Gradient-based Learning of Simple yet Accurate Rule-Based Classifiers

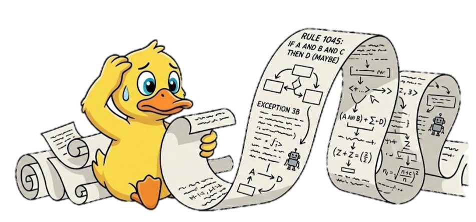
Javier Fumanal-Idocin[1], Raquel Fernandez-Peralta[2], Javier Andreu-Perez[1]

[1]University of Essex [2]Slovak Academy of Sciences

Abstract

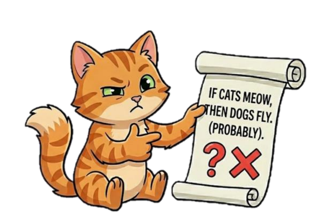
We present the **Fuzzy Rule-based Reasoner** (FRR), a novel **gradient-based rule learning** system that supports strict user constraints over rule-based complexity while achieving **competitive performance**. To maximize **interpretability**, the FRR uses semantically meaningful fuzzy logic partitions and single-rule decision-making. Through evaluation across 40 datasets, FRR demonstrates superior performance to traditional rule-based methods (**5% improvement over RIPPER**), **comparable** accuracy to tree-based models (**CART**) using 90% **more compact rule bases**, and 96% of state-of-the-art additive rule-based accuracy using only 3% of their rule base size.

Good Rule-based Systems are…



Easy to Understand

Hard to mine!



Accurate

Gradient Learning for Rule-based Classification

The FRR has many **indicator functions** to select discrete components. These causes problems because they are **not differentiable** and we need to use the straight through estimator:

Select linguistic label:

$$u_j^{(2)} = \sum_{v=1}^{V} f(W_{j,v}^{(2)}) W_{j,v}^{(2)} \mu_{j,v}(X_{i,j})$$

Select feature:

$$A_k = \sum_{j=1}^{M} f(W_{k,j}^{(3)}) W_{k,j}^{(3)} u_j^{(2)}$$

Winner Takes All (WTA) function

However, WTA makes **gradients very sparse**. We propose **Restricted additions** to substitute indicators in training:

$$f_\beta(M_{i,j}) = \begin{cases} \frac{1}{1+\beta(m-1)} & \text{if } j = \arg\max_k M_{i,k}, \\ \frac{\beta}{1+\beta(m-1)} & \text{otherwise,} \end{cases}$$

$$r(X) = w_r \prod_{a \in A_r} \mu_a(X)$$

Rule truth values are a productory. This causes important **vanishing gradients.**

- **Residual-like connections:**

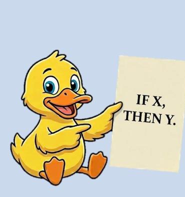
$$\tilde{u}^{(3)} = u^{(3)} + \gamma \sum_{k=1}^{A} A_k$$

- **Projection:** the root makes small values larger.

$$\mathbb{P}(R_r) = \sqrt[r]{R_r}$$

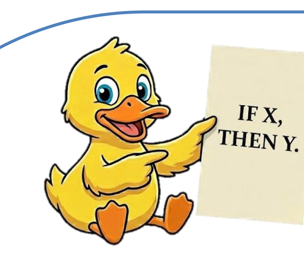
Conclusions

The FRR is a novel explainable neural classifier with a custom architecture that learns **interpretable rules** through **gradient-based optimization**.

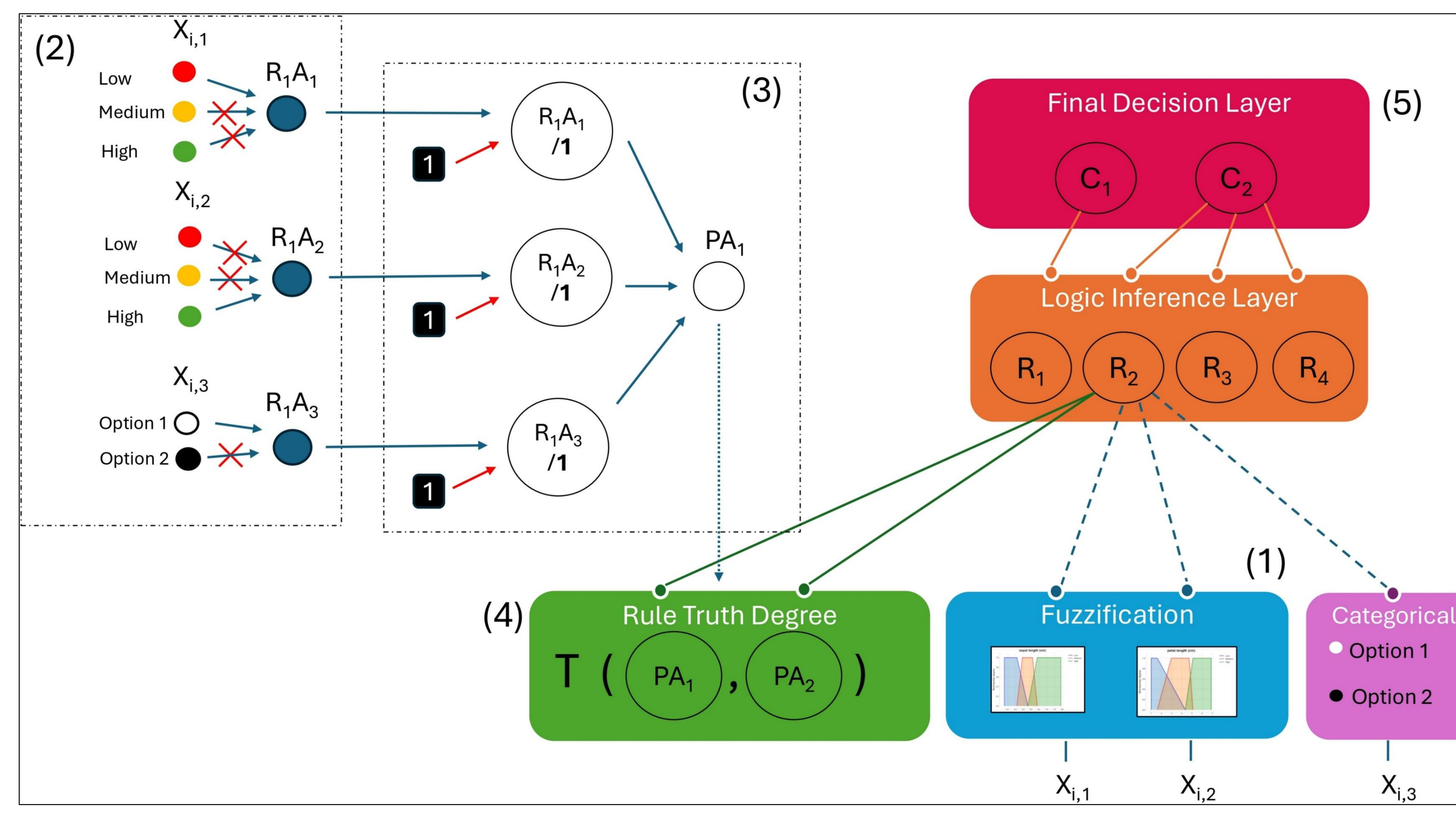


In the FRR the user can set the maximum number of rules and the length of their antecedents, allowing the user to **control the semantics** of the conditions.

The FRR obtains a very good **balance** in terms of **accuracy** and **complexity** using sufficient rules.



Complexity-controlled Architecture

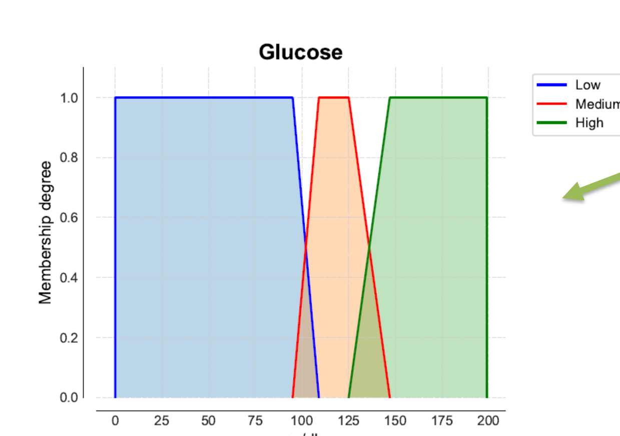


Fuzzy Rule-based Reasoner (1) We **fuzzify** the input for each variable (2) We **select** the **linguistic label** for each condition. (3) We **reduce** the size of the rule. (4) We compute the **truth value** for each **rule**. (5) We select the **output class** indicated by the rule with the maximum truth degree.

The **User Controls Complexity**:

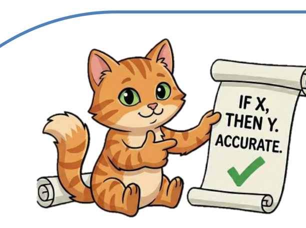
- Number of **rules**
- **Rule length**
- **Semantics** of conditions

Rules for Non-Diabetic Patients
IF Diabetes Pedigree IS Low **AND** Age IS Low
IF Skin Thickness IS Low **AND** Insulin IS Medium
IF Body Mass Index (BMI) IS Low
IF Blood Pressure IS High **AND** Insulin IS Low
IF Times Pregnant IS Medium **AND** Blood Pressure IS High **AND** Age IS Medium
Rules for Diabetic Patients
IF Glucose Level IS High



FRR can discard useless conditions as well!

$$u_k = \underbrace{\prod_{k=1}^{A} (\underbrace{\alpha_{k,1} f(\alpha_{k,1}) A_k}_{\text{Real condition}} + \underbrace{\alpha_{k,2} f(\alpha_{k,2})}_{\text{Static value}})}_{\text{WTA}}$$



Results

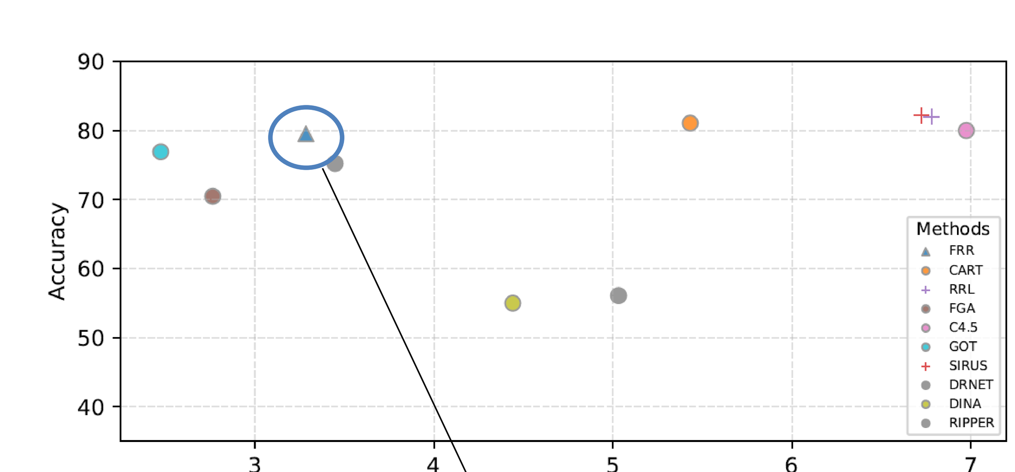
Method	FRR	FGA	DRNet	DINA	RIPPER	CART	C4.5	GOT	SIRUS	RRL	LR	GB
	Sufficient Rule-based					Tree-based			Additive Rule-based			
Accuracy	79.51	70.46	56.08	54.99	75.22	81.06	79.99	76.91	82.17	81.99	82.12	86.04
Number of Rules	13.77	7.12	24.04	18.48	16.04	39.75	131.92	5.23	286.71	99.35	–	–
Conditions/Rule	1.94	2.23	6.37	4.59	1.96	5.75	8.10	2.27	2.90	8.85	–	–
Rule base Size	26.71	15.87	153.13	84.82	31.43	228.56	1068.55	11.87	831.45	879.24	–	–
Unique Conditions	10.78	10.52	16.26	9.18	21.30	34.72	68.56	11.00	357.05	125.16	–	–

Better accuracy than:

- RIPPER
- Genetic Optimization
- Other gradient-based sufficient rule learners.

Comparable to:

- C4.5 with less complexity.



3rd less complex – Better performing than the two less complex ones

University of Essex
YUFE4Postdocs
Funded by the European Union NextGenerationEU
RECOVERY AND RESILIENCE PLAN
DEPUTY PRIME MINISTER'S OFFICE FOR THE RECOVERY PLAN AND THE KNOWLEDGE ECONOMY
EurIPS

Corresponding author: j.fumanal-Idocin@essex.ac.uk