Multi-Objective Counterfactuals for Counterfactual Fairness in User-Centered AI

Rifat Mehreen Amin
rifat.amin@ifi.lmu.de
LMU Munich
Munich, Bavaria, Germany

ABSTRACT

This position paper emphasizes the role of user-centered artificial intelligence in critical decision-making domains in machine learning models. In this paper, I introduce MOCCF (Multi-Objective Counterfactuals for Counterfactual Fairness) as an extended method that generates realistic counterfactuals by leveraging multiple objectives. Furthermore, to increase transparency, I propose two fairness metrics, Absolute Mean Prediction Difference (AMPD), and Model Biasness Estimation (MBE). I argue that these metrics enable the detection and quantification of unfairness in binary classification models both at the individual and holistic levels consecutively and contribute to user-centered artificial intelligence.

KEYWORDS

machine learning, fairness, counterfactuals, user-centered AI

1 INTRODUCTION

Counterfactuals play a crucial role in user-centered artificial intelligence (AI) by aiding in detecting and mitigating bias in machine learning (ML) models. This assists in ensuring that AI technologies are developed and deployed in a manner that is more equitable and aligned with human values. Counterfactuals are used for model interpretability by identifying input changes that lead to different predictions. However, generating realistic counterfactuals is crucial to avoid unrealistic conclusions [3]. To address this concern, researchers have proposed methods that specifically aim to generate plausible counterfactuals [2, 4].

This work is based on Multi-Objective Counterfactuals (MOC) [2] for counterfactual generation at its core for its unique approach to generating counterfactuals with multiple objectives. As an extension to MOC, MOCCF makes adjustments to one of these objectives in order to ensure the goal of fairness. With this idea of generating multi-objective counterfactuals, MOCCF contributes to the advancement of user-centered AI by addressing fairness concerns in ML models. For detecting unfairness in ML models, I propose two fairness metrics, AMPD and MBE. These metrics are evaluated using the counterfactuals generated by MOCCF. Additionally, I have also compared the performances of different counterfactual generation methods in terms of their number of generated counterfactuals, execution time, and quality of the counterfactuals. The entire process of this benchmark study is implemented in R and is available on GitHub [1]. With this implementation, my experiments answer the following research question:

> RQ: How do the proposed fairness metrics perform in terms of determining and explaining the unfairness of an ML model on an individual and a holistic level?

2 APPROACH: ILLUSTRATION WITH LAW SCHOOL DATASET

In this section, the methodology of MOCCF is explained using the Law School Admission dataset, which contains information about 163 law schools in the US [5], including features such as LSAT scores, undergraduate GPA, expected average grade for the first year, and more. To demonstrate the process, a random forest model is used to predict first-year performance. An instance is randomly selected from the dataset where the student is *black* and has low first-year performance. The model is trained without this instance, and it predicts with a 90% probability that the student will not perform highly in the first year.

The goal is to investigate if changing the student's sensitive attribute, race, from *black* to *white* would alter the prediction. The process involves specifying the instance, the sensitive attribute (race), the desired attribute status (white), and the desired probability interval for the counterfactuals' predicted probabilities. In this case, 230 counterfactuals are generated. These counterfactuals are then evaluated using a second prediction model that predicts the probabilities of the sensitive attribute. Out of the 230 counterfactuals, 44 are classified as *white* with a high probability. Next, the initial model is tested on these counterfactuals to observe if the decision changes from the initial prediction of '*No*' for high first-year performance. The prediction differences between the counterfactuals and the original instance are calculated, and the percentages of predicted classes are determined.

In Figure 1, I present a t-SNE plot that demonstrates the plausibility and effectiveness of the generated counterfactuals in testing bias. The actual instance is represented by the circled green triangle data point, while the big magenta triangles represent the generated counterfactuals. The plot reveals that the majority of the counterfactuals are in close proximity to the actual instance. This visualization supports the selection of plausible counterfactuals, which is crucial in testing model bias.

From the generated counterfactuals, 88.64% predicted a higher probability of high first-year performance when the students were *white*, indicating a potential disparity in the predictions based on race.

Veröffentlicht durch die Gesellschaft für Informatik e.V.
in P. Fröhlich & V. Cobus (Hrsg.):
Mensch und Computer 2023 – Workshopband, 03.-06. September 2023, Rapperswil (SG)

https://doi.org/10.18420/muc2023-mci-ws16-397

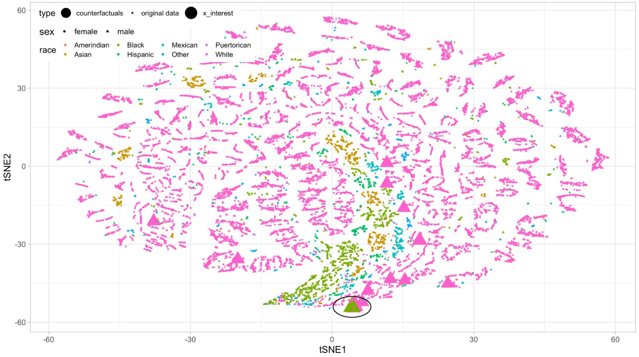

Figure 1: tSNE plot for the combined data (original data, actual instance in circled green triangle, and counterfactuals in magenta triangles) of the Law School Admission Dataset.

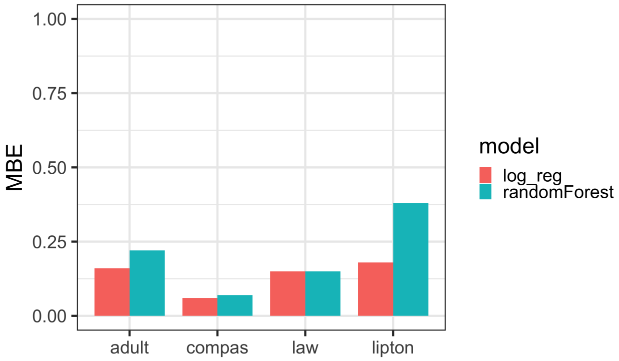

Figure 2: MBE comparison across different ML models and datasets.

3 FAIRNESS METRICS

I use MOCCF to generate plausible counterfactuals to evaluate the fairness metrics. I compare two different ML models (random forest and logistic regression) on four different datasets (COMPAS, Law School Admission, UCI adult, and Lipton hiring). I have compared the performance of the binary classifiers in terms of unfairness and used AMPD and MBE for the unfairness estimation. The motivation is to formulate a more intuitive metric and quantify the extent of the unfairness of a model against a single instance and multiple instances.

AMPD. Absolute Mean Prediction Difference measures the absolute value of how the prediction probabilities $f^*(\mathbf{x}'_i)$ of the generated counterfactuals $\mathbf{x'}_i$ differ from prediction probability $f^*(\mathbf{x}^*)$ of the actual observation $\mathbf{x}^*$ on average.

$$AMPD(\mathbf{x}^*) = \frac{1}{n_{cf}} | \sum_{i=1}^{n_{cf}} f^*(\mathbf{x}^*) - f^*(\mathbf{x}'_i)| \tag{1}$$

Here n_{cf} denotes the number of counterfactuals generated for the actual instance $\mathbf{x}^*$. AMPD provides us with a value between 0 and 1 for a single instance. We can say that if the AMPD of a model for an instance is 0, that means it is very likely that the model is not unfair. Any value greater than 0 implies the presence of unfairness in the model for the specific instance.

MBE. Model Biasness Estimation is calculated by averaging over the *AMPD*s of the n data points in the following way:

$$MBE = \frac{\sum_{i=1}^{n} AMPD(\mathbf{x}_i^*)}{n} \tag{2}$$

Where $\mathbf{x}_i^*$ is i^{th} observation of the dataset. As MBE is derived from AMPD, it, therefore, provides an estimate of the model’s fairness on average for the given dataset, ranging from 0 to 1. A value of 0 signifies no unfairness on a holistic level, while values greater than 0 indicate the presence of unfairness for the given dataset.

In Figure 2, I present the MBE values of the random forest and logistic regression classifiers across four datasets. The random forest classifier exhibits the highest MBE in all datasets except for the Law School Admission dataset, where it shares the same MBE value with logistic regression. This indicates that the random forest is generally unfairer than the logistic regression model in three datasets, as it has higher MBE values.

This comparison of models with respect to their MBE values helps answer the research question of measuring the performance of the proposed metric MBE in terms of model fairness testing. We can see that the higher the value of MBE for an ML model, the greater the chance of a model being unfair. Hence, we can conclude that with a reliable counterfactual generation method, MBE is capable of capturing the unfairness in ML models for an entire dataset.

4 CONCLUSION AND FUTURE WORK

In this paper, I present MOCCF, a fairness testing tool, as a contribution to user-centered AI. MOCCF utilizes multi-objective counterfactuals to detect unfairness in ML models. The introduction of the fairness metrics, AMPD and MBE enables the detection of unfairness at the individual and holistic level, expanding the scope of counterfactual-based fairness testing, which would help users understand the ML models better. Additionally, future improvements include generalizing the testing mechanism for regression tasks. Future work may also involve developing user interfaces for improving the explainability of black-box ML models based on MOCCF.

REFERENCES

[1] 2022. MOC for Counterfactual Fairness. https://github.com/RifatMehreen/MOC-for-CounterfactualFairness.

[2] Susanne Dandl, Christoph Molnar, Martin Binder, and Bernd Bischl. 2020. Multi-Objective Counterfactual Explanations. In *Parallel Problem Solving from Nature – PPSN XVI*. Springer International Publishing, 448–469. https://doi.org/10.1007/978-3-030-58112-1_31

[3] Susanne Dandl, Florian Pfisterer, and Bernd Bischl. 2022. Multi-objective counterfactual fairness. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*. 328–331.

[4] Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. 2020. FACE: Feasible and Actionable Counterfactual Explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (New York, NY, USA) *(AIES ’20)*. Association for Computing Machinery, New York, NY, USA, 344–350. https://doi.org/10.1145/3375627.3375850

[5] Linda F Wightman. 1998. LSAC National Longitudinal Bar Passage Study. LSAC Research Report Series. (1998).