Consequence-aware Sequential Counterfactual Generation

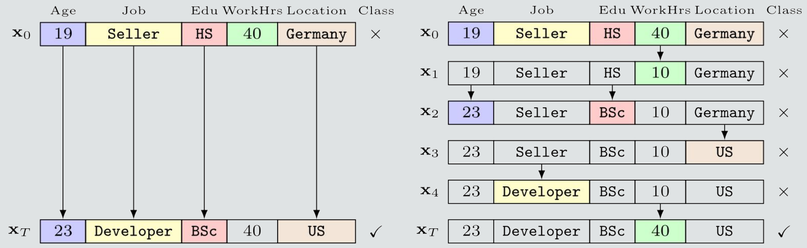
Philip Naumann, Eirini Ntoutsi

ECML PKDD 2021
VIRTUAL
13-17 September

Counterfactuals have become a popular technique nowadays for interacting with black-box machine learning models and understanding how to change a particular instance to obtain a desired outcome from the model. However, most existing approaches assume instant materialization of these changes, ignoring that they may require effort and a specific order of application. Recently, methods have been proposed that also consider the order in which actions are applied, leading to the so-called sequential counterfactual generation problem.
In this work, we propose a model-agnostic method for sequential counterfactual generation. We formulate the task as a multi-objective optimization problem and present a genetic algorithm approach to find optimal sequences of actions leading to the counterfactuals. Our cost model considers not only the direct effect of an action, but also its consequences. Experimental results show that compared to state-of-the-art, our approach generates less costly solutions, is more efficient and provides the user with a diverse set of solutions to choose from.

BACKGROUND

The growing use of machine learning algorithms in sensitive areas such as law, finance, and labor has increased the need for transparent explanations and countermeasures for algorithmic decisions. Counterfactual explanations have emerged as a method to address this problem. These involve making changes (expressed here as actions) to a given input so that the classification would subsequently be in one's favor. Conventional methods consider these changes as instantaneous materializations, ignoring the fact that multiple changes in reality usually require multiple steps (see Fig. 1). Moreover, these steps may not be independent, but interrelated. An action taken now may have consequences for future actions. This problem setting leads to the so-called sequential counterfactual generation problem, which is the focus of this paper.



(a) Traditional counterfactual (b) Sequential counterfactual

Fig. 1: Difference of applying changes in traditional vs. sequential counterfactuals.

APPROACH

In our work, we propose a novel, model-agnostic, method for generating sequential counterfactuals. Our method finds the actions and their tweaking values to achieve the counterfactual by treating the problem as a sequential process that considers possible consequences which are also modeled in the objective space. In this way, we realize more meaningful sequences that provide the user with additional information about the order in which changes should ideally be made. In addition, our objective space formulation allows us to provide diverse sequential counterfactual options for accomplishing the class change. We consider three optimization sub-problems for this task:

1. Find an **optimal subset of actions**
2. Find **optimal tweaking values** for the actions
3. Find the **optimal order** to apply the actions

These can be encoded into a genotype representation and efficiently solved at once by using a *Biased Random-Key Genetic Algorithm (BRKGA)* (see Fig. 2).

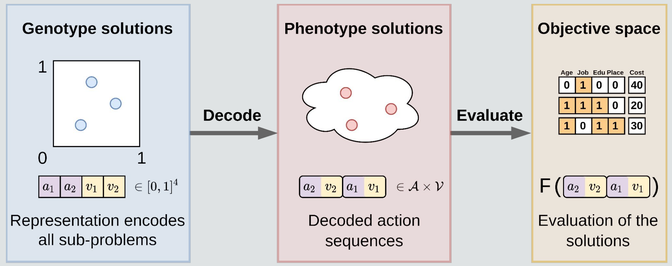


Fig. 2: Steps of the proposed method and different solution representations.

MAIN IDEAS

1. Multiple options: Proposing multiple solution sequences gives the user more freedom of choice and an overview of the options available to them. We achieve this by formulating the problem as a multi-objective one. The main idea for covering multiple options is to treat the changes in each feature as a separate objective. In addition, there is the cost of the sequence and the Gower's distance to the original input instance. Applying non-dominated sorting to this objective space (see Fig. 3) then leads to solutions that change a different subset of features while still respecting the objective with the lowest cost. (For simplicity, the Gower's distance objective is left out in Fig. 3).

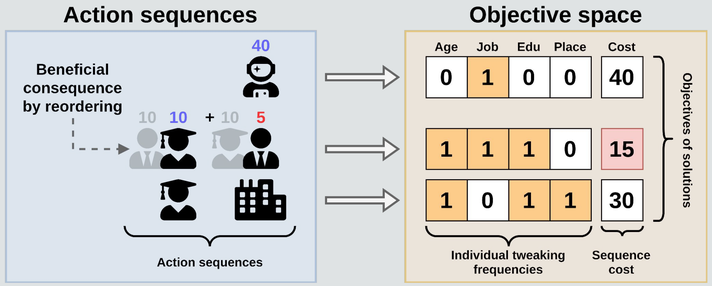


Fig. 3: Consequence-aware, diversity-enforcing, objective space formulation.

2. Consequence-awareness: Actions are interrelated and can affect future ones. We model these relationships in a feature relationship graph (see Fig. 4), where the edges contain the relationship functions used to discount the cost of an action. In this way, we obtain more meaningful sequences that take into account the (beneficial) consequences. Moreover, sequences can be found that, in the right order, can lead to an improvement in costs that would not have been found otherwise.

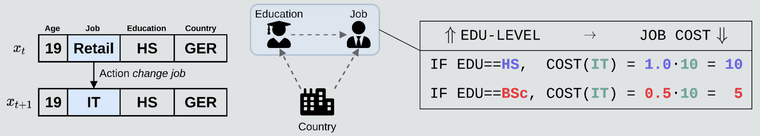


Fig. 4: Derivation of a consequential discount through feature relationships.

RESULTS

- More efficient than SOTA[1] and finds sequences of any length.
- Find multiple sequences compared to only a single one of SOTA.
- Sequence costs are on par with SOTA.
- Considering the consequences produces more meaningful sequence orderings.

[1]Ramakrishnan, Goutham, Yun Chan Lee, and Aws Albarghouthi. "**Synthesizing action sequences for modifying model decisions.**" *Proceedings of the AAAI Conference on Artificial Intelligence.* Vol. 34. No. 04. 2020.

Freie Universität Berlin

Philip Naumann **philip.naumann@fu-berlin.de** **@ PhilipNaumann.com** **Artificial Intelligence and Machine Learning group** at **Freie Universität Berlin**