Gradient-based Optimization for Compact and Explainable Fuzzy Rule-based Classification

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

[1]School of Computer Science and Electronic Engineering, University of Essex
[2]Mathematical Institute, Slovak Academy of Sciences, Bratislava, Slovakia
j.fumanal-idocin@essex.ac.uk, raquel.fernandez@mat.savba.sk, j.andreu-perez@essex.ac.uk

Abstract

Rule-based models are valued in high-stakes decision-making for their transparency, but their discrete nature limits optimization and scalability. We propose the Fuzzy Rule-based Reasoner (FRR), a gradient-based rule learner that enforces user-defined complexity constraints while maintaining strong performance. FRR combines interpretable fuzzy logic partitions with sufficient (single-rule) decision-making, avoiding the combinatorial growth of additive ensembles. Across 40 datasets, FRR outperforms traditional rule-based methods (by about 5% over RIPPER), matches the accuracy of tree-based models like CART with rule bases 90% smaller, and achieves 96% of the accuracy of additive rule-based models while using only 3% of their rule base size.

1 Introduction

Deep neural networks excel in handling large volumes of unstructured data, such as images and videos [1], but their lack of transparency limits their use in high-stakes domains like medicine and finance [2]. Rule-based algorithms, by contrast, are inherently interpretable, as they reveal explicit decision patterns and their relevance, offering more trustworthy explanations than many post-hoc XAI methods [3, 4]. Studying learned rules allows practitioners to validate or challenge model insights and even use them to approximate complex models like deep networks [5, 6]. However, rule-based classifiers often face a trade-off between interpretability and accuracy: larger rule sets improve performance but reduce transparency [7]. While fuzzy logic and genetic optimization have been used to balance this trade-off [8], they struggle to scale, and gradient-based approaches [9–13] often produce overly complex rule bases. To address these issues, we propose the Fuzzy Rule-Based Reasoner (FRR), a fully differentiable rule-based classifier that integrates user-defined complexity constraints (maximum rules and conditions per rule) with interpretable fuzzy partitions, achieving a balance between performance, simplicity, and transparency.

*Corresponding Author.

2 Fuzzy Rule-Based Reasoner

The FRR is a hierarchical model composed of four layers of matrix operations that mimic fuzzy logical reasoning (Figure 1). It receives an input vector X_i and produces a class prediction through a sequence of fuzzification, inference, and decision steps.

1. **Fuzzification layer**: The fuzzification layer transforms input features into interpretable fuzzy membership values. Continuous variables are mapped to linguistic terms such as "low," "medium," and "high" using trapezoidal fuzzy sets defined according to the data's quantile distribution, while categorical variables are encoded using a one-hot representation.

2. **Logic inference layers**: The logic inference process in the FRR has two main steps. First, for each feature, the model selects the most relevant fuzzy label, the one with the highest weight, representing which linguistic partition is activated. Next, it determines which features form the rule's antecedent, keeping a fixed number of conditions per rule. The truth value of a rule is then computed as the product of the selected condition contributions, ensuring the result remains within the $[0, 1]$ range.

3. **Decision layer**: Implements sufficient-rule prediction by selecting the consequent of the highest-scoring activated rule.

Beyond these layers, the FRR also incorporates a parsimony mechanism that automatically prunes irrelevant rule conditions during training. Through a competitive cancellation process, conditions that contribute little to the prediction are deactivated and penalized via a regularization term, ensuring compact and efficient rule bases without sacrificing accuracy.

3 Training

Training the FRR addresses three main challenges: non-differentiability, gradient sparsity, and vanishing gradients. The non-differentiable arg max used for rule and feature selection is approximated with the

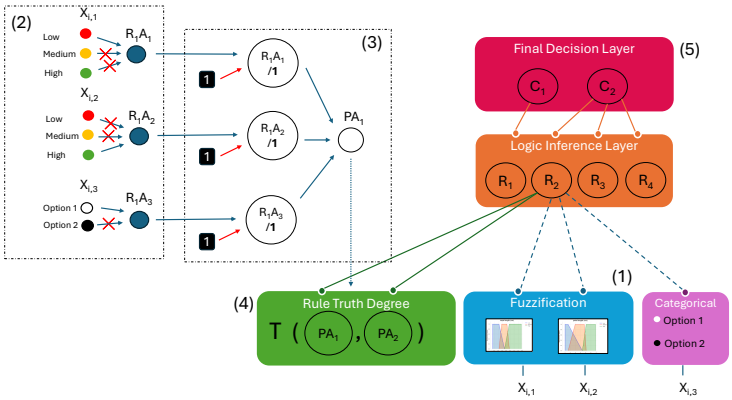

Figure 1. FRR scheme, using an input X_i with three features, four rules, and two target classes.

Table 1. 5-fold results for all the datasets considered.

	Sufficient Rule-based					Tree-based			Additive Rule-based			
Method	FRR	FGA	DRNet	DINA	RIPPER	CART	C4.5	GOT	SIRUS	RRL	LR	GB
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	–	–

Straight-Through Estimator (STE) [14, 15], enabling gradient flow through discrete choices. To reduce gradient sparsity, a relaxed indicator function with parameter β allows partial updates to non-selected features; β gradually decreases during training to recover discrete behavior. Finally, to combat vanishing gradients from multiplicative rule inference, two strategies are applied: root-normalized activation, which rescales small truth values, and residual connections [16], which preserve gradient flow early in training and fade out over time. Together, these mechanisms ensure stable and efficient optimization of interpretable fuzzy rules.

4 Experiments

We evaluated the proposed Fuzzy Rule-based Reasoner (FRR) across 40 widely used classification datasets, ranging from 80 to 19,020 samples and 2 to 85 features, using 5-fold cross-validation and accuracy as the primary metric. Statistical differences between classifiers were assessed using the Friedman Test and Post-hoc Nemenyi procedure [17]. The FRR was compared against three groups of baselines: (i) rule-based methods, including the Fuzzy Genetic Algorithm classifier (FGA) [18], RIPPER [19], SIRUS [20], DRNet [9], DINA [21], and Rule-based Representation Learning (RRL) [22]; (ii) tree-based models, such as CART [23], C4.5 [24], and the Generalized Optimal Sparse Decision Tree (GOT) [25]; and (iii) non-rule-based classifiers, namely Logistic Regression (LR) and Gradient Boosting (GB) [26]. While GB achieved the highest average accuracy (86.04%), FRR maintained competitive performance (79.51%) with significantly lower complexity. Indeed, its rule base was only 3% the size of SIRUS and 11% that of CART. FRR was statistically superior to other sufficient rule-based classifiers, particularly RIPPER, and demonstrated scalability across datasets of varying dimensionality. Gradient-based rule learners (DRNet and DINA) performed worse, likely due to their sensitivity to hyperparameters and smaller dataset sizes. Overall, FRR achieved a strong balance between interpretability and accuracy, offering an efficient and explainable alternative to traditional and gradient-based rule learning methods.

5 Conclusion

We introduced the Fuzzy Rule-based Reasoner (FRR), an explainable classifier that learns interpretable rules through gradient-based optimization while allowing users to control model complexity by setting limits on rule count and length. This design maintains interpretability without sacrificing performance, achieving a strong balance between accuracy and simplicity. FRR outperforms other gradient-based rule learners and reaches accuracy comparable to tree-based models with far lower complexity. Future work will integrate FRR into deep learning frameworks to provide rule-based explanations within gradient flows and explore how its learned rules relate to epistemic and aleatoric uncertainty.

References

[1] Y. LeCun, Y. Bengio, and G. Hinton. "Deep learning". In: *Nature* 521.7553 (2015), pp. 436–444. DOI: 10.1038/nature14539.

[2] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". In: *Information fusion* 58 (2020), pp. 82–115. DOI: 10.1016/j.inffus.2019.12.012.

[3] R. Tomsett, D. Harborne, S. Chakraborty, P. Gurram, and A. Preece. "Sanity checks for saliency metrics". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 34. 04. 2020, pp. 6021–6029. DOI: 10.5555/3327546.3327621.

[4] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong. "Interpretable machine learning: Fundamental principles and 10 grand challenges". In: *Statistic Surveys* 16 (2022), pp. 1–85. DOI: 10.1214/21-SS133.

[5] J. Fumanal-Idocin, J. Andreu-Perez, O. Cord, H. Hagras, H. Bustince, et al. "Artxai: Explainable artificial intelligence curates deep representation learning for artistic images using fuzzy techniques". In: *IEEE Transactions on Fuzzy Systems* (2023). DOI: 10.1109/TFUZZ.2023.3337878.

[6] R. Li, Q. Li, Y. Zhang, D. Zhao, Y. Jiang, and Y. Yang. "Interpreting unsupervised anomaly detection in security via rule extraction". In: *Advances in Neural Information Processing Systems* 36 (2024). DOI: 10.5555/3666122.3668840.

[7] L. Breiman. "Random forests". In: *Machine learning* 45 (2001), pp. 5–32. DOI: 10.1023/A:1010950718922.

[8] J. Alcalá-Fdez, R. Alcala, and F. Herrera. "A fuzzy association rule-based classification model for high-dimensional problems with genetic rule selection and lateral tuning". In: *IEEE Transactions on Fuzzy systems* 19.5 (2011), pp. 857–872. DOI: 10.1109/TFUZZ.2011.2147794.

[9] L. Qiao, W. Wang, and B. Lin. "Learning accurate and interpretable decision rule sets from neural networks". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 5. 2021, pp. 4303–4311. DOI: 10.5555/AAI27997821.

[10] W. Zhang, Y. Liu, Z. Wang, and J. Wang. "Learning to Binarize Continuous Features for Neuro-Rule Networks". In: *IJCAI*. 2023, pp. 4584–4592. DOI: 10.24963/ijcai.2023/510.

[11] Z. Wang, W. Zhang, N. Liu, and J. Wang. "Learning interpretable rules for scalable data representation and classification". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024). DOI: 10.1109/TPAMI.2023.3328881.

[12] S. Xu, N. P. Walter, J. Kalofolias, and J. Vreeken. "Learning Exceptional Subgroups by End-to-End Maximizing KL-Divergence". In: *International Conference on Machine Learning*. PMLR. 2024, pp. 55267–55285. DOI: 10.5555/3692070.3694348.

[13] Y. Yang, W. Ren, and S. Li. "Hyperlogic: Enhancing diversity and accuracy in rule learning with hypernets". In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 3564–3587. DOI: 10.5555/3737916.3738034.

[14] P. Yin, J. Lyu, S. Zhang, S. J. Osher, Y. Qi, and J. Xin. "Understanding Straight-Through Estimator in Training Activation Quantized Neural Nets". In: *International Conference on Learning Representations*. 2019.

[15] M. Schoenbauer, D. Moro, L. Lew, and A. Howard. *Custom Gradient Estimators are Straight-Through Estimators in Disguise*. 2024. DOI: 10.48550/arXiv.2405.05171. arXiv: 2405.05171 [cs.LG].

[16] K. He, X. Zhang, S. Ren, and J. Sun. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

[17] J. Demšar. "Statistical Comparisons of Classifiers over Multiple Data Sets". In: *Journal of Machine Learning Research* 7 (2006), pp. 1–30. ISSN: 1532-4435. DOI: 10.5555/1248547.1248548.

[18] J. Fumanal-Idocin and J. Andreu-Perez. "ExFuzzy: A library for symbolic explainable AI through fuzzy logic programming". In: *Neurocomputing* (2024), p. 128048. ISSN: 0925-2312. DOI: 10.1016/j.neucom.2024.128048.

[19] W. W. Cohen et al. "Fast effective rule induction". In: *Proceedings of the twelfth international conference on machine learning*. 1995, pp. 115–123. DOI: 10.1016/B978-1-55860-377-6.50023-2.

[20] C. Bénard, G. Biau, S. Da Veiga, and E. Scornet. "Interpretable random forests via rule extraction". In: *International conference on artificial intelligence and statistics*. PMLR. 2021, pp. 937–945.

[21] N. P. Walter, J. Fischer, and J. Vreeken. "Finding interpretable class-specific patterns through efficient neural search". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 8. 2024, pp. 9062–9070. DOI: 10.1609/aaai.v38i8.28756.

[22] Z. Wang, W. Zhang, N. Liu, and J. Wang. "Scalable rule-based representation learning for interpretable classification". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 30479–30491. DOI: 10.5555/3540261.3542593.

[23] R. Timofeev. "Classification and regression trees (CART) theory and applications". In: *Humboldt University, Berlin* 54 (2004).

[24] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993. DOI: 10.1007/BF00993309.

[25] H. McTavish, C. Zhong, R. Achermann, I. Karimalis, J. Chen, C. Rudin, and M. Seltzer. "Fast sparse decision tree optimization via reference ensembles". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 36. 9. 2022, pp. 9604–9613. DOI: 10.1609/aaai.v36i9.21194.

[26] J. H. Friedman. "Greedy function approximation: a gradient boosting machine". In: *Annals of statistics* (2001), pp. 1189–1232. DOI: 10.1214/aos/1013203451.