The Impact of Using Constraints on Counterfactual Explanations

Maciej Falbogowski, Jerzy Stefanowski, Zuzanna Trafas, and Adam Wojciechowski

Poznan University of Technology, Poznań, Poland
jstefanowski@cs.put.poznan.pl

Abstract. Constraints of attributes in counterfactual explanations of machine learning models are studied. The experiments showed that adding them leads to small, acceptable decreases of the evaluation measures while ensuring the correctness of the explanations.

1 Introduction

In the last decade, intensive development of the *explainable artificial intelligence* (XAI) has been observed. It supports explanations of the predictions of machine learning models, i.e. answering the questions of why the model made such a decision for the given instance [3].

In our work we focus on *counterfactual explanations* (briefly counterfactuals). Unlike other explanation methods, they provide recommendations on how to change the description of an example to achieve the desired prediction. For instance, consider a case where someone applies for a loan and gets rejected by the system. A counterfactual explanation of this decision provides the information on minimal changes of attributes' values that will alter the decision of the system to accept this loan application instead of rejecting it. This kind of explanation is appreciated by many researchers and users, as counterfactuals are quite intuitive and may indicate to people what to do in order to achieve the desired outcome [1]. The reader is referred to [5] for a recent and comprehensive review of algorithms used for generating counterfactual explanations.

However, some limitations of these methods and their implementations could be observed. While generating a counterfactual example, most algorithms minimize certain loss functions, which may lead to changes in attribute values that are unrealistic or unacceptable for a given problem, e.g. a recommendation to change their gender or race or to reduce their age. Therefore, we share the postulates of defining additional constraints on such attributes that should be taken into account by algorithms when searching for the best explanation.

We have implemented a specialized library including new algorithms used for generating counterfactual explanations with such constraints. The aim of this paper is to experimentally evaluate the impact of using these constraints on measures of the quality of explanations for benchmark tabular datasets.

2 Counterfactual Methods and Constraints

Our library allows defining several constraints on changes of attribute values[1]. In the experiment, we explore three of the most practical ones: *freeze* – prohibits any change of the value of a sensitive attribute (e.g., gender or race of a person); *monotonicity* – indicates the preferred direction of the attribute value change (e.g., only an increase in age is allowed), *one-hot encoding* – in case of nominal attributes, changes of encoding to arbitrary real numbers are prevented and a change should preserve an appropriate zero-one encoding.

Implementations of two methods, CADEX and FIMAP, were chosen for the experiments as they were extended to address the limitations defined above. We briefly describe them below. Let $f(x) = \hat{y}$ where f is the model to be explained, x is a specific input example and $\hat{y}$ is the output of the model. Both methods aim to find a counterfactual explanation x' for which $f(x') = y'$, where $\hat{y} \neq y'$ and the attribute description of x' is as similar as possible to x.

Constrained Adversarial Examples (CADEX) [4] is a method looking for examples that change the model prediction with a minimal perturbation on attribute values. To find the best perturbations it uses the gradient descent, calculating the loss between the actual output $\hat{y}$ and the desired one y'. Following the gradient with the optimizer allows to gradually update the values of input attributes towards the decision boundary. This method is naturally suited to handle constraints by using a special mask vector for selecting attributes and limiting the direction of change.

The main idea of the Feature Importance by Minimal Adversarial Perturbation (FIMAP) method [2] is to generate minimal perturbations of the example x by a neural network g (i.e. the counterfactual of x is $x + g(x)$). As the training of g requires a differentiable model f, FIMAP uses the auxiliary neural network s which approximates predictions of the model f. The main network g and its parameters are trained in such a way that $s(x + g(x))$ should change the prediction of the model. It is achieved by optimizing the objective loss function with regularization terms by gradient descent.

3 Experiments

The constraints introduced in the previous section ensure the creation of more valid and reliable counterfactuals. For example, running CADEX for credit data without these constraints for a certain application will result in a proposal to change the nominal attribute (*Other debtors/guarantors*) from a code of 0 (*none* in one-hot encoding) to a real value (-0.459) that has no interpretation; whereas applying the constraints will result in realistic changes to these attributes.

Nevertheless the open research question is: how much the use of these constraints affects the quality of generated counterfactuals. To test this we will use four measures. For the sake of brevity, we name them by the following Greek letters:

[1] Consult https://github.com/LoGosX/CFEC for more details

- $\alpha - metric$: The distance between an example and its counterfactual [2]. The minimum difference in distance between the original example and the explanation is desired.
- $\beta - metric$: The distance between counterfactuals originated from an example and its nearest neighbor. Intuitively closest pairs of original instances should yield close pairs of counterfactuals.
- $\gamma - metric$: The fraction of attributes that were changed during the counterfactual generation process. Due to using neural networks, attributes are compared with the indiscernibility threshold of 0.1. The values of this metric are between 0 and 1. As simplicity (scarcity) of the explanation is desired, the lower value is better.
- $\delta - metric$: The binary test, which inspects if the model's decision differs between prediction for a chosen example and its counterfactual. $True = 1$ value is desired.

The experiments were conducted using two datasets from the UCI ML repository – German Credit Data (Statlog) and Adult – as they were often used in related studies and contain different types of attributes. The results are presented in Table 1.

Dataset	**Method**	**Constraints**	**Metric**			
			$\boldsymbol{\alpha}$	$\boldsymbol{\beta}$	$\boldsymbol{\gamma}$	$\boldsymbol{\delta}$
Adult	FIMAP	None	0,0201	0,0029	0,6892	0,9600
		Only One-hot	0,0930	0,0754	0,2630	0,8000
		One-hot + Monotonicity	0,1648	0,0740	0,2585	0,8750
		One-hot + Freeze	0,1724	0,0719	0,2526	0,9600
	CADEX	None	0,0180	0,0042	0,1284	1,0000
		Only One-hot	0,0497	0,0035	0,1297	1,0000
		One-hot + Monotonicity	0,0517	0,0023	0,1313	1,0000
		One-hot + Freeze	0,0495	0,0021	0,1313	1,0000
Statlog	FIMAP	None	0,0197	0,0388	0,7087	0,9300
		Only One-hot	0,1079	0,0705	0,4144	0,9400
		One-hot + Monotonicity	0,0825	0,0709	0,3921	0,4050
		One-hot + Freeze	0,0877	0,0709	0,4197	0,6750
	CADEX	None	0,0206	0,0389	0,0803	1,0000
		Only One-hot	0,0397	0,0362	0,0978	1,0000
		One-hot + Monotonicity	0,0432	0,0292	0,1188	1,0000
		One-hot + Freeze	0,0367	0,0302	0,1026	1,0000

Table 1. Evaluation metrics calculated as a mean for randomly selected 100 examples for each dataset and different constraint definitions. Monotonicity constraints were defined as increasing on *age* column for both datasets. Freeze constraints were defined on *native.country*, *sex* and *race* columns for Adult and *credit* for Statlog.

[2] Distances are calculated with HOEM metric to handle heterogeneous types of attributes. The original values of attributes are always normalized.

4 Results and Conclusions

The experimental results show that adding constraints decreases values of some metrics, but the differences are not big and they depend mainly on the method – they are negligible for CADEX while a bit more visible for FIMAP. There is no strong dependence on the type of constraint. Adding either monotonic direction or freezing an attribute led to nearly the same values of the measures in the case of CADEX. Below each metric is discussed in more detail.

The values of the α metric are generally smaller for CADEX. Defining the constraints only slightly affects the results of this method, but more strongly affects the results of FIMAP. In our opinion, it results from the sampling mechanism for discrete attributes.

The values of β metric are virtually the same for CADEX, and even slightly better with the constraints defined. For FIMAP, the values are worse but no dependence on the type of constraint is seen.

In the case of γ metric CADEX changes less attributes than FIMAP, as this method (in the default version) allows changing at most 5 attributes. It is particularly visible when constraints are not defined.

The δ metric values for CADEX are always equal to 1, since the method has access to the model and can check if its decision has been changed. In contrast, FIMAP does not have access to the original model, so it makes no such guarantee. The decrease of this metric is stronger for German data when adding constraints.
Summary: The presented experiments show that adding constraints to the methods considered does not deteriorate the explicitly considered metrics, regardless of the type of defined constraints. Declines in value are acceptable. The CADEX method achieves better evaluation values than FIMAP and, in particular, always leads to a change of the model decision into the desired value.
Ack: The research of J. Stefanowski was partially supported by the TAILOR project (EU Horizon 2020 No 952215) and 0311/SBAD/0726, while A. Wojciechowski's by AI Tech project (POPC.03.02.00-00-0001/20).

References

1. Artelt, A., Hammer, B.: Efficient computation of counterfactual explanations and counterfactual metrics of prototype-based classifiers. Neurocomputing **470**, 304–317 (2022)
2. Chapman-Rounds, M., Bhatt, U., Pazos, E., Schulz, M., Georgatzis, K.: FIMAP: feature importance by minimal adversarial perturbation. In: Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021. pp. 11433–11441 (2021)
3. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM computing surveys (CSUR) **51**(5), 1–42 (2018)
4. Moore, J., Hammerla, N., Watkins, C.: Explaining deep learning models with constrained adversarial examples. In: PRICAI 2019: Trends in Artificial Intelligence - 16th Pacific Rim Int. Conf. on Artificial Intelligence. pp. 43–56 (2019)
5. Verma, S., Dickerson, J., Hines, K.: Counterfactual explanations for machine learning: A review. arXiv: abs/2010.10596 (2020)