ABSTRACT

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. 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 explanation method is integrated into X-PLAIN, an interactive tool that allows human-in-the-loop inspection of the reasons behind model predictions. Its support for the local analysis of individual predictions enables users to inspect the local behavior of different classifiers and compare the knowledge different classifiers are exploiting for their prediction. The interactive exploration of prediction explanation provides actionable insights for both trusting and validating model predictions and, in case of unexpected behaviors, for debugging and improving the model itself.

BIOGRAPHY

Eliana Pastor received the M.Sc. degree in Computer Engineering from Politecnico di Torino, Italy, in 2017. She is currently a Ph.D. candidate of the SmartData@Polito center. Her main research interests concern machine learning, big data and data mining. Her current research focuses on Explainable AI, Fairness in Machine Learning and predictive maintenance.