

A MULTI-CRITERIA APPROACH FOR SELECTING AN EXPLANATION FROM THE SET OF COUNTERFACTUALS PRODUCED BY AN ENSEMBLE OF EXPLAINERS

IGNACY STEPKA ^a, MATEUSZ LANGO ^a, JERZY STEFANOWSKI ^{a,*}

^a Institute of Computing Sciences
Poznan University of Technology
ul. Piotrowo 2, 60-965 Poznań, Poland
e-mail: ignacy.stepka@put.poznan.pl,
{mateusz.lango, jerzy.stefanowski}@cs.put.poznan.pl

Counterfactuals are widely used to explain ML model predictions by providing alternative scenarios for obtaining more desired predictions. They can be generated by a variety of methods that optimize various, sometimes conflicting, quality measures and produce quite different solutions. However, choosing the most appropriate explanation method and one of the generated counterfactuals is not an easy task. Instead of forcing the user to test many different explanation methods and analysing conflicting solutions, in this paper we propose to use a multi-stage ensemble approach that will select a single counterfactual based on the multiple-criteria analysis. It offers a compromise solution that scores well on several popular quality measures. This approach exploits the dominance relation and the ideal point decision aid method, which selects one counterfactual from the Pareto front. The conducted experiments demonstrate that the proposed approach generates fully actionable counterfactuals with attractive compromise values of the quality measures considered.

Keywords: counterfactual explanations, ensemble of explainers, ideal point method, multiple criteria analysis, explainable artificial intelligence.

1. Introduction

Despite incredible progress in machine learning (ML), wide adoption of its algorithms, especially in critical domains such as finance or medicine, often encounters obstacles related to the lack of their interpretability. This is due to the fact that the majority of currently used machine learning methods are black-box models that do not provide information about the reasons behind taking a certain decision, nor do they explain the logic of an algorithm leading to it. Therefore, there is a growing research interest in explainable artificial intelligence methods (Bodria *et al.*, 2021) offering explanations for the predictions of black-box models.

Counterfactual explanations (briefly *counterfactuals*) are one particular type of such explanations that provide information about how feature values of an example should be changed to obtain a desired prediction of the model (change its decision). On

the one hand, by interacting with the model using counterfactuals, the user can better understand how the system works by exploring “what would have happened if ...” questions. This approach to building human understanding of machine learning models has some psychological justifications (Miller, 2019). On the other hand, a good counterfactual provides a clear recommendation to the user about what changes are needed in order to achieve the desired outcome.

There are many practical applications for counterfactual explanations, including loan decisions (Wachter *et al.*, 2017), recruitment processes (Pearl *et al.*, 2016), the discovery of chemical compounds with similar structures but different properties (Wellawatte *et al.*, 2022), analysis of medical diagnosis results (Mertes *et al.*, 2022), and many others (see, e.g., the recent survey by Guidotti (2022)). For example, consider a scenario where an individual submits a purchase offer for a property, initially rejected by a model assessing such proposals. A

*Corresponding author

counterfactual explanation for this situation reveals the minimal adjustments or enhancements the offeror could implement to ensure the acceptance of their offer (e.g., increase the offered price, relax contingencies). Another practical scenario involves a company training a neural network to assist in the recruitment process for a specific job position, automating the shortlisting of resumes. In this context, counterfactual explanations can be employed to verify that the black-box model does not discriminate against candidates who only differ in terms of a sensitive and non-actionable feature. Furthermore, for a candidate facing rejection, a counterfactual explanation can provide valuable insights into the specific qualifications that were lacking compared with the most similar candidates who were shortlisted.

A counterfactual explanation is expected to be similar to the example that the ML model was queried with, but it should change a class prediction. This leads to a situation where one instance can be explained by many different counterfactuals. This has resulted in the introduction of several desired properties, optimization strategies and *quality measures* that a counterfactual explanation should possess. Such properties include:¹ *proximity* (a counterfactual should be as similar as possible to a given instance), *sparsity* (the number of modified features should be low), *actionability* (the counterfactual should not modify immutable features, such as race, or violate monotonic constraints, e.g., decrease one's age), *discriminative power* (the generated counterfactual examples should be in the region of the feature space dominated by the expected class), and others. Even though some of these measures are at least related to each other (e.g., *proximity* and *sparsity*), many of them are not aligned and even contradictory. For example, *proximity* encourages the generation of a counterfactual that is as close to the decision boundary as possible, whereas the *discriminative power* favours explanations in dense areas dominated by the other class, which are most likely to be far from the decision boundary. Related user studies (Spreitzer et al., 2022; Förster et al., 2021) show that human users prefer counterfactuals which, on the average, score well on the various criteria.

Generating an appropriate counterfactual explanation is, therefore, quite a challenging task that involves finding a trade-off between divergent aspects of explanation quality. Nevertheless, the vast majority of counterfactual generation methods optimize only one or two measures, usually aggregated by a weighted sum in the optimized loss function, where the weight is a hyperparameter of the method (see, e.g., Wachter et al., 2017; Chapman-Rounds et al., 2021; Mothilal et al., 2020; Van Looveren and Klaise, 2021). The

choice of a satisfactory weight value, which controls the trade-off between different aspects of explanation quality, is a non-trivial task. The few methods (Rasouli and Chieh Yu, 2022; Dandl et al., 2020) that consider multiple quality criteria in the process of obtaining counterfactual explanations generate a large set of explanations, but leave the task of selecting a final one to the user. We argue that it is not enough to present too many explanations to the final user due to the choice overload phenomenon (Iyengar and Lepper, 2000; Stefanowski, 2023). Iyengar and Lepper (2000) or Inbar et al. (2011) show that presenting a limited number of alternatives is superior to posing too many alternatives, when it comes to the human ability to analyze these options and make optimal choices. In order to mitigate these issues, we postulate needs for developing approaches that would reduce the number of the proposed counterfactuals and support the user in choosing the most compromise solution.

In this paper, instead of presenting yet another method for generating counterfactuals, we claim that the already existing methods should be sufficient to provide a diversified set of explanations. However, the more challenging problem of selecting a suitable, compromise counterfactual while taking multiple quality criteria into account is still open. Inspired by the research on *classifier ensembles*, which achieve better classification performance by exploiting the predictions from a diversified set of base classifiers (Kuncheva, 2004), we propose to use an *ensemble of multiple base explainers* to provide a richer set of counterfactuals, each of which establishes a certain trade-off between values of different quality measures (often referred to as criteria). We also put forward an approach to significantly reduce the number of explanations considered to a smaller and concise set by constructing a Pareto front, i.e., a subset of explanations that are no worse than others with respect to at least one criterion. Then we propose to select a final counterfactual from this front by applying the multiple criteria ideal point method (Steuer, 1986; Skulimowski, 1990), which does not require additional preference elicitation from the user and is computationally efficient. To sum up, the main contributions of this paper are as follows:

1. Proposal of a new approach, integrating an ensemble of explainers with the multiple criteria analysis to select a counterfactual representing a suitable trade-off between the quality measures.
2. Experimental evaluation of the proposed approach, demonstrating that it provides the best trade-off between different quality measures for a wide range of user preferences for these criteria.
3. Multi-criteria analysis of counterfactuals generated by various methods, which provides insights into the dominance relationship between them.

¹ See Section 2 for more details and precise definitions.

2. Related works

Following Wachter *et al.* (2017), a counterfactual explanation is defined as a perturbation of the instance x , denoted as x' , that results in a different prediction from the same black box model b , i.e., $b(x) \neq b(x')$.

2.1. Counterfactual explanation methods.

Numerous methods for generating counterfactual explanations have already been introduced. According to Guidotti (2022), they can be divided into four categories based on their methodological paradigms. Instance-based explainers select the most similar examples with a desired class from the dataset (e.g., FACE (Poyiadzi *et al.*, 2020)). Decision tree approaches approximate the behaviour of a black box model with a decision tree and exploit its structure to generate counterfactual explanations (e.g., Tolomei *et al.*, 2017). The next approaches optimize a certain loss function by adopting specific optimization algorithms (e.g., Wachter (Wachter *et al.*, 2017), CEM (Dhurandhar *et al.*, 2018), Dice (Mothilal *et al.*, 2020), Fimap (Chapman-Rounds *et al.*, 2021), CFProto (Van Looveren and Klaise, 2021), ActionableRecourse (Ustun *et al.*, 2019)). Heuristic search strategies find counterfactuals through iterative heuristic choices minimizing the chosen cost function (e.g., Cadex (Moore *et al.*, 2019), GrowingSpheres (Laugel *et al.*, 2018)). This taxonomy is only one of many; however, it exhibits different approaches to the process of obtaining counterfactual explanations. Another thorough review with an alternative categorization is provided by Verma *et al.* (2020).

Most of the papers introducing counterfactual generation methods do not report experiments that compare them with previous methods on data benchmarks. In general, the field of counterfactual explanations suffers from a lack of comparative studies examining multiple methods, and the work by Guidotti (2022) is the rare exception.

2.2. Most related methods for generating several counterfactual explanations. So far, only one paper has considered combining multiple methods for the same data, namely, Guidotti and Ruggieri (2021) has studied the committee of counterfactual explainers. In their proposal base explainers may produce several explanations which are, then, combined (selected up to the required number) by optimizing a simple two criteria distance-driven aggregation function.

The multi-criteria approach to generating a set of counterfactuals has only been explored in a few papers (Dandl *et al.*, 2020; Rasouli and Chieh Yu, 2022). They utilize genetic algorithms to generate a large (tens to hundreds of explanations) set of non-dominated counterfactual explanations based on multiple criteria.

Table 1. Quality measures of counterfactual explanations: x' is the counterfactual for instance $x \in X$, $neigh_k$ is the k -th nearest neighbour of x' in X .

Measure	Definition
$proximity(x, x')$	$distance(x, x')$
$feasibility(x')$	$\frac{1}{k} \sum_{i=1}^k distance(x', neigh_i)$
$sparsity(x, x')$	number of features changed in x to get x'
$discriminative_power(x')$	$\frac{1}{k} \sum_{i=1}^k \mathbf{1}[[b(x') = b(neigh_i)]]$
$instability(x')$	$dist(x', x'_1)$ where x'_1 is a counterfactual obtained for example $x_1 \in X$ being the closest neighbour of x

However, these approaches do not tackle the problem of reducing the choice overload, and the selection of the best counterfactual explanation is left to the final user or decision-maker. The needs for adapting the multi-criteria decision analysis to support such users is also postulated by Stefanowski (2023).

2.3. Quality measures for counterfactuals. The research on counterfactual explanations and their psychological analysis show that they should fulfil some expectations that a human decision-maker might have. Here we list the properties most frequently mentioned in the literature (Guidotti, 2022; Wachter *et al.*, 2017): (i) *validity*: a counterfactual x' has to change the prediction, i.e., $b(x) \neq b(x')$; (ii) *sparsity*: a valid counterfactual should change as few features as possible; (iii) *proximity*: a counterfactual should be a result of the smallest perturbation, i.e., x' should be as similar to original x as possible. It is also postulated that counterfactuals need to be (iv) *distributionally faithful*, i.e., they should be located in feature space regions that ensure their *feasibility/plausibility* (as some generation methods may produce examples of out-of-data distributions or with unrealistic feature values). In order to take into account fairness and real world utility, many works also introduce (v) *actionability*: a counterfactual should not alter any attributes from x that are sensitive and immutable in certain scenarios (e.g., changing race in a loan application setting). Other proposed properties are discussed by Guidotti (2022).

The above-mentioned properties have led to defining different quality measures evaluating either a single counterfactual or a set of counterfactuals. For our further experiments, we choose the most commonly used measures in the literature; see their specific definitions in Table 1.

In the rest of the paper we use the following notation: x is the original instance, classified by the

black box model b ; x' is the generated counterfactual that corresponds to x ; *distance* stands for the distance between two examples calculated with a chosen metric; $neigh_k$ is the k -th closest neighbour to the instance x in the training data X .

3. Proposed multi-criteria approach

We propose a new multi-stage approach that integrates an *ensemble of different methods for generating counterfactuals* with a *multi-criteria approach* for selecting the counterfactual that provides a compromise solution with respect to conflicting criteria. It consists of four consecutive steps illustrated in Fig. 1. Firstly, each explainer included in the ensemble is queried to generate counterfactuals for a given instance x (Section 3.1). Next, all explanations are combined to form a set of candidate solutions. This resulting set is filtered to remove invalid and non-actionable instances (Section 3.2). In the third step (Section 3.3), we employ the dominance relation to reduce the set of remaining explanations without loss of quality for any criterion considered. As the final step, we use the ideal point method to select the best solution (Section 3.4). The pseudocode of the proposed approach can be found in Algorithm 1. To ease the comprehension of the proposed approach, we present an illustrative example in Section 3.5.

3.1. Constructing an ensemble of explainers. In order to obtain a set of diversified counterfactuals, we construct an ensemble of different methods chosen under the following premises: they are based on different paradigms and thus generate quite diverse explanations, they have positive literature recommendations, and their stable open source implementations are available.

In Section 4.1, we list the specific methods used in our ensemble, although we argue that our approach is general enough to employ different sets of base explainers. While some methods may generate few solutions and others produce a single counterfactual, experiments have shown that none of the examined methods is superior to others with respect to any of the criteria considered (see Stepka et al., 2023). This supports our hypothesis that a set of solutions obtained by using different explanation methods provides a broader and richer set of possible counterfactuals than searching for a single best method.

3.2. Enforcement of validity and actionability constraints. Some of the chosen base explainers may generate counterfactuals which do not change the black-box prediction to the desired value, and also some generated solutions may violate actionability constraints. Therefore, in this step we perform an additional filtering of counterfactuals generated by the ensemble, i.e.,

Algorithm 1. Pseudocode for the proposed approach.

```

 $b \leftarrow$  black-box classifier
 $X \leftarrow$  training data
 $x \leftarrow$  query instance s.t.  $x \in X$ 
 $E \leftarrow$  set of base explainers

STEP 1 (Ensemble of explainers):
 $C \leftarrow \emptyset$ 
for  $explainer \in E$  do
     $c \leftarrow explainer(x, X, b)$  {run the base explainer}
     $C \leftarrow C \cup \{c\}$ 
end for

STEP 2 (Enforcement of validity and actionability):
for  $c$  in  $C$  do
    if  $\neg is\_valid(c) \vee \neg is\_actionable(c)$  then
         $C \leftarrow C \setminus \{c\}$ 
    end if
end for

STEP 3 (Filtering dominated solution):
 $D \leftarrow \emptyset$  {empty set of dominated solutions}
for  $c$  in  $C$  do
    for  $d \in (C \setminus \{c\})$  do
        if  $d$  is better than  $c$  on one criterion and better or
        equal on all criteria then
             $D \leftarrow D \cup c$ 
        end if
    end for
end for
 $ND \leftarrow C \setminus D$  {set of non-dominated solutions}

STEP 4 (Selection with ideal point)
 $p \leftarrow vector()$ 
for each criterion  $i$  do
     $p_i \leftarrow \max_{c \in ND} c_i$ 
end for
 $s \leftarrow \arg \min_{c \in ND} distance(c, p)$ 

return  $s$ 

```

we discard explanations which are not *valid* or *non-actionable*. *Actionability* is examined with respect to restricted attributes that have to be distinctly pre-defined for any given dataset.

3.3. Filtering of dominated solutions. The ensemble of base explainers produces a large number of counterfactuals characterized by different and sometimes conflicting quality criteria (see the discussion of *proximity* and *discriminative power* in Section 1). The problem of selecting the best solution is challenging, since the objective comparison of two explanations that excel

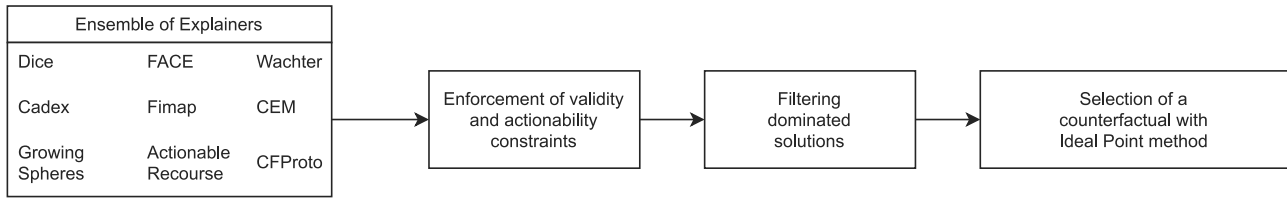


Fig. 1. Visualization of the algorithm steps in our approach.

on different criteria is impossible and largely depends on the preferences of the user/decision-maker. Such problems are of interest to the field of multi-criteria decision-aid (MCDA) (Ehrgott, 2005), which deals with sets of alternative solutions/decisions x evaluated by many criteria $g_i(x)$. In brief, for each criterion, the direction of the user's preference is defined as increasing (gain criteria) or decreasing (cost criteria). The values of criteria for contemplated alternatives may be contradictory, but some of them are more preferred than others because of the preference directions. MCDA methods can solve the trade-off between them.

This leads us to exploiting the *dominance relation* (Ehrgott, 2005), which is defined as follows. Let us assume the gain direction of preferences for all criteria $g(x)$ and consider two alternatives: counterfactuals x' and y' . We say that x' *dominates* y' if for each criterion i holds $g_i(x') \geq g_i(y')$ and exists at least one i for which $g_i(x') > g_i(y')$. In other words, criteria values of x' are better than or equal to the corresponding values of y' . The dominated counterfactuals can be removed from the set of solutions as they are objectively worse than the *non-dominated* alternatives.

Then, our approach constructs a *Pareto front*, i.e., it builds a set of all non-dominated counterfactuals obtained by applying the dominance relation on all explanations in a given set. In other words, each counterfactual explanation is examined to see whether there exists any other counterfactual that has better or equal scores in relation to all criteria considered. Only if such examples do not exist is the counterfactual non-dominated and therefore included in the *Pareto front*. Our experiments show that, thanks to exploiting this relation, the size of the candidate solution set can be significantly reduced, on the average by approximately 80% (see the experiments in Section 4.2).

3.4. Selection of a counterfactual with the ideal point method. As the final set of non-dominated alternatives on the Pareto front may still be too large for a user to be analysed manually, it is necessary to support the choice of the counterfactual that represents the trade-off best suited to the user. There exist many MCDA methods that allow this by acquiring the global model of user preferences for these criteria by interacting with them (Ehrgott, 2005). However, we will follow a simpler approach which is, in

our opinion, more suited for our problem and better for carrying out our automatic experiments.

Note that the criteria that characterize different counterfactuals are quality measures (e.g., *proximity*). A typical decision situation considered in MCDA assumes that the user/decision maker is able to compare the alternatives based on the criteria values (i.e., in the feature space). However, in our case the quality measures of counterfactuals are not so intuitive for non-expert humans to interpret and compare.

This also limits the possibility of using typical interactive MCDA methods to elicit the decision maker's preferences, which require accurate assessments about the relative importance of criteria or comparing different variants usually in the original features.

Therefore, we propose to use the *ideal point method*, which is a simple and computationally-efficient approach recommended in the literature for the case of equally important criteria (Branke *et al.*, 2008; Skulimowski, 1990). We briefly explain this method below:

- Let D be a set of non-dominated counterfactuals in the Pareto front with c criteria. The ideal solution is artificially created in the criteria space by selecting the best possible value for every criterion. Assuming that all criteria $g_i(x)$ have an increasing direction of preference (gain), the ideal point $z = [z_1, z_2, \dots, z_c]$ can be formally defined as having $z_i = \max\{g_i(x) : x \in D\}$ for all $i = 1, \dots, c$. Usually z is an abstract point which does not belong to D .
- Then, for each counterfactual $x \in D$, its distance measure to the ideal point z is computed.
- The closest counterfactual to z is selected as the final best solution.

The ideal point method originates from the works on multi-objective mathematical programming, proposed in the previous century (Steuer, 1986).² It has been further extended for different scalarized distances with criteria weights or to so-called reference points (see the discussions by Ehrgott (2005) and Skulimowski (1990)). Nevertheless, in our paper we want to show that even the

²It can be easily extended to use other ways of calculating distances with the hyper-plane connecting the ideal point with the anti-ideal/nadir point (Ehrgott and Tenfelde-Podehl, 2003) (which we will consider in Section 4.4).

Table 2. Characteristics of datasets. From left: *size*—the total number of instances, *test size*—the size of the holdout test set, *continuous/categorical*—the number of continuous/categorical features, *immutable*—the names of features which were designated as non-actionable.

Dataset	Size	Test size	Continuous	Categorical	Immutable
Adult	32561	250	5	6	race, sex, native-country
German	1000	100	7	13	foreign-worker
Compas	7214	250	7	3	age, sex, race, charge-degree
Fico	10459	250	23	0	external-risk-estimate

simplest version is sufficient to demonstrate the usefulness of the MCDA approach in the proposal of the ensemble of explainers.

3.5. Approach walkthrough with a toy example.

In order to help understand the proposed method, we will go step by step through the computations for the following instance taken from the Adult³ dataset: $x = \{\text{age: } 24, \text{education.num: } 10, \text{capital.gain: } 0, \text{capital.loss: } 0, \text{hours.per.week: } 30, \text{workclass: Self-emp-not-inc, marital.status: Never-married, occupation: Prof-specialty, race: Asian-Pac-Islander, sex: Male, native.country: United-States, income} > 50K\}$.

First, we run all the explainers from the ensemble to construct counterfactuals. Assuming the set of methods used in our experiments (see the experimental setup in Section 4.1), 82 counterfactuals are returned. We provide the full list of returned counterfactuals and further details in the online appendix.⁴

We then feed these counterfactuals into the next step of our approach, which focuses on enforcing validity and actionability. The validity filter eliminates the counterfactuals that do not change the predicted class from $> 50K$ to $\leq 50K$, reducing the candidate set by 5. Subsequently, actionability enforcement excludes all the alternatives that alter the values of non-mutable features, in this case: *race*, *sex*, and *native.country*. This actionability test removes 18 additional counterfactuals, reducing the total number of candidates to 59. Nine of the removed counterfactuals suggested changing the person's native country from the United States to other countries such as France, the Philippines, or China. All but one of the non-actionable counterfactuals suggested changing the person's race, and eight of them suggested changing both race and country of birth. This demonstrates the need for actionability and validity testing.

In the next step, the multi-criteria analysis of the remaining 59 counterfactuals is already in progress. Initially, we apply the dominance relation to reduce the candidate set to the Pareto front, which consists of 13 alternatives, discarding 46 dominated alternatives. This

means that we remove all counterfactuals with criteria values objectively worse in regard to all criteria than a set of other counterfactuals. We visualize this process in Fig. 2, where dominated alternatives are denoted with squares, and alternatives forming the Pareto front are marked with dots.

Subsequently, we calculate the coordinates of the ideal point (marked with an 'x' in the figure) from the Pareto front. The coordinates of the ideal point correspond to the best criteria values obtained by any of the alternatives considered. In this case, these values are 0.04 for *proximity*, 0.11 for *feasibility*, and 1 for *discriminative power*.

The final step involves calculating the distance between the alternatives from the Pareto front and the ideal point, and selecting the closest one (indicated by the dotted red line) as the final counterfactual. The distance is calculated using the Euclidean metric, preceded by feature min-max normalization. In the analysed example, the selected counterfactual is $c = \{\text{age: } 24.0, \text{education.num: } 10.0, \text{capital.gain: } 17327.0, \text{capital.loss: } 0.0, \text{hours.per.week: } 30.0, \text{workclass: Self-emp-not-inc, marital.status: Never-married, occupation: Prof-specialty, race: Asian-Pac-Islander, sex: Male, native.country: United-States}\}$, and it has scores of 0.173 for *proximity*, 0.962 for *feasibility*, and 0.11 for *discriminative power*.

4. Experimental evaluation

The aims of the experiments are to assess the utility of the proposed approach and to compare the quality of counterfactuals obtained by different methods.

Firstly, we analyse the Pareto front and utility of different steps in our approach. To investigate if constructing an ensemble is justified, we examine the impact that each step has on the size of a set of candidate counterfactual explanations, as well as the contribution of each of the methods incorporated into the ensemble (Section 4.2).

Then, we analyse the results obtained by different explanation methods and our ensemble, by comparing them on different quality metrics (Section 4.3).

We also study the impact of individual components of our approach on the final result (Section 4.4).

³See Section 4.1 for dataset details.

⁴<https://www.cs.put.poznan.pl/mlango/publications/amcs24.pdf>.

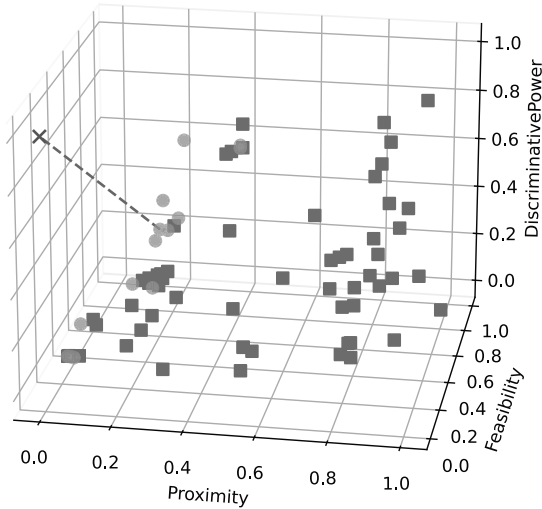


Fig. 2. Example demonstrating the application of dominance relations and the ideal point method. Note that the optimization directions for *proximity* and *feasibility* are min, and for *discriminative power*—max.

Finally, using a utility function model of user preferences, we verify whether the returned counterfactuals represent useful trade-offs between quality criteria considered (Section 4.5).

In order to ensure reproducibility and allow future benchmarking, we publicly release the code used for the experiments.⁵

4.1. Experimental setup. Experiments were conducted on four widely adopted datasets in the related literature, namely: Adult, German, Compas and Fico, the characteristics of which can be found in Table 2. For every dataset, we defined a set of immutable attributes which were later used to evaluate the *actionability* of provided explanations. The information about immutability of certain attributes was passed as an additional input to all methods capable of handling it. Note that these subsets could be different depending on the perspective of a particular stakeholder.

As a black-box classification method, we employed an artificial neural network consisting of two hidden layers with 16-128 neurons in each layer. The number of neurons was optimized for each dataset separately to maximize the accuracy on the validation set. We used the ReLU activation and a dropout between layers. More details of its topology and hyperparameters are provided in the electronic appendix.

To compute the measures described in Section 2.3, it is necessary to select a distance function. We chose the

HEOM distance (Wilson and Martinez, 1997) due to its ability to handle both nominal and continuous variables than other measures.

The evaluated ensemble uses the collection of nine popular and diversified counterfactual generation algorithms as base explainers, namely, Wachter (Wachter *et al.*, 2017), CEM (Dhurandhar *et al.*, 2018), Dice (Mothilal *et al.*, 2020), Fimap (Chapman-Rounds *et al.*, 2021), Cadex (Moore *et al.*, 2019), FACE (Poyiadzi *et al.*, 2020), CFProto (Van Looveren and Klaise, 2021), ActionableRecourse (Ustun *et al.*, 2019) and GrowingSpheres (Laugel *et al.*, 2018). In addition to diversity, we were also guided by their popularity in related works and the availability of their implementations. We employed the following open source libraries: CARLA (Pawelczyk *et al.*, 2021), ALIBI (Klaise *et al.*, 2021), CFEC (Falbowski *et al.*, 2022). The methods were mostly used with default parameters in their implementations; however, to construct a more extensive set of possible explanations, we exploited the possibilities of generating multiple counterfactuals from these methods.

To obtain several counterfactuals from some of the methods, we apply different strategies; because only Dice natively supports generating a set of explanations (in our case $k = 20$).

The first adopted strategy involves random sampling of counterfactuals discovered during the optimization process, but not selected as the final solution by the method. Therefore, it not only gathers the explanation that the method ultimately selects, but also incorporates randomly sampled counterfactuals discovered during optimization that may perform better on some criteria that the method does not directly optimize. We apply this strategy to CFProto ($k = 10$) and Wachter ($k = 10$).

The second strategy is to restart the method with a different set of hyperparameters to slightly change the optimization process and obtain different explanation. We use this strategy for GrowingSpheres ($k = 20$ restarts with a different random seeds), FACE ($k = 10$ restarts with different random seeds), Cadex ($k = 15$ different numbers of features to change ranging from 1 to 14), and Fimap ($k = 7$ different combinations of parameters for the objective function⁶).

While selecting the final counterfactual with our approach, we analysed three criteria: *proximity*, *feasibility* and *discriminative power*. Recall that in MCDA criteria must form a coherent family of diverse views on the problem. Indeed, these three criteria were relatively poorly correlated in our preliminary experiments (Stepka *et al.*, 2023). As a distance function for the ideal point

⁵<https://github.com/istepka/MCSECE>.

⁶The following combinations of (Gumbel-softmax temperature τ , L1 regularization, L2 regularization) were applied: (0.1, 0.001, 0.01), (0.1, 0.05, 0.5), (0.2, 0.01, 0.1), (0.2, 0.08, 0.8), (0.5, 0.001, 0.01), (0.5, 0.01, 0.5).

Table 3. Comparison of the explainers used in the ensemble according to their average number of generated counterfactuals per instance per dataset. The columns, listed from left to right, indicate the number of counterfactuals left after applying consecutive steps from our approach: *all* (no filters applied), *val* (after applying the validity requirement), *act* (after applying both the validity and actionability requirements), *front* (after exploiting the dominance relation), and *ideal* (after using the ideal point method for selection). The best results are bolded, the second best are underlined, and the third best are in italics.

Dataset	Adult					German				
	all	val	act	front	ideal	all	val	act	front	ideal
Dice	20.00	20.00	20.00	0.52	<i>0.16</i>	20.00	20.00	20.00	0.94	<u>0.33</u>
FACE	10.00	9.68	8.16	3.48	0.69	10.00	9.94	0.37	4.30	0.01
Cadex	6.62	6.47	6.47	<i>1.18</i>	<u>0.34</u>	13.99	11.50	11.50	<u>2.64</u>	0.28
Fimap	6.00	4.70	4.70	0.37	0.07	6.00	5.02	3.04	0.57	0.09
Wachter	9.60	5.03	5.03	<u>1.54</u>	0.02	10.00	7.15	7.15	1.40	0.14
CEM	1.00	0.66	0.66	0.26	0.00	1.00	1.00	1.00	0.09	0.43
CFProto	7.63	7.57	0.78	0.06	0.01	5.63	3.98	2.25	<i>1.48</i>	0.08
GrowingSpheres	20.00	15.54	15.54	0.14	0.01	20.00	10.54	10.54	1.00	0.02
ActionableRecourse	0.40	0.10	0.10	0.20	0.01	0.00	0.00	0.00	0.00	0.00

Dataset	Compas					Fico				
	all	val	act	front	ideal	all	val	act	front	ideal
Dice	20.00	20.00	20.00	1.83	0.35	20.00	20.00	20.00	<u>2.77</u>	0.33
FACE	10.00	9.93	0.45	0.42	0.14	10.00	9.94	0.37	0.37	0.01
Cadex	6.00	4.78	4.74	1.08	<i>0.19</i>	13.99	11.50	11.50	3.72	<u>0.28</u>
Fimap	6.00	5.58	3.30	<i>1.26</i>	<u>0.33</u>	6.00	5.02	3.04	1.13	<i>0.09</i>
Wachter	10.00	6.07	6.07	1.03	0.03	10.00	7.15	7.15	<i>1.65</i>	0.14
CEM	1.00	1.00	1.00	0.33	0.07	1.00	1.00	1.00	0.30	0.04
CFProto	5.75	3.65	0.73	0.13	0.00	5.63	3.98	2.25	0.95	0.08
GrowingSpheres	20.00	8.80	8.80	<u>1.73</u>	0.06	20.00	10.54	10.54	0.71	0.02
ActionableRecourse	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

method, we considered the Manhattan distance (L_1), Euclidean distance (L_2) and Chebyshev distance (L_∞) (following, e.g., Branke *et al.*, 2008; Skulimowski, 1990).

4.2. Analysis of the Pareto front of counterfactuals generated by the base explainers in the ensemble. In the first experiment, we take a closer look at the degree of the base explainers' contribution to the Pareto front. For each dataset and base explainer method, we computed the average number of counterfactuals generated for an instance (all), the number of valid counterfactuals, i.e., those that change the model's prediction (val), the number of actionable and valid counterfactuals (act), the number of non-dominated, valid, actionable counterfactuals on the Pareto front (front), and finally the percentage of cases for which the counterfactual generated by a given method was selected as the final answer (ideal). The results are reported in Table 3 and some additional data visualizations are provided in the online appendix.

Results show that there is no single method that is superior to the others considered in the experiments. Some of the methods clearly contribute much more frequently to the Pareto front, but there is no method

whose contribution is negligible on all examined datasets. Contribution to the Pareto front can be treated as a good indicator of the performance of the method across different measures, as it shows that counterfactuals generated from some methods dominate others on all quality measures. Building on that we justify the utility of incorporating all these methods into the ensemble, as all of them produce best explanations, although with differing frequencies.

Further analysis indicates that all steps of our approach eliminate many counterfactuals. First, enforcing *validity* constraints reduces the number of candidates considered by 17%. Second, examining *actionability* eliminates 16% of examples from the previous step. Third, the use of the dominance relation reduces the remaining set by another 83%. Finally, the ideal point method selects one compromise counterfactual from the remaining set of explanations.

Employing consecutive steps of our approach, with an exclusion of the last one, it shrinks the original set of explanations, without the loss of quality, on the average by 88%.

Table 4. Results obtained for the German dataset. The best results are marked in bold, the second best are underlined, and the third best are in italics.

Method	prox ↓	feas ↓	dpow ↑	spars ↓	instab ↓	cover ↑	act ↑	rank ↓
Dice	1.69	3.92	0.44	<u>1.93</u>	4.15	1.00	1.00	3.14
FACE	5.05	1.91	0.60	8.12	3.82	1.00	0.98	3.14
Cadex	<i>1.38</i>	3.74	0.41	2.64	3.87	0.97	0.97	5.43
Fimap	6.85	3.01	0.60	9.91	3.71	0.97	0.97	4.57
Wachter	11.67	7.29	0.64	14.65	5.91	0.37	0.37	5.14
CEM	0.62	4.18	0.31	<i>2.15</i>	3.99	0.13	0.13	7.00
CFProto	3.56	4.40	0.48	4.79	4.53	0.99	0.91	5.29
GrowingSpheres	7.65	5.79	0.60	10.73	5.42	1.00	1.00	2.71
ActionableRecourse	<u>1.01</u>	3.55	0.44	1.39	<u>3.60</u>	0.23	0.23	6.29
Random selection	4.39	3.95	0.50	6.28	4.61	1.00	0.98	6.86
Our approach (Manhattan)	3.83	<u>2.15</u>	0.85	6.06	3.50	1.00	1.00	1.86
Our approach (Euclidean)	3.21	<i>2.46</i>	<u>0.80</u>	4.99	<i>3.68</i>	1.00	1.00	<u>2.00</u>
Our approach (Chebyshev)	2.90	2.70	<i>0.74</i>	4.38	3.71	1.00	1.00	<i>2.14</i>

Table 5. Results obtained for the Adult dataset.

Method	prox ↓	feas ↓	dpow ↑	spars ↓	instab ↓	cover ↑	act ↑	rank ↓
Dice	1.03	0.77	0.37	<u>1.65</u>	1.13	1.00	1.00	3.00
FACE	0.98	0.90	0.36	<i>1.92</i>	1.10	0.10	0.10	3.43
Cadex	<u>0.20</u>	0.30	0.17	2.27	0.65	0.99	0.99	4.86
Fimap	2.12	0.35	0.59	5.75	1.17	0.99	0.99	4.00
Wachter	4.36	1.23	0.84	7.76	3.07	0.52	0.20	5.29
CEM	0.13	0.32	0.17	1.16	0.67	0.66	0.66	6.14
CFProto	1.45	1.13	0.25	7.00	2.97	0.98	0.06	6.14
GrowingSpheres	2.87	1.39	0.47	6.17	1.70	0.99	0.99	4.14
ActionableRecourse	1.14	0.10	0.70	3.72	0.56	1.00	0.84	6.57
Random selection	1.55	0.81	0.38	3.56	1.17	1.00	0.88	3.86
Our approach (Manhattan)	1.02	<u>0.17</u>	0.94	3.34	<u>0.63</u>	1.00	1.00	1.86
Our approach (Euclidean)	0.99	<i>0.21</i>	<u>0.93</u>	3.22	<u>0.63</u>	1.00	1.00	<u>2.00</u>
Our approach (Chebyshev)	<i>0.96</i>	0.26	<i>0.89</i>	2.97	0.66	1.00	1.00	<i>2.14</i>

4.3. Evaluating counterfactuals with different quality measures. In the second experiment, we compare the performance of the proposed approach with other methods for generating counterfactuals. The quality of the generated counterfactual explanations is evaluated using a wide spectrum of measures employed in the related works: *proximity* (prox), *feasibility* (feas), *discriminative power* (dpow), *sparsity* (spars), *instability* (instab), and *actionability* (act), all of which are discussed in Section 2.3. Additionally, to further assess the reliability of the methods in finding counterfactuals, we report the *coverage* (cov) metric, which represents the ratio of instances for which a counterfactual is found. The results of the experiments for the German, Adult, Compas, Fico datasets can be found in Tables 4, 5, 6 and 7, respectively.

Among the methods under consideration, the proposed approach was the only one that was always able to generate an actionable counterfactual for all instances of tested datasets. Looking at the remaining five quality aspects (other than *actionability* and *coverage*),

the proposed approach rarely obtains the highest score, however, for the vast majority of cases it arrives at one of the top three results. This was expected since our method looks for a trade-off between multiple quality measures; therefore it is not surprising that in the individual ranking for each measure it may not obtain superior results.

We also ranked the data for each quality measure and computed the average rank for each method (the lower the better), and the results can be observed in the last column of Tables 4–7. Taking all the metrics into account, the proposed approach always achieves the best (lowest) rank for all datasets under study. The lowest rank for Adult, German and Compas datasets is achieved by the variant of our approach employing the Manhattan distance. Only for the Fico dataset is the order of three best methods different. Nevertheless, for all datasets included in the experiments, variants of our approach take consistently the best three ranks, meaning that they offer the best compromise between all the seven quality measures considered.

Table 6. Results obtained for the Compas dataset.

Method	prox ↓	feas ↓	dpow ↑	spars ↓	instab ↓	cover ↑	act ↑	rank ↓
Dice	0.93	0.80	0.34	<i>1.71</i>	1.33	1.00	1.00	3.00
FACE	0.53	0.04	0.69	3.55	0.15	1.00	0.07	4.14
Cadex	<u>0.29</u>	0.23	0.15	2.72	0.36	0.98	0.97	5.43
Fimap	0.52	<u>0.11</u>	0.69	3.53	<i>0.17</i>	0.99	0.58	5.00
Wachter	0.69	<i>0.12</i>	0.73	2.92	0.36	0.67	0.67	5.00
CEM	0.33	0.32	0.29	<u>1.57</u>	0.38	1.00	1.00	3.14
CFProto	0.91	0.21	0.28	3.17	0.41	0.54	0.13	6.29
GrowingSpheres	<u>0.29</u>	0.19	0.18	3.09	0.32	0.80	0.80	5.57
ActionableRecourse	0.07	0.32	0.89	1.00	0.66	0.00	0.00	5.29
Random selection	0.63	0.37	0.40	2.74	0.61	1.00	0.70	3.86
Our approach (Manhattan)	0.54	<i>0.12</i>	<u>0.87</u>	2.40	<u>0.28</u>	1.00	1.00	2.00
Our approach (Euclidean)	0.55	0.14	<i>0.86</i>	2.45	<u>0.28</u>	1.00	1.00	<u>2.14</u>
Our approach (Chebyshev)	0.55	0.17	0.84	2.44	0.30	1.00	1.00	2.29

Table 7. Results obtained for the Fico dataset. Note that the ActionableRecourse method is missing, as it failed to provide any valid counterfactual.

Method	prox ↓	feas ↓	dpow ↑	spars ↓	instab ↓	cover ↑	act ↑	rank ↓
Dice	1.11	2.15	0.38	1.88	2.53	1.00	1.00	3.00
FACE	2.34	0.82	<u>0.70</u>	17.88	<u>1.78</u>	1.00	0.03	3.57
Cadex	0.92	1.72	0.38	7.66	2.04	1.00	1.00	3.14
Fimap	1.55	1.78	0.62	16.01	<i>1.87</i>	0.98	0.62	4.71
Wachter	6.66	3.50	0.81	18.39	6.45	0.49	0.49	4.71
CEM	1.12	2.08	0.50	<u>5.80</u>	2.47	1.00	1.00	2.71
CFProto	0.82	<u>1.53</u>	0.35	10.82	1.76	0.58	0.39	6.29
GrowingSpheres	1.44	1.90	0.35	16.48	2.39	0.97	0.97	5.29
Random selection	1.30	1.84	0.40	9.67	2.21	1.00	0.84	3.86
Our approach (Manhattan)	<u>0.87</u>	<u>1.51</u>	0.60	7.33	1.89	1.00	1.00	2.57
Our approach (Euclidean)	<i>0.90</i>	1.59	<i>0.63</i>	7.48	1.93	1.00	1.00	2.14
Our approach (Chebyshev)	0.99	1.66	<i>0.63</i>	<i>7.11</i>	2.03	1.00	1.00	2.14

As a form of ensemble baseline, we also report the results of the ensemble that includes all the base explainers of the proposed method but chooses the final counterfactual at random (random selection) without employing our algorithm. This ensemble was never better than any of the three tested variants of our approach for any dataset and any quality measure, demonstrating that the performance of our ensemble is a result of choosing an appropriate final counterfactual with the multi-criteria analysis (i.e., applying the dominance relation and ideal point method selection) rather than just using an ensemble of different methods.

Regarding the comparison of various variants of our method employing different distance measures, the discrepancies between the obtained scores are usually very small (except for *sparsity* in regard to the German dataset, where the difference between the extreme scores is 1.68). The variant employing the Manhattan distance obtains a slightly lower average rank for most datasets, hence we choose this variant to represent our method in the next experiment.

4.4. Analyzing the impact of the elements in our approach. In this section, we provide a concise analysis of how the components of our approach influence the selection of the final counterfactuals and their corresponding evaluation measures. The experimental data used for this analysis is available in the online appendix.

The first element in our approach emphasizes validation and actionability enforcement. Our experiments demonstrate the critical role this element plays in achieving fully actionable and valid counterfactuals. Without it, the scores for the *validity* and *actionability* criteria range from 24% to 54% and 78% to 98%, respectively (varies by dataset). Such performance falls short of suitability for most real-world scenarios, proving the usefulness of this operation.

The second element in our approach involves applying the dominance relation to filter out dominated alternatives. Omitting this component does not lead to a decrease in the values of various quality measures. This is because, in our approach, the ideal point

selects the counterfactual closest to the $(0,0,0)$ point in the *proximity-feasibility* – $(1 - \text{discriminative power})$ ⁷ min-max normalized criteria space, which is attained through min-max normalization performed on the dataset. Consequently, the dominance relation do not enhance the performance of the ideal point method. Nevertheless, removing dominated, i.e., worse, solutions is necessary for considering other multi-criteria methods. Moreover, as we have already shown in experiments (see Table 3), this step significantly reduced the number of solutions.

The final component of our approach involves selecting the counterfactual by means of the ideal point method. On the average, this method yields significantly better-scoring counterfactuals compared with random selection from the Pareto front. An interesting comparison lies in contrasting the ideal point method with the simple unweighted sum ($s = \text{proximity} + \text{feasibility} + (1 - \text{discriminative power})$) of the criteria scores. Our experiments (see the online appendix) reveal that the variant employing the Manhattan distance achieves the best average scores across the examined datasets. This variant is equivalent to the unweighted sum ($s = c_{\text{Manhattan}}$), as, in the previously mentioned min-max normalized criteria space, the ideal point coordinates are $p = (0, 0, 0)$. Therefore, the distance to the ideal point $c_{\text{Manhattan}} = |p_0 - \text{proximity} + p_1 - \text{feasibility} + p_2 - (1 - \text{discriminative power})|$ is just the sum of scores on all criteria, equivalent to the unweighted sum: $s = c_{\text{Manhattan}}$. It is worth noting that other variants of our approach using the Euclidean or the Chebyshev distance are not equivalent and select different alternatives.

While this paper primarily utilizes the simplest multi-criteria selection method, we also conducted preliminary experiments (included in the online appendix) with a slightly more advanced selection method based on the distance to the nadir-ideal plane (Ehrgott and Tenfelde-Podehl, 2003), showing slightly superior results compared with unweighted sum selection. This underscores the value of employing multi-criteria selection methods. Even this simple distance methods are also useful for possible further extensions of the methods towards an interactive dialogue with the decision-maker.

4.5. Using user preference models to evaluate selected trade-offs between different quality measures. The main aim of our work is to generate a counterfactual for a given instance, which represents a suitable trade-off between various aspects of explanation quality. Let us recall that in our approach we do not use criterion weights due to the difficulty of being estimated by humans. However, for the purpose of examining possible trade-off criteria in potential

models of decision-makers' preferences, we decided to experimentally simulate a simple utility function model, where we analyzed the impact of all possible weight configurations. Since an objective comparison of different non-dominated trade-offs is impossible without additional information about user preferences towards optimizing particular quality precisely, we employ a simple mathematical model of user preference to verify the utility of counterfactuals under all possible configuration of preferences. Specifically, we model the utility of an explanation as a weighted sum of three selected quality criteria:

$$U(x') = w_p \cdot \text{proximity}(x') + w_d \cdot \text{discriminative power}(x') + w_f \cdot \text{feasibility}(x'), \quad (1)$$

where $w_p, w_d, w_f \geq 0$, $w_p + w_d + w_f = 1$ are parameters which control the importance of individual quality measures for a potential user. It is assumed that the user should select the solution which maximizes the utility function.

Since the space of all possible utility functions has only three parameters (w_p, w_d, w_f) which sum up to 1, it can be easily visualized on a 2-dimensional plot (see Fig. 3) with the Barycentric coordinate system. The barycentric plot takes the form of an equilateral triangle with each vertex associated with one weight of the utility function. Each point inside the triangle represents one possible user utility function, i.e., one possible setup of w_p, w_d, w_f weights. For instance, a point at the vertex corresponding to *proximity* (w_p) represents the utility function with $w_p = 1$ and $w_d = w_f = 0$. A point in the middle of the edge connecting the vertices corresponding to *proximity* (w_p) and *feasibility* (w_f) represents the utility function with $w_p = w_f = 0.5$ and $w_d = 0$. Finally, the central point in the middle of the triangle corresponds to a user having equal preferences for all quality criteria ($w_p = w_d = w_f = 1/3$).

We computed the utilities of the counterfactuals returned by all the methods under study using utility functions for all⁸ possible combinations of weight values. Later, we verified which counterfactual method would be selected as the preferred one by a potential user with such preferences. The results of this experiment are visualized in Fig. 3.

For the Adult dataset, despite being uninformed about user preferences and always selecting the same counterfactual with the ideal point method, our proposed approach would be selected as the best one according to approximately 85% of all possible utility functions. It is only when the user has a strong preference towards one criterion that the proposed approach loses to CADEX on

⁷For simplicity, we invert the *discriminative power* criterion to have the min optimization direction.

⁸More precisely, we evaluated a grid of 136 weight combinations, approaching weight values from 0 to 1 with a step of 1/16.

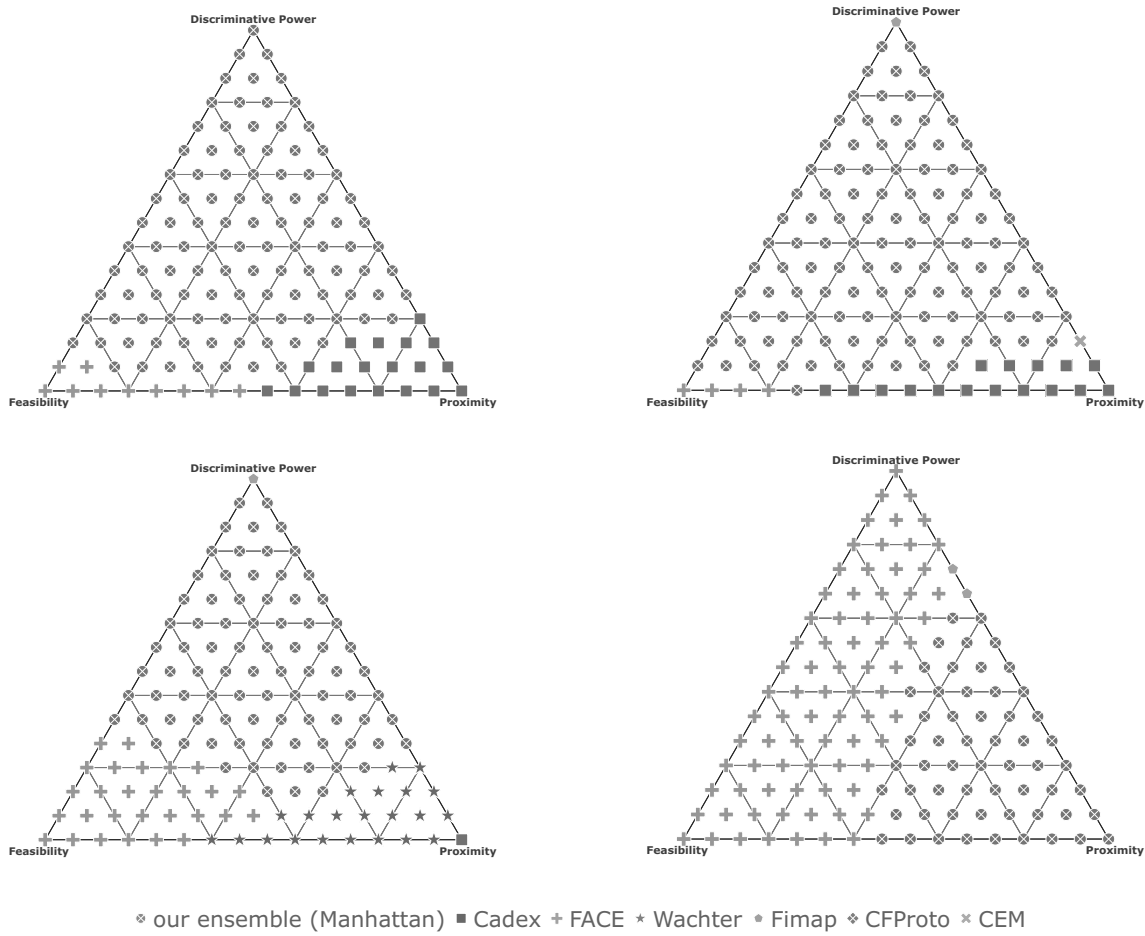


Fig. 3. Barycentric plots depicting the best method for a given dataset according to the utility functions with different weights assigned to quality criteria. The datasets are: German (upper-left), Adult (upper-right), Compas (lower-left), Fico (lower-right).

proximity and to FACE on *feasibility*. For the user solely interested in *discriminative power* the counterfactuals produced by Fimap would be the most appropriate. Similar observations can be made for the German and Compas datasets, where the results of our method constitute the best trade-off for all users who are not too strongly biased towards one particular criterion.

It is only for the Fico dataset that the counterfactuals produced by FACE seem to represent a good trade-off between *feasibility* and *discriminative power* for many cases. However, it is important to note that for this dataset FACE generates actionable explanations only for 3% of examples (see Table 7). Therefore, this result does not demonstrate inefficiency of our selection procedure since we automatically discard all non-actionable explanations beforehand. Nevertheless, our proposed method achieves good trade-offs for users who are more inclined towards obtaining explanations similar to the instance being explained (*proximity*) but who are also interested in reasonable *feasibility* and *discriminative power*. In other words, unless the decision-maker is interested solely in

counterfactual explanations that score high only on one criterion, our proposed method provides more preferred counterfactuals.

5. Conclusions and future work

The main message of our work is to promote the use of the multi-criteria decision analysis (MCDA) to select contractual explanations for predictions made by black-box machine learning models.

The proposed approach has at least two original contributions to current research. Firstly, we propose to construct an ensemble of various explanatory methods that are effective in generating a fairly large (in our experiments about 80–90) set of diverse solutions. Secondly, we employ further multi-criteria analysis to first reduce the number of counterfactuals considered to only non-dominated explanations, and then, to support the selection of a solution that offers a compromise between the values of the evaluation measures considered.

In the present work, we have consciously and

deliberately chose relatively simple and well-known MCDA proposals: (i) the use of the dominance relation, which is the only approach that is fully objective to filter out the worse evaluated solutions; (ii) the ideal point method, which is also a no preference approach (i.e., it does not require any acquisition of user preferences on the relative importance of criteria—weights or comparisons of alternatives). It is also computationally simple to run automated experiments.

Additional experiments (Section 4.4) also showed that all steps of our approach are essential. Moreover, even a relatively simple multi-criteria ideal point method leads to good choices of counterfactuals and can be easily further extended, e.g., to the ideal-nadir version.

Despite the simplicity of these multi-criteria methods, we believe that the presented experiments demonstrated the utility of our approach. Indeed, the selected counterfactuals are competitive with the solutions offered by the best of the single methods used, which often optimize only one criterion. Furthermore, our additional experiments simulating the utility model of a potential decision maker show that when they do not have a strong preference for a single criterion, the proposed multi-criteria approach is highly beneficial and provides a good trade-off between criteria.

Furthermore, our comparative study reveals that no method is superior to others in all criteria. Therefore, we conclude that more attention should be given to comparing different counterfactual generation methods using multi-criteria analysis to highlight the various trade-offs made by these methods.

We also argue that looking at explanations from different perspectives should be studied more extensively, and we believe that multi-criteria analysis is the natural choice for this type of investigation. As a next step, we suggest that research should focus on whether the proposed framework aligns with the human perspective, and whether humans make trade-offs between criteria when selecting the best explanations or whether they prefer only one of them. Moreover, it is worth considering adaptations of interactive, dialogue multi-criteria selection methods, which will be the subject of further research.

Acknowledgment

The research has been partially supported by the 0311/SBAD/0743 and 0311/SBAD/0740 grants of the Poznan University of Technology as well as TAILOR, a project funded by the EU Horizon 2020 research and innovation programme under the GA no. 952215.

References

Bodria, F., Giannotti, F., Guidotti, R., Naretto, F., Pedreschi, D. and Rinzivillo, S. (2021). Benchmarking and survey of

explanation methods for black box models, *Data Mining and Knowledge Discovery* **37**(5): 1719 – 1778.

Branke, J., Deb, K., Miettinen, K. and Slowiński, R. (2008). *Multiobjective Optimization: Interactive and Evolutionary Approaches*, Springer, Berlin/Heidelberg.

Chapman-Rounds, M., Bhatt, U., Pazos, E., Schulz, M.-A. and Georgatzis, K. (2021). Fimap: Feature importance by minimal adversarial perturbation, *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 11433–11441, (virtual).

Dandl, S., Molnar, C., Binder, M. and Bischl, B. (2020). Multi-objective counterfactual explanations, in T. Bäck et al. (Eds), *Parallel Problem Solving from Nature, PPSN XVI*, Springer, Cham, pp. 448–469.

Dhurandhar, A., Chen, P.-Y., Luss, R., Tu, C.-C., Ting, P., Shanmugam, K. and Das, P. (2018). Explanations based on the missing: Towards contrastive explanations with pertinent negatives, *32nd International Conference Neural Information Processing Systems, Montreal, Canada*, pp. 590–601.

Ehrgott, M. (2005). *Multicriteria Optimization*, Springer-Verlag.

Ehrgott, M. and Tenfelde-Podehl, D. (2003). Computation of ideal and nadir values and implications for their use in MCDM methods, *European Journal of Operational Research* **151**(1): 119–139.

Falbogowski, M., Stefanowski, J., Trafas, Z. and Wojciechowski, A. (2022). The impact of using constraints on counterfactual explanations, *Proceedings of the 3rd Polish Conference on Artificial Intelligence, PP-RAI 2022, Gdynia, Poland*, pp. 81–84.

Förster, M., Hühn, P., Klier, M. and Kluge, K. (2021). Capturing users' reality: A novel approach to generate coherent counterfactual explanations, *Proceedings of the 54th Hawaii International Conference on System Sciences, Maui, USA*, pp. 1274–1284.

Guidotti, R. (2022). Counterfactual explanations and how to find them: Literature review and benchmarking, *Data Mining and Knowledge Discovery*, DOI: 10.1007/s10618-022-00831-6.

Guidotti, R. and Ruggieri, S. (2021). Ensemble of counterfactual explainers, *24th International Conference on Discovery Science, Halifax, Canada*, pp. 358–368.

Inbar, Y., Botti, S. and Hanks, K. (2011). Decision speed and choice regret: When haste feels like waste, *Journal of Experimental Social Psychology* **47**(3): 533–540.

Iyengar, S. and Lepper, M.R. (2000). When choice is demotivating: Can one desire too much of a good thing?, *Journal of Personality and Social Psychology* **79**(6): 995–1006.

Klaise, J., Looveren, A. V., Vacanti, G. and Coca, A. (2021). Alibi explain: Algorithms for explaining machine learning models, *Journal of Machine Learning Research* **22**(1): 1–7.

Kuncheva, L.I. (2004). *Combining Pattern Classifiers: Methods and Algorithms*, Wiley, Hoboken.

- Laugel, T., Lesot, M.-J., Marsala, C., Renard, X. and Detynecki, M. (2018). Comparison-based inverse classification for interpretability in machine learning, in J. Medina et al. (Eds), *Information Processing and Management of Uncertainty in Knowledge-Based Systems: Theory and Foundations*, Springer, Cham, pp. 100–111.
- Mertes, S., Huber, T., Weitz, K., Heimerl, A. and André, E. (2022). GANterfactual—Counterfactual explanations for medical non-experts using generative adversarial learning, *Frontiers in Artificial Intelligence* 5.
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences, *Artificial Intelligence* 267: 1–38.
- Moore, J., Hammerla, N. and Watkins, C. (2019). Explaining deep learning models with constrained adversarial examples, in A.C. Nayak and A. Sharma (Eds), *PRICAI 2019: Trends in Artificial Intelligence*, Springer, Cham, pp. 43–56.
- Mothilal, R.K., Sharma, A. and Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations, *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona, Spain*, pp. 607–617.
- Pawelczyk, M., Bielawski, S., van den Heuvel, J., Richter, T. and Kasneci, G. (2021). Carla: A Python library to benchmark algorithmic recourse and counterfactual explanation algorithms, *arXiv* 2108.00783.
- Pearl, J., Glymour, M. and Jewell, N. (2016). *Causal Inference in Statistics: A Primer*, Wiley, Hoboken.
- Poyiadzi, R., Sokol, K., Santos-Rodriguez, R., De Bie, T. and Flach, P. (2020). FACE: Feasible and actionable counterfactual explanations, *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, Association for Computing Machinery, New York, USA*, pp. 344–350.
- Rasouli, P. and Chieh Yu, I. (2022). CARE: Coherent actionable recourse based on sound counterfactual explanations, *International Journal of Data Science and Analytics* 17(1): 1–26.
- Skulimowski, A. (1990). Applicability of ideal points in multicriteria decision-making, *Proceedings of the 9th International Conference on Multiple Criteria Decision-Making, Fairfax, USA*, pp. 5–8.
- Spreitzer, N., Haned, H. and van der Linden, I. (2022). Evaluating the practicality of counterfactual explanations, *Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS 2022, New Orleans, USA*.
- Stefanowski, J. (2023). Multi-criteria approaches to explaining black box machine learning models, *Asian Conference on Intelligent Information and Database Systems ACIIDS 2023, Phuket, Thailand*, pp. 195–208.
- Stepka, I., Lango, M. and Stefanowski, J. (2023). On usefulness of dominance relation for selecting counterfactuals from the ensemble of explainers, *Proceedings of the 4th Polish Conference on Artificial Intelligence, PP-RAI 2023, Łódź, Poland*, pp. 125–130.
- Steuer, R. (1986). *Multiple Criteria Optimization: Theory, Computation, and Application*, Wiley, Hoboken.
- Tolomei, G., Silvestri, F., Haines, A. and Lalmas, M. (2017). Interpretable predictions of tree-based ensembles via actionable feature tweaking, *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, Canada*.
- Ustun, B., Spangher, A. and Liu, Y. (2019). Actionable recourse in linear classification, *Proceedings of the Conference on Fairness, Accountability, and Transparency, Atlanta, USA*, pp. 10–19.
- Van Looveren, A. and Klaise, J. (2021). Interpretable counterfactual explanations guided by prototypes, in N. Oliver et al. (Eds), *Machine Learning and Knowledge Discovery in Databases: Research Track*, Springer, Cham, pp. 650–665.
- Verma, S., Boonsanong, V., Hoang, M., Hines, K.E., Dickerson, J.P. and Shah, C. (2020). Counterfactual explanations and algorithmic recourses for machine learning: A review, *arXiv* 2010.10596.
- Wachter, S., Mittelstadt, B. and Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR, *Harvard Journal Law & Technology* 31(2): 841–887.
- Wellawatte, G.P., Seshadri, A. and White, A.D. (2022). Model agnostic generation of counterfactual explanations for molecules, *Chemical Science* 13(13): 3697–3705.
- Wilson, D.R. and Martinez, T.R. (1997). Improved heterogeneous distance functions, *Journal of Artificial Intelligence Research* 6: 1–34.



Ignacy Stepka is an undergraduate research assistant in the Institute of Computing Science at the Poznan University of Technology. His research interests focus on machine learning and human-centered, explainable AI.



Mateusz Lango is an assistant professor in the Institute of Computing Science at the Poznan University of Technology. His research interests include explainable AI methods, in particular for natural language processing applications, as well as learning from complex and imbalanced data.



Jerzy Stefanowski is a full professor in the Institute of Computing Science of the Poznan University of Technology and in the Polish Academy of Sciences. His research interests include machine learning, data mining and intelligent decision support, in particular rule induction, ensembles, class imbalance, concept drift, classification of data streams and explainable AI.

Received: 6 July 2023

Revised: 1 November 2023

Accepted: 5 December 2023