Case-based Explanation for Black-Box Time Series and Image Models with Applications in Smart Agriculture

Eoin Delaney[1,2,3,*,†]

[1]School of Computer Science, University College Dublin, Belfield, Dublin 4, Ireland
[2]VistaMilk SFI Research Centre, Dublin, Ireland
[3]Insight Centre for Data Analytics, Dublin, Ireland

Abstract

Black-box models are frequently deployed for high stakes prediction tasks in a variety of domains (e.g., disease diagnosis and agricultural prediction). The predictions of these opaque systems are often plagued by a lack of transparency, motivating novel research in eXplainable AI (XAI) aiming to understand why a certain prediction was made. One increasingly promising form of explanation is counterfactual explanation where the aim is to elucidate how a prediction could change, given some change in the input space. While the majority of existing work has focused on producing counterfactual explanations for tabular data, significantly less focus has been placed on generating and evaluating counterfactual explanations for time series and image data. Explaining predictions for these data types, arguably, presents a whole new set of issues for XAI, due to the complex and multi-dimensional nature of the data. In this research, we examine how leveraging case-based reasoning (CBR) techniques such as Nearest-Unlike-Neighbors (NUNs) can aid the generation and evaluation of explanations in these domains. We also demonstrate the inadequacies of many traditional techniques that are used to evaluate explanations and highlight the promise of CBR and user studies in the evaluation of explanations.

Keywords

Explainable AI, Counterfactual, Time Series, Prefactual, XCBR, Smart-Agriculture, User Study

1. Introduction

In recent years, the predictive prowess of machine learning systems has been undermined by a worrying lack of interpretability, fairness, accountability and transparency [1, 2]. These challenges have resulted in major research efforts in Explainable AI (XAI) where the core objective is to offer insights into the predictions of black-box models that are commonly deployed in high stakes scenarios. One such scenario that is of particular interest to our research is in smart agriculture. Previous CBR research has already shown immense promise in both grass growth and grasshopper infestation prediction [3, 4]. While the majority of XAI research focus has been on tabular data, less attention has been attributed to time series data, introducing a new set of complex issues for XAI due to high data dimensionality and strong feature dependencies [5].

ICCBR DC'22: Doctoral Consortium at ICCBR-2022, September, 2022, Nancy, France
*Corresponding author.
eoin.delaney@insight-centre.org (E. Delaney)
https://e-delaney.github.io/ (E. Delaney)
0000-1111-2222-3333 (E. Delaney)

A variety of eXplainable CBR (XCBR) methods have shown immense promise for XAI (see [6] for a review). These XCBR techniques provide factual, example-based explanations (e.g., [7, 8]), feature-weighting explanations (CBR-LIME; [9]), and counterfactual explanations [10] with a focus typically on tabular and sometimes image data.

Counterfactual explanations aim to elucidate how a prediction could change if some input was different. There is growing evidence from psychology, philosophy and sociology indicating that they provide more human friendly and GDPR compliant explanations in comparison to other popular forms of explanations [11, 12, 10]. While there are over 100 techniques proposed to generate counterfactual explanations [13], very few of these methods focus on image data, and even fewer on time series data (see e.g., [14, 15] for closest works). In a similar fashion, it is unclear if the proposed properties of *good* counterfactual explanations for tabular data such as proximity, sparsity, and plausibility [12, 10] will extend to other data types. Moreover, evaluating these properties is non trivial and there is growing evidence to suggest that user studies are desperately needed in order to reliably evaluate explanations [16, 10].

2. Research Plan and Objectives

The overall goal of this research is to develop techniques that can be used to generate and evaluate explanations for time series and image data through leveraging case-based reasoning. Building on evidence from psychology, philosophy and social science [11, 12, 13], a core focus of this research is in the generation and evaluation of counterfactual explanations.

I have identified several research questions that underpin the goal of generating and evaluating counterfactual explanations for time series and image data;

- Can case-based reasoning be leveraged to generate *good* explanations for applied time series prediction tasks both in terms of (i) counterfactual explanations for classification and (ii) explanations for applied agricultural forecasting problems?
- What are the properties of good counterfactual explanations for time series and image data, and do they mirror the properties of good counterfactual explanations for tabular data (e.g., proximity, sparsity and plausibility)?
- Do explanations that are automatically generated by computational techniques align with explanations that are informative for human users?

In previous work, Keane and Smyth [10] designed a novel case based technique to generate counterfactual explanations for tabular data through leveraging existing counterfactual instances in the training data (i.e., nearest unlike neighbors (NUNs) [17]). So, exploring the role of NUNs in the generation of counterfactual explanations is a promising line of research in the context of time series and image data. The combination of CBR with Deep Learning feature weighting techniques (e.g., class activation mapping [18]) in a Twin-Systems framework [19] is another promising area of research for the development of counterfactual explanations for time series and image data. Feature weighting techniques are perhaps the most common XAI method in time series classification [5], and the availability of open source data on the UCR archive [20] readily facilitates the development and experimental comparison of XAI techniques.

In terms of time series forecasting, one untapped line of work from our review of the psychological literature is in *prefactual explanation.* Prefactual explanations describe conditional (if-then) propositions about, as yet not undertaken, actions and the corresponding outcomes that may (or may not) take place in the future [21, 22]. While counterfactuals focus on the past, prefactuals look to the future, capturing the idea of something that is not yet a fact, but could become a fact [21]. Such explanations could also be leveraged in other challenging domains such as reinforcement learning [23].

One applied area that is of particular interest to our research is in smart and sustainable agriculture. We have a data set from an industry partner containing information about milk yield from over 2000 commercial dairy herds. One of our goals is to accurately provide long term milk supply forecasts to farmers, supplementing the predictions with explanations that indicate different actions they could take to boost milk yield in future years. Related CBR work in goal-based recommendation has shown how different training plans can be recommended to runners to produce new personal best times [24], so relating this CBR research to producing prefactual explanations for farmers to improve their output is a promising line for novel research.

Finally, it is unclear if the properties of good explanations for tabular data will extend to time series and image data. For example, when generating explanations one popular technique is to minimize the distance between the query and the counterfactual [12]. However, this runs the risk of generating adversarial explanations that may not be noticeably different for users in relation to the query instance. In time series and image data, discriminative and semantically meaningful information is often contained in localized regions of the time series or image. So, it is clear that user evaluation and rigorous testing of explanation evaluation metrics are needed in this research.

3. Progress Summary

We developed a novel CBR technique, *Native-Guide*, to generate counterfactual explanations for time series classification tasks [5]. The technique leverages both in-sample counterfactual explanations (e.g., Nearest Unlike Neighbors [17, 10]) and feature weight vectors from techniques such as class-activation mapping [18] to create explanations. This work was presented at ICCBR'21 where it received a best-student paper award. More recently, we developed a novel forecasting technique and a method to provide prefactual explanations with applications in milk supply prediction [25]. Specifically, we highlighted how producing explanations through comparatively contrasting high performing exemplar herds and low performing herds (retrieved using class prototypes) could boost future on-farm performance - *"Your projected milk supply for next year is 250'000 litres. However, if you reduced the calving period (In a similar fashion to farmer Y), your projected supply would be 300'000 litres and your milk would likely have a higher protein content".* This work will appear in the main proceedings of ICCBR'22.

In terms of counterfactual evaluation, we conducted a literature review of over 100 papers and discussed five key deficits to rectify in the evaluation of counterfactual XAI techniques. This review paper was presented at IJCAI'21 [13]. We noted the over-reliance on computational proxy measures for proximity, sparsity and plausibility without any conclusive evidence from user studies. In work presented at the ICML Workshop on Algorithmic Recourse we identified

the utility of case-based evaluation methods in determining how well a counterfactual fit the data distribution, and highlighted that optimizing for proximity often generated adversarial explanations that would not be noticeably different than the query for a human user [26]. Results in our work in time series classification also demonstrated similar results [5]. So, a natural avenue for current and future work is to focus on evaluating counterfactual explanations through conducting user studies.

Currently we are focusing on addressing some of the central issues presented in our IJCAI review paper and we are conducting large scale user studies to evaluate explanations and critically assess the suitability of computational evaluation techniques. In our latest experiments human users (N=42) created counterfactuals through correcting misclassifications of a convolutional neural network on the MNIST and Google Quickdraw data sets using a drawing tool. This data represents the first ground truth explanation data set for counterfactual visual explanations. By comparing explanations generated by humans with those that are generated automatically by computational techniques, we aim to provide novel insights into (i) the properties of good explanations according to humans and (ii) the unreliability of many popular evaluation metrics (e.g., L_1 and L_2 distances for proximity). Contrary to popular belief, our initial results indicate that people do not minimally edit instances when creating counterfactual visual explanations. Instead they modify a larger, and often semantically meaningful region when creating an explanation, often pushing the explanation towards a class prototype. So, leveraging psychologically grounded models of similarity such as Tversky's contrast model of similarity [27] in counterfactual generation and evaluation may result in more informative explanations and is an interesting avenue for future work.

References

[1] D. Gunning, D. Aha, Darpa's explainable artificial intelligence (xai) program, AI Magazine 40 (2019) 44–58.

[2] A. Adadi, M. Berrada, Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI), IEEE Access 6 (2018) 52138–52160.

[3] E. M. Kenny, E. Ruelle, A. Geoghegan, L. Shalloo, M. O'Leary, M. O'Donovan, M. T. Keane, Predicting grass growth for sustainable dairy farming: A cbr system using bayesian case-exclusion and post-hoc, personalized explanation-by-example (xai), in: International Conference on Case-Based Reasoning, Springer, 2019, pp. 172–187.

[4] J. Hastings, K. Branting, J. Lockwood, Carma: A case-based rangeland management adviser, AI Magazine 23 (2002) 49–49.

[5] E. Delaney, D. Greene, M. T. Keane, Instance-based counterfactual explanations for time series classification, in: International Conference on Case-Based Reasoning, Springer, 2021, pp. 32–47.

[6] J. M. Schoenborn, R. O. Weber, D. W. Aha, J. Cassens, K.-D. Althoff, Explainable case-based reasoning: A survey, in: AAAI-21 Workshop Proceedings, 2021.

[7] M. T. Keane, E. M. Kenny, 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: Proc. ICCBR'19, Springer, 2019, pp. 155–171.

[8] F. Sørmo, J. Cassens, A. Aamodt, Explanation in case-based reasoning–perspectives and goals, Artificial Intelligence Review 24 (2005) 109–143.
[9] J. A. Recio-García, B. Díaz-Agudo, V. Pino-Castilla, CBR-LIME: A Case-Based Reasoning Approach to Provide Specific Local Interpretable Model-Agnostic Explanations, in: ICCBR, Springer, 2020, pp. 179–194.
[10] M. T. Keane, B. Smyth, Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable ai (xai), in: Proc. ICCBR'20, Springer, 2020, pp. 163–178.
[11] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence 267 (2019) 1–38.
[12] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: automated decisions and the gdpr, Harv.J.Law Tech. 31 (2017) 841.
[13] M. T. Keane, E. M. Kenny, E. Delaney, B. Smyth, If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual xai techniques, in: IJCAI-21, 2021.
[14] E. Ates, B. Aksar, V. J. Leung, A. K. Coskun, Counterfactual explanations for machine learning on multivariate time series data, arXiv preprint arXiv:2008.10781 (2020).
[15] Y. Goyal, Z. Wu, J. Ernst, D. Batra, D. Parikh, S. Lee, Counterfactual visual explanations, in: ICML, PMLR, 2019, pp. 2376–2384.
[16] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning, arXiv preprint arXiv:1702.08608 (2017).
[17] C. Nugent, D. Doyle, P. Cunningham, Gaining insight through case-based explanation, Journal of Intelligent Information Systems 32 (2009) 267–295.
[18] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: IEEE CVPR, 2016, pp. 2921–2929.
[19] E. M. Kenny, M. T. Keane, Twin-systems to explain artificial neural networks using case-based reasoning: comparative tests of feature-weighting methods in ann-cbr twins for xai, in: IJCAI-19, 2019, pp. 2708–2715.
[20] H. A. Dau, A. Bagnall, K. Kamgar, C.-C. M. Yeh, Y. Zhu, S. Gharghabi, C. A. Ratanamahatana, E. Keogh, The UCR time series archive, IEEE/CAA Journal of Automatica Sinica 6 (2019) 1293–1305.
[21] K. Epstude, A. Scholl, N. J. Roese, Prefactual thoughts: Mental simulations about what might happen, Review of General Psychology 20 (2016) 48–56.
[22] R. M. Byrne, S. M. Egan, Counterfactual and prefactual conditionals., Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale 58 (2004) 113.
[23] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, C. Zhong, Interpretable machine learning: Fundamental principles and 10 grand challenges, Statistics Surveys 16 (2022) 1–85.
[24] B. Smyth, P. Cunningham, A novel recommender system for helping marathoners to achieve a new personal-best, in: Proceedings of the eleventh ACM conference on recommender systems, 2017, pp. 116–120.
[25] E. Delaney, D. Greene, L. Shalloo, M. Lynch, M. T. Keane, Forecasting for sustainable dairy produce: Enhanced long-term, milk-supply forecasting using k-nn for data augmentation, with prefactual explanations for xai., in: To appear in ICCBR'22, 2022.

[26] E. Delaney, D. Greene, M. T. Keane, Uncertainty estimation and out-of-distribution detection for counterfactual explanations: Pitfalls and solutions, arXiv preprint arXiv:2107.09734 (2021).
[27] A. Tversky, Features of similarity., Psychological review 84 (1977) 327.