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

Sara Narteni¹,†
Alberto Carlevaro¹,†
Fabrizio Dabbene¹
Marco Muselli¹,²
Maurizio Mongelli¹

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 $x \in \mathcal{X}$ and any label $y \in \mathcal{Y}$, given a prediction set at level of confidence $1 - \varepsilon$, $\varepsilon \in (0, 1)$,

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

(1)
where \( s_\varepsilon \) 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:

\[
S_\varepsilon = \{ \mathbf{x} \mid \Pr\{y \in C(\mathbf{x})\} \geq 1 - \varepsilon, \forall y \in \mathcal{Y}\} = \{ \mathbf{x} \mid s(\mathbf{x}, y) \leq s_\varepsilon, \forall y \in \mathcal{Y}\}.
\] (2)

**Novel Score Function.** An innovative score function suitable to find \( S_\varepsilon \) for rule-based models is designed as follows:

\[
s(\mathbf{x}, y) = \sum_{r_k \in R^y_\mathbf{x}} \tau(\mathbf{x}, r_k)(1 - R(r_k)) \quad \text{where} \quad \tau(\mathbf{x}, r_k) = \frac{1}{1 + e^{-1/\gamma}}
\] (3)

with \( \gamma = \gamma(\mathbf{x}, r_k) \) the minimum of the Euclidean distances between the point \( \mathbf{x} \) covered by the rule \( r_k \) and each side of the rule boundary. Values of \( \tau \) 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 \( R^y_\mathbf{x} \) of rules covered by \( \mathbf{x} \) predicting class \( y \).

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

![Figure 1: Risk Factors for Cardiovascular Heart Disease (CHD) dataset](https://www.kaggle.com/datasets/thedevastator/exploring-risk-factors-for-cardiovascular-disease)

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.
CONFIDERAI

References


