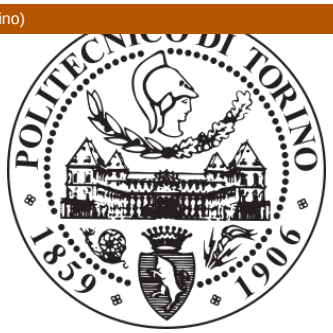

Enhancing Interpretability of Black Box Models by means of Local Rules

Enhancing Interpretability of Black Box Models by means of Local Rules

Eliana Pastor
eliana.pastor@polito.it
Dipartimento di Automatica e Informatica
Politecnico di Torino
Torino, Italy

Elena Baralis
elena.baralis@polito.it
Dipartimento di Automatica e Informatica
Politecnico di Torino
Torino, Italy

ABSTRACT

We propose a novel rule-based method that explains the prediction of any classifier on a specific instance by analyzing the joint effect of feature subsets on the classifier prediction. The relevant subsets are identified by learning a local rule-based model in the neighborhood of the prediction to explain. While local rules give a qualitative insight of the local behavior, their relevance is quantified by using the concept of prediction difference.

CCS CONCEPTS

• **Information systems → Data mining**; • **Computing methodologies → Machine learning**; • **Human-centered computing** → *Human computer interaction (HCI).*

KEYWORDS

Interpretability, Prediction Explanation, Local Model

1 INTRODUCTION AND BACKGROUND

Many high performance machine learning methods produce black box models, which do not disclose their internal logic yielding the prediction. However, in many application domains understanding the motivation of a prediction is becoming a requisite to trust the prediction itself. The demand for transparency comes also from institutions. The European Union approved the GDPR, a regulation for ensuring personal data protection. It states that individuals have the right to receive "meaningful information about the logic involved" in case of automated decision-making. For some authors, this requirement legally mandates a "right to explanation" [2].

We propose LACE (Local Agnostic attribute Contribution Explanation), a novel method to explain classifier predictions on single instances [3]. This methodology is model-agnostic. Hence, it is applicable to any classification method without making any assumption on its internal logic.

2 THE LACE EXPLANATION APPROACH

Being x the instance to be explained, LACE first step is capturing the model local behavior in the neighborhood of the prediction. Firstly, its K neighbors in the training set are computed. Next, the model is exploited to label the selected instances. The K neighbors are used as training data for training a local interpretable model. The local model is L^3 [1], an associative classifier which provides a set of local rules. They capture the labelling behavior of the original black box model in the neighborhood of instance x to be explained. Local rules give a qualitative insight of the reason for a prediction.

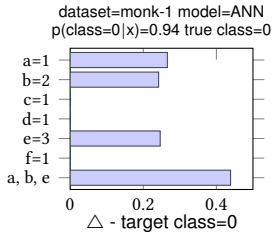

Figure 1: Example of LACE explanation. The bottom term is a relevant subset extracted by local rules.

The relevance of each attribute value and local rule is quantified by using the concept of prediction difference. One or more attribute values at a time are omitted and the prediction change is measured. A change of the prediction probability implies that the considered attributes are relevant to perform the prediction for that particular instance. The omission of sets of attribute values is performed only for relevant subsets of feature values extracted from the local rules. They represent the subsets of attribute values that jointly determine the prediction for instance x. The prediction difference is finally visualized by means of a bar plot representation. Figure 1 shows an example of explanation provided by LACE.

3 EXPLANATION RESULTS

Experiments performed both on artificial and real-world data sets highlighted the ability of the LACE explanation method to capture all relevant attribute subsets that jointly contribute to a single instance prediction. LACE explanations provide both qualitative and quantitative understanding of individual predictions. Qualitative insight of local behavior is captured by local rules. The importance of attribute values and local rules importance is quantified in terms of prediction difference. The model-agnostic nature of our method and the uniform presentation of explanations allow an easy comparison of explanations.

As future work, we will extend our approach to (i) deal with Big data and (ii) devise an automatic adaptation of the number of neighbors K to different data distributions.

ACKNOWLEDGMENTS

This work is partially funded by SmartData@PoliTO.

REFERENCES

[1] Elena Baralis, Silvia Chiusano, and Paolo Garza. 2008. A lazy approach to associative classification. *IEEE TKDE* 20, 2 (2008), 156–171.
[2] Bryce Goodman and Seth Flaxman. 2016. European Union regulations on algorithmic decision-making and a" right to explanation". *arXiv:1606.08813* (2016).
[3] Eliana Pastor and Elena Baralis. 2019. Explaining Black Box Models by means of Local Rules. In *34rd ACM SAC 2019, Limassol, Cyprus, April 8-12, 2019.*

ACM womENcourage 2019, September 16-18, 2019, Rome, Italy