Should counterfactual explanations always be data instances?

Junqi Jiang[0000−0002−7007−0560], Antonio Rago[0000−0001−5323−7739], and Francesca Toni[0000−0001−8194−1459]

Imperial College London, UK
{junqi.jiang20, a.rago, ft}@imperial.ac.uk

Abstract. *Counterfactual explanations* (CEs) are an increasingly popular way of explaining machine learning classifiers. Predominantly, they amount to data instances pointing to potential changes to the inputs that would lead to alternative outputs. In this position paper we question the widespread assumption that CEs should always be data instances, and argue instead that in some cases they may be better understood in terms of special types of *relations* between input features and classification variables. We illustrate how a special type of these relations, amounting to *critical* influences, can characterise and guide the search for data instances deemed suitable as CEs. These relations also provide compact indications of which input features - rather than their specific values in data instances - have counterfactual value.

Keywords: counterfactual explanations · machine learning classifiers · relation-based explanations

1 Introduction

As modern machine learning methods have become increasingly complex and more involved in our daily lives, ways to explain models have gained more attention. One popular way to explain models' predictions focuses on the extraction of *counterfactual explanations* (CEs), which attempt to answer the question "which changes would have to be made to an input in order to alter a prediction made by the model?" Generally, CEs are data instances with modified feature values, wrt the input instance whose output needs explaining, such that the prediction model will give a different output when given the modified input instance. CEs can be very useful when applied in human-centered areas like finance and healthcare with the aim of providing information on further actions that users can take in order to improve their outlook. For example, if a person is denied a loan, a CE may indicate that an increase (e.g. of £5000) in annual income could flip the denial, thus suggesting a concrete action (increasing one's income) towards obtaining the loan. CEs have been identified as crucial, e.g. in [13, 9], to meet the requirements of regulations, such as the GDPR [5]. However, many relevant data instances satisfying this core requirement of counterfactuality may exist for

each given input (e.g. when applying for a loan, different income increments, including large, unachievable ones), potentially exceeding the cognitive capacities of users. This calls into questioning the widespread assumption that CEs should be provided to users as data instances, since other formats may more compactly and effectively display this information to users.

Currently, various approaches exist for finding instance-based CEs, including formulating the CEs' search directly into gradient-based optimisation problems [13]; integer programming problems [7]; encoding into logic formulae and solving integer programming problems [4] or satisfiability modulo theories (SMT) problems [8], etc. We refer readers to [6] for a recent survey.

In this position paper, building upon [1], we argue that, in some cases, particularly those where *any* change in a feature's value (assuming constant values for the other features) will give a different result, CEs may be better understood in terms of *relations* amounting to sets of influences between input features and output classifications, and that these relations can guide the search for data instances as CEs. So, in the loan example, let *any* change in the value of employment status (e.g. from unemployed to full-time, part-time or retired) result in a classification change; then, we may compactly identify the influence between employment status and obtaining a loan as *critical*. The use of relations (which, as in [1], may be critical, as indicated above, or *potential*, if the feature is critical when the other features are in some other states) renders CEs more compact, and arguably more customisable to users' needs (in that users can choose to act upon the specific value of the features identified, e.g. as in a critical relation with the output variable). Whereas [1] focuses on CEs for Bayesian classifiers for discrete variables, we argue that CEs built from (critical and potential) influences can be used to obtain CEs for any classifier for heterogeneous data types (with discrete and/or continuous variables).

2 A Motivating Illustration

We demonstrate the underpinning use of relations for CE generation with the Parole Violation dataset [12]. Consider some classifier[1] trained on this dataset to predict whether a person violated the conditions of parole and was confirmed as a parole violator, using eight input features. These include three continuous features *age*, *time.served*, and *max.sentence* and five discrete features, with *male*, *race*, *multiple.offenses* being binary discrete and *state* and *crime* having four possible values each (below we map the values of discrete features to integers in quotes starting from '0'). The classification variable is *violator* for which value 1 indicates a violation and value 0 a non-violation. Now consider the following data instance:

Here, the data subject is determined a violator by the classifier. The following process exemplifies how relations could be used to counterfactually explain the prediction result for this input, based on an extension of the method of [1]

[1] We omit to indicate here the specific type of classifier and how it is obtained from the dataset, as this plays no role in the considerations we make about CEs.

male	*race*	*age*	*state*	*time.served*	*max.sentence*	*multiple.offenses*	*crime*	*violator*
'1'	'1'	41.2	'2'	6	12	'1'	'2'	1

to an unspecified classifier and any data types (beyond discrete variables only). First, special relations amounting to *critical* or *potential* influences between input features and classification variables (where each influence is an element of the respective relation) could be identified, as shown in the following:

male	*race*	*age*	*state*	*time.served*	*max.sentence*	*multiple.offenses*	*crime*
critical	critical	potential	critical	neither	potential	critical	critical

Here, for example, the influence between *crime* and *violator* is critical (and thus *crime* can be deemed a critical influencer of *violator*). Intuitively, critical influencers are such that *any* change in their value would result in a change in the output. In our case, for example, using the fact that *crime* is a critical influencer, we know, in particular, that the following instances changing only *crime* are valid CEs (below, changes from the input instance are highlighted in **bold**):

male	*race*	*age*	*state*	*time.served*	*max.sentence*	*multiple.offenses*	*crime*	*violator*
'1'	'1'	41.2	'2'	6	12	'1'	'2'	1
'1'	'1'	41.2	'2'	6	12	'1'	**'0'**	**0**
'1'	'1'	41.2	'2'	6	12	'1'	**'1'**	**0**
'1'	'1'	41.2	'2'	6	12	'1'	**'3'**	**0**

All these CEs are so for the same reason, namely that *crime* is a critical influencer. This way, the use of relations (amounting to all pairs of variables such that the first variable in the pair is a critical influencer of the second) can provide more compact information than other methods producing only data instances as CEs. For comparison, in this example, Proto [10], a popular method for identifying CEs, would only return the first instance in the table as a CE (with a random choice amongst all alternative *crime* values). The indication that *crime* is a critical influencer, underpinning our envisaged notion of relation-based CEs, compactly indicates that any alternative values for *crime* would also lead to different results for *violator*. In addition to critical influencers, CEs could also use potential influencers, which are such that *some* change in their value would result in a change in the output. In our setting, for example, *age* is a potential influencer of *violator*.

Without an indication of these influencers, users would miss important information when shown a CE. In the current literature, CE generation methods produce either one data instance (like Proto [10]), or a *diverse* set (like [11]) as CEs depending on the design. Though the diverse methods could provide more

information than the single-instance methods, such *diversity*, which implies that each counterfactual instance should be sufficiently different wrt some distance metrics, might not be helpful in scenarios like the above *crime* example where the three counterfactual instances are not a diverse set, but in other settings could equally be important to the users.

Overall, our vision for CEs, amounting to data instances with an indication of critical and potential influencers within them, includes potentially useful additional information (with respect to existing methods for CE extraction) for human consumption.

3 Conclusions

In this position paper we have outlined how a different style of counterfactual explanations (CEs), namely those based on relations between variables rather than data instances, could be more useful and intuitive in cases such as those we have described (particularly those where any change in a feature's value, all else being equal, will give a different result). Much remains to be investigated with regards to this proposal: the effectiveness of the search process for CEs of the kind we envisage, especially for data with continuous variables; the amenability of such CE to various models used in AI beyond Bayesian classifiers, e.g. tree-based classifiers or reinforcement learning models; possible links to existing (particularly graphical) formalisms such as *computational argumentation* (see [2, 3] for recent overviews) used in explainable AI. First and foremost in our opinion, however, would be an investigation into how well these relation-based explanations perform from a user standpoint, as extensive studies examining the various evaluation measures along these lines, e.g. quantifying user trust and comprehension, are needed to determine conclusively advantages and disadvantages of these explanations.

Acknowledgements

Toni was partially funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 101020934). All authors were partially funded by J.P. Morgan and by the Royal Academy of Engineering under the Research Chairs and Senior Research Fellowships scheme. Any views or opinions expressed herein are solely those of the authors.

References

1. Emanuele Albini, Antonio Rago, Pietro Baroni, and Francesca Toni. Relation-based counterfactual explanations for bayesian network classifiers. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI*, 2020.

2. Katie Atkinson, Pietro Baroni, Massimiliano Giacomin, Anthony Hunter, Henry Prakken, Chris Reed, Guillermo Ricardo Simari, Matthias Thimm, and Serena Villata. Towards artificial argumentation. *AI Magazine*, 38(3):25–36, 2017.
3. Pietro Baroni, Dov Gabbay, Massimiliano Giacomin, and Leendert van der Torre, editors. *Handbook of Formal Argumentation*. College Publications, 2018.
4. Zhicheng Cui, Wenlin Chen, Yujie He, and Yixin Chen. Optimal action extraction for random forests and boosted trees. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 179–188, 2015.
5. Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a "right to explanation. *AI Magazine 38 50–57*, 2017.
6. Riccardo Guidotti. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, pages 1–55, 2022.
7. Kentaro Kanamori, Takuya Takagi, Ken Kobayashi, and Hiroki Arimura. Dace: Distribution-aware counterfactual explanation by mixed-integer linear optimization. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20, Christian Bessiere (Ed.). International Joint Conferences on Artificial Intelligence Organization*, pages 2855–2862, 2020.
8. Amir-Hossein Karimi, Gilles Barthe, Borja Balle, and Isabel Valera. Model-agnostic counterfactual explanations for consequential decisions. In *International Conference on Artificial Intelligence and Statistics*, pages 895–905. PMLR, 2020.
9. Anton Korikov, Alexander Shleyfman, and J. Christopher Beck. Counterfactual explanations for optimization-based decisions in the context of the GDPR. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 4097–4103, 2021.
10. Arnaud Van Looveren and Janis Klaise. Interpretable counterfactual explanations guided by prototypes. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 650–665. Springer, 2021.
11. Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 607–617, 2020.
12. United States Department of Justice. National corrections reporting program. https://doi.org/10.3886/ICPSR26521.v2, 2004.
13. Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.