Generation and Evaluation of Factual and Counterfactual Explanations for Decision Trees and Fuzzy Rule-based Classifiers

Título Generation and Evaluation of Factual and Counterfactual Explanations for Decision Trees and Fuzzy Rule-based Classifiers

Autores Ilia Stepin, Jose M. Alonso, Alejandro Catala and Martin Pereira

Tipo Comunicación para congreso

Fonte IEEE World Congress on Computational Intelligence, Glasgow (Scotland), IEEE, pp. 1-8 , 2020.

Abstract Data-driven classification algorithms have proven highly effective in a range of complex tasks. However, their output is sometimes questioned, as the reasoning behind it may remain unclear due to a high number of poorly interpretable parameters used when training. Evidence-based (factual) explanations for single classifications answer the question why a particular class is selected in terms of the given observations. On the contrary, counterfactual explanations pay attention to why the rest of classes are not selected. Accordingly, we hypothesize that providing classifiers with a combination of both factual and counterfactual explanations is likely to make them more trustworthy. To test our hypothesis, we introduce a rule-based method to generate factual and counterfactual explanations for the output of pretrained decision trees and fuzzy rule-based classifiers. Experimental results show that unification of factual and counterfactual explanations under the paradigm of fuzzy inference systems proves promising for explaining the reasoning of classification algorithms.

Palabras chave Explainable Artificial Intelligence, Counterfactuals, Decision Trees, Fuzzy Inference Systems, Natural Language Generation

DESCARGAS

Referencia BibTex

PROXECTOS DE INVESTIGACIÓN

eXplica-IA: Diseñando Sistemas Inteligentes Explicables que Interaccionan con Personas y Generan Explicaciones en...

PROGRAMAS CIENTÍFICOS

Tecnoloxías da Linguaxe Natural