Proceedings of Machine Learning Research 204:1–3, 2023 Conformal and Probabilistic Prediction with Applications

CONFIDERAI: CONFormal Interpretable-by-Design score function for Explainable and Reliable Artificial Intelligence

Sara Narteni[1,†] SARA.NARTENI@IEIIT.CNR.IT
Alberto Carlevaro[1,†] ALBERTO.CARLEVARO@IEIIT.CNR.IT
Fabrizio Dabbene[1] FABRIZIO.DABBENE@IEIIT.CNR.IT
Marco Muselli[1,2] MARCO.MUSELLI@IEIIT.CNR.IT
Maurizio Mongelli[1] MAURIZIO.MONGELLI@IEIIT.CNR.IT
[1] *CNR-IEIIT, 10129, Turin, Italy*
[2] *Rulex Innovation Labs, Rulex Inc., 16122 Genoa, Italy*
† *S. Narteni and A. Carlevaro contributed equally to the development of the article. (Corresponding authors: S.Narteni, A. Carlevaro.)*

Editor: Harris Papadopoulos, Khuong An Nguyen, Henrik Boström and Lars Carlsson

Abstract

The concept of trustworthiness has been declined in different ways in the field of artificial intelligence, but all its definitions agree on two main pillars: explainability and conformity. In this extended abstract, our aim is to give an idea on how to merge these concepts, by defining a new framework for conformal rule-based predictions. In particular, we introduce a new score function for rule-based models, that leverages on rule relevance and geometrical position of points from rule classification boundaries.
Keywords: XAI, conformal safety sets, novel score function, conformal prediction.

1. Introduction

Literature around combination of eXplainable AI (XAI) and conformal prediction (CP) has recently gained popularity but it still remains little investigated in research. Some relevant approaches proposed so far investigated conformal prediction for XAI models (Bhattacharyya, 2011; Johansson et al., 2014, 2018, 2022), but to the best of our knowledge, no previous study of this type addressed score functions and quantile, tailored for rule-based models.

For this reason, we propose CONFIDERAI, an innovative approach, based on a new score function, to build conformal prediction of rule-based models. The rationale behind the approach is the combination of the global properties of decision rules (i.e., their covering and error) and the geometrical position of the points inside rule boundaries. The resulting prediction set leads to a restricted *conformal safety set*, i.e., the set of points for which the underlying XAI model performs with probabilistic guarantees.

2. CONFIDERAI

Conformal Safety Set. CSS allows to insert CP in a more safety-based context. For for any input feature $\boldsymbol{x} \in \mathcal{X}$ and any label $y \in \mathcal{Y}$, given a *prediction set* at *level of confidence* $1-\varepsilon$, $\varepsilon \in (0,1)$,

$$\mathcal{C}(\boldsymbol{x}) = \{y \mid s(\boldsymbol{x}, y) \leq s_\varepsilon\} \in 2^{\mathcal{Y}}, \tag{1}$$

© 2023 S. Narteni, A. Carlevaro, F. Dabbene, M. Muselli & M. Mongelli.

where s_ε is the $1-\varepsilon$ *quantile* of the score values computed on a *calibration set*, CSS is defined as a subset of the input feature space in which probabilistic safety guarantees can be provided to the machine learning (ML) model:

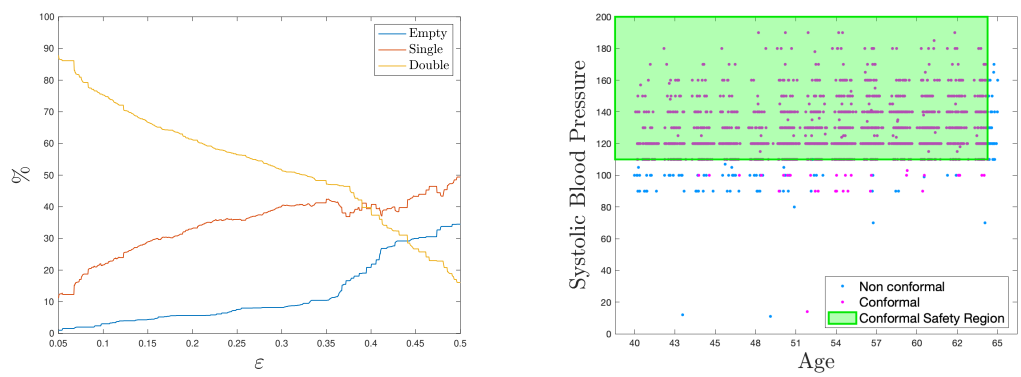
$$\mathcal{S}_\varepsilon = \{\boldsymbol{x} \mid \Pr\{y \in \mathcal{C}(\boldsymbol{x})\} \geq 1-\varepsilon,\ \forall y \in \mathcal{Y}\} = \{\boldsymbol{x} \mid s(\boldsymbol{x},y) \leq s_\varepsilon,\ \forall y \in \mathcal{Y}\}\,. \tag{2}$$

Novel Score Function. An innovative score function suitable to find $\mathcal{S}_\varepsilon$ for rule-based models is designed as follows:

$$s(\boldsymbol{x},y) \doteq \sum_{r_k \in \mathcal{R}^y_{\boldsymbol{x}}} \tau(\boldsymbol{x},r_k)(1-R(r_k)) \quad \text{where} \quad \tau(\boldsymbol{x},r_k) = \frac{1}{1+e^{-1/\gamma}}, \tag{3}$$

with $\gamma = \gamma(\boldsymbol{x},r_k)$ the minimum of the Euclidean distances between the point $\boldsymbol{x}$ covered by the rule r_k and each side of the rule boundary. Values of τ closer to 1 thus encode higher proximity to rule boundary and probability of misclassification. Rule relevance $R(r_k)$ is also accounted Ferrari et al. (2022). The sum is over rules belonging to the set $\mathcal{R}^y_{\boldsymbol{x}}$ of rules covered by $\boldsymbol{x}$ predicting class y.

Preliminary Results. A further classifier is trained to distinguish *conformal* and *non conformal* points, i.e., individuate a *Conformal Safety Region*.



(a) Size of $C(\boldsymbol{x})$ varying ε. (b) Conformal Safety Region

Figure 1: Risk Factors for Cardiovascular Heart Disease (CHD) dataset *.

Acknowledgements

This work was supported in part by REXASI-PRO H-EU project, call HORIZON-CL4-2021-HUMAN-01-01, Grant agreement ID: 101070028. The work was also supported by Future Artificial Intelligence Research (FAIR) project, Recovery and Resilience Plan ("Piano Nazionale di Ripresa e Resilienza"), Spoke 3 - Resilient AI.

*https://www.kaggle.com/datasets/thedevastator/exploring-risk-factors-for-cardiovascular-diseas

References

Siddhartha Bhattacharyya. Confidence in predictions from random tree ensembles. In *2011 IEEE 11th International Conference on Data Mining*, pages 71–80, 2011. doi: 10.1109/ICDM.2011.41.

Enrico Ferrari, Damiano Verda, Nicolò Pinna, and Marco Muselli. A novel rule-based modeling and control approach for the optimization of complex water distribution networks. In *Advances in System-Integrated Intelligence: Proceedings of the 6th International Conference on System-Integrated Intelligence (SysInt 2022), September 7-9, 2022, Genova, Italy*, pages 33–42. Springer, 2022.

Ulf Johansson, Rikard König, Henrik Linusson, Tuve Löfström, and Henrik Boström. Rule extraction with guaranteed fidelity. In *Artificial Intelligence Applications and Innovations: AIAI 2014 Workshops: CoPA, MHDW, IIVC, and MT4BD, Rhodes, Greece, September 19-21, 2014. Proceedings 10*, pages 281–290. Springer, 2014.

Ulf Johansson, Henrik Linusson, Tuve Löfström, and Henrik Boström. Interpretable regression trees using conformal prediction. *Expert systems with applications*, 97:394–404, 2018.

Ulf Johansson, Cecilia Sönströd, Tuwe Löfström, and Henrik Boström. Rule extraction with guarantees from regression models. *Pattern Recognition*, 126:108554, 2022.