neural network optimization

- ADAM optimizer
- Cross-entropy loss
- Callback on validation loss
- Balanced loss

Simple multi-label adaptation

- Activation: softmax $\longrightarrow$ sigmoid
- Loss: cross-entropy $\longrightarrow$ binary cross-entropy

Overview

- Objective
- Model
- **Experiments**
- Results
- Conclusion

Experiments

Binary classification

- Datasets: 7
- Comparison models: DR-Net, RIPPER, CART

Multi-class classification

- Datasets: 6
- Comparison models: RIPPER, CART

Multi-label classification

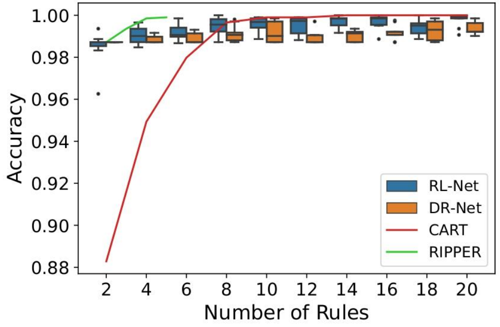
- Datasets: 2
- Comparison models: CART, baseline

Overview

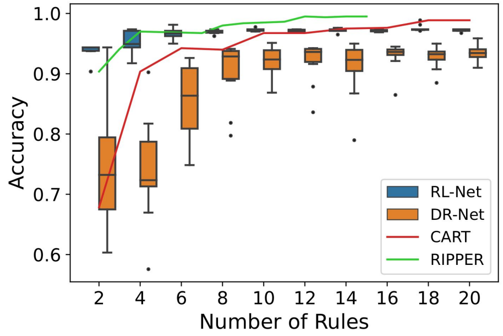
- Objective
- Model
- Experiments
- **Results**
- Conclusion

Results: Binary Classification

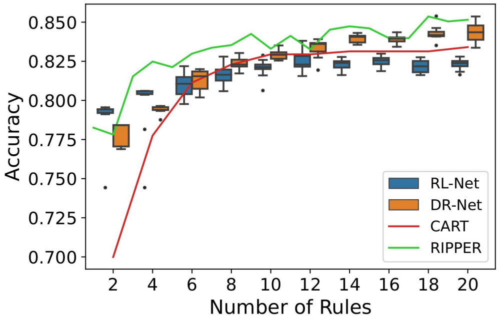
Mushroom



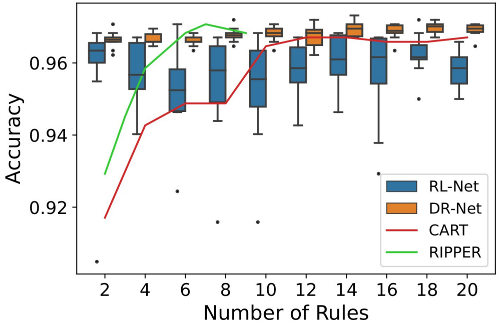
Chess



Magic

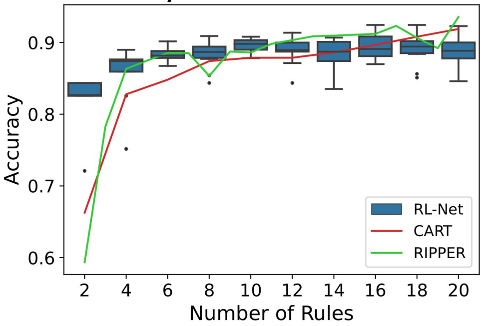


Ads

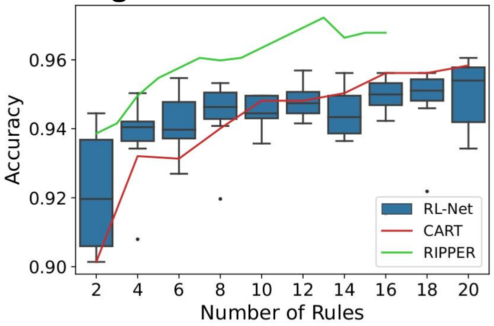


Results: Multi-Class Classification

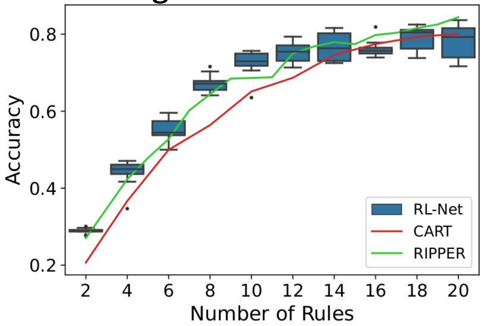
Nursery



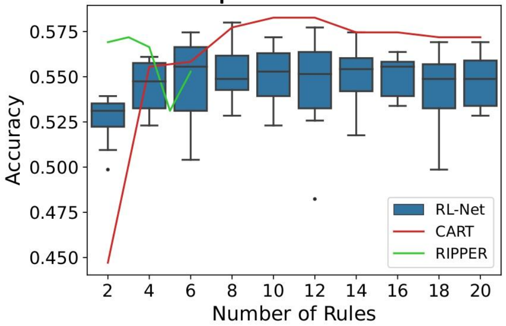
Pageblocks



Pendigits

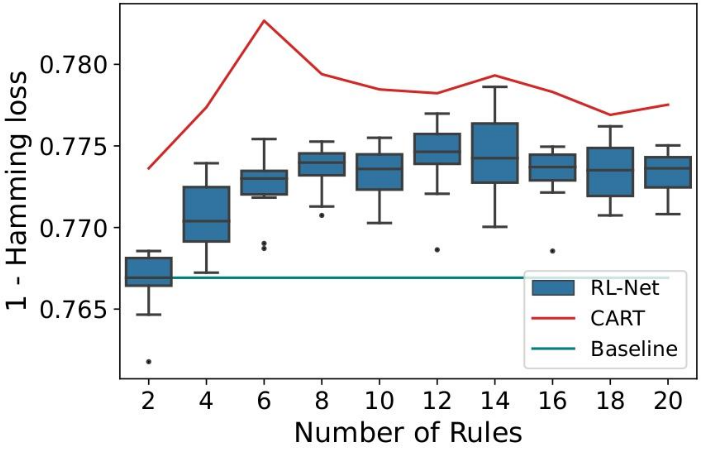


Contraceptivemc

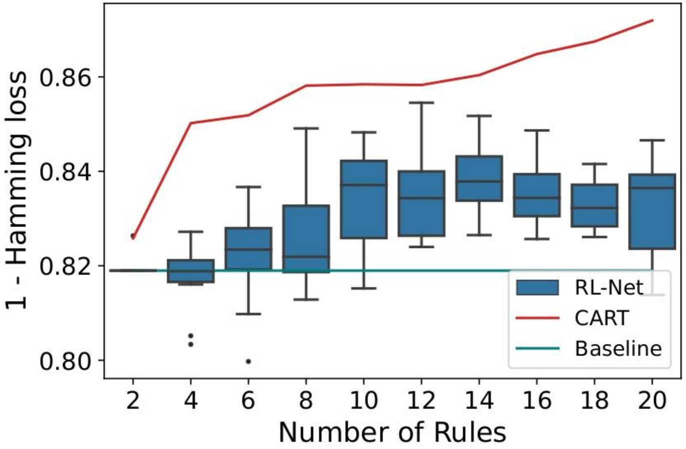


Results: Multi-Label Classification

Yeast



Scene



Overview

- Objective
- Model
- Experiments
- Results
- **Conclusion**

Conclusion

RL-Net

- Interpretable model
- Neural network learning
- Learns binary and multi-class classification
- Minor adaptations for multi-label

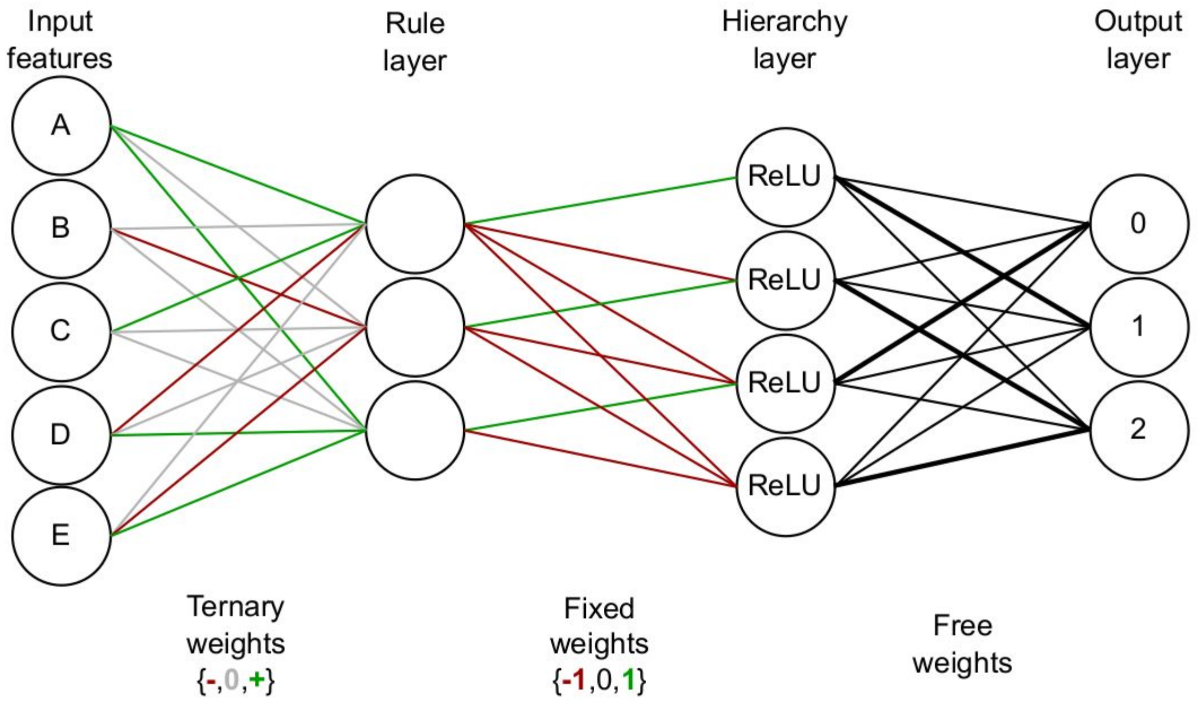
Performance

- Binary and multi-class: close to classical algorithms
- Multi-label: further work needed

⟶ Future work: Multi-label, initialization, integration to NN research

RL-Net Overview

https://github.com/luciledierckx/RLNet



IF A and C and not D THEN 1

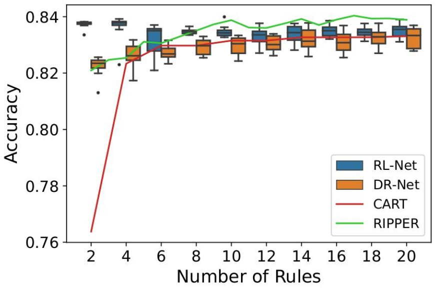
ELSE IF not B and not E THEN 2

ELSE IF A and D and E THEN 0

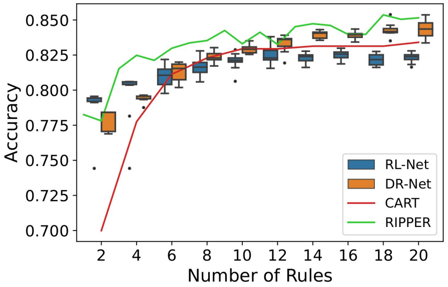
ELSE 2

Results: Binary Classification

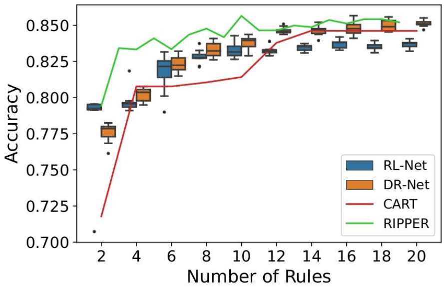
Adult



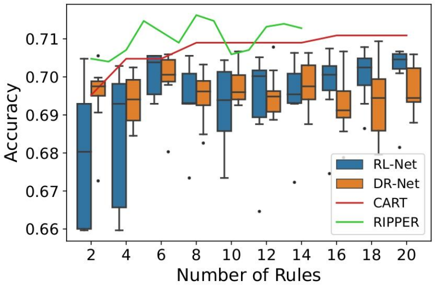
Magic



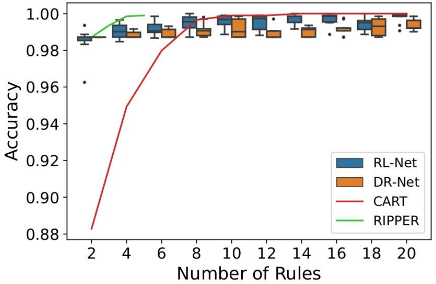
House



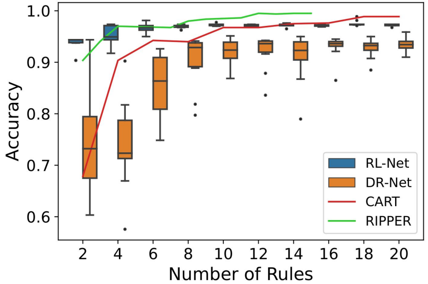
Heloc



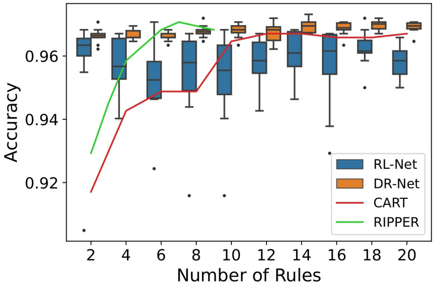
Mushroom



Chess

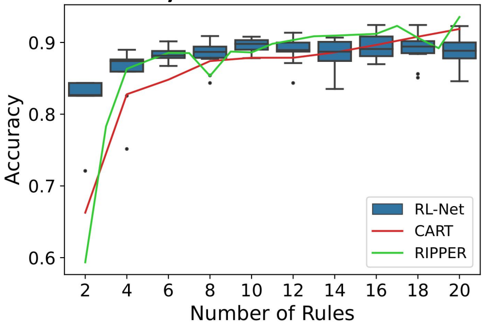


Ads

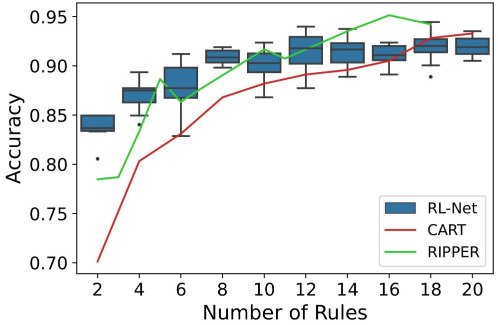


Results: Multi-Class Classification

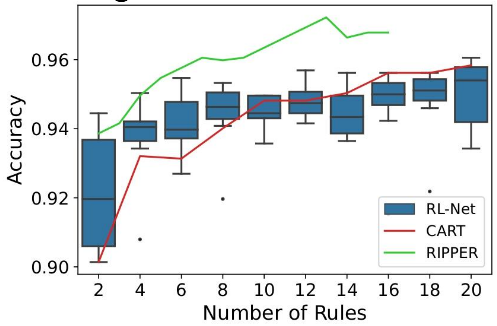
Nursery



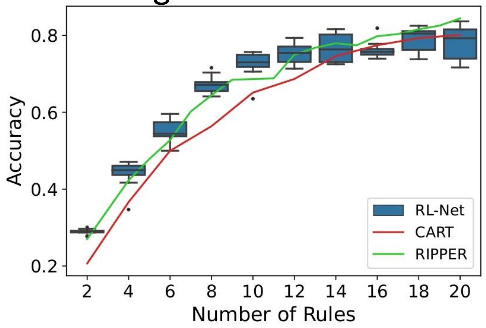
Car



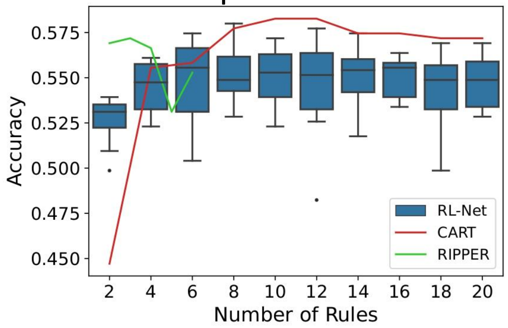
Pageblocks



Pendigits



Contraceptivemc



Drive