Estimating Quality of Approximated Shapley Values Using Conformal Prediction

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

[1]KTH Royal Institute of Technology

[2]Jönköping University

COPA 2024

A Need for Explanation

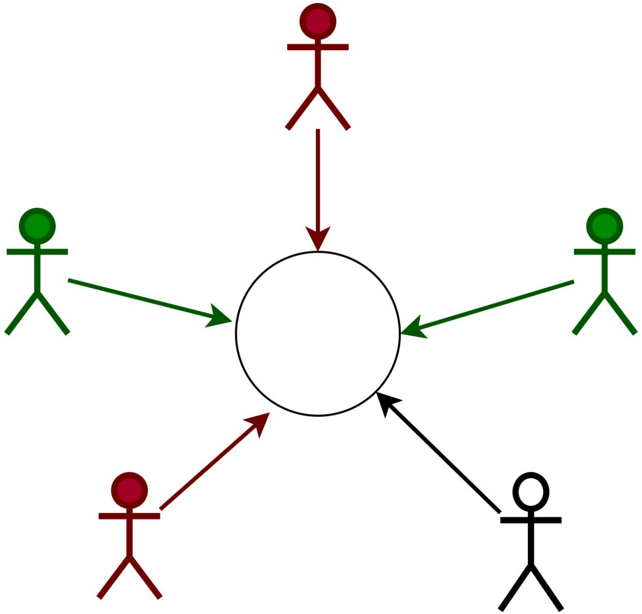
- Building trust in machine learning models

- Ethical and legal considerations

Explanation Methods

- Surrogate Models
- Important Features Selection
- Generation of Adversarial Examples
- The Shapley Value

The Shapley Value



The Shapley Value

- **Local accuracy:** the explanation matches the model
- **Missingness:** a missing feature is attributed a value of zero
- **Consistency:** if the contribution of a feature increases or remains unchanged, the Shapley value increases or remains unchanged

Efficient Approximation of the Shapley Value

- KernelSHAP
- TreeSHAP for a tree-based model
- FastSHAP learns to approximate the Shapley values
- Hierarchical Shap, H-Shap, for image classification

Fast approximations are not always accurate!

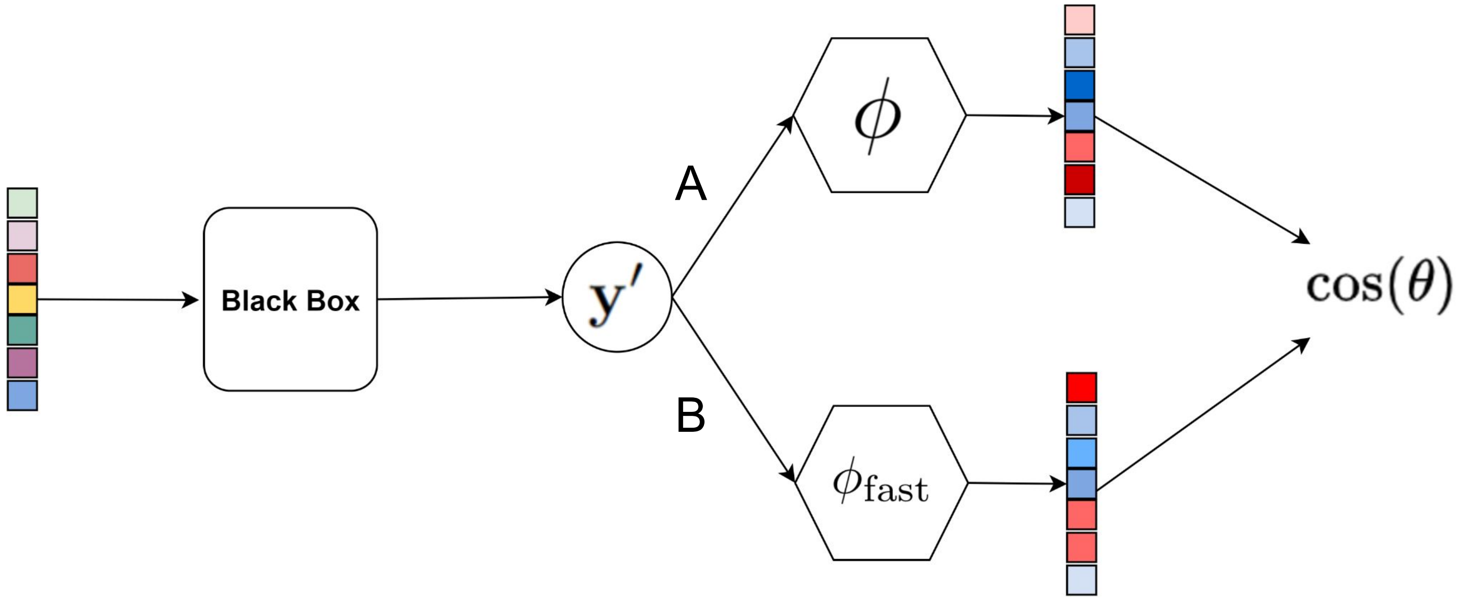
Dataset	FastSHAP
Abalone	0.81
Bank32nh	0.598
Churn	0.311
Delta Ailerons	0.867
Electricity	0.625
Elevators	0.828
Higgs	0.678
JM1	0.781
MC1	0.198
PC2	0.299

The Main Contributions

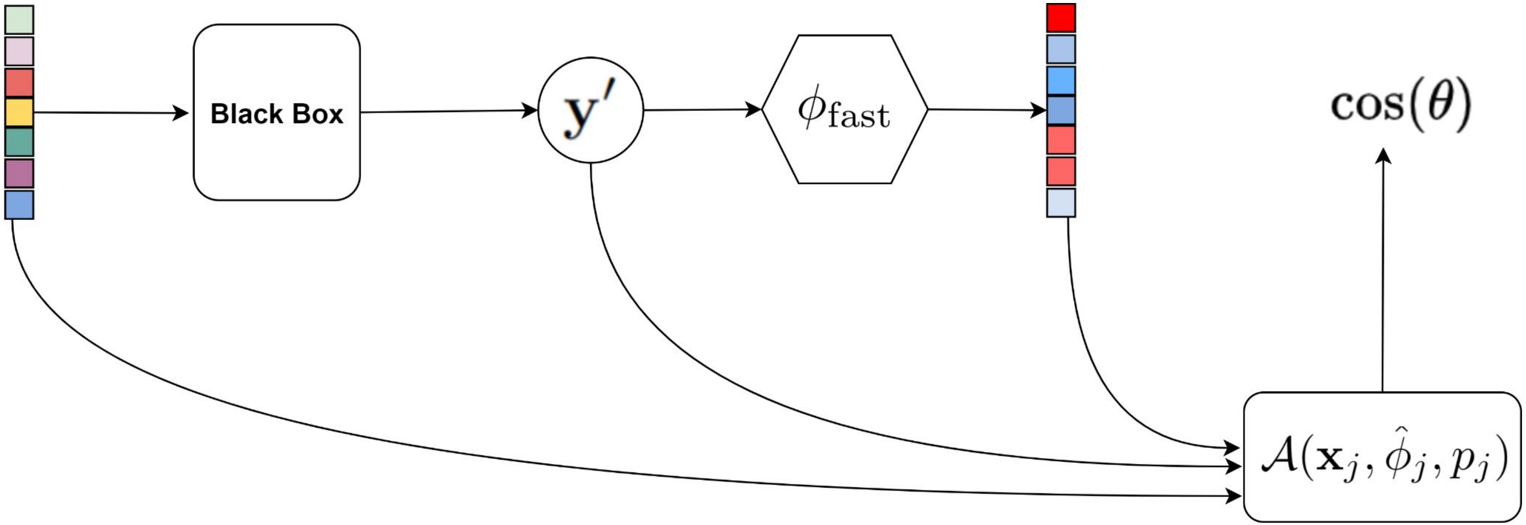
- An approach for quantifying the fidelity of Shapley value approximations accompanied with validity guarantees

- A set of non-conformity measures for the conformal prediction framework

The Proposed Method



The Proposed Method



The Proposed Difficulty Estimation Functions

- Probability of the explanation:

$$\varphi_i = 0.5 - \left| \frac{1}{1 + e^{-(\sum \hat{\phi})}} - 0.5 \right|$$

- Probability difference:

$$\varphi_i = \left| \frac{1}{1 + e^{-(\sum \hat{\phi})}} - \mathcal{B}(x_i; \theta) \right|$$

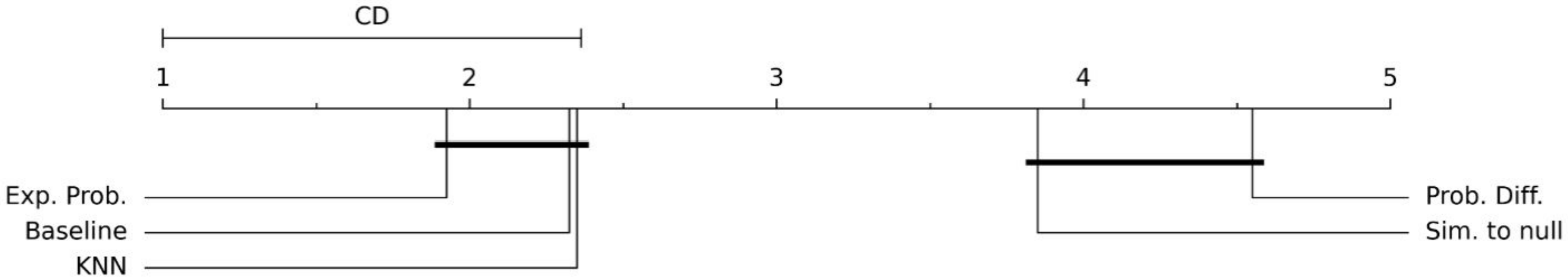
- Similarity to null:

$$\varphi_i = 1 - \left| \frac{\phi^{(null)} \hat{\phi}_i}{\|\phi^{(null)}\| \|\hat{\phi}_i\|} \right|$$

Experimental Setup

- The experiments were conducted on 20 public datasets available on Openml.org
- The data was split into training, development, calibration, and test subsets
  - 60% training, 20% calibration, and 20% test
- The black-box models were generated using the XGBoost algorithm
- The regression models are gradient boosting regressors with 600 estimators

Experimental Results



Experimental Results

Dataset	Baseline	KNN	Prob. Diff.	Exp. Prob.	Sim. to Null
Delta Ailerons	0.146	0.138	0.187	0.143	0.218
Electricity	0.136	0.126	0.186	0.129	0.194
Elevators	0.023	0.021	0.032	0.023	0.03
JM1	0.23	0.188	0.596	0.225	0.24
Heloc	0.314	0.322	0.98	0.309	0.346
MagicTelescope	0.077	0.069	0.107	0.072	0.079

Concluding Remarks

- We proposed an efficient method to estimate the quality of Shapley value approximations while providing validity guarantees using the conformal prediction framework
- We proposed difficulty estimates targeting explanations
- We have presented results from a large-scale empirical evaluation, comparing the proposed difficulty estimates

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.