Flexible Counterfactual Explanations with Generative Models

Hellemans Stig[1,2][0000−0001−9441−3882], Algaba Andres[1][0000−0002−0532−3066], Sam Verboven[1][0000−0002−1742−5561], and Ginis Vincent[1,3][0000−0003−0063−9608]

[1] Vrije Universiteit Brussel
[2] Universiteit Antwerpen
[3] Harvard University
stig.hellemans@student.uantwerpen.be andres.algaba@vub.be
sam.verboven@vub.be ginis@seas.harvard.edu

Abstract. Counterfactual explanations are crucial for enabling users to comprehend and engage with machine learning models, particularly in high-stakes domains such as finance and healthcare, where decisions have significant impacts on individuals' lives. However, existing methods often lack the flexibility to accommodate users' unique constraints and preferences. We present a framework for Flexible Counterfactual Explanations, with an implementation using Generative Adversarial Networks (FCE-GAN). Our approach introduces counterfactual templates, allowing users to dynamically specify mutable and immutable features at inference time, to generate flexible counterfactual explanations for any black-box model. For instance, in heart disease risk prediction, patients may be unable or unwilling to change certain factors like age or family history, while being open to modifying diet or exercise habits. Our framework employs a two-stage process: first generating candidate counterfactuals, then selecting those meeting predefined quality measures. We demonstrate our framework's effectiveness on the Adult UCI income and Heart Disease Risk Prediction datasets, showing improved performance in generating diverse, realistic, and actionable counterfactual explanations compared to existing methods. Our approach offers a powerful, generalizable tool for enhancing model interpretability and fairness in critical decision-making systems, with the flexibility to accommodate various generative models beyond GANs.

Keywords: counterfactual explanations · generative models · interpretability · tabular data.