"Better" Counterfactuals, Ones People Can Understand: Psychologically-Plausible Case-Based Counterfactuals Using Categorical Features for Explainable AI (XAI)

Greta Warren[1[0000-0002-3804-2287]], Barry Smyth[1,2[0000-0003-0962-3362]] and Mark T. Keane[1,2,3[0000-0001-7630-9598]]

[1] School of Computer Science, University College Dublin, Dublin, Ireland
[2] Insight Centre for Data Analytics, University College Dublin, Dublin, Ireland
[3] VistaMilk SFI Research Centre, University College Dublin, Dublin, Ireland
greta.warren@ucdconnect.ie, {mark.keane, barry.smyth}@ucd.ie

Abstract. A recent surge of research has focused on counterfactual explanations as a promising solution to the eXplainable AI (XAI) problem. Over 100 counterfactual XAI methods have been proposed, many emphasising the key role of features that are "important" or "causal" or "actionable" in making explanations comprehensible to human users. However, these proposals rest on intuition rather than psychological evidence. Indeed, recent psychological evidence [22] shows that it is abstract feature-types that impact people's understanding of explanations; *categorical features* better support people's learning of an AI model's predictions than *continuous features*. This paper proposes a more psychologically-valid counterfactual method, one extending case-based techniques with additional functionality to transform feature-differences into categorical versions of themselves. This enhanced case-based counterfactual method, still generates good counterfactuals relative to baseline methods on coverage and distances metrics. This is the first counterfactual method specifically designed to meet identified psychological requirements of end-users, rather than merely reflecting the intuitions of algorithm designers.

Keywords: CBR, Explanation, XAI, Counterfactuals, Contrastive

1. Introduction

In recent years, a significant effort has been made to develop methods that can explain to people why/how an automated decision was made by some black-box AI system (see [1–3]). Indeed, given the requirements of GDPR to provide such explanations when automated decisions are made without human intervention, there is an added urgency to solve this eXplainable AI (XAI) problem [3,4]. Many XAI strategies proposed in recent years have echoed long-standing work in case-based reasoning (CBR) where the provision of explanations for model predictions has always been a motivating concern (see e.g., [5–7]). Hence, case-based explanations proposed over two decades ago have been revisited and extended to be applied to deep learning (see e.g., [8–10]). Similarly, *a-fortori* explanations proposed in CBR [11,12] have been revived as semi-factuals to

explain deep learners [13] and Nearest Unlike Neighbours (NUNs; [11,14]), have been re-cast as counterfactual explanations realised in 100+ different algorithms in the literature [15]. This last explanation strategy is the focus for the current paper, as it has become one of the most researched post-hoc solutions to the XAI problem.

The classic example of a counterfactual explanation is one for an automated banking application where, on foot of being refused a loan, the customer asks for an explanation and is told "if you asked for a lower loan of $30,000 over a shorter term of two years, you would have been granted the loan". Importantly, the counterfactual tells the end-user about what feature changes will result in a different decision outcome, giving the user some insight into how they might reverse the outcome of an automated decision (so-called *algorithmic recourse* [16]). Counterfactual explanations have attracted a lot of attention in XAI because they appear to be psychologically comprehensible [17], GDPR-compliant [18], and invite a wide variety of computational solutions (see [15,16]). One family of solutions adopts a case-based approach, where NUNs are used in a variety of different ways to explain some current target (query) case [14,19–21].

In this paper, we extend the case-based approach of [20] to make it more psychologically-valid by modifying the feature-types its uses in its explanations. The next section places this work in the context of recent relevant research. Recent psychological research shows that, where possible, categorical features should be preferred as the basis for the generated counterfactuals used in explanations [22]. In section 3 we perform a computational study to determine the "natural" occurrence of counterfactuals with categorical feature-differences in several representative datasets; this study shows they are rare and motivates our method that transforms all continuous feature-values into categorical ones (see section 4). Then in section 5, we report a second study to test these feature-transforming methods using an extension of case-based counterfactual generation methods on representative datasets; we assess if there is any decrement in the quality of counterfactuals produced under these transformations. In section 6, we conclude by discussing the implications of these results for counterfactuals in XAI.

2. Background: Computation & Psychology of Counterfactuals

Though it is not always recognised in the AI literature, the computation of counterfactuals has been with us for some time. In earlier guises, it was cast as finding NUNs [23,24] or inverse classifications [25]. For instance, in a binary classification problem we could have two cases that are very close to one another, but one feature change flips the class of the cases; in the loan domain, a reduction in the value of the *loan-amount* feature, might change the decision from "refuse" to "grant". Depending on the dataset, these two instances could be NUNs, the closest pair of instances in the dataset where specific feature-changes modify the class predicted. Several early papers on NUNs considered their use in the context of explanation for domains using tabular [11,23] and textual data [14], emphasising the use of instances in the existing dataset.

However, recently, a very different optimisation approach has been proposed that uses similarity constraints to generate synthetic instances that balance similarity to the query against distance to the decision boundary. Wachter et al. [18] propose that counterfactuals can be computed using the loss function, L:

$$L(x,\ x', y', \lambda) = \ \lambda(f(x') - y')^2 + d(x, x') \qquad (1)$$

$$arg \min_{x'} \max_{\lambda} L(x, x', y', \lambda) \quad (2)$$

where x is the vector for the query case and x' is the counterfactual vector, with y' being the desired (flipped) prediction from $f(..)$ the trained model, where λ acts as the balancing weight. In formula (2), λ balances the closeness of the counterfactual to the query case against making minimal changes to the query case while delivering a prediction change, using the ℓ_1 norm weighted by median absolute deviation. In implementations, this method generates a space of feature perturbations for the original query and then uses gradient descent to settle on a minimally-perturbed (aka the *best*) counterfactual. This method tends to generate low-sparsity counterfactuals, that is, counterfactuals with few feature differences; a property seen as attractive as it allows people to better understand them[1] (e.g., Table 1 shows "good" and "bad" counterfactuals for explaining a blood-alcohol-level prediction based on sparsity). Unfortunately, this original method has been shown to have several limitations. First, it cannot handle categorical features, as it only addresses features with continuous values. Second, it sometimes generates invalid, out-of-distribution data-points and, therefore, invalid explanations in the domain [26]. Many subsequent papers have aimed to rectify these deficits. For example, DiCE [27] handles categorical features using one-hot encoding and adds constraints for diversity, while other models supplement the constraints to be more sensitive to the data [28,29]. So, there is now a whole family of these optimisation methods that claim to foster better computation of counterfactuals.

Table 1. A query case paired with a "Good", "Better" and "Bad" Counterfactual from the Blood Alcohol Content (BAC) case-base (the feature-differences shown in bold-italics).

Features	**Query Case**	**"Bad" Counterfactual**	**"Good" Counterfactual**	**"Better" Counterfactual**
Weight	90 kg	90 kg	***100 kg***	90 kg
Duration	1 hr	***3 hrs***	***1.5 hr***	1 hr
Gender	Female	***Male***	Female	***Male***
Stomach	Empty	***Full***	Empty	***Full***
Units	5	***4***	5	5
BAC Level	Over	Under	Under	Under

One of the persistent themes in this literature has been around the differential importance of features. In the psychological literature on counterfactual thinking, Nobel laureate Daniel Kahneman noted that only some features are "mutable" [30]; some features cannot be changed in creating a counterfactual (e.g., *age* is not a feature that can change to improve one's loan chances). So, counterfactual explanations need to change "plausible" features, namely ones that are "actionable" (i.e., that the user can action; [31–33]), "causally important" (i.e., that play a key role; [34]) or "predictively important" [35,36]. Overall, this concern with features tries to ensure that these methods produce counterfactual explanations that make sense to people, that people can act on and that increase their understanding of the domain (e.g., how the model makes its decisions). However, there are many issues around the identification and use of the "right"

[1] Keane and Smyth [19] argued that, for tabular data, counterfactuals should be sparse; no more than 2 feature-differences, to allow people to understand them. Recent user studies show that people prefer counterfactuals with 2-3 feature differences [45].

features. Firstly, these featural proposals often end in ad-hoc solutions, such as end-users interacting to mark features as important or to define ranges on feature-values [27]. Secondly, for causal importance, methods assume causal models that are not often or always available. Thirdly, what constitutes the "right" feature seems to be very context-sensitive; for example, if I earn $300k a year, then increasing income by $5k may be actionable, but if I earn $30k a year, earning an additional $5k may be impossible. Perhaps one of the main attractions of case-based approaches is that they exploit implicit dependencies in the data and accordingly, by definition, rely on plausible/mutable/important features and avoid producing out-of-distribution counterfactuals.

Furthermore, as we shall see in the next section, this AI literature on feature importance has overlooked one key aspect of features shown to be psychologically critical. Recent user studies have revealed that abstract feature-types – categorical versus continuous features – are understood very differently by human users [22]. In the next sub-section, we consider related work on the psychology of counterfactuals in XAI and their implications for models of counterfactual generation.

2.1 User Studies of Counterfactual XAI: Mixed Results

Although the AI literature on counterfactual methods has exploded in the last few years, user studies examining how people understand and use counterfactuals have lagged considerably; indeed, the user studies that have been done tend to be too general, report mixed results, or both. Keane et al. [15] report that of 100+ distinct counterfactual methods reported in the XAI literature (of which ~5 are case-based methods) only ~20% report any user tests; even fewer papers test specific aspects of specific techniques (~7%). Many of these studies report quite general findings, showing that counterfactuals broadly improve people's responding in some way [21,37,38]. More focused user studies tend to report mixed results on people's performance with counterfactual explanations. Van der Waa et al. [39] tested people's performance with a simulated blood-sugar-prediction app using either contrastive-rule or example-based explanations and found that neither strategy did better than no-explanation controls. Lage et al. [40] found when people were given counterfactual tasks, in which users were asked if a system's recommendation would change given a perturbation of some input feature, they reported greater difficulty; also longer response times and lower accuracy in prediction tasks were recorded. Taken together, these studies present a confusing picture, in which counterfactuals seem to sometimes help and other times hinder. They also suggest a focus on how people really understand the features in counterfactual explanations, to uncover when and how they really work, psychologically.

Though many counterfactual methods have emphasised featural aspects, none of these papers user-test their proposals. To the best of our knowledge only two papers have specifically user-tested feature-types in counterfactual methods [22,41]. Kirfel and Liefgreen [41] found that people's perceptions of the quality and comprehensibility of explanations was affected by whether they involved actionable and mutable features, as opposed to immutable ones. However, they also pointed out that the actionable/mutaibility distinctions made by AI researchers were not as clear-cut for laypeople. This raises the question about whether there are more fundamental feature-categories that could impact people understanding. Indeed, longstanding evidence from human reasoning suggests that people do not spontaneously change continuous variables (e.g., the speed or timing of vehicles involved in a road accident) when generating counterfactuals [42]. To address this issue, Warren et al. [22] examined the role of abstract feature-

types – categorical or continuous – in people's understanding of counterfactual explanations for a black-box model's predictions. They examined the effects of different counterfactual (and causal) explanations on people's understanding of a blood-alcohol-content (BAC) domain. They presented people with a simulated AI model that predicts if someone is over/under the legal alcohol limit for driving, based on five features: weight, duration of drinking, gender, stomach-fullness, and number of units drunk. In the training phase of the experiment, people were presented with query cases with different values for these features and asked to predict the outcome as under/over the BAC limit. After responding, they were shown the model's prediction along with a counterfactual explanation (e.g., see the "good counterfactual" in Table 1). In the training phase, they saw 40 such cases with equally-balanced cases involving counterfactuals for the five features (i.e., 8 cases for each feature). Then, in the testing phase, they received 40 new cases and were asked to predict the outcome while focusing on a specific feature (8 instances per feature), without receiving explanations or feedback, to determine how accurate their predictions were. This experimental paradigm tests a critical aspect of explanation use, namely, if experience with the model's predictions combined with explanations improve people's understanding of the domain, as measured by accuracy in the test phase.

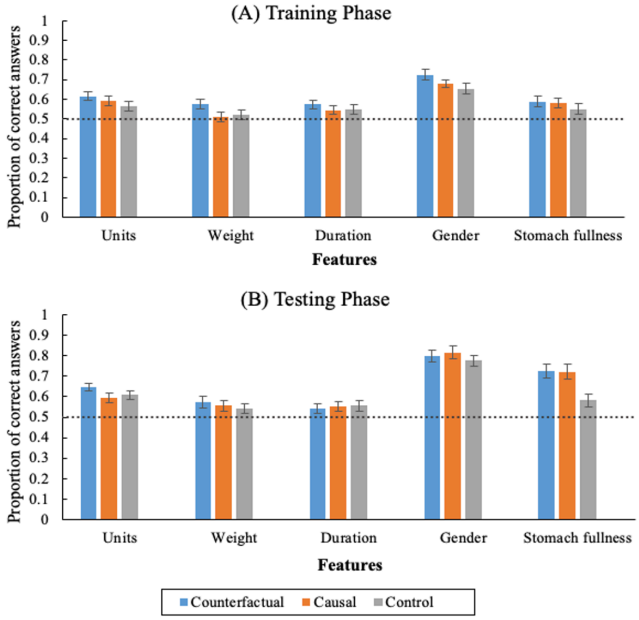

Fig. 1. Mean accuracy for three conditions (counterfactual, causal and control) by each feature in the (A) training and (B) testing phases of Warren et al. [22] (error bars represent standard error of the mean; dashed line represents chance accuracy).

Overall, the results, shown in Figure 1, show an effect of explanation, where people given counterfactual explanations were more accurate than no-explanation controls. However, the results also showed an independent effect of feature, that also interacted with phase; in the testing phase, accuracy for cases using categorical features improved, whereas those using continuous features did not; note the increase in Figure 1(B) relative to Figure 1(A), for *gender* and *stomach-fullness*. Indeed, this feature-type factor

accounts for almost all the improvement in accuracy seen between the experiment's training and test phases. These results show that improvements in accuracy were solely due to the presence of categorical features over continuous ones. So, counterfactuals with categorical-differences should be "better" (see Table 1). This finding motivates the current work, to extend counterfactual methods to take feature-type into account.

3. Study 1: Plotting Counterfactuals That Have Categoricals

Given the importance of categorically-based explanations, in this study we examined a number of UCI datasets [43] – based on (i) prior use in testing counterfactual methods, (ii) their use of categorical features – to determine their potential to yield *native counterfactuals* (i.e., existing pairs of counterfactually-related instances in the dataset that can be used to generate synthetic counterfactuals) that rely on categorical feature-differences (as opposed to continuous ones). Seven datasets were selected: the blood-alcohol-content, contraceptive-choice, cleveland-heart, horse-colic, credit, german credit, and thyroid datasets. Note, datasets containing categorical features may be relatively rare: of the 622 UCI datasets publicly available as of May 2022, 38 (.06%) contain only categorical features, while 55 (.09%), contain mixed features (some of which contain only a single categorical feature, such as the abalone, diabetes datasets). We used the case-based counterfactual method, CB2-CF (see section 4), to compute all pairs of cases either side of a decision boundary (i.e., native counterfactuals) noting the number of feature differences in each, and if they had *at least one* categorical feature. This method uses a *tolerance* to identify feature-differences so small differences [e.g., ±20% of 1 standard deviation (SD)] in continuous features are treated as essentially identical; varying this tolerance (≤1 SD) did not materially change the results.

Table 2. Study 1 Results: Frequencies of native categorical counterfactuals (≥1 categorical feature-difference) over 7 datasets, for 1-5 feature-differences, as a % of potential counterfactuals.

Dataset	***N cases***	***N feats***	***N cat. feat.***	***1-diff CFs (% tot.)***	***2-diff CFs (% tot.)***	***3-diff CFs (% tot.)***	***4-diff CFs (% tot.)***	***5-diff CFs (% tot.)***
Blood Alcohol.	4748	5	2	19 (0.4%)	1302 (27.4%)	4574 (96.3%)	4736 (99.8%)	0 (0%)
Contracept.	1425	9	7	236 (16.6%)	1050 (73.7%)	1345 (94.5%)	1377 (96.7%)	1379 (96.8%)
Cleveland Heart	303	13	7	0 (0%)	0 (0%)	0 (0%)	8 (2.65%)	93 (30.8%)
Colic	300	26	19	1 (0.33%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Credit	690	15	9	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
German Credit	1000	20	13	0 (0%)	0 (0%)	3 (0.3%)	20 (2.0%)	108 (10.8%)
Thyroid	2753	28	22	0 (0%)	0 (0%)	0 (0%)	0 (0%)	69 (2.5%)

3.1 Results & Discussion

Table 2 shows frequencies of native counterfactuals involving at least one categorical feature-difference (and their percentage in the total set of counterfactuals). Note, this is a low bar for categorical counterfactuals, as it admits one with 5 feature-differences where, perhaps, only one of those feature-differences were categorical. Even with this low bar, counterfactuals based on categorical features are rare. For 5 datasets, none of the *good* counterfactuals (i.e., those with 1-3 feature differences) involved categorical features. Of the 2 datasets – blood-alcohol-content and contraceptive – that yield more categorical counterfactuals, there are still very few 1-feature-difference counterfactuals that are categorical (respectively ~0.4% and ~16.6%). We found no relationship between the number of categorical features in a dataset and its propensity to generate categorically-based counterfactuals; for instance, the contraceptive dataset has 7/9 categorical features (77%) whereas the thyroid dataset has 22/28 categorical features (78%) but they both show very different results.

Clearly the occurrence of categorical features in counterfactuals must depend on the underlying domain theory, on the presence/absence of dependencies between categorical and continuous variables in the domain. For instance, in the BAC domain *gender* has a big impact on outcomes; females have different metabolic rates to males and this factor impacts other variables in the BAC formula. From these results, it is hard to escape the conclusion that categorically-based, counterfactual explanations do not naturally occur in many datasets. We also note that, unlike those evaluated here, a significant proportion of the datasets widely used in the machine learning literature (e.g., see UCI repository [43]) do not contain any categorical features whatsoever, and hence will not yield native categorical-counterfactuals. So, to present such explanations to end-users, we will need to transform feature-differences involving continuous features into categorical representations of themselves. In the next section, we consider how an existing instance-based counterfactual method can be re-designed to do such transformations in a post-processing step, before an explanation is presented to users.

4. Transforming Case-Based Counterfactuals, Categorically

In AI, NUNs have been considered for some time (see [24] for a review), though the idea of using a NUN as a counterfactual explanation is more recent [11,21]. However, NUNs on their own are not a general solution to counterfactual explanation; even if available, they may be too distant from the query to provide a good explanation. Hence, most current techniques try to generate synthetic counterfactuals that are close to the query and within-distribution [19,20,27]. Case-based counterfactual techniques tend to use NUNs as templates for generating synthetic counterfactuals either by selecting specific features from the NUN in some constrained way [19,20] or by perturbing the NUN towards the query [14,36]. In the next subsection, we quickly describe the case-based counterfactual method used in the current experiments, before describing two algorithmic extensions to it, that perform categorical transformations to explanatory cases.

4.1 Case-Based Counterfactual Methods: CB1-CF and CB2-CF

Keane and Smyth [19] proposed a case-based counterfactual method (CB1-CF here) designed to generate plausible and informative counterfactual explanations for any presented query case. Unlike optimisation methods, CB1-CF uses historical counterfactual-pairs in the dataset – so-called *native counterfactuals* – as templates for building new, synthetic counterfactual cases for the query case. As Figure 2 shows, to explain the outcome of *p*, CB1-CF identifies a nearby pair of cases, *cf(x, x')*, where *x* has the same class as *p* and *x'* is a good counterfactual for *x*; *x* and *x'* differ by a small number of features and these features are adjusted in *p* to obtain a new explanation case, *p'*, that is counterfactually related to *p* (see [19] for details on how these *difference-features* are adjusted). There may be other counterfactual pairs in the case-base, such as *cf(q, q')*, but CB1-CF only uses the closest to build a single explanatory counterfactual.

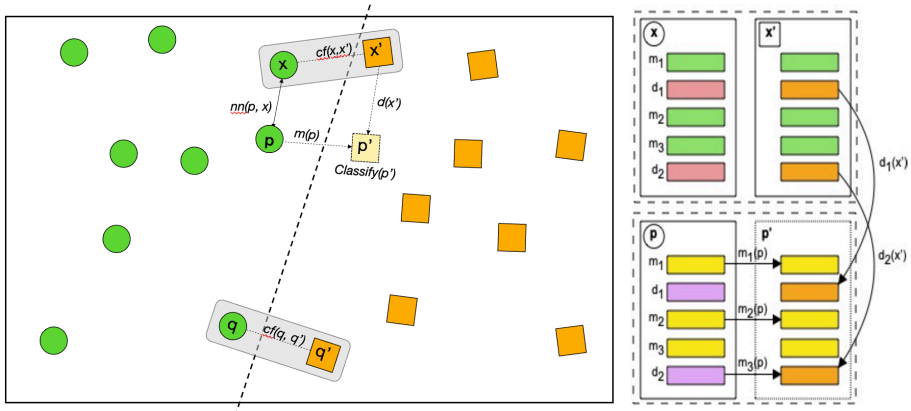

Fig. 2. An illustration of (a) a two-class case-base with 2 native counterfactuals [i.e., *(x, x')* and *(q, q')*], where one *(x, x')* is the nearest-neighbour native to the query, *p*, and *x'* is used to create the explanatory counterfactual case, *p'*; (b) how a synthetic, counterfactual case, *p'*, is generated from the values in the match-features of *p* and the difference-features of *x'*.

Recently, Smyth and Keane [20] generalised CB1-CF to go beyond just considering a single, native counterfactual (k=1); this new method (which we refer to as CB2-CF) can arbitrarily vary the number of natives considered (any k >1). The authors show that CB2-CF generates better counterfactual cases (i.e., ones closer to the query) with better coverage (i.e., it can find good counterfactuals for most queries) for k=10-30 in representative datasets. In the current tests, we use a simplified version of CB2-CF to test for the effects of categorical transformations on the generation of explanations[2]. Of course, CB2-CF does not consider whether categorical/continuous features are used in the counterfactual. So, in the next section, we extend CB2-CF, using variants that transform feature-difference values to be categorical (see section 5 for tests).

[2] As well as considering multiple natives, CB2-CF also considers nearest-like-neighbours of the native's *x'* (e.g., the three closest, same-class datapoints to *x'*) to expand on the variations of natives considered. This second step in not implemented in our version of CB2-CF.

CAT-CF^{global} *(q, CB):*

1. cfs, dists ⟵ FindCFs$_k$(q, CB)
2. for each cf ∈ cfs:
3. for each f_i ∈ DiffFeatures(q, cf):
4. cf(f_i) ⟵ BinaryBin(CB, f_i, cf(f_i))
5. if cf(f_i) = q(f_i)
6. cf ⟵ conflicting
7. valid_cfs ⟵ remove conflicting CFs
8. valid_cfs ⟵ Sort(valid_cfs, by=[‘sparsity', ‘dist’])
9. Return valid_cfs

CAT-CF^{Local} *(q, CB):*

1. cfs, dists ⟵ FindCFs$_k$(q, CB)
2. for each cf ∈ cfs:
3. for each f_i ∈ DiffFeatures(q, cf):
4. if cf(f_i)<q(f_i)
5. cf(f_i) ⟵ lower
6. else:
7. cf(f_i) ⟵ higher
8. cf(f_i) ⟵ Direction(q, cf, f_i)
9. valid_cfs ⟵ Sort(valid_cfs, by=[‘ sparsity', ‘direction']
10. Return valid_cfs

Algorithms: Two methods for transforming feature differences in counterfactuals. Notes: *BinaryBin(CB, f, v)* converts a feature value *f(v)* into a binary categorical value *v'* based on the binning of features values fo*r f* in *CB*. This process can be done once for a given *CB* so that *BinaryBin(CB, f, v)* can be as a simple lookup.

4.2 Counterfactuals With Categorical Transforms #1: Global Binning

Study 1 showed that many datasets produce little or no categorical counterfactuals or only produce them in counterfactuals with poor sparsity (i.e., >2 feature differences). These findings led us to conclude that all continuous-type feature-differences in generated counterfactuals need to be transformed into categorical versions of themselves in the explanation-generation process (as in Table 3). Hence, we propose a post-processor that considers alternative counterfactual explanations for a given query, transforming them into categorical versions and, after some minimal checking, produces the best one as an explanation. Based on the psychological evidence [22], we apply binary transformations to continuous features; though should future work identify similar benefits of categorical features with more than two possible values, our approach can be adapted to reflect this. The first method we considered takes the dataset as is and performs a global binary binning on all the continuously-valued features. For instance, in the dataset the weight feature varies between 40kg and 191kg with a median of 94kg; so, all values greater than the median are labelled as *high-weight* and all those equal-to-or-below the median are labelled as *low-weight* (see Table 3). This binning step is computed at the outset for the dataset. Using the CB2-CF method, for a given query *k* counterfactual-candidates are produced (assume *k*=20) and the difference-feature-values found in these candidates are all transformed using the binning-labels (obviously categorical-feature-differences are left as is). Note, after the categorical-transformation, some of these candidates will need to be removed because they are *conflicting*; that is, the continuous feature-values in difference-pairs are transformed into the “same” categorical feature (e.g., two weights of 100kg and 115kg which were a difference both become labelled as *high-weight*). This means that, after the categorical transformation, the original counterfactual has not been preserved appropriately; therefore, these conflicting candidates are removed. Indeed, as we shall see, this step is probably the main source of performance decrements for this method. See ***Algorithms*** for the steps in this

global binning method, called CAT-CBR^{global}. After the conflicting counterfactuals have been removed, it sorts the candidates by sparsity (lowest to highest) and then by distance (i.e., ℓ_2 norm on original untransformed, values of the counterfactual; lowest to highest) choosing the one with the best sparsity and distance score.

Table 3. A query case with its paired explanatory case, before and after it is transformed using the Global Categorical (CF-CAT^{global}) and Local Categorical (CF-CAT^{local}) methods.

Features	**Query Case**	**Original Explanatory Counterfactual**	**Global Categorical Transformation**	**Local Categorical Transformation**
Weight	90 kg	***100 kg***	***high***	***higher***
Duration	1 hr	***1.5 hrs***	***high***	***higher***
Gender	Female	Female	Female	Female
Stomach	Empty	Empty	Empty	Empty
Units	5	5	5	5
BAC Level	Over	Under	Under	Under

4.3 Counterfactuals With Categorical Transforms #2: Local Direction

On the face of it, the CAT-CF^{global} method looks like a plausible solution to the problem of transforming continuous features into categorical ones for psychologically comprehensible counterfactual explanations. However, we have also seen that it can produce conflicts when a continuous feature-difference is not preserved after categorical transformation, which may limit its potential to produce counterfactuals for certain query cases, depending on how a given feature's values are distributed. Hence, we developed and tested a more local method, called CAT-CF^{local}. This method, as its name suggests, works locally within the candidate counterfactual by re-labelling the continuous values in *c(p, p')* as being higher/lower, depending on the direction of difference. For instance, if the *p* query had a value of 110kg and the candidate *p'* counterfactual instance has a value of 120kg, then the former would be labelled *lower-weight* and the latter *higher-weight* (and vice versa if the weight values were reversed; see Table 3). This approach avoids the global binning of values (and conflicts that arise) using instead a more relative transformation; it tells the user that one feature was significantly higher/lower than the other and the direction of the difference that produced the counterfactual outcome. Arguably, this method is simpler and easier to compute, though it does give users more relativistic explanations (e.g., people will not know whether *higher* is *high* in some absolute sense, just that the value is high relative to the paired case). The method also prioritises counterfactuals with relative difference-features that are most representative of the set of potential counterfactuals, by assigning each candidate counterfactual a *direction-consistency* score. For example, where there are 20 candidate counterfactuals for a given query case, a certain difference-feature may be relatively *higher* in 15 of these candidates and relatively *lower* in 5. The proportion of candidates with each relative categorical value for that feature is calculated (e.g., .75 for *higher*, .25 for *lower*), and the mean direction-consistency score for each candidate is obtained by averaging this score for all difference-features. The candidates are ordered in terms of (i) sparsity (lowest to highest), and then by (ii) direction (highest to lowest), selecting the one with the best sparsity and direction-consistency score. In the next section, we report our tests

of these two methods, implementing variants of the CB2-CF method to test each on those datasets that we saw to be most important in Study 1.

5. Study 2: Evaluating CAT-CF Methods

We evaluated the performance of the two methods described above using seven UCI datasets [43] and in comparison to two baseline techniques (from [20]): (i) $CBR_{Proximity}$ selects the best candidate counterfactual using the ℓ_2 distance between the query and counterfactual candidate; and (ii) $CBR_{Sparsity\ \&\ Proximity}$ prioritises sparsity between the query and counterfactual case before selecting the most proximal candidate. All methods were evaluated using a *k*-NN model for varying *k*s, but here we report the results for k=20, as different values of k yield similar results.

5.1 Method: Data & Procedure

The evaluation datasets vary in the number of features, classes, and overall size (see Table 2), but all contain some categorical features and are used in classification tasks. In order to evaluate the two CAT-CF approaches against baseline CBR methods, we focus on two metrics: (i) *explanation competence* or *coverage*, that is, the proportion of query cases for which at least one counterfactual case can be generated; and (ii) *relative counterfactual distance*, that is, the ratio of distance between a query case and its selected counterfactual case, to the distance between the query case and explanation-case generated for the counterfactual (n.b., ℓ_2 is a standard measure used to evaluate counterfactual methods, with low distance seen as an indicator of better or more plausible explanations). For each dataset we used a tolerance of ±20% 1 SD for a given feature. Ten-fold cross-validation was used in evaluation, randomly selecting 10% of instances as query cases, and the remainder as the basis for the explanation cases. The means for each dataset across all 10 folds are reported here.

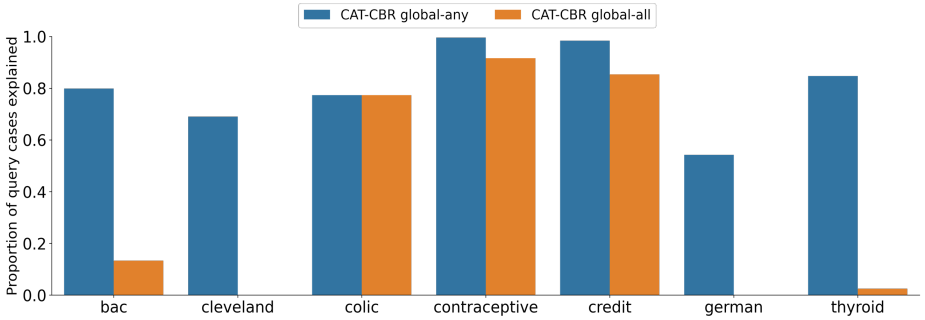

Fig. 3. Study 2 Coverage results: The explanatory competence of $CAT\text{-}CBR^{global\text{-}any}$ compared to $CAT\text{-}CBR^{global\text{-}all,}$ for seven datasets.

To determine the impact of the removal of conflicting-candidates in CAT-CF$^{\text{Global}}$, and its feasibility, we first compared a version of CAT-CF$^{\text{Global}}$ that requires *all* feature differences be preserved following the categorical-binning step (CAT-CF$^{\text{Global-all}}$) to one that relaxes this constraint to require that *at least one* feature difference is preserved (CAT-CF$^{\text{Global-any}}$). Figure 3 shows the explanatory competence of these variants, in which CAT-CF$^{\text{Global-all}}$ performs poorly on all datasets relative to the CAT-CF$^{\text{Global-any}}$ ($M_{\text{global-all}}$≈39%; $M_{\text{global-any}}$≈80%; z=-7.01, p<.001), failing completely on several. Clearly, the former is too conservative, so we adopt CAT-CF$^{\text{Global-any}}$ in subsequent tests. Though this means the global method will not necessarily transform all continuous feature-differences, a counterfactual explanation with one categorical difference-feature and one continuous difference-feature still holds a distinct psychological advantage over an explanation in which both difference-features are continuous.

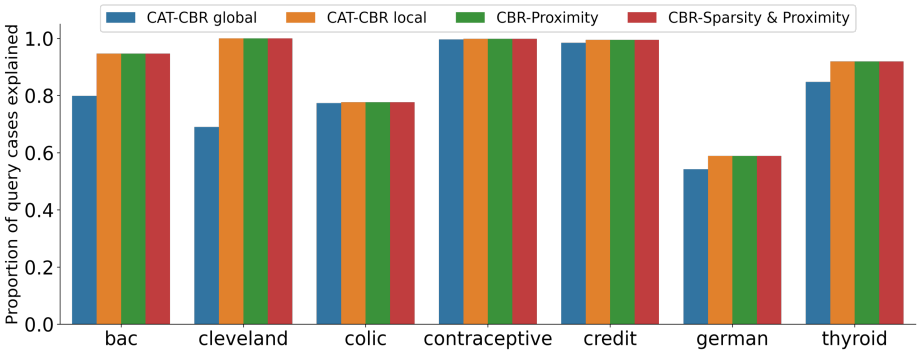

Fig. 4. Study 2 Coverage Results: The explanatory competence of both categorical methods in comparison to two baseline methods, across 7 datasets.

5.3 Results & Discussion: Counterfactual Distance

Having shown the explanatory competence of the two proposed counterfactual methods, we now move to assessing their explanatory power. *Relative counterfactual distance* (RCF) is used as proxy measure for the quality of generated counterfactuals and is computed by dividing the distance of generated counterfactual pairs, by the baseline distance between native-counterfactuals in the dataset. If RCF is <1, the generated counterfactuals are closer to the query case than the mean baseline distance. RCF for each of the methods and datasets are shown in Figure 5. Analyses with t-tests showed that both baselines (RCF$_{\text{Proximity}}$≈.93; RCF$_{\text{Sparsity \& Proximity}}$≈.96) performed better than CAT-CF$^{\text{Global}}$ (RCF$_{\text{global}}$≈1.05), t(70)=4.17, p<.001; t(70)=3.33, p<.001 respectively. There was no significant difference between the mean RCF of the counterfactuals produced by CAT-CF$^{\text{Global}}$ and CAT-CF$^{\text{Local}}$ (RCF$_{\text{local}}$≈.99), t(70)=1.91, p=.058 (note that this metric only captures the distance between those query-counterfactual pairs that were successfully produced, so the poorer coverage of CAT-CF$^{\text{Global}}$ is not accounted for here). CAT-CF$^{\text{Local}}$ does not score as well as CBR-Proximity, t(70)=2.54, p=.012, but is not significantly different to CBR-Sparsity & Proximity, t(70)=1.65, p=.09. This suggests that selecting the best counterfactual by prioritising direction-consistency does not sacrifice similarity any more than prioritising sparsity, which is widely accepted to be psychologically important.

Overall, from these results it is clear we can be confident about transforming features into more psychologically-acceptable variants using CAT-CF^{Local}; though, there is a slight hit on the distance measure, this decrement should be compensated for by the improved psychological comprehensibility of the explanations generated using this method. Depending on the task context, as well as the domain-knowledge and the goals of users, their requirements of an explanation are likely to vary [6, 44]. For example, in applications where the user aims to develop a general understanding of how features affect a system's decision (e.g., auditing system fairness), explanations focused on categorical features are likely to be highly effective. They may also be appropriate where there are concerns regarding model extraction or breaches of sensitive personal data, where it may be desirable to avoid explicitly providing raw data points. Even where users require or request more precise information, the categorical explanations proposed here can be easily supplemented with reference to the original data; bearing in mind that developing a basic understanding of the features and how they contribute to a system's decision is clearly a fundamental first step towards actionable recourse.

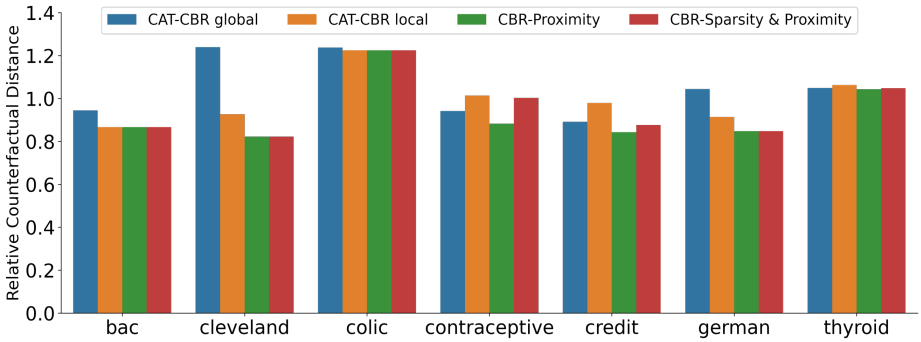

Fig 5. Study 2 Distance Results: The counterfactual distance of good counterfactuals produced for 7 datasets, relative to baseline counterfactual distance (between query case and its NUN)

6. Conclusions

In recent years, the XAI literature has been replete with many counterfactual methods that claim to generate plausible explanations based on the "right" features [15] but with little or no psychological evidence to support the claims made. Recent studies have shown that people learn from counterfactuals involving categorical features rather than those using continuous features [22], which motivates the case-based, counterfactual method proposed here; it produces categorical counterfactuals by transforming continuous features into categorical alternatives. We have tested two variants of this transforming approach and found that CAT-CF^{Local}, which performs local transformations, works well on coverage and relative distance measures, compared to standard non-transforming methods. This means we can retain the benefits of current methods but boost them psychologically with categorical transformations of their proposed explanations. The main novelty of this work is that it is the first counterfactual method that has been specifically designed to meet identified psychological requirements of end-users, rather than merely reflecting the intuitions of algorithm designers.

7. Acknowledgments

This research was supported by (i) the UCD Foundation, (ii) UCD Science Foundation Ireland via the Insight SFI Research Centre for Data Analytics (12/RC/2289) and (iii) the Department of Agriculture, Food and Marine via the VistaMilk SFI Research Centre (16/RC/3835)

References

1. Gunning, D., & Aha, D. W.: DARPA's Explainable Artificial Intelligence Program. AI Magazine, 40(2), 44-58 (2019)
2. Adadi, A. and Berrada, M.: Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). IEEE Access, 6, 52138-52160 (2018)
3. Miller, T.: Explanation in artificial intelligence. Artificial Intelligence, 267, 1-38 (2019)
4. Goodman, B. and Flaxman, S.: European Union regulations on algorithmic decision-making and a "right to explanation". AI Magazine, 38(3), 50-57 (2017)
5. Leake, D. and McSherry, D.: Introduction to the special issue on explanation in case-based reasoning. Artificial Intelligence Review, 24(2), 103-108 (2005)
6. Sørmo, F., Cassens, J. and Aamodt, A.: Explanation in case-based reasoning–perspectives and goals. Artificial Intelligence Review, 24(2), 109-143 (2005)
7. Schoenborn, J. M., & Althoff, K. D. Recent Trends in XAI: In Case-Based Reasoning for the Explanation of intelligent systems (XCBR) Workshop (2019).
8. Kenny, E.M. and Keane, Mark T.: Twin-systems to explain neural networks using case-based reasoning. IJCAI-19, pp. 326-333 (2019)
9. Keane, M. T., & Kenny, E. M.: How case-based reasoning explains neural networks. In Bach, K. and Marling, C.(eds.). In ICCBR-19, Springer, Berlin (2019).
10. Kenny, E.M. and Keane, Mark T.: Explaining deep learning using examples: Optimal feature weighting methods for twin systems using post-hoc, explanation-by-example in XAI, Knowledge-Based Systems, 233, 1-14, 107530 (2021)
11. Nugent, C., and Cunningham, P.: Gaining insight through case-based explanation. Journal of Intelligent Information Systems, 32(3), 267-295 (2009)
12. Cummins, L. and Bridge, D.: KLEOR: A knowledge lite approach to explanation oriented retrieval. Computing and Informatics, 25(2-3), 173-193 (2006).
13. Kenny, E.M. and Keane, Mark T.: On generating plausible counterfactual and semi-factual explanations for deep learning. AAAI-21, 11575-11585 (2021)
14. Martens, David, and Provost, F.: Explaining data-driven document classifications. Mis Quarterly, 38, 73-100 (2014)
15. Keane, M.T., Kenny, E.M., Delaney, E. and Smyth, B.: If only we had better counterfactual explanations. In IJCAI-21, pp. 4466-4474 (2021)
16. Karimi, A-H., Barthe, G., Schölkopf, B. and Valera, I.: A survey of algorithmic recourse. arXiv preprint arXiv:2010.04050 (2020).
17. Byrne, R.M.J.: Counterfactuals in explainable artificial intelligence (XAI): evidence from human reasoning. In IJCAI-19, pp. 6276-6282 (2019)
18. Wachter, S., Mittelstadt, B. and Russell, C.: Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harv. JL & Tech., 31, p.841 (2018)
19. Keane, M. T. and Smyth, B.: Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable AI (XAI). In ICCBR-20, pp. 163-178. Springer (2020).
20. Smyth, B. and Keane, M.T.: A few good counterfactuals: Generating interpretable, plausible and diverse counterfactual explanations. In ICCBR-22, Springer, Berlin (2022)

21. Wexler, J., Pushkarna, M., Bolukbasi, T., Wattenberg, M., Viégas,F., & Wilson, J.: The what-if tool: Interactive probing of machine learning models. IEEE TVCG, 26(1), 56-65 (2019)
22. Warren, G., Keane, M.T. and Byrne, R.M.J.: Features of explainability: How users understand counterfactual and causal explanations for categorical and continuous features in XAI. In IJCAI-22 Workshop on Cognitive Aspects of Knowledge Representation (2022)
23. Nugent, C., and Cunningham, P.: A case-based explanation system for black-box systems. Artificial Intelligence Review, 24(2), 163-178 (2005)
24. Kumar, R.R., Viswanath, P. and Bindu, C.S.: Nearest neighbor classifiers: a review. Int. J. Comput. Intell. Res., 13(2), pp.303-311 (2017)
25. Aggarwal, C. C., Chen, C., & Han, J. (2010). The inverse classification problem. Journal of Computer Science and Technology, 25(3), 458-468.
26. Laugel, T., Lesot, M. J., Marsala, C., Renard, X., & Detyniecki, M.: The dangers of post-hoc interpretability. In: IJCAI-19. 2801-2807 (2019)
27. Mothilal, R.K., Sharma, A. and Tan, C.: Explaining machine learning classifiers through diverse counterfactual explanations. In FAT*20, pp. 607-617 (2020)
28. Van Looveren, A. and Janis Klaise, J.: Interpretable counterfactual explanations guided by prototypes. In EMCL PKDD-19, pp. 650-665. Springer, Cham. (2019)
29. Russell, C.: Efficient search for diverse coherent explanations. In FAT-19, pp. 20-28 (2019).
30. Kahneman, D. and Miller, D.T.: Norm theory. Psychological Review, 93(2), 136-153 (1986)
31. Ustun, B., Spangher, A. and Liu, Y.: Actionable recourse in linear classification. In FAT-19, pp. 10-19 (2019)
32. Karimi, A.H., Barthe, G., Balle, B. & Valera, I.: Model-agnostic counterfactual explanations for consequential decisions. In: AISTATS-20. Palermo, Italy. PMLR: Volume 108 (2020)
33. Wiratunga, N., Wijekoon, A., Nkisi-Orji, I., Martin, K., Palihawadana, C. and Corsar, D.: Actionable feature discovery in counterfactuals using feature relevance explainers. CEUR Workshop Proceedings (2021).
34. Karimi, A.H., von Kügelgen, J., Schölkopf, B. & Valera, I.: Algorithmic recourse under imperfect causal knowledge. In NeurIPS-20, 33 (2020)
35. Ramon, Y., Martens, D., Provost, F. and Evgeniou, T.: A comparison of instance-level counterfactual explanation algorithms for behavioral and textual data: SEDC, LIME-C and SHAP-C. Advances in Data Analysis and Classification, 14(4), 801-819 (2020)
36. Delaney, E., Greene, D. and Keane, M.T.: Instance-based counterfactual explanations for time series classification. In ICCBR-21, pp. 32-47. Springer, Cham. (2021)
37. Dodge, J., Liao, Q.V., Zhang, Y., Bellamy, R.K. & Dugan, C.: Explaining models: An empirical study of how explanations impact fairness judgment. In: IUI-19. pp. 275-285 (2019)
38. Lucic, A., Haned, H. and de Rijke, M.: Contrastive local explanations for retail forecasting. In FAT*20, pp. 90-98 (2020).
39. Van der Waa, J., Nieuwburg, E., Cremers, A. and Neerincx, M.: Evaluating XAI: A comparison of rule-based and example-based explanations. Artificial Intelligence, 291, (2021)
40. Lage, I., Chen, E., Jeffrey He, J., Narayanan, M., Kim, B., Gershman, S.J. and Doshi-Velez, F.: Human evaluation of models built for interpretability. In HCOMP-19, pp. 59-67 (2019).
41. Kirfel, L. and Liefgreen, A.: What if (and how...)? Actionability shapes people's perceptions of counterfactual explanations in automated decision-making. In ICML-21 Workshop on Algorithmic Recourse (2021).
42. Kahneman, D. and Tversky, A.: The simulation heuristic. In D. Kahneman, P. Slovic, A. Tversky (Eds.), Judgment Under Uncertainty: Heuristics and Biases, pp. 201–8. CUP (1982).
43. Dua, D. and Graff, C.: UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science. (2019)
44. Keil F. C.: Explanation and understanding. Annual Rev. Psychol., 57, pp.227–254 (2006)
45. Förster, M., Klier, M., Kluge, K. and Sigler, I.: Evaluating explainable artificial intelligence: What users really appreciate. In ECIS-2020 (2020)