FRI Data Science Project Competition 2024
Machine Learning report

CnC Counterfactual Visualiser

Cristian Bassotto, Camile Lendering, and Nikolay Kormushev

Abstract

Counterfactuals are a crucial tool in explainable AI, offering insights into model decisions by presenting alternative scenarios. While many papers focus on generating counterfactuals, few address visualization. Existing works often feature rudimentary interfaces, rely on specific generators, or lack user validation.
In this paper, we propose a model-agnostic counterfactual UI emphasizing simplicity. We aim for a streamlined user experience, informed by psychology, and validated through a user study. Our goal is to enhance the accessibility and interpretability of counterfactual explanations, advancing explainable AI.

Keywords
XAI, UX, Counterfactuals

Advisors: prof. dr. Erik Štrumbelj

Introduction

Counterfactuals are established as one of the main approaches in explainable AI (XAI). They offer the user insight on how model decisions are made, by providing counterfactual examples of a sample, showing the changes in feature values needed for that sample to be labeled differently.

There are many papers written that focus on the process of generating counterfactuals with a wide variety of approaches [1], but far less papers that focus on the visualisation aspects. Most visualisation methods found in literature present either underdeveloped or overly complex user interfaces, rely on a specific counterfactual generator or do not validate their method with a user study. This research aims to address these shortcomings by developing a user-friendly, model agnostic counterfactual visualization method.

Related works

From the counterfactual visualisation methods we looked at, we really liked the approach of SDA-Vis [2] which is a great visualisation and unlike most other papers they did a simple user study. Still we feel can the UI be a bit convoluted and that it is focused in a specific domain. It does have the potential to generalise but we believe it needs some getting used to and is more tailored somewhat more expert users which is not what our goal of simplicity is.

AdViCE [3] and DECE [4] are two other examples of tools that visualise counterfactuals but are either by design or ended up being not suitable for end users.

Simpler UIs we looked at are ViCE [5] which only works with numeric values and an attempt at an improvement [6] which has a similar interface but supports categorical values. The issue we found with their approach is that it does not scale well if we increase the number of features as the interface will start displaying too much information and this becomes less interpretable. The densities of variables used as coloring in ViCE can also be confusing and hard to understand.

In the visualisation tool proposed by Guyomard et al. [6] counterfactuals need to be generated beforehand and input as a json format which makes the solution counterfactual model-agnostic but can require an expert to use.

An issue with all the UIs we mentioned so far is that they do not take into account psychological aspects. The authors of [7] argue that based on a study users have an easier time interpreting categorical values compared to continuous numbers so it might be a better idea to divide continuous values into bins or ranges..

Purpose

Based on our research in our paper we plan to develop a counterfactual model-agnostic UI with a focus on end-users and with as little expert intervention as possible for initial configuration. We want a streamlined user experience that makes the process of generating and understanding counterfactuals as easy as possible. We will work with tabular data containing both continuous and categorical features and we test and improve on the psychological approach suggested in [7] by using configurable binning and performing a user study to

validate the results.

Methods

In this section we outline our methods for developing and evaluating a user-centered, model-agnostic counterfactual explanation visualisation system for tabular data.

Visualisation

In Figure 1, we have created an initial sketch based on our ideas so far. On the right, the generation procedure and selection of the counterfactual choosen; in the center the comparison between initial instance and final counterfactual prediction and features. Model, dataset and method are selectable in the menu on the left and displayed as title on the top.

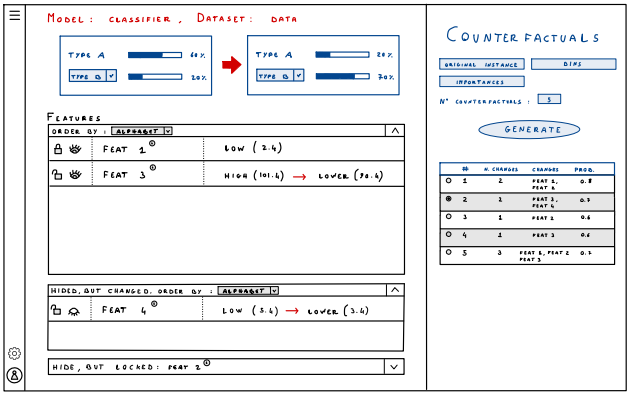

Figure 1. Initial sketch of the final visualization.

Model-agnostic Counterfactual Generation

A Django back-end is implemented to handle counterfactual generation requests. The back-end accepts `GET` requests containing an instance in JSON format and returns corresponding counterfactual examples. The system is model-agnostic, meaning that users can upload their own trained models and datasets for generating counterfactual instances. The only requirement is that the model is differentiable. But, it can be argued that most non-differentiable models (Decision Trees, Rule based systems) are inherently more interpretable.

Model-Agnostic Binning

To make the counterfactual explanations more interpretable, a model-agnostic binning method will be implemented. This transforms continuous feature values into discrete bins, aligning with the findings of Warren et al. [8] that show improved user understanding of counterfactual explanations with categorical features. We will use the local transformation method: CAT-CFLocal as described in [7], where continuous feature values in the counterfactual instance are re-labelled as being "higher" or "lower" relative to the values in the original instance.

User Study

While there exist many numerical evaluation metrics for counterfactual explanations, such as sparsity and proximity [1], these measures are not suitable for evaluating the discretised counterfactual instances. Hence, a user study will be conducted to evaluate the effectiveness our proposed visualization method.

References

[1] Mark T. Keane and Barry Smyth. Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable ai (xai), 2020.

[2] Germain Garcia-Zanabria, Daniel A Gutierrez-Pachas, Guillermo Camara-Chavez, Jorge Poco, and Erick Gomez-Nieto. Sda-vis: A visualization system for student dropout analysis based on counterfactual exploration. *Applied Sciences*, 12(12):5785, 2022.

[3] Oscar Gomez, Steffen Holter, Jun Yuan, and Enrico Bertini. Advice: Aggregated visual counterfactual explanations for machine learning model validation. In *2021 IEEE Visualization Conference (VIS)*, pages 31–35. IEEE, 2021.

[4] Furui Cheng, Yao Ming, and Huamin Qu. Dece: Decision explorer with counterfactual explanations for machine learning models. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1438–1447, 2020.

[5] Oscar Gomez, Steffen Holter, Jun Yuan, and Enrico Bertini. Vice: Visual counterfactual explanations for machine learning models. In *Proceedings of the 25th international conference on intelligent user interfaces*, pages 531–535, 2020.

[6] Victor Guyomard, Françoise Fessant, Thomas Guyet, Tassadit Bouadi, and Alexandre Termier. Interactive visualization of counterfactual explanations for tabular data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 330–334. Springer, 2023.

[7] Greta Warren, Barry Smyth, and Mark T Keane. "better" counterfactuals, ones people can understand: psychologically-plausible case-based counterfactuals using categorical features for explainable ai (xai). In *International conference on case-based reasoning*, pages 63–78. Springer, 2022.

[8] Greta Warren, Ruth M. J. Byrne, and Mark T. Keane. Categorical and continuous features in counterfactual explanations of ai systems. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23, page 171–187, New York, NY, USA, 2023. Association for Computing Machinery.