# Good Counterfactuals and Where to Find Them: A Case-Based Technique for Generating Counterfactuals for Explainable AI (XAI)

Mark T. Keane[1,2,3] and Barry Smyth[1,2](✉)

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

**Abstract.** Recently, a groundswell of research has identified the use of counterfactual explanations as a potentially significant solution to the Explainable AI (XAI) problem. It is argued that (i) *technically*, these counterfactual cases can be generated by permuting problem-features until a class-change is found, (ii) *psychologically*, they are much more causally informative than factual explanations, (iii) *legally*, they are GDPR-compliant. However, there are issues around the finding of "good" counterfactuals using current techniques (e.g. *sparsity* and *plausibility*). We show that many commonly-used datasets appear to have few "good" counterfactuals for explanation purposes. So, we propose a new case-based approach for generating counterfactuals, using novel ideas about the *counterfactual potential* and *explanatory coverage* of a case-base. The new technique reuses patterns of good counterfactuals, present in a case-base, to generate analogous counterfactuals that can explain new problems and their solutions. Several experiments show how this technique can improve the counterfactual potential and explanatory coverage of case-bases that were previously found wanting.

**Keywords:** CBR · Explanation · XAI · Counterfactuals · Contrastive

## 1 Introduction

In recent years, there has been a tsunami of papers on Explainable AI (XAI) reflecting concerns that recent advances in machine learning may be limited by a lack of transparency (see e.g., [1, 2]) or by government regulation (e.g., GDPR in the EU, see [3, 4]; for reviews [5–7]). Historically, Case-Based Reasoning (CBR) has always given a central role to explanation, as predictions can readily be explained by cases, akin to human reasoning from precedent/example [8–12]). Indeed, Kenny & Keane's [13, 14] *twin systems approach*, explicitly maps black-box deep-learning systems into CBR systems to find *post-hoc* explanatory cases for their predictions. Typically, CBR uses "factual cases"; nearest *like* neighbors that explain why a prediction was made [14]. But, recently, another class of explanatory cases is attracting interest, *counterfactual cases*; nearest *unlike* neighbors that explain how a prediction might be changed. For example, a loan

I. Watson and R. Weber (Eds.): ICCBR 2020, LNAI 12311, pp. 163–178, 2020.
https://doi.org/10.1007/978-3-030-58342-2_11

application system might explain its decision to refuse a loan by presenting a factual case: "*you were refused because a previous customer had the same salary as you and they were refused a loan for this amount"*. In contrast, the same loan system might, arguably, provide a better explanation by presenting a counterfactual case; effectively saying "*if you asked for a slightly lower amount you would have been granted the loan*". Researchers championing the use of counterfactual explanations, argue that they provide better solutions to the XAI problem [7, 15–18] (see Sect. 2).

In this paper, we consider counterfactual explanations from a CBR perspective. Though any CBR system can explain its predictions directly using counterfactual cases, here, we assume a twin-system context [13, 14]; where some opaque machine-learning model (e.g., deep learning model) generating predictions to be explained by finding case-based explanations from a twinned CBR[1]. We assess how many "good" counterfactuals are available in a given case-base (i.e., ones that are easily comprehended by people). So, we systematically map the topology of "good" counterfactuals in different case-bases, what we call their *counterfactual potential* (see Sect. 2). Initially, we perform an analysis of 20 frequently-used case-bases from the ML/CBR literature (see Sect. 3). To presage our results, to our surprise, we find that in most case-bases "good" counterfactuals are quite rare. This leads us to the novel notion of *explanatory coverage* by analogy to *predictive coverage* [19–21], from which we develop and evaluate a new case-based technique for counterfactual generation in XAI (Sects. 4 and 5).

## 2 Counterfactual Explanation: Promise, Problems and Prospects

Intuitively, counterfactual explanations seem to provide better explanations than factual ones; nearest-unlike-neighbor (NUN) explanations are better than nearest-like-neighbor (NLN) explanations[2]. Imagine you using the drink-&-drive app, DeepDrink, that can predict whether you are under/over the alcohol limit for driving. DeepDrink knows your physical profile and when you tell it (i) how many drinks you have taken, (ii) your recent food intake and (iii) when you started drinking, it predicts you are *over the limit* explaining it with a *factual case*; saying that a person with a similar profile to you was also over the limit when they were breathalysed (see Table 1). This explanation is reasonable but perhaps less informative than a *counterfactual case*; which would tell you that someone with your profile who drank a similar amount over a longer period, ended up being under the limit (see "good" counterfactual in Table 1). The counterfactual directly tells you more about the causal dependencies in the domain and, importantly, provides you with "actionable" information (i.e., if you stopped drinking for 30 min you could be under the limit). Technically, counterfactuals can tell you about the feature

[1] This context assumes an existing (albeit opaque) model to which cases can be presented to find predictions/labels; all counterfactual-generation techniques make this assumption, though there is some discussion around whether the training data would also always be accessible (obviously, we assume the training-data/case-base is available).

[2] Though NUNs have been studied in CBR (e.g., [22, 23]), few consider counterfactual cases (aka NUNs) for explanation; [24, 25] are exceptions but they viewed NUNs as being more important as confidence indicators with respect to decision boundaries.

differences that affect the decision boundary around a prediction. Accordingly, [20] define counterfactual explanations as statements taking the form:

*Score y was returned because variables V had values (v1, v2, ..). If V had values (v1′, v2′ …), and all others remain constant, score y′ would have been returned.*

where, in our example, score $y$ would be the class "over the limit" and $y'$ the class "under the limit". Recently, researchers championing counterfactual cases for XAI have argued that psychologically, technically and legally they provide better explanations than other techniques for XAI [7, 16, 17, 26–29].

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

| Features | Query case | "Good" Counterfactual | "Bad" Counterfactual |
|---|---|---|---|
| *Weight* | **80 kg** | 80 kg | 80 kg |
| *Duration* | **1 h** | ***1.5 h*** | ***3 h*** |
| *Gender* | **Male** | Male | ***Female*** |
| *Meal* | **Empty** | Empty | ***Full*** |
| *Units* | **6** | 6 | ***6.5*** |
| *Bac Level* | **Over** | Under | Under |

## 2.1 Counterfactual Promise

Many have argued that counterfactual thinking has a promising role to play in explanation from philosophical, psychological, computational and legal perspectives. Philosophers of science have argued that true causal explanation only emerges from contrastive propositions, using counterfactuals [30, 31]. Psychologists have also shown that counterfactuals play a key role in human cognition and emotion, eliciting spontaneous causal thinking about what might have been the case [15, 16, 32]. Byrne [16] has explicitly related this literature to the XAI problem, laying out the different ways in which counterfactuals could be used (see also [7, 33]). For example, as counterfactuals engender more active causal thinking in people, they are more likely to facilitate "human in the loop" decision making [16]. Recently, Dodge et al. [34] assessed explanations of biased classifiers using four different explanation styles and found counterfactual explanations to be the most effective. In AI, Pearl [27] has proposed an influential structural Bayesian approach to counterfactuals that can test the fairness of AI systems, but it has been less used in explanation generation (e.g., see [35, 36]). In the XAI literature, the use of counterfactuals has been used to counter popular post-hoc perturbation approaches (e.g., LIME; [37, 38]), with many researchers arguing that counterfactuals provide more robust and informative post-hoc explanations [18, 26, 38–40]; these "counterfactualists" have also argued that counterfactual explanations are GDPR compliant [4, 39].

## 2.2 Counterfactual Problems

However, the promise of counterfactuals for XAI comes with a number of problems; the three main ones being prolixity, sparcity and plausibility.

**Prolixity.** Currently, most XAI systems generate counterfactuals using random perturbation and search, making them somewhat *prolix* [4, 17]; that is, many counterfactuals may be produced for a given prediction from which a "good" one must be filtered (e.g., in the loan system, one could be shown counterfactuals for every $10 incremental change in one's salary). Stated simply, this prolixity is handled by filtering counterfactuals on the minimally-changed features to the query case that flip the prediction (i.e., the nearest unlike neighbor). So, [20] propose the following loss function, $L$:

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

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

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 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 L1 norm weighted by median absolute deviation (MAD). This technique claims to find minimally-mutated counterfactuals, solving the prolixity problem (see [17, 39, 40] and [39] for *diversity* between counterfactuals).

**Sparcity.** These methods also profess to solve the *sparcity* problem. All commentators argue that good explanatory counterfactuals need to be *sparse*; that is, they need to modify the *fewest* features of the query case. For example, Table 1 shows, for the blood alcohol domain, two different counterfactuals, one with a 1-feature change and another with a 4-feature change, with the sparcity of the former making it better than the latter. Wachter et al. [4] argue that the L1 norm delivers sparse counterfactuals, though many of these appear to still involve high numbers of feature-differences (e.g., >4, see [40]). Importantly, the argument for sparcity is a psychological one that has not been specifically tested in the XAI literature. Typically, AI researchers propose sparcity is important because of human working memory limits [41, 42], but we argue that people prefer sparse counterfactuals because of limits on human category learning. For example, [43] have shown that when people are learning categories for unfamiliar items they prefer single-feature changes between to-be-learned items over multiple-feature changes, because it makes the learning task easier (unless there is additional domain knowledge on dependencies between features). Based on this evidence, we operationalize the *sparcity* of "good" counterfactuals (as items with 1 or 2 feature differences) versus "bad" counterfactuals (those with >2 feature changes). This definition helps us develop the novel idea of the *counterfactual potential* of case-bases, based on quantifying the "good" counterfactuals they contain (see Sects. 4 and 5).

**Plausibility.** The final problem is that of *plausibility*; that is, the counterfactuals generated by these methods may not be valid data-points in the domain or they may suggest feature-changes that are difficult-to-impossible. Classic examples of such counterfactuals in loan decisions, are explanations that propose increasing one's salary by an implausible amount (i.e., *if you earned $1M, you would get the loan*) or radically altering oneself (i.e., *if you changed gender, you would get the loan*). Plausibility is the least-solved problem facing counterfactual generation; many researchers propose to "lock" features (e.g., not allow *gender* change) or to get users to provide inputs on feature weights [40] (e.g., using interface sliders on salary boundaries). However, automated solutions to the plausibility problem are thin on the ground[3]. Our proposal is to *directly* generate counterfactuals analogically from the dataset, rather than producing them by "blind", random perturbation followed by filtering. As counterfactuals generated in this way are based on "real experiences" in the problem domain, they should be inherently plausible. However, this raises another question: namely, how many good counterfactuals "naturally" occur in case-bases, what is their *counterfactual potential*.

### 2.3 CBR's Prospects for Counterfactuals

Most techniques for generating counterfactuals for XAI perform random perturbations of a query case followed by a search to find minimally-different items that are close to the decision boundary (i.e., a NUN). These perturbation techniques can encounter problems, notably in meeting *sparcity* and *plausibility,* which may benefit from a case-based approach. Just as CBR has successfully explained predictions using factual cases [10, 25], perhaps it can also deliver counterfactual cases that are *sparse* and *plausible*. However, if CBR is to be used, we need to establish whether case-bases/datasets actually contain good counterfactuals, whether they have high counterfactual potential. We define a *good counterfactual* to be a NUN that differs from the query case by no more than 2 features. So, *counterfactual potential* can be computed from the feature-differences for all pairwise comparisons of cases in the case-base. If these comparisons find many "good" counterfactuals then the potential is high, if not then it is low. So, in our first experiment, we computed the counterfactual potential of 20 classic ML/CBR datasets, from the UCI repository [45]. From this analysis we develop the idea of *explanatory coverage* before proposing a novel case-based technique for counterfactual generation (Sect. 4). Finally, in Sect. 5, we report a set of experiments on five representative datasets to show how our technique can enhance counterfactual potential.

## 3 Experiment 1: Plotting Counterfactual Potential

In this experiment, we computed the counterfactual potential of 20 classic datasets from the UCI repository [45], ones that have been commonly used in many CBR papers. This analysis was done by computing the number of feature differences between all pairwise comparisons of cases in the case-base, noting the proportion of "good" counterfactuals found (i.e., $\leq 2$ feature difference counterfactuals). This analysis provides us with an

[3] Rare recent attempts include Laugel et al.'s [44] method to "justify" generated counterfactuals using nearest neighbors in the training data, and [29] finding "feasible paths" to counterfactuals in the dataset; both methods attempt to *ground* counterfactuals in prior experience.

upper/lower bound on the potential of a case-base to deliver good counterfactuals. Obviously, in any specific CBR system, one might be able to adjust weights, how features are matched or *k*-values to find such counterfactuals, but such fine-tuning will not improve matters hugely if good counterfactual-cases are just not there.

### 3.1 Method: Data Sets and Procedure

Twenty UCI datasets were used in the experiment, selected on the basis of their common usage in CBR. We compared all pairings of query cases (one side of a decision boundary) to training cases (on the other side of a decision boundary) calculating the number of feature differences found in each.

### 3.2 Results and Discussion

Table 2 shows the counterfactual potential of the UCI datasets, as the percentage of counterfactuals from 1 to >5 feature-differences. The results show that "good" counterfactuals are rare[4]; in nearly every dataset, the 1-diff and 2-diff counterfactual categories

**Table 2.** Percent counterfactuals for feature-differences in 20 UCI datasets (expt.1)

| DataSets | N of cases | Feat. no. | Class no. | N of pairs | 1-diff | 2-diff | 3-diff | 4-diff | >5-diff |
|---|---|---|---|---|---|---|---|---|---|
| Abalone | 4177 | 10 | 8 | 15.6M | **0%** | **0%** | 0% | 0% | 99.9% |
| Auto MPG | 398 | 8 | 5 | 52.3k | **0%** | **0%** | 0% | 0.4% | 99.6% |
| BAC | 9291 | 7 | 2 | 19M | **0%** | **1.5%** | 23% | 3% | 72% |
| Bupa liver | 345 | 6 | 2 | 29k | **0%** | **0%** | 0.1% | 3.1% | 96.8% |
| Credit | 653 | 15 | 2 | 105.7k | **0%** | **0%** | 0% | 0% | 99.9% |
| Cleveland heart | 303 | 13 | 5 | 32.9k | **0%** | **0%** | 0% | 0.1% | 99.9% |
| Ecoli | 336 | 7 | 7 | 41k | **0%** | **0%** | 0% | 0.2% | 99.8% |
| Glass | 214 | 9 | 7 | 21.9k | **0%** | **0%** | 0% | 0% | 99.9% |
| German credit | 914 | 20 | 2 | 177k | **0%** | **0%** | 0% | 0% | 99.9% |
| Horse colic | 300 | 22 | 2 | 20.8k | **0%** | **0%** | 0% | 0% | 99.9% |
| Indian liver | 583 | 10 | 2 | 69.5k | **0%** | **0%** | 0% | 0% | 99.9% |
| Ionosphere | 351 | 34 | 2 | 28.3k | **0%** | **0%** | 0% | 0% | 100% |
| Iris | 150 | 4 | 3 | 7.5k | **0%** | **0.3%** | 8.8% | 91% | n/a |
| Sonar | 208 | 60 | 2 | 10.8k | **0%** | **0%** | 0% | 0% | 100% |
| Soybean (large) | 307 | 26 | 19 | 43k | **0%** | **0%** | 0.2% | 0.6% | 99.2% |
| Thyroid | 2800 | 27 | 3 | 355.8k | **0%** | **0%** | 0% | 0% | 99.9% |
| Votes | 435 | 17 | 2 | 44.8k | **0%** | **0.3%** | 0.9% | 1.9% | 88.8% |
| Wine-Italian | 178 | 13 | 3 | 10.4k | **0%** | **0%** | 0% | 0% | 100% |
| Wisconsin breast | 699 | 9 | 2 | 110k | **0%** | **0%** | 0% | 0.4% | 99.5% |
| Yeast | 1484 | 8 | 10 | 855.3k | **0%** | **0%** | 0.3% | 4.8% | 94.9% |

[4] We extensively tested this Blood Alcohol Content (BAC) case-base [24, 25], but cannot report it for reasons of space. Using a mechanical model for estimating BAC, we generated several master-case-bases from which we sampled 50+ specific case-bases; across all of these case-bases, to our astonishment, we repeatedly found the same absence of good counterfactuals.

account for $<1\%$ of total counterfactuals. Most of the counterfactuals found have poor sparcity (i.e., $>5$ feature-differences) and would likely be hard for people to understand.

It should be noted that in the above, we determine feature differences using an exact match. Such an approach is inherently conservative with real-valued features. In practice, a matching tolerance could be used by, for example, treating two feature-values as equivalent if they are within 1% of each other. While this tolerance-matching improves the results (albeit in a somewhat *ad hoc* fashion), the fraction of good counterfactuals ($\leq 2$ feature differences) still typically remains very low (see Sect. 5 for tests).

On the face of it, these results suggest that a case-based approach to counterfactual generation is a bad idea; if most datasets do not deliver good counterfactuals then case-based techniques seem bound to fail? However, as we shall see in the following sections, there are additional steps that can be used to meet and resolve this challenge.

## 4 A Case-Based Technique for Good Counterfactuals

Ironically, the above analysis suggests that CBR seems to have little to offer in using counterfactuals for XAI. For most case-bases good counterfactuals are rare, few query cases have associated good counterfactual cases. This may explain why the dominant counterfactual XAI techniques use perturbation, where synthetic counterfactuals are generated "blindly" from problem-cases and labelled using a machine-learning model, without reference to other known cases in the training set [18, 26, 38–40]. In contrast to these approaches, we believe that counterfactuals need to be explicitly grounded in known cases (*aka* the training data) to ensure plausibility. Hence, we developed a novel case-based technique for counterfactual-XAI which reuses patterns of good counterfactuals that already exist in a case-base, to generate analogous counterfactuals (as new datapoints) that can explain new target problems and their solutions. In generating new counterfactuals, these existing good counterfactuals provide 'hints' about what features can and should be adapted and plausible feature-values to use in them. This new technique relies on the notion of *explanatory competence* (see Sect. 4.1). Note, the context for the use of this method is a twin-system approach to XAI, where an opaque ML model is "explained" by twinning it with a more transparent CBR-system to find explanatory cases [13, 14]; hence, along with all other counterfactual-generation techniques, we assume an ML model is available to assign labels for any newly-generated synthetic case.

### 4.1 Explanatory Competence

The notion of *predictive competence* or simply *competence* (i.e., an assessment of an ML/CBR system's potential to solve a range of future problems) has proved to be a very useful development for AI systems [19–21]. For example, in CBR, predictive competence can assess the overall problem-solving potential of a system, to help avoid the utility problem as a case-base grows, to maintain case-bases and so on [19, 20]. A parallel notion of *explanatory competence* can also be applied to any case-base.

Just as the fundamental unit of (predictive) competence is a relation of the form *solves(c, c′)* to indicate that case/example *c* can be used to solve some target/query *c′*, the basic unit of explanatory competence is *explains(c, c′)* indicating that some case *c* can be used to explain the solution of *c′*; where the explanatory cases (*c*) are the counterfactuals of *c′*. So, the explanatory competence of a case-base *C* can be represented by a *coverage set* (Eq. 3) and explanatory competence can be estimated as the size of the coverage set as a fraction of the case-base (Eq. 4):

$$XP\_Coverage\_Set(C) = \{c' \in C \,|\, \exists c \in C - \{c'\} \,\&\, explains(c, c')\} \tag{3}$$

$$XP\_Coverage(C) = |XP\_Coverage\_Set(C)|/|C| \tag{4}$$

## 4.2 Leveraging Counterfactual Cases for Explanation

Although good counterfactuals are rare, in practice most case-bases should offer some examples where a query/problem-case can be associated with a good counterfactual, with or without some matching tolerance (as mentioned above). For example, in the Abalone dataset, even though there are few good counterfactuals (<1%), with a similarity tolerance of 0.02, ~20% of cases are found to have good counterfactuals; for the Liver dataset a tolerance of 0.025 results in ~4% of cases having associated good counterfactuals. Can these query-counterfactual case-pairs guide the search for novel (good) counterfactuals for new target problems that otherwise lack a good counterfactual?

Below, we refer to the pairing of a case and its corresponding good counterfactual as an *explanation case* (*XC*). For any given case-base, we can generate a corresponding case-base of these explanation cases for use during counterfactual generation; see Eqs. 5 and 6. By definition explanation cases are symmetric; either of the cases can be viewed as the query or counterfactual, which, in practice, means that each pair of unlike neighbours, which differ by ≤2 features, contributes two XCs to the XC case-base.

$$xc\left(c, c'\right) \Leftrightarrow class(c) \neq class(c') \,\&\, diffs(c, c') \leq 2 \tag{5}$$

$$XC(C) = \{(c, c') : c, c' \in C \,\&\, xc(c, c')\} \tag{6}$$

Each XC is associated with a set of *match-features (m)*, the features that are the same between the query and counterfactual (using a specified tolerance), and a set of *difference-features (d)*, the ≤2 features that differ between the query and counterfactual.

Figure 1(a) shows a two-class case-base of cases (*C*) with its corresponding *XCs* – *xc(x, x′)*, *xc(y, y′)*, and *xc(z, z′)* – along with two query cases (*p* and *q*), which have been classified by the underlying ML-model, and which now need to be explained. For our purposes, we assume that there are no existing good counterfactuals for *p* or *q* in *C*, hence the need to generate new good counterfactuals for them.

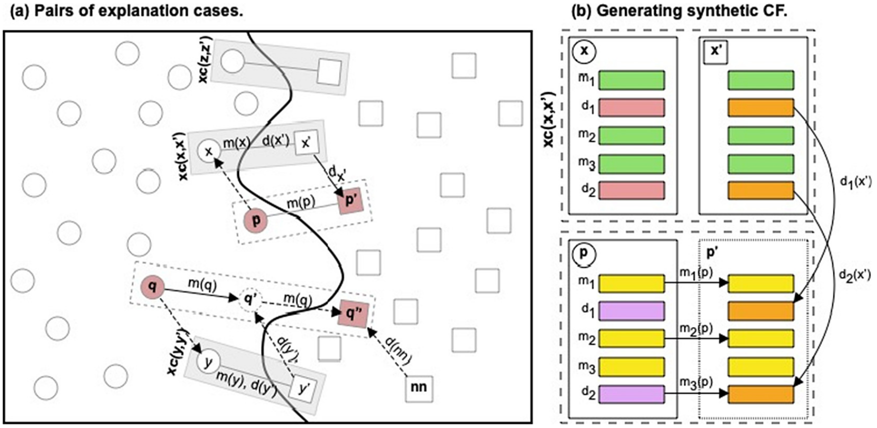

**Fig. 1.** An illustration of (a) a two-class case-base with 3 explanation cases; (b) how a synthetic counterfactual, *(p, p′),* is generated from an existing explanation-case, *xc(x, x′).*

### 4.3 A Case-Based Approach to Generating Good Counterfactuals

We propose a classical case-based reasoning approach to generating good counterfactuals by *retrieving*, *reusing*, and *revising* a nearby explanation case as follows:

1. First, we identify the XC case whose query is most similar to *p* while sharing p's class; this is *xc(x, x′)* in Fig. 1. Since *xc(x, x′)* has a good counterfactual, *x′*, and because the *p* is similar to *x*, then the intuition is that *x′* is a suitable basis for a new counterfactual *p′* to explain *p*. The *difference-features* between *x* and *x′*, which are solely responsible for the class change between *x* and *x′*, should play a critical role in constructing *p'*.
2. For each of the *match-features* in *xc(x, x′),* we copy the *values* of these features in *p* to the new counterfactual *p′*. Similarly, for each of the *difference-features* in *xc(x, x′)* we copy their *values* from *x′* into *p′*. In this way, *p′* is a combination of feature values from *p* and *x′*. It differs from *p* in a manner that is similar to the way in which *x′* differs from *x* and, by construction, *p′* is a *candidate good counterfactual* because these differences amount to no more than two features. This transfer of values from *p* and *x′* into *p′* is illustrated in Fig. 1(b).
3. For *p′* to be *actually a* good counterfactual, it has to be a different class from *p*, which is not yet guaranteed. We determine the class of *p′* by using the underlying ML-model (from the twin-system) and, if it is different from *p*, then *p′* can be used directly as a good counterfactual to explain *p* (see Fig. 1(a)).
4. Sometimes, however, the class of the new counterfactual, after retrieval/reuse, is not different from the target query. For example, the new counterfactual *q′*, which is generated for *q* by reusing *xc(y, y′)* in Fig. 1(a), has the same class as *q*, because the combination of the match-feature values (from *q*) and difference-features (from *y′*) are not sufficient to change its class from that of *q*.
5. Since *q′* is not a valid counterfactual, we perform an *adaptation* step to *revise* the values of the difference-features in *q′* until there is a class change; note, we cannot

change the match-features in $q'$ without increasing the number of feature differences with $q$. We can revise the values of the difference-features in $q'$ in various ways, for example, by perturbing them to further increase their distance from $q$. However, we instead iterate over the *ordered nearest neighbours* of $q$ with the same class as $y'$, until there is a class change[5]. The values of the difference-features from each *nearest neighbour* leads to a new candidate, $q''$, and adaptation terminates successfully when the class of *q”* differs from that of *q;* if none of the *neighbours* produce a class change, then adaptation fails. In Fig. 1(a), when the difference-feature values from the neighbour, *nn,* are used to produce $q''$, the result is a class change, and so $q''$ can be used as a good counterfactual for $q$.

Note that the primary contribution of explanation cases is to identify and distinguish between common combinations of features (match-features and difference-features) that tend to participate in good counterfactuals. Depending on the domain this may reflect important relationships (causal or otherwise) that exist within the feature-space. In other words, the XCs tell us about which features *should* be changed (or held constant) when generating new counterfactuals in the feature-space near a query case.

Another advantage of this approach is that, because it reuses *actual feature-values* from *real cases,* it should lead to more plausible counterfactuals and, better explanations. This contrasts with perturbation approaches, which rely on arbitrary values for features (and may even produce invalid data-points) and is consistent with approaches that try to ground counterfactuals in the training data [28, 44]. However, [28, 44] still use prior experience in a less direct way; they justify/link the generated counterfactual to known data-points rather directly using those data-points to directly create the counterfactual, as we do here. Notably, this method reminds one of analogical extrapolation methods in CBR [46] and structural analogical transfer [47, 48].

Finally, though our approach may succeed in finding a suitable counterfactual without the need for the adaptation/revision step, it may be desirable to proceed with this step, nonetheless. This is because the adaptation step has the potential to locate a suitable counterfactual that is *closer to the query* than the candidate counterfactual produced by the retrieval step alone and finding counterfactuals that are maximally similar to the query is an important factor when it comes to explanation [17].

## 5 Experiment 2: Evaluating Explanation Competence

A *preliminary* evaluation of the above approach was carried out using five popular ML/CBR datasets to demonstrate how explanatory competence can be improved over the baseline level of good counterfactuals naturally occurring in a dataset. For the ML model used to validate the generated counterfactuals, we used a *k*-NN model, to determine whether the predicted counterfactual class differs from the test/query case, but other classifiers could also be used if available.

[5] More generally, for multi-class datasets, this adaptation can be modified to iterate over all ordered nearest neighbours with a *different class* to $q$, not just those with the same class as $y'$. This provides a larger pool of difference-feature values and increase the likelihood of locating a good counterfactual for $q$.

### 5.1 Method: Data and Procedure

Each of the datasets represent a classification task of varying complexity, in terms of the number of classes, features, and training examples. The task of interest, however, is not a classification one but an explanation one. As such we are attempting to generate good counterfactuals in order to *explain* target/query cases and their classes. The key evaluation metrics will be: (a) the fraction of target/query cases than can be associated with good counterfactuals (*explanatory competence*); and (b) the distance from the target/query case to the newly-generated good counterfactual (*counterfactual distance*).

As a baseline for explanatory competence we use the fraction of cases that can be associated with a good counterfactual in each case-base. In each dataset we use a matching tolerance of 1–2% with normalized features and the Minkowski similarity metric was used throughout; variations in these settings will increase the fraction of existing and generated good counterfactuals and future work will need to explore such matters more completely. As a corresponding baseline for counterfactual distance, we use the average distance between these cases and their good counterfactuals. A 10-fold cross-validation was used to evaluate the newly-generated counterfactuals, selecting 10% of the cases at random to use as queries, and building the XC case-base from the remaining cases. Then, we use the above technique to generate good counterfactuals for the queries, noting the fraction of the queries that can be associated with good counterfactuals, and the corresponding counterfactual distances, after the retrieval/reuse and adaptation steps. Results reported are the averages for the 10 folds for each dataset.

### 5.2 Results and Discussion: Explanatory Competence

The explanatory competence results are presented in Fig. 2, showing the explanatory competence (fraction of queries that can be explained) for the dataset (baseline), and for the synthetic counterfactuals generated after the retrieval and adaptation steps of our approach. The results show how explanatory competence can be significantly increased by our case-based-counterfactual technique. For example, on average only about 11% of the cases in these datasets can be associated with good counterfactuals (the average baseline competence when a tolerance is applied) but by retrieving and re-using explanation cases we can reach an average explanatory competence of just over 40%. Implementing the adaptation step further increases the explanatory competence just under 94%, on average. Notably, even datasets with very low baseline explanatory competence benefit from significant improvements in explanatory competence particularly when the adaptation step is used. For example, the 6,400 case Wine dataset (12 features and 7 classes) has a baseline explanatory competence of just 6%, but its 559 XC-cases can be used to achieve almost 90% in explanatory competence.

### 5.3 Results and Discussion: Counterfactual Distance

Of course, just because it is possible to generate a counterfactual for a query that has no more than 2 feature-differences, does not necessarily mean that the counterfactual will make for an ideal explanation, in practice. To test this would require a succession of live user-trials (currently planned), that are beyond the scope of the present work. As a proxy

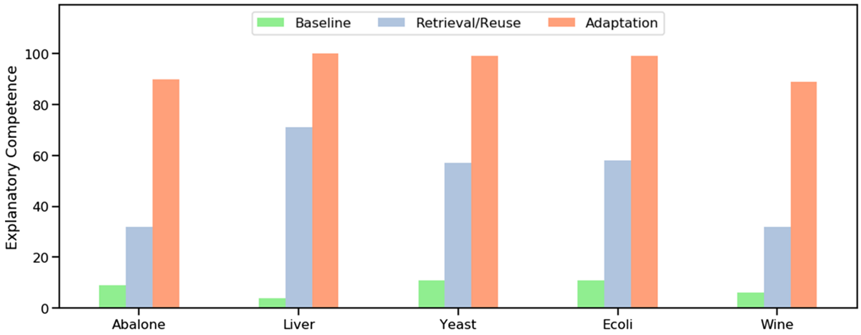

**Fig. 2.** The explanatory competence (XP_Coverage) of five case-bases/datasets, showing baseline competence and how competence increases by reusing and adapting explanation cases.

for the utility of the explanation, however, we can use the distance between the query and the generated counterfactual, on the grounds that counterfactuals which are closer to a query are more likely to serve as more useful explanations. Since counterfactual distance will vary from dataset to dataset, reflecting the nature of the feature space, we use a *relative counterfactual distance* (RCF) measure by dividing the counterfactual distances of the synthetic counterfactuals by the baseline counterfactual distance for the dataset. Thus, if RCF $>1$, then it indicates that the synthetic counterfactual is farther from the query that the average baseline counterfactual distance.

The results are presented in Fig. 3, which include the relative distance of the good counterfactuals produced by the retrieval/reuse and the adaptation steps for each dataset. We also show the relative distance results for an additional condition, *Closest*, which is defined as follows: when both the retrieval/reuse and adaptation steps lead to a good counterfactual, then choose the one with the lower counterfactual distance, otherwise if only one good counterfactual is produced then use its distance.

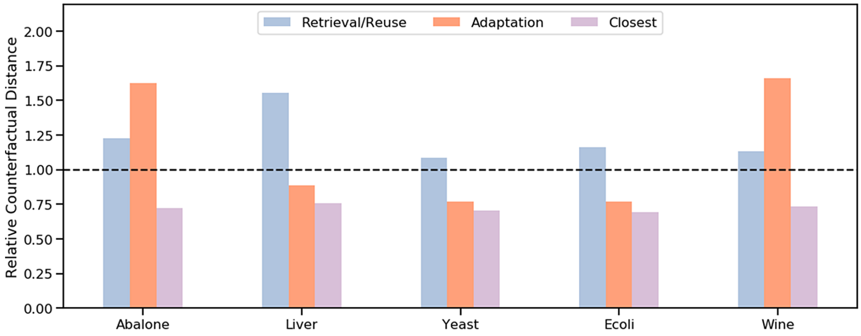

**Fig. 3.** The counterfactual distance of good counterfactuals produced for five case-bases/datasets, relative to the baseline counterfactual distance (between a query case and its counterfactual)

On average, good counterfactuals produced by the retrieval/reuse step are farther from the test query than the baseline counterfactual distance (RCF $\approx$ 1.2). In most cases the additional distance beyond the baseline is modest with the exception of the Liver dataset, where the retrieval/reuse step produces good counterfactuals that are 55% (RCF $\approx 1.55$) more distant from the query than the baseline distance. The good counterfactuals produced by the adaptation step are closer to the test queries – the average RCF $\approx 1.1$, and in 3 out of the 5 datasets the generated counterfactuals are closer than the baseline (RCF $< 1$). If we select the closest counterfactual, when both retrieval/reuse and adaptation produce one, then the RCF $< 1$ for all of the datasets. This further validates the need for, and quantifies the benefits of, the adaptation step: it provides an opportunity to choose a counterfactual that is significantly closer to the query.

## 6 Conclusions and Future Directions

In the last three years, there has been a significant upsurge in XAI research arguing for the computational, psychological and legal advantages of counterfactuals. Most of this work generates synthetic counterfactuals without reference to the training-data in the domain and, as such, can suffer from *sparsity* and *plausibility* deficits. In short, these methods do not guarantee the production of good counterfactuals and, may indeed, sometimes generate invalid data points. This state of affairs invites a case-based solution to counterfactual generation that leverages the prior experience of the case-base, adapting known counterfactual associations between query-problems and known cases. In this paper, we advance just such a technique and show how it can improve the counterfactual potential of many datasets. In developing this technique, we have (i) clarified the definition of good counterfactuals, (ii) proposed the new idea of explanation competence, (iii) reported significant new evidence for the utility of this novel technique.

This approach is *model agnostic*, in that it can operate with any underlying classifier (e.g., deep learner, decision tree, k-NN) once it has access to the features of training data, an agreed distance metric, and the dataset (see [49, 50] for a discussion of this issue). However, the approach makes some assumptions that might limit its utility beyond the datasets discussed. It assumes the availability of at least some explanation cases, which is typically feasible; even though good counterfactuals are rare they are seldom so rare as to exclude a minimally-viable explanation case-base, at least when a degree of matching tolerance is allowed for when computing feature similarities and differences; note, different degrees of matching tolerance, similarity metrics, and feature normalization strategies may have an impact on outcomes. The approach also assumes the availability of sufficiently accurate underlying ML-model (e.g., in a twin system) for the purpose of counterfactual validation, though this is an accepted assumption in all approaches. Finally, though previous psychological work supports our operational definition of good counterfactuals, more user testing is required; notions of *goodness* in general (see [2]) need to be squared with *sparcity goodness*. Notwithstanding this future research, from the current findings, it is clear that a CBR approach to counterfactuals has much to offer the explainable AI (XAI) problem.

**Acknowledgements.** This paper emanated from research funded by (i) Science Foundation Ireland (SFI) to the *Insight Centre for Data Analytics* under Grant Number 12/RC/2289_P2 and (ii) SFI and the Department of Agriculture, Food and Marine on behalf of the Government of Ireland to the *VistaMilk SFI Research Centre* under Grant Number 16/RC/3835.

## References

1. Gunning, D.: Explainable artificial intelligence (XAI). Defense Advanced Research Projects Agency (DARPA), Web, vol. 2 (2017)
2. Gunning, D., Aha, D.W.: DARPA's explainable artificial intelligence program. AI Mag. **40**(2), 44–58 (2019)
3. Goodman, B., Flaxman, S.: European Union regulations on algorithmic decision-making and a "right to explanation". AI Mag. **38**(3), 50–57 (2017)
4. Wachter, S., Mittelstadt, B., Floridi, L.: Why a right to explanation of automated decision-making does not exist in the general data protection regulation. Int. Data Priv. Law **7**(2), 76–99 (2017)
5. Adadi, A., Berrada, M.: Peeking inside the black-box: a survey on Explainable Artificial Intelligence (XAI). IEEE Access **6**, 52138–52160 (2018)
6. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM Comput. Surv. **51**(5), 93 (2018)
7. Miller, T.: Explanation in artificial intelligence. Artif. Intell. **267**, 1–38 (2019)
8. Leake, D.B.: CBR in context: the present and future. In: Case-Based Reasoning: Experiences, Lessons, and Future Directions, pp. 3–30 (1996)
9. Leake, D., McSherry, D.: Introduction to the special issue on explanation in case-based reasoning. Artif. Intell. Rev. **24**(2), 103–108 (2005)
10. Sørmo, F., Cassens, J., Aamodt, A.: Explanation in case-based reasoning–perspectives and goals. Artif. Intell. Rev. **24**(2), 109–143 (2005)
11. Schoenborn, J.M., Althoff, K.D.: Recent trends in XAI: In: Case-Based Reasoning for the Explanation of intelligent systems (XCBR) Workshop (2019)
12. Lipton, Z.C.: The Mythos of model interpretability. Queue **16**(3), 30 (2018)
13. Kenny, E.M., Keane, M.T.: Twin-systems to explain neural networks using case-based reasoning. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence, IJCAI 2019, pp. 326–333 (2019)
14. Keane, M.T., Kenny, E.M.: How case-based reasoning explains neural networks: a theoretical analysis of XAI using *Post-Hoc* explanation-by-example from a survey of ANN-CBR twin-systems. In: Bach, K., Marling, C. (eds.) ICCBR 2019. LNCS (LNAI), vol. 11680, pp. 155–171. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-29249-2_11
15. Byrne, R.M.J.: The Rational Imagination. MIT Press, Cambridge (2007)
16. Byrne, R.M.J.: Counterfactuals in explainable artificial intelligence (XAI): evidence from human reasoning. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, pp. 6276–6282 (2019)
17. Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box: automated decisions and the GDPR. Harv. J. Law Tech. **31**, 841 (2018)
18. Guidotti, R., Monreale, A., Giannotti, F., Pedreschi, D., Ruggieri S., Turini. F.: Factual and counterfactual explanations for black box decision making. IEEE Intell. Syst. **34**(6), 14–23 (2019)
19. Smyth, B., Keane, M.T.: Remembering to forget. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence, IJCAI 1995, pp. 377–382 (1995)
20. Smyth, B., McKenna, E.: Modelling the competence of case-bases. In: Smyth, B., Cunningham, P. (eds.) EWCBR 1998. LNCS, vol. 1488, pp. 208–220. Springer, Heidelberg (1998). https://doi.org/10.1007/BFb0056334
21. Juarez, J.M., Craw, S., Lopez-Delgado, J.R., Campos, M.: Maintenance of case-bases: current algorithms after fifty years. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, pp. 5457–5463 (2018)

22. Delany, S.J., Cunningham, P., Doyle, D., Zamolotskikh, A.: Generating estimates of classification confidence for a case-based spam filter. In: Muñoz-Ávila, H., Ricci, F. (eds.) ICCBR 2005. LNCS (LNAI), vol. 3620, pp. 177–190. Springer, Heidelberg (2005). https://doi.org/10.1007/11536406_16
23. Kumar, R.R., Viswanath, P., Bindu, C.S.: Nearest neighbor classifiers: a review. Int. J. Comput. Intell. Res. **13**(2), 303–311 (2017)
24. Cunningham, P., Doyle, D., Loughrey, J.: An evaluation of the usefulness of case-based explanation. In: Ashley, K.D., Bridge, D.G. (eds.) ICCBR 2003. LNCS (LNAI), vol. 2689, pp. 122–130. Springer, Heidelberg (2003). https://doi.org/10.1007/3-540-45006-8_12
25. Nugent, C., Cunningham, P.: A case-based explanation system for black-box systems. Artif. Intell. Rev. **24**(2), 163–178 (2005)
26. Mittelstadt, B., Russell, C., Wachter, S.: Explaining explanations in AI. In: Proceedings of Conference on Fairness, Accountability, and Transparency, FAT 2019 (2019)
27. Pearl, J.: Causality, Cambridge University Press, Cambridge (2000)
28. Sokol, K., Flach, P.: Desiderata for interpretability: explaining decision tree predictions with counterfactuals. In: AAAI 20119, Doctoral Consortium, pp. 10035–10036 (2019)
29. Poyiadzi, R., Sokol, K., Santos-Rodriguez, R., De Bie, T., Flach, P.: FACE: feasible and actionable counterfactual explanations. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pp. 344–350 (2020). https://doi.org/10.1145/3375627.3375850
30. Woodward, J.: Making Things Happen. Oxford University Press, Oxford (2003)
31. Van Fraassen, B.C.: The Scientific Image. Oxford University Press, Oxford (1980)
32. Kahneman, D., Miller, D.T.: Norm theory: comparing reality to its alternatives. Psychol. Rev. **93**(2), 136–153 (1986)
33. Mueller, S.T., Hoffman, R.R., Clancey, W.J., Emery, A.K., Klein, G.: Explanation in human-AI systems. Florida Institute for Human and Machine Cognition (2019)
34. Dodge, J., Liao, Q.V., Zhang, Y., Bellamy, R.K., Dugan, C.: Explaining models: an empirical study of how explanations impact fairness judgment. In: Proceedings of the 24th International Conference on Intelligent User Interfaces, pp. 275–285 (2019)
35. Miller, T.: Contrastive explanation. arXiv preprint arXiv:1811.03163 (2018)
36. Russell, C., Kusner, M.J., Loftus, J., Silva, R.: When worlds collide: integrating different counterfactual assumptions in fairness. In: Advances in Neural Information Processing Systems, pp. 6414–6423 (2017)
37. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should I trust you?. In: Proceedings of the 22nd ACM SIGKDD, pp. 1135–1144. ACM (2016)
38. Pedreschi, D., Giannotti, F., Guidotti, R., Monreale, A., Ruggieri, S., Turini, F.: Meaningful explanations of Black Box AI decision systems. In: Proceedings of AAAI 2019 (2019)
39. Mothilal, R.K., Sharma, A., Tan, C.: Explaining machine learning classifiers through diverse counterfactual explanations. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT 2020, pp. 607–617 (2020)
40. McGrath, R., et al.: Interpretable credit application predictions with counterfactual explanations. In: NIP Workshop on Challenges and Opportunities for AI in Financial Services, Montreal, Canada (2018)
41. Miller, G.A.: The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychol. Rev. **63**(2), 81 (1956)
42. Alvarez, G., Cavanagh, P.: The capacity of visual STM is set both by visual information load and by number of objects. Psychol. Sci. **15**, 106–111 (2004)
43. Medin, D.L., Wattenmaker, W.D., Hampson, S.E.: Family resemblance, conceptual cohesiveness, and category construction. Cogn. Psychol. **19**(2), 242–279 (1987)
44. Laugel, T., Lesot, M.J., Marsala, C., Renard, X., Detyniecki, M.: The dangers of post-hoc interpretability: unjustified counterfactual explanations. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence, IJCAI 2019, pp. 2801–2807 (2019)

45. Dua, D., Graff, C.: UCI Machine Learning Repository University of California, School of Information and Computer Science, Irvine, CA. http://archive.ics.uci.edu/ml (2019)
46. Lieber, J., Nauer, E., Prade, H.: Improving analogical extrapolation using case pair competence. In: Bach, K., Marling, C. (eds.) ICCBR 2019. LNCS (LNAI), vol. 11680, pp. 251–265. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-29249-2_17
47. Veale, T., Keane, M.T.: The competence of sub-optimal theories of structure mapping on hard analogies. In: International Joint Conference on Artificial Intelligence, pp. 232–237 (1997)
48. Keane, M.T.: Analogical asides on case-based reasoning. In: Wess, S., Althoff, K.D., Richter, M.M. (eds.) EWCBR 1993. LNCS, vol. 837, pp. 21–32. Springer, Heidelberg (1994). https://doi.org/10.1007/3-540-58330-0_74
49. Karimi, A.H., Barthe, G., Balle, B., Valera, I.: Model-agnostic counterfactual explanations for consequential decisions. In: Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, Palermo, Italy, vol. 108. PMLR (2020)
50. Sokol, K., Flach, P.: Explainability fact sheets: a framework for systematic assessment of explainable approaches. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT 2020, pp. 56–67 (2020)