

Global Explanations with Local Scoring

Mattia Setzu¹ Riccardo Guidotti² Anna Monreale¹ Franco Turini¹
¹University of Pisa, Italy ²ISTI-CNR, Pisa, Italy

The Local-to-Global Problem (L2G)

Goals

Given

- a black box classifier b ,
- a set of records $X = \{x_1, \dots, x_n\} \in \mathbb{R}^{n \times m}$
- a set of *local* decision rules $R = \{r_1, \dots, r_n\}$, with rule r_i explaining the decision $b(x_i)$,

subsume the local explanations into a set of global ones to explain the behavior of the black box on the whole input space.

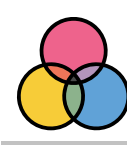


- **Low complexity** Small ruleset, few premises per rule.
- **High fidelity** Accuracy w.r.t. the black box's predictions.
- **High coverage** Every prediction can be explained.

The Rule Relevance Score (RRS)

Key idea: Assign a *local* score to each rule, rank them, select the best ones. The scoring formula is based on a weighted sum of *score vectors*:

$$RRS_{R,X} = \alpha_1 \cdot c + \alpha_2 \cdot s + \alpha_3 \cdot a + \alpha_4 \cdot \tilde{c} + \alpha_5 \cdot \tilde{a},$$

each holding a score for the rules' coverage (c, \tilde{c}), association score (a, \tilde{a}), sparsity (s) and accuracy.

	Coverage	$c(r_i, X) = \{x_i \in X \mid x_i \text{ satisfies } r_i\} $	How many records can the rule explain?
	Sparsity	$s(r_i, X) = K^{-2} \sum_{x_i, x_j \in \Gamma(r_i, X)} \ x_i - x_j\ ^2$	How generic is the rule in the input space?
	Association	$a(r_i, X) = \langle c(r, X)(c^{-1}(r, X))^{-1}, \dots \rangle$, where $c^{-1}(r, X)$ is the vector holding, for each record, the number of rules satisfying it.	Are the covered records "easy" to explain?

The $\tilde{\cdot}$ indicates that the score is computed only on the covered records on which the rule has the correct prediction. Score vectors are summed, and the rules with fidelity lower than a given β percentile are discarded.

Results

RRS shows higher harmonic score and complexity comparable to CPAR, the best-performing rule-mining competitor. The ruleset size varies across datasets, with RRS showing smaller, hence more interpretable, rulesets in 3/4 datasets, and comparable in 1/4 datasets.

	RRS			CPAR			CORELS			SBRL		
	H	size	len	H	size	len	H	size	len	H	size	len
Adult	.93	331	2.74	.90	299	4.15	.00	1	1	.84	24	1
Compas	.99	51	1.90	.83	94	2.37	.91	3	1	.99	6	1
Churn	.81	58	2.62	.77	>2k	3.08	.53	1	1	.08	7	1
German	.92	21	1.80	.75	26	2.19	-	-	-	.00	2	1

Table 1. Harmonic mean between fidelity and coverage (H), average rule length and ruleset size for a Deep Neural Network with $\alpha_i = 0.2$

The effects of the pruning factor β on the fidelity of the ruleset are shown in Figure 1.

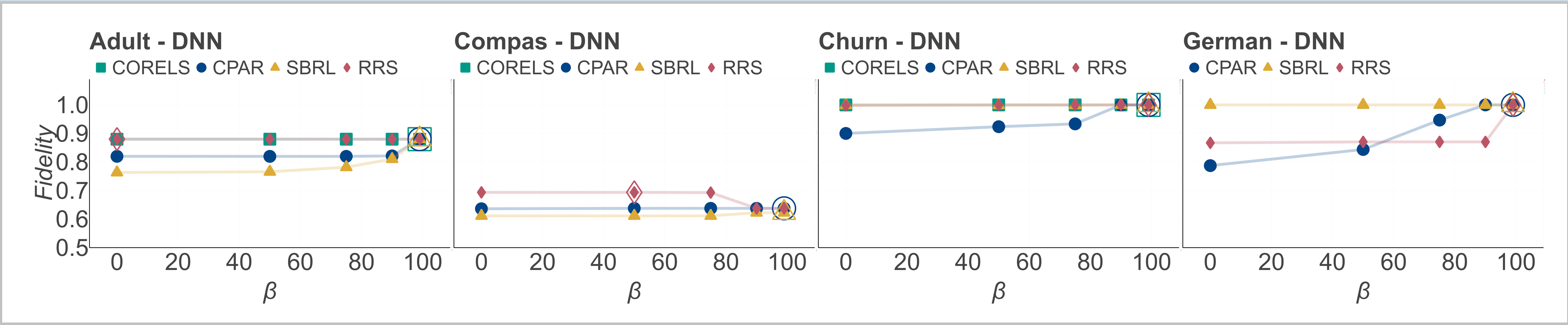


Figure 1. Fidelity for RRS on local explanations and global methods for NN, for the different datasets varying the pruning percentile threshold β . The highest score is highlighted by a double marker.

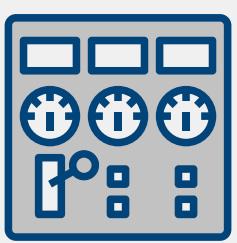
Pros

- › Fast computation $\in \mathcal{O}(\overline{m}^2)$, \overline{m} largest cardinality coverage
- › High fidelity

Cons

- › May need hyperparameter tuning
- › Slow on highly general local rules (large \overline{m})

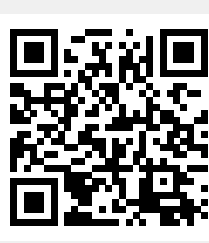
Future development & Sources



Effects of each hyperparameter



Finer-grained filtering for smaller rulesets



Source code on Github