Assessing Explanation Quality by Venn Prediction

Amr Alkhatib[1], Henrik Boström[1], Ulf Johansson[2]

[1]KTH Royal Institute of Technology

[2]Jönköping University

Sweden

COPA 2022

Explainable Machine Learning

- Many state-of-the-art machine learning algorithms produce black boxes

- A large number of techniques for explaining black-box models have been proposed; model-specific and model-agnostic, local or global, and with several different formats for the explanations, e.g., feature scores, partial-dependency plots, and rules

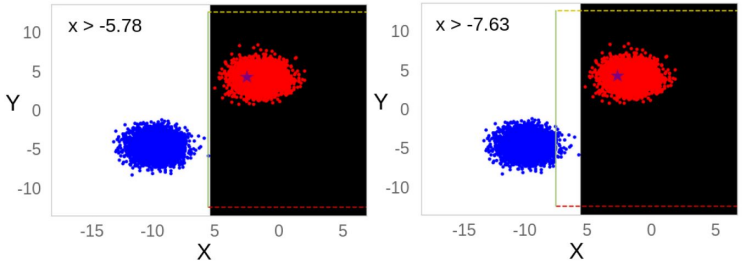
Explanations in the Form of Rules

- Rules are easy to understand, e.g., Anchors and LORE

(Total night calls ≤ 101 & Account length ≤ 128 & total day calls > 87) ⟶ (Churn)

Limitations of Explanatory Rules

- Explanations are affected by the method's design
  - Discretization of continuous features is one examples:



(a) A good discretization (b) A suboptimal discretization

  - Some methods use features perturbations:

    (different perturbations = different explanations)

Limitations of Explanatory Rules

- Sometimes explanatory rules are overly specific

- Fidelity can vary between different explanations

The Problem

- We cannot always trust the explanations

The Problem

- We cannot always trust the explanations

Solution?

- There is a need to quantify the uncertainty of the explanatory rules

The main contributions

- A novel method for quantifying the uncertainty of rule-based explanations
- A set of metrics designed to measure the uncertainty of the explanatory rules
- An empirical investigation comparing the uncertainty of explanations as produced by two different methods

Inductive Venn Predictors (IVPs)

- Split data into a proper training set to train the underlying model and calibration set to estimate label probabilities
- Divide the calibration data into categories
- Use the frequency of label Y_j in a category to estimate label probabilities for new instances in the same category

Inductive Venn Predictors (IVPs)

- The category of new objects is determined by the model in the same way as the calibration objects
- The label frequencies in the calibration data (in the same category) are used to calculate the label probabilities
- All possible labels are used since the true label is unknown (label probability distributions)

Inductive Venn Predictors (IVPs)

➢ A more compact representation can be obtained using the lower and upper probability estimates for each label

$$L(Y_j) = \frac{|\{(x_m, y_m) \in Z_k \mid y_m = Y_j\}|}{|Z_k| + 1}$$

$$U(Y_j) = \frac{|\{(x_m, y_m) \in Z_k \mid y_m = Y_j\}| + 1}{|Z_k| + 1}$$

- *Where Y_j is the label, x_m is the instance, y_m is intance's label, and Z_k is the calibration set*

The Proposed Method

1. The dataset is split into development and calibration sets

2. For each rule, a subset of the calibration set is selected where the items in the antecedent are true

3. The labels of the calibration subset objects are obtained through the black-box model

The Proposed Method

4. The upper and lower probability estimates are computed for each class label

$$L(Y_j) = \frac{|\{(x_i, y_i) \in Z_r | y_i = Y_j\}|}{|Z_r| + 1}$$

$$U(Y_j) = \frac{|\{(x_i, y_i) \in Z_r | y_i = Y_j\}| + 1}{|Z_r| + 1}$$

Evaluation Metrics

- The Average Lower Bound (ALB):

$$ALB = \frac{1}{n}\sum_{i=1}^{n} L(Y_{ji}),$$ *where n is the number of rules and Y_{ji} is the correct label of rule i*

- The Interval Size:

$$Interval\ Size = U(Y_j) - L(Y_j),$$ *where Y_j is a class label of rule*

- The absolute lower bounds difference:

$$\triangle LB = |L(Y_1) - L(Y_0)|,$$ *where Y_1 and Y_0 are the labels of the positive and negative classes respectively*

Evaluation Metrics Extensions

- ALB-P penalizes a method for explanations with low coverage
  - ALB is averaged over objects with applicable rules, while ALB-P adds zero for objects not covered by any rules
- Weighted interval size and △LB averages
  - Each value is multiplied by the number of instances that the rule covers, then summed and averaged by the total count of covered objects

Experimental Setup

- The experiments were conducted on 12 public datasets available on Openml.org
  - The datasets are ada, Bank Marketing, BNG breast-w, Compas, Internet Advertisements, Jungle Chess 2pcs, mc1, Mushroom, Phishing Websites, Spambase, Telco Customer Churn
- The data was split into training, development, calibration, and test subsets
  - 40% training, 20% development, 20% calibration, and 20% test
- The black-box models were generated using the XGBoost algorithm
  - The learning rate, num of estimators, and the regularization parameter were tuned by grid search

Compared Methods

1. Anchors with the default hyperparameters, and 0.9 confidence threshold

2. Association rule mining using the Apriori algorithm

Experiments

1. Rule-Based Evaluation:
   - The interval size, $\triangle$LB, and the ALB are computed

2. Instance-Based Evaluation:
   - The weighted averages of the interval size and $\triangle$LB, in addition to ALB-P, are computed

Example Output

Top 4 rules of each class output by Anchors for the BNG breast dataset

Conditions	Label	L(Y)	U(Y)	Iterval Size	△LB
Cell Size Uniformity ≤ 1 & Cell Shape Uniformity ≤ 1	Benign	0.997	0.998	0.001	0.995
Cell Size Uniformity ≤ 1 & Clump Thickness ≤ 2.22	Benign	0.998	1	0.002	0.998
Cell Shape Uniformity ≤ 1 & Clump Thickness ≤ 2.22	Benign	0.997	1	0.003	0.997
Cell Shape Uniformity ≤ 1 & Clump Thickness ≤ 4	Benign	0.998	0.999	0.001	0.996
Clump Thickness > 5.66 & Cell Size Uniformity > 4.49	Malignant	0.997	1	0.003	0.997
Clump Thickness > 5.66 & Normal Nucleoli > 3.55	Malignant	0.989	0.993	0.004	0.982
Cell Size Uniformity > 4.49 & Bare Nuclei > 6.61	Malignant	0.997	1	0.003	0.997
Cell Size Uniformity > 4.49 & Single Epi Cell Size > 2	Malignant	0.961	0.963	0.002	0.924

Results

1. Rule-Based Evaluation:

	Association Rules					Anchors				
	ALB	Interval Size	△LB	#Rules*	Cov.**	ALB	Interval Size	△LB	#Rules	Cov.
Average values	0.95	0.015	0.95	175	0.84	0.93	0.1	0.81	204	1.0
Average rank	1.08	1.08	1			1.92	1.92	2		

2. Instance-Based Evaluation:

Dataset	Association Rules					Anchors				
	ALB-P	Interval Size	△LB	#Rules	Cov.	ALB-P	Interval Size	△LB	#Rules	Cov.
Average values	0.80	0.0062	0.97	175	0.84	0.90	0.0064	0.92	204	1.0
Average rank	1.92	1.375	1.125			1.08	1.625	1.875		

Conclusions

- We proposed a method to quantify the uncertainty of the explanatory rules
- We provided a set of metrics of rule quality based on uncertainty
- We presented results from an empirical evaluation

Future Work

- Investigate uncertainty quantification for other explanation Forms

Thank You!

References

- Henrik Boström, Ram B. Gurung, Tony Lindgren, and Ulf Johansson. Explaining random forest predictions with association rules. Archives of Data Science, Series A (Online First), 5(1):A05, 20 S. online, 2018. ISSN 2363-9881. doi: 10.5445/KSP/1000087327/05.

- Julien Delaunay, Luis Gal´arraga, and Christine Largou¨et. Improving Anchor-based Explanations. In CIKM 2020 - 29th ACM International Conference on Information and Knowledge Management, pages 3269–3272, Galway / Virtual, Ireland, October 2020. ACM. doi: 10.1145/3340531.3417461

- A. Gammerman, V. Vovk, and V. Vapnik. Learning by transduction. In Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, UAI'98, page 148–155, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc. ISBN 155860555X

- Christoph Molnar. Interpretable Machine Learning. 2022.