

Addressing Trust and Mutability Issues in XAI utilising Case Based Reasoning

Pedram Salimi^{1,*}

¹Robert Gordon University, Garthdee House, Garthdee Rd, Garthdee, AB10 7AQ, Aberdeen, United Kingdom

Abstract

Explainable AI (XAI) research is required to ensure that explanations are human readable and understandable. The present XAI approaches are useful for observing and comprehending some of the most important underlying properties of any Black-box AI model. However, when it comes to pushing them into production, certain critical concerns may arise: (1) How can end-users rely on the output of an XAI platform and trust the system? (2) How can end-users customise the platform's output depending on their own preferences In this project, we will explore how to address these concerns by utilising Cased-based Reasoning. Accordingly, we propose to exploit the neighbourhood to improve end-user trust by offering similar cases and confidence scores and using different retrieval strategies to address end-user preferences. Additionally, this project will also look at how to leverage Conversational AI and Natural Language Generation approaches to improve the interactive and engaging user experience with example-based XAI systems.

Keywords

Explainable AI, Cased-based Reasoning, Conversational AI, Natural Language Generation


1. Introduction


Due to recent breakthroughs in Artificial intelligence (AI) such as deep learning approaches, AI models are getting more accurate and powerful while also becoming more complicated [1]. However, because of their complexity, comprehending how these models work and making judgments has proven difficult. Earlier AI systems were build on approaches that are fundamentally explainable (i.e. white-box) where Rule-based methods, Decision Trees, Hidden Markov Models, and Logistic Regressions are some examples. Thanks to recent breakthroughs, novel AI techniques such as deep learning are more accurate and powerful than traditional approaches; however, they are also more complicated [1, 2]. Due to complexity of the models (i.e. black-box), comprehending how they work and make decision is difficult. Consequently they reduce model explainability.

Accordingly, there is an armory of Explainable AI (XAI) methods developed in recent literature to explain black-box AI models and the decisions they make. Example-based explanations assist people in developing mental models of the machine learning method and the data on which the machine learning method was trained[3]. Literature shows that humans tend to provide contrastive explanations when explaining their decisions to one another. Accordingly, explaining an AI decision using counterfactual examples can be most understandable to humans

ICCBR DC'22: Doctoral Consortium at ICCBR-2022, September, 2022, Nancy, France

 p.salimi@rgu.ac.uk (P. Salimi)

 © 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

because they both have the same conceptual model as human explanations [4]. Therefore, In this project, we are focusing on example-based XAI and we propose to use CBR approaches to address two existing challenges in this domain.

Example-based XAI systems, similar to other methods have been good at explaining the current user problem. Also, they are able to guide the users to solve their problems. However, there are several limits to these XAI methods. In this project we are focusing on two of these limitations. One of them is that current approaches are static, which means they provide explanations based on the user query but cannot react to user modifying the query based on their own personal preferences [5]. There is also a lack of trust between the user and the XAI system that is yet to be addressed. User studies by [4] and [6] are few those who highlight this issue of trust in XAI system with respect to applications in speech recognition, forest coverage prediction and recidivism.

Case-Based Reasoning (CBR) is a methodology that emulate how humans reason from precedent and examples and it has a central role in XAI systems [7]. CBR has been the underpinning of many example-based XAI methods providing explanations ranging from factual to counterfactuals [8]. Accordingly, we ask the following research questions:

- **RQ1:** How can we approach the issue of trust in an interactive example-based XAI system using the CBR system?
- **RQ2:** How can a case-based approach assist us in dealing with mutability of features when generating counterfactual explanations?

2. Background

In this section, we will first study example-based XAI approaches before briefly discussing the CBR methodology.

2.1. Example-based Explanations

Example-based approaches are classified into three categories: factual, semi-factual, and counterfactual.

Factual Explanations provide information about why a certain outcome was received based on query features [9]. Using nearest neighbours is an example-based approach for finding factual explanations. For example, in loan application a factual explanation using nearest neighbors could be “Your loan got rejected because there is another person whose situation is quite similar to you and had their loan declined”. Explanations-by-example is a factual explainer algorithm where nearest-neighbours are found using Critical Classification Regions in images[10].

Semi-factual Explanations present the maximum distance an instance may go without changing the black-box outcome. A semi-factual explanation for a reject loan application would be, “Even if the installment amount is increased, loan would be still rejected”. PIECE is a case-based method for generating semi-factual explanations which uses a convolutional model to detect important features and to generate semi-factual explanations [11].

Counterfactual Explanations define a causal, synthetic or past event with the smallest change in feature values that causes the prediction to shift to a desired outcome. A counterfactual explanation for a rejected loan application would be, “If the loan amount is reduced, loan would have been accepted”. Some of the state-of-the-art counterfactual methods are as follows:

- **NICE** approach is divided into two steps. First, the nearest unlike neighbour (NUN) is retrieved, which leads to the finding of non-overlapping features against the query. Then the algorithm iteratively attempts to determine the optimal counterfactual using a reward function that consider properties like [12].
- **DisCERN** is a case-based counterfactual explanation method. Here, counterfactuals are created by substituting feature values of the query from the NUN until an outcome change is detected. Features to substitute are selected based on feature attributions. [13].

2.2. Case-based Reasoning

Most XAI methods, including ones discussed above, fail to establish trust with the end-user as discussed in the introduction. Their one-shot nature (instead of being interactive) also fail to incorporate user preferences. To address these challenges, in this project, we explore techniques from Case-based Reasoning (CBR). A CBR methodology consists of four stages: retrieve, reuse, revise, and retain [14]. The first stage involves providing an input that describes the present user query and retrieving similar cases in the case base by employing similarity metrics. The second stage utilises retrieved cases and use adaption knowledge to present the user with a solution to their query. The user may accept or reject the solution for a variety of reasons, for example, user could be unable to accept the entire proposed solution, based on their own preferences. In the case of rejection, the following step is to revise. In most cases, the revise stage includes incorporating feedback acquired from testing the suggested solution. During the final step, the new case may be retained in the case base for future use.

3. Approach / Methodology

In this section we are going to explore the research questions identified using CBR techniques.

3.1. RQ1: How to address trust in an example-based XAI system using CBR?

This RQ explore how to establish trust between an XAI system and a user once they are presented with an explanation. Specifically we want to identify what additional information or explanation will help the user to better believe the recommendations provided by an XAI system in terms of reliability of the XAI system. Following CBR approaches may assist us in addressing the trust issues:

- **Nearest Neighbors** may be to retrieve the nearest neighbours of the provided solution which previously were successful examples
- **Coverage** and population density of instances which are similar to the provided counterfactual in a case base[15].

For evaluating the impact of proposed methods on trust, a user study is proposed. This work is informed by XAI evaluation methods like the Hoffman Trust scale when seeking feedback from users [16].

3.2. RQ2: How to address feature mutability preferences using CBR?

The interactive system should allow the user to modify the criteria on which an explanation is generated by considering mutability of features. In another words, mutability in an XAI system is about to giving the control to the user in terms of the degree of complexity and difficulty of what they can change. To address this we explore following techniques from CBR:

- **Collaborative Filtering** It is a mechanism for proposing alternative solutions to a user based on similarities (similarity assessment [17]) in the user's prior behaviour and that of other users.
- **Adaptation** When the best partial-matching case from the case repository does not perfectly match the new case, the previous solution must be altered to fit the new case solution more accurately. There are several adaptation methods such as null adaptation, structural adaptation, or a combination of methods[18].

Figure 1 depicts an example interaction between a user and AI that demonstrates how such a system might deal with trust and mutability issues.

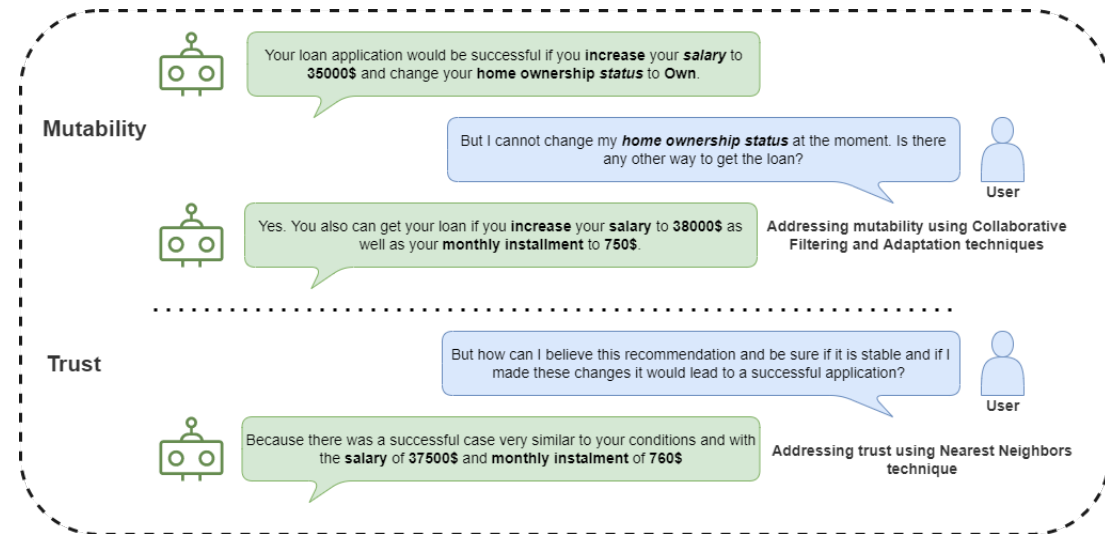


Figure 1: Employing Nearest Neighbors and Collaborative Filtering in order to deal with trust and mutability issues correspondingly.

4. Progress Summary

Recent work has explored several ways for data-to-text generation in terms of numerical reasoning, with the objective of mapping counterfactuals to a natural language representation

which would facilitate a more engaging interaction with the end user. We have designed two distinct template-based text generation algorithms, one with features grouped based on attribution change and the other without. This is illustrated in Figure 2 with two presentations of the counterfactual with and without the feature grouping template for a loan application.

	Loan Amount	Recoveries	Installment	Interest Rate	Home Ownership
Query	2200	1985.75	1200	10.78	OWN
Counterfactual	3700	1310	1700	4.78	RENT
Without Grouping	The loan application would be successful if you increase your loan amount by 1500\$, decrease your recoveries by 676.75\$, increase your installment by 500\$, decrease the interest rate by 6%, and change your home ownership status to rent in the exact order of priorities				
Grouping	The loan application would be successful if you increase your loan amount by 1500\$, your installment by 500\$, and decrease your recoveries by 676.75\$, interest rate by 6%, and change your home ownership status to rent in the exact order of priorities				

Figure 2: Different template based text generation based on feature grouping. This classification is part of the user research to determine which one is more plausible.

Our immediate next task is to design a user study to assess such generation templates and to understand to what extent it could impact end-user engagement and trustworthiness. A questionnaire will be prepared to gather feedback on several counterfactual explanation scenarios. In our user study we are going to consider three framing concept in order to prevent potential biases in our user study[19]. For example, we are going to employ NASA Task Load Index questionnaire. But we are going to modify them with respect to positive framing concept in a manner that instead of asking how much the user got frustrated during the task, we are going to ask how much the task was easy to do. Results of the user study will inform us to identify best template generation strategies in terms of the quality of generated textual explanation and also provide insights for addressing the project’s research questions.

Having input from a DC mentor to help improve the user study design will be very valuable; as would directions for integrating case-based strategies for improving user trust.

References

- [1] W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, K.-R. Müller, Explainable AI: interpreting, explaining and visualizing deep learning, volume 11700, Springer Nature, 2019.
- [2] M. Danilevsky, K. Qian, R. Aharonov, Y. Katsis, B. Kawas, P. Sen, A survey of the state of explainable ai for natural language processing, arXiv preprint arXiv:2010.00711 (2020).
- [3] C. Molnar, Interpretable machine learning, Lulu. com, 2020.
- [4] X. Wang, M. Yin, Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making, in: 26th International Conference on Intelligent User Interfaces, 2021, pp. 318–328.
- [5] K. Sokol, P. A. Flach, Counterfactual explanations of machine learning predictions: opportunities and challenges for ai safety, SafeAI@ AAI (2019).
- [6] K. Weitz, D. Schiller, R. Schlagowski, T. Huber, E. André, ” do you trust me?” increasing user-trust by integrating virtual agents in explainable ai interaction design, in: Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents, 2019, pp. 7–9.

- [7] M. T. Keane, B. Smyth, Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable ai (xai), in: *International Conference on Case-Based Reasoning*, Springer, 2020, pp. 163–178.
- [8] M. T. Keane, E. M. Kenny, M. Temraz, D. Greene, B. Smyth, Twin systems for deepcbr: A menagerie of deep learning and case-based reasoning pairings for explanation and data augmentation, *arXiv preprint arXiv:2104.14461* (2021).
- [9] I. Stepin, J. M. Alonso, A. Catala, M. Pereira-Fariña, A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence, *IEEE Access* 9 (2021) 11974–12001.
- [10] E. M. Kenny, E. D. Delaney, M. T. Keane, Advancing nearest neighbor explanation-by-example with critical classification regions (2021).
- [11] E. M. Kenny, M. T. Keane, On generating plausible counterfactual and semi-factual explanations for deep learning, *AAAI-21* (2021) 11575–11585.
- [12] D. Brughmans, D. Martens, Nice: an algorithm for nearest instance counterfactual explanations, *arXiv preprint arXiv:2104.07411* (2021).
- [13] N. Wiratunga, A. Wijekoon, I. Nkisi-Orji, K. Martin, C. Palihawadana, D. Corsar, Discern: Discovering counterfactual explanations using relevance features from neighbourhoods, in: *2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)*, IEEE, 2021, pp. 1466–1473.
- [14] A. Aamodt, E. Plaza, Case-based reasoning: Foundational issues, methodological variations, and system approaches, *AI communications* 7 (1994) 39–59.
- [15] A. Lawanna, J. Daengdej, Methods for case maintenance in case-based reasoning, *International Journal of Computer and Information Engineering* 4 (2010) 82–90.
- [16] R. Hoffman, S. T. Mueller, G. Klein, J. Litman, Metrics for explainable ai: Challenges and prospects, *ArXiv abs/1812.04608* (2018).
- [17] R. Burke, A case-based reasoning approach to collaborative filtering, 2000. doi:10.1007/3-540-44527-7_32.
- [18] S. Craw, N. Wiratunga, R. C. Rowe, Learning adaptation knowledge to improve case-based reasoning, *Artificial intelligence* 170 (2006) 1175–1192.
- [19] S. Schoch, D. Yang, Y. Ji, “this is a problem, don’t you agree?” framing and bias in human evaluation for natural language generation, in: *Proceedings of the 1st Workshop on Evaluating NLG Evaluation*, 2020, pp. 10–16.