L2loRe: a method for explaining the reject option

Clara Punzi[1,*], Roberto Pellungrini[1] and Fosca Giannotti[1]

[1]Scuola Normale Superiore, via Roma, 3, 56126, Pisa, Italy

Abstract

Research on human-computer interaction emphasise the importance of reliability in hybrid decision-making systems. Trust hinges on the performance and trustworthiness of AI, achievable through accuracy metrics, confidence scores, eXplainable AI, and abstention mechanisms. This study presents an explainable abstaining classifier named Learning to Reject via Local Rule-based Explanations (L2loRe), a novel approach that leverages the distance between data points and counterfactuals to evaluate the confidence of predictions, thus facilitating the formulation of a rejection policy and generating clear explanations for the reasoning behind predictions or rejections.

Keywords

Explainable AI, Learning to Reject, Learning to Defer, AI Transparency, AI Reliability

1. Introduction

Research on human-computer interaction emphasizes that a safe and effective utilization of Artificial Intelligence (AI) in decision-making requires human agents to properly rely on AI systems, which in turn is achieved by building the hybrid system for appropriate trust [1]. Key factors influencing trust in automation include performance and transparency, typically conveyed via accuracy metrics, confidence scores, and explanations [2]. The inclusion of an abstention mechanism can further improve the reliability of the system [3]. When the AI system lacks sufficient confidence in its predictions or the impact of errors could be serious, it may be more prudent to refrain from making a prediction and instead direct the input to a more advanced system or a human agent. Several approaches, referred to as "Learning to Abstain" (L2A, [4]), have adopted this mechanism. Nonetheless, limited attention has been directed towards a significant drawback of L2A, namely, the opaqueness of the rejection policy, which may ultimately undermine human trust and satisfaction with the automated system.

This research extends prior work concerning the opaqueness of abstaining algorithms [5, 6, 7, 8]. It presents an explainable abstaining classifier named *Learning to Reject via LOcal Rule-based Explanations* (L2loRe, 1). This novel algorithm incorporates an interpretable abstention mechanism, allowing for the extraction of the rationale behind the decision to abstain. L2loRe draws from the L2R framework, where the rejection policy depends on input features and machine outcomes, while presuming the presence of a downstream human agent. In general, a key element of an L2R algorithm is to identify a suitable confidence metric to quantify the (un)certainty of the prediction. In L2loRe, we define an *interpretable* confidence score described by the distance between data points and their corresponding counterfactual instances. L2lore leverages this distance as a proxy to quantify the certainty level of a prediction and to reject it accordingly.

Overall, the contributions of this work can be summarized as follows: *i)* We present a novel strategy to fine-tune an existing pre-trained classifier with local rule-based explanations; *i)* we propose to use the distance between data points and their corresponding counterfactuals as a confidence metric to define a rejection policy for a given classifier; *i)* we generate human-understandable explanations to enrich the outcome of the classifier in case of rejection.

Discovery Science - Late Breaking Contributions 2024

*Corresponding author.

clara.punzi@sns.it (C. Punzi)

2. Proposed methodology

Problem setting. The goal of L2R algorithms is to learn a model f_ρ that consists of two components, namely a predictor f and a rejection policy ρ. The former is defined as a function $f : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X}$ denotes the feature space and $\mathcal{Y}$ the target space, while the latter is generally defined (at inference time) as $\rho : \mathcal{X} \to \{0, 1\}$. In the case of *dependent* rejectors, ρ depends also on the predictor f through a confidence function $c_f : \mathcal{X} \to \mathbb{R}_+$ and a certain threshold τ. The composed system is then defined by a function $f_\rho : \mathcal{X} \to \mathcal{Y} \bigcup \{\emptyset\}$ such that:

$$f_\rho(x) = \begin{cases} \emptyset & \text{if } \rho(x; c_f, \tau) = 1, \\ f(x) & \text{otherwise.} \end{cases}$$

In other words, given an instance x, if the rejection policy ρ rejects it, then no prediction is made and the instance is directly deferred to a downstream agent (e.g., a human decision-maker); conversely, if ρ accepts x, then the prediction function f is applied to x and the outcome $f(x)$ is observed. In this study, we assume f to be a binary or multiclass classifier. Ideally, ρ should be able to accurately capture the decision boundary of f in order to reject the examples on which f is prone to make mistakes, while accepting those where a correct prediction is more probable. In accordance with the terminology in [4], our proposed method implements a *dependent* and *staged* (i.e., learnt post-hoc with respect to the classifier) abstention policy for the rejection of *ambiguous* samples. L2loRe is *model agnostic* for classification models on tabular data.

Architecture. L2loRe implements the following sequential functions:

1. **Learning a confidence function c_f that quantifies the uncertainty associated with the prediction of a test example by measuring its proximity to its closest counterfactual.** Intuitively, the distance between an instance and its counterfactual might serve as a proxy for the confidence of the original classifier. The first step to learn c_f is the generation of a set of counterfactual rules $C = \{r_c^1, \dots r_c^m\}$ and a set of counterfactual instances $X_c(x) = \{x_c^1, \dots, x_c^m\}$ obtained by applying the counterfactual rule r_c^i to the original instance x to get x_c^i. This step largely relies XAI method LORE$_{sa}$ [9]. Successively, the confidence function $c_f : \mathcal{X} \times \mathcal{X}^m \to \mathbb{R}_+$ is estimated as:
$$c_f(x, X_c(x)) := d\Big(x - \min_{x_c^i \in X_c} x_c^i\Big).$$
2. **Learning the rejection policy.** The rejection policy ρ is formulated in such a way that data points whose distance from their closest counterfactual is less than a specified threshold τ_r are deemed too uncertain to be predicted by the classifier f and are thus rejected. Formally:
$$\rho(x; c_f, \tau) = \begin{cases} 1 & \text{if } c_f(x, X_c(x)) < \tau_r, \\ 0 & \text{otherwise.} \end{cases}$$

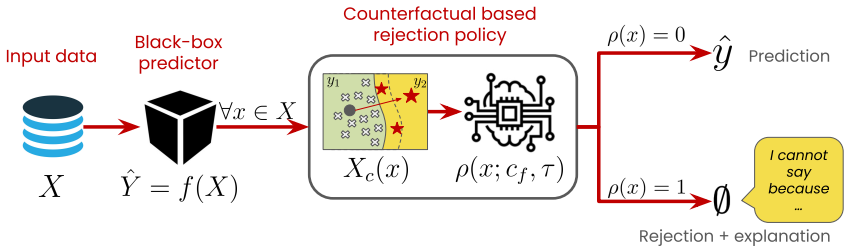
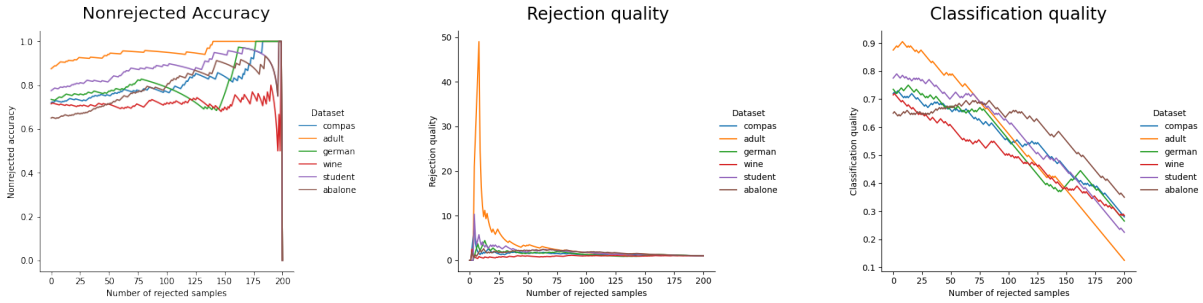


Figure 1: Overview of L2loRe.



(a) Non-rejection accuracy (b) Rejection Quality (c) Classification quality

Figure 2: Classification-rejection performance evaluation as a function of the rejection rate. The metrics are defined in [10] and are computed on 200 test samples.

The rejection threshold τ_r is chosen as the solution of the following constrained optimization problem:

$$\begin{aligned} \text{minimize} \quad & \arg\max_{\tau_r \in [0,1]} \mu(\tau_r) \\ \text{subject to} \quad & r(\tau_r) < r_{\max} && \textit{(rejection rate)} \\ & \tau_r > 1 - \frac{e_{\max}}{1 - \mu(\tau_r)} && \textit{(misprediction error)} \end{aligned}$$

where $\mu : [0,1] \to \mathbb{R}_+$ is a measure defined as a function of the rejection threshold (e.g., one of the three L2R evaluation metrics suggested in [10]), $r_{\max}$ is an upper bound for the rejection rate, and $e_{\max}$ is an upper bound for the proportion of misprediction. In order to determine the optimal value for τ_r, we employed a heuristic approach that involved assessing the performance of μ in correspondence of a number of candidate rejection thresholds (i.e., distance values). These are sampled from a gamma distribution fitted on the vector of confidence scores $\{c_f(x, X_c(x))\}_{x \in X_{train}}$.

3. **Generating explanations.** L2loRe additionally provides a textual explanation for the decision/abstention. This can be achieved by directly exploiting the underlying LORE$_{sa}$ architecture that provides counterfactual explanations regarding the existence of high confident counterfactual points in the very close proximity of the decision boundary of the classifier.

Experiments. In order to perform experiments with L2loRe, we selected three binary (Compas [11], Adult [12], German Credit [13]) and three multiclass (Wine [14], Abalone [15], Student [16]) tabular datasets, most of which contained a combination of categorical and continuous variables. As a first preliminary analysis, we examined how the performance of L2loRe varies as a function of the rejection rate r over the aforementioned datasets. As it can been seen in Figure 2, L2loRe improved the performance of the classification task over all datasets. Secondly, we fine-tuned the rejection threshold τ. Given the the optimal choice of τ for all datasets, we then computed and analyzed the classification-rejection scores, as reported in Table 1. Specifically, we set $r_{\max} = 0.6$ and $e_{\max} = 0.3$ as upper bounds to the rejection and misprediction rate, respectively, and optimized τ with respect to each of the three L2R metrics defined in [10]. Figure 3 displays the explanation provided by L2loRe in the a case of an rejected instanced selected from the Compas dataset.

3. Conclusion and future works

Based on these preliminary findings, the next steps will involve conducting in-depth experiments to assess L2loRe from both quantitative and qualitative perspectives. In addition, we suggest that a potentially promising avenue for future research could be extending L2loRe within the L2D framework. By incorporating a deferral policy that learns from and accounts for human expected performance, this scenario should also explain why a specific human agent or machine was considered more suitable for making a prediction.

```
The instance x = 3680 of the dataset COMPAS has been REJECTED.
In fact, it would have been classified by the black-box as did_recid=1 with 0.75 predicted probability,
but 1 counterfactuals of different classes were found in close proximity:
1) did_recid=0:
        Predicted probabilities:  0: 0.77,  1: 0.23,
        Changes needed:
                - race=African-American <= 0.50 (original value: 1 )
```

Figure 3: Example of explanation provided by L2loRe for a case of rejection.

	compas				adult				german_credit			
	base	rej rate	τ	value	base	rej rate	τ	value	base	rej rate	τ	value
classification quality	0.72	0.015	0.415	0.725	0.875	0.005	0.304	0.880	0.735	0.060	3.580	0.745
nonrejected accuracy	0.72	0.595	3.213	0.827	0.875	0.005	0.304	0.879	0.735	0.340	3717.094	0.803
rejection quality	0	0.025	0.488	3.857	0	0	0.304	0	0	0.060	3.516	3.883

	wine				student				abalone			
	base	rej rate	τ	value	base	rej rate	τ	value	base	rej rate	τ	value
classification quality	0.715	0	0	0.715	0.775	0.005	0	0.78	0.65	0.315	0.096	0.695
nonrejected accuracy	0.715	0.01	0	0.717	0.775	0.455	0.543	0.89	0.65	0.59	0.169	0.854
rejection quality	0	0.01	0	2.509	0	0.035	0.092	4.593	0	0.345	0.102	2.414

Table 1
Evaluation of L2loRe with optimal choice of the rejection threshold τ.

Acknowledgments

This work has been supported by the European Union under Horizon Europe Programme Grant Agreement no. 101120763 (TANGO). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them. This work has been realised also thanks to ERC-2018-ADG GA 834756 (XAI), by HumanE-AI-Net GA 952026, by the Partnership Extended PE00000013 - "FAIR - Future Artificial Intelligence Research" - Spoke 1 "Human-centered AI", by "SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics" (http://www.sobigdata.eu), GA 871042 and by NextGenerationEU - National Recovery and Resilience Plan, PNRR) - Project: "SoBigData.it - Strengthening the Italian RI for Social Mining and Big Data Analytics" - Prot. IR000001 3 - Notice n. 3264 of 12/28/2021.

References

[1] J. D. Lee, K. A. See, Trust in automation: Designing for appropriate reliance, Hum. Factors 46 (2004) 50–80.
[2] J. Zerilli, U. Bhatt, A. Weller, How transparency modulates trust in artificial intelligence, Patterns 3 (2022) 100455.
[3] K. Hendrickx, L. Perini, D. V. der Plas, W. Meert, J. Davis, Machine learning with a reject option: a survey, Mach. Learn. 113 (2024) 3073–3110.
[4] C. Punzi, R. Pellungrini, M. Setzu, F. Giannotti, D. Pedreschi, AI, meet human: Learning paradigms for hybrid decision making systems, CoRR abs/2402.06287 (2024).
[5] A. Artelt, B. Hammer, "even if ..." - diverse semifactual explanations of reject, in: SSCI, IEEE, 2022, pp. 854–859.
[6] A. Artelt, J. Brinkrolf, R. Visser, B. Hammer, Explaining reject options of learning vector quantization classifiers, in: IJCCI, SCITEPRESS, 2022, pp. 249–261.
[7] A. Artelt, R. Visser, B. Hammer, Model agnostic local explanations of reject, in: ESANN, 2022.
[8] S. Singla, N. Murali, F. Arabshahi, S. Triantafyllou, K. Batmanghelich, Augmentation by counterfactual explanation -fixing an overconfident classifier, in: WACV, IEEE, 2023, pp. 4709–4719.

[9] R. Guidotti, A. Monreale, S. Ruggieri, F. Naretto, F. Turini, D. Pedreschi, F. Giannotti, Stable and actionable explanations of black-box models through factual and counterfactual rules, Data Min. Knowl. Discov. 38 (2024) 2825–2862.
[10] F. Condessa, J. M. Bioucas-Dias, J. Kovacevic, Performance measures for classification systems with rejection, Pattern Recognit. 63 (2017) 437–450.
[11] J. Dressel, H. Farid, The accuracy, fairness, and limits of predicting recidivism, Science Advances 4 (2018). doi:10.1126/sciadv.aao5580.
[12] B. Becker, R. Kohavi, Adult, UCI Machine Learning Repository, 1996. URL: https://doi.org/10.24432/C5XW20.
[13] H. Hofmann, Statlog (German Credit Data), UCI Machine Learning Repository, 1994. URL: https://doi.org/10.24432/C5NC77.
[14] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, J. Reis, Wine Quality, UCI Machine Learning Repository, 2009. URL: https://doi.org/10.24432/C56S3T.
[15] W. Nash, T. Sellers, S. Talbot, A. Cawthorn, W. Ford, Abalone, UCI Machine Learning Repository, 1995. URL: https://doi.org/10.24432/C55C7W.
[16] V. Realinho, M. V. Martins, J. Machado, L. M. T. Baptista, Predict students' dropout and academic success, https://doi.org/10.24432/C5MC89, 2021. Accessed on YYYY-MM-DD.