GLocalX - From Local to Global Explanations of Black Box AI Models

Mattia Setzu [1] Riccardo Guidotti [1] Anna Monreale [1] Franco Turini [1] Dino Pedreschi [1] Fosca Giannotti [2]

[1]University of Pisa, Italy [2]ISTI-CNR, Italy

The Local to Global explanation problem

Local explanations tend to be faithful yet complex, while global explanations tend to be more general yet less faithful. **Can we retain the faithfulness of local explanations and the simplicity of global ones?** In a *Local to Global* setting [2] we are given:

black-box model set of local explanations handful of data

and aim to infer a set of global explanations from the local ones.

The GLocalX algorithm

GLocalX takes after aggregating hierarchical clustering algorithms. It yields global explanations by iteratively

1. identifying similar sets of explanations,
2. merging them,
3. validating their merge.

The resulting dendrogram is a binary tree whose nodes hold local explanations in its leaves, and increasingly global explanations as the nodes near the root. We extract a set of explanations of size α from the root node(s).

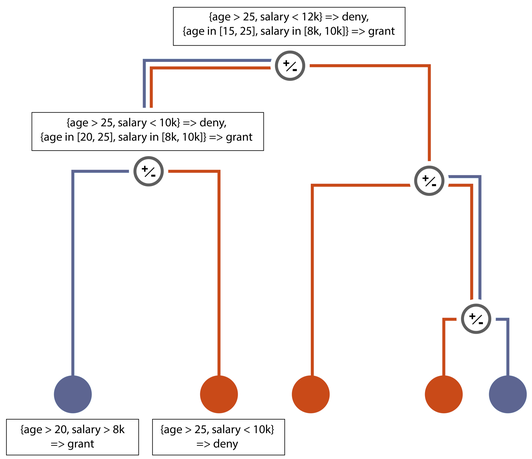


Figure 1. An example of merge in the dendrogram.

Here we focus on *axis-parallel decision rules* as local explanations, locally inferred by LORE [1].

Finding Similar Sets of Explanations

Similarity is measured as Jaccard similarity of explanation *coverage*, that is the larger the ratio of records explained by two explanations, the higher their similarity and viceversa.

$$sim(E_1, E_2) = \frac{cov(E_1, X) \cap cov(E_2, X)}{cov(E_1, X) \cup cov(E_2, X)}$$

When dealing with sets of explanations, the coverage of a set of explanations E can be:

linkage	explanation(s) selected	coverage
min	explanation e with minimum coverage	$cov(e, X), e : \arg\min_{e \in E} cov(e, X)$
max	explanation e with maximum coverage	$cov(e, X), e : \arg\max_{e \in E} cov(e, X)$
full	the union of all explanations	$\bigcup_{e \in E} cov(e, X)$

Merging explanations

An explanation is a polyhedra in the feature space, each premise $P_i : [a, b)$ identifying a continuous interval $[a, b)$ in the feature space of feature i.

$$\underbrace{age \geq 20}_{P_{\text{age}=[20,+\infty)}}, \underbrace{salary \geq 8k}_{P_{\text{salary}=[8k,+\infty)}} \rightarrow grant$$

Merge is bipartite:

- the join operator ($\oplus$) generalizes concordant explanations, increasing their coverage and possibly decreasing their fidelity;
- the cut operator ($\ominus$) localizes discordant explanations, decreasing their coverage, possibly increasing their fidelity.

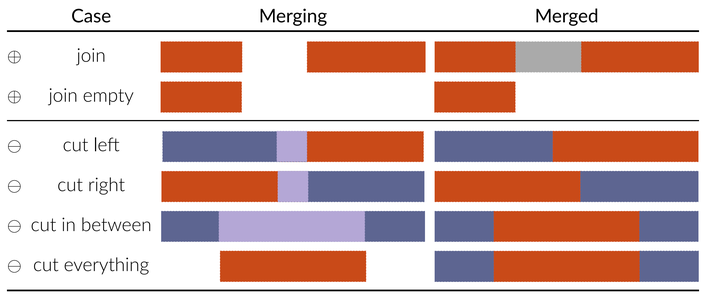


Table 1. Visual representation of the join ($\oplus$) and cut ($\ominus$) operators.

Accepting and Rejecting Merges

Not all merges are created equal. A merge is accepted if it yields a faithful and simple explanations [3]:

$$BIC(E) = fid(E, X, \tilde{Y}) + \frac{1}{|E|} \sum_{e \in E} len(e) + \frac{|E|}{n}$$

Local to Global vs Natively Global and Natively Local

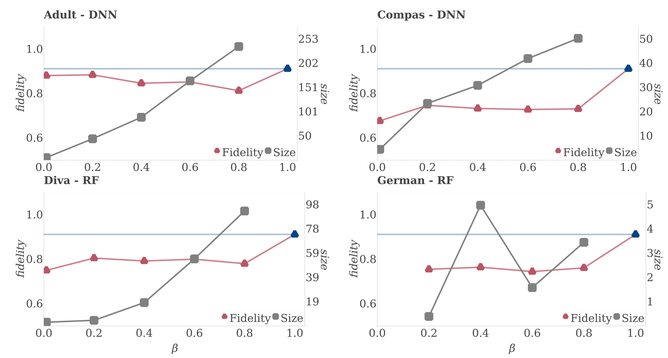
Table 2. GLocalX (trained on synthetic data) vs natively global models and a baseline of local explanations union.

	Adult, $\alpha = 10$			Compas, $\alpha = 5$		
Deep Network	Fidelity	Size	Length	Fidelity	Size	Length
GLocalX	.912	5	6.00 ± 1.0	.759	3	2.33 ± 0.4
GLocalX*	.880	10	6.30 ± 2.5	.756	6	4.50 ± 0.9
CPAR	.929	100	3.78 ± 2.4	.821	69	3.11 ± 1.4
Decision Tree	.917	1068	7.22 ± 1.9	.789	1014	6.00 ± 1.8
Pruned Decision Tree	.908	28	2.71 ± 0.8	.780	30	2.33 ± 0.6
Union	.880	6838	3.86 ± 2.2	.627	1515	4.65 ± 1.7
Random Forest	Diva, $\alpha = 25$			German, $\alpha = 2$		
GLocalX	.854	26	3.26 ± 0.9	.786	2	3.00 ± 1.0
GLocalX*	.848	26	3.88 ± 1.3	.766	2	5.87 ± 2.4
CPAR	.850	221	2.03 ± 1.0	.773	18	2.33 ± 1.4
Decision Tree	.853	976	10.63 ± 3.9	.830	76	4.50 ± 1.7
Pruned Decision Tree	.836	28	3.21 ± 0.9	.796	28	2.78 ± 0.8
Union	.794	2013	2.90 ± 1.0	.766	210	5.62 ± 2.4

What if I have less rules available?

Extracting local explanations can be time-consuming, can we reduce the input size?

Figure 2. GLocalX performance with $\beta \in 1, 20, 40, 60, 80$ percentage of input rules, randomized in 10 trials, and $\alpha_q = 50$.



References

[1] R. Guidotti, A. Monreale, F. Giannotti, D. Pedreschi, S. Ruggieri, and F. Turini. Factual and counterfactual explanations for black box decision making. *IEEE Intelligent Systems*, 34(6):14–23, 2019.

[2] D. Pedreschi, F. Giannotti, R. Guidotti, A. Monreale, S. Ruggieri, and F. Turini. Meaningful explanations of black box ai decision systems. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9780–9784, 2019.

[3] E. Wit, E. v. d. Heuvel, and J.-W. Romeijn. 'all models are wrong...': an introduction to model uncertainty. *Statistica Neerlandica*, 66(3):217–236, 2012.