Factual and Counterfactual Explanation of Fuzzy Information Granules

Título Factual and Counterfactual Explanation of Fuzzy Information Granules

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

Tipo Capítulo de libro

Fonte Interpretable Artificial Intelligence: A perspective of Granular Computing, Springer-Verlag, 2021.

DOI 10.1007/978-3-030-64949-4_6

Abstract In this chapter, we describe how to generate not only interpretable but also self-explaining fuzzy systems. Such systems are expected to manage information granules naturally as humans do. We take as starting point the Fuzzy Unordered Rule Induction Algorithm (FURIA for short) which produces a good interpretability-accuracy trade-off. FURIA rules have local semantics and manage information granules without linguistic interpretability.With the aim of making FURIA rules self-explaining, we have created a linguistic layer which endows FURIA with global semantics and linguistic interpretability. Explainable FURIA rules provide users with evidence-based (factual) and counterfactual explanations for single classifications. Factual explanations answer the question why a particular class is selected in terms of the given observations. In addition, counterfactual explanations pay attention to why the rest of classes are not selected. Thus, endowing FURIA rules with the capability to generate a combination of both factual and counterfactual explanations is likely to make them more trustworthy. We illustrate how to build self-explaining FURIA classifiers in two practical use cases regarding beer style classification and vehicle classification. Experimental results are encouraging. The generated classifiers exhibit accuracy comparable to a black-box classifier such as Random Forest. Moreover, their explainability is comparable to that provided by white-box classifiers designed with the Highly Interpretable Linguistic Knowledge fuzzy modeling methodology (HILK for short) in terms of explainability.

Palabras chave Interpretable Artificial Intelligence, Granular Computing, Counterfactual Reasoning, Fuzzy Rule-based Classifiers

LIGAZÓNS

Versión da editorial

DESCARGAS

Referencia BibTex

DATOS ADICIONAIS

Datos e software adicionais

PROXECTOS DE INVESTIGACIÓN

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

PROGRAMAS CIENTÍFICOS

Tecnoloxías da Linguaxe Natural