

Beyond Post-Hoc Instance-Based Explanation Methods

Betül Bayrak¹

Norwegian University of Science and Technology (NTNU), Høgskoleringen 1, Trondheim, 7034, Norway

Abstract

With the increasing demand for understanding the decision-making processes of artificial intelligence applications, explainable AI (XAI) systems have become increasingly important. Counterfactual explanations, a promising approach in XAI, take advantage of human counterfactual reasoning mechanisms to offer intuitive explanations of how a model's predictions could have been different. This Ph.D. project focuses on the design of post-hoc XAI techniques to generate counterfactual explanations that utilize case-based reasoning. It highlights the benefits of post-hoc explanation systems in improving our understanding of black-box models and explores the unique advantages of counterfactual explanations as an instance-based method. Furthermore, this report presents an overview of my doctoral studies and current state. It contributes to the growing body of research on XAI by presenting novel insights into the design of post-hoc XAI systems. Additionally, the report identifies areas in the existing literature that require further investigation and suggests potential directions for future research. Overall, this report offers valuable insights for researchers and practitioners interested in the design of XAI systems and highlights the importance of transparency and interpretability in artificial intelligence.

Keywords

Counterfactual Explanation Generation, Explainable Artificial Intelligent (XAI), Explainable Case-Based Reasoning (XCBR)

1. Introduction

The increasing prevalence of artificial intelligence models in various aspects of our daily lives has created a growing need to understand how these models make decisions. However, the complexity of these models has made it challenging to comprehend the factors that contribute to their predictions. For instance, consider two individuals with similar backgrounds applying for a home loan from a bank that uses a black-box model to assess loan applications. As shown in Figure 1, one applicant is declined, while the other is approved, leaving the rejected applicant wondering about the reasons behind their application's rejection. Counterfactual explanations, which are a type of post-hoc explanation, can provide highly satisfactory explanations in such situations [1, 2].

Counterfactual explanations aim to answer the "what if?" question by presenting hypothetical examples that demonstrate how a model's prediction can be changed with minimal effort. For example, counterfactual explanations answer the "What if the rejected applicant had earned


ICCBR DC'23: Doctoral Consortium at ICCBR2023, July 17 – 20, 2023, Aberdeen, Scotland

✉ betul.bayrak@ntnu.no (B. Bayrak)

🌐 <https://www.ntnu.edu/employees/betul.bayrak/> (B. Bayrak)

🆔 0000-0002-0554-9823 (B. Bayrak)

© 2023 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)

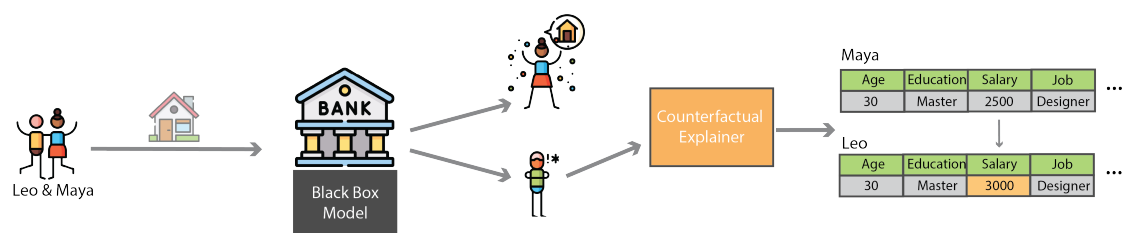


Figure 1: Illustration of how a counterfactual explainer can provide insight into a loan application decision.

“\$500 more? Would their application have been accepted?” questions. These explanations offer valuable insights that can help users understand the predictions made by black-box models, particularly when the factors influencing the model’s decision are unclear [3].

Explanation generation in Explainable AI (XAI) applications has different purposes and approaches. Nunes and Jannach’s study [4] listed 17 different explanation types under four categories and emphasized the importance of considering the aim of the explanation in the explanation generation process. From another perspective, Arrieta et al.’s paper [5] discusses two different approaches for creating explanations, model-dependent and model-agnostic explanation systems. Model-dependent explanation systems are designed for specific models, whereas model-agnostic explanation systems can perform with any model, regardless of its structure or complexity.

Another challenge in XAI applications is generating or selecting the best explanation for a case, which requires essential quality metrics, such as trustworthiness, understandability, informativeness, sufficiency, and unbiasedness. There are various approaches from different fields to meet these quality requirements, including the Case-Based Reasoning (CBR) methodology. CBR is a problem-solving methodology that has four steps (retrieve, reuse, revise, retain) and benefits from past experiences with high interpretability [6]. The CBR methodology is often used to explain AI models since it concentrates on open-ended, often changing, uncertain, and incomplete problems. Thus, the concept of XCBR emerged as a sub-field of XAI [7], offering flexible, interpretable, sustainable, and evolving explanation systems using CBR.

This report focuses on post-hoc counterfactual XCBR techniques and identifies areas in the existing literature that require further investigation while presenting an overview of the author’s doctoral studies and current state.

2. Research Objectives

This section serves as a foundation for outlining the various points of contribution and aims of my doctoral work. The points that are already published and in progress currently are elaborated upon in detail in the next section (See Section 3).

1. *Literature review:* A comprehensive literature review to establish the theoretical foundation for the research in XAI and counterfactual explanation generation.

2. *Combining global and local explanations:* Global explanations help to provide an overall understanding of how a model is working and can identify important features that are driving its predictions. This can be particularly useful for identifying biases or areas where the model could be improved. Local explanations, on the other hand, provide insight into individual predictions and can help to build trust in the model by giving users a clear understanding of how the model arrived at a particular decision. Without local explanations, it can be difficult for users to understand why a particular decision was made and this may lead to mistrust in the model. By combining both global and local explanations, machine learning models can be made more transparent, trustworthy, and ultimately more useful to their intended users.
3. *High-quality counterfactuals:* Counterfactuals allow hypothetical scenarios to be generated and evaluated, thus promoting model transparency and interpretability. Moreover, counterfactuals can be used to test model robustness and evaluate the effect of input changes on model output. Therefore, the generation of robust, diverse, trustworthy counterfactuals is essential for the XAI systems that provide counterfactual explanations.
4. *Flexible XAI system:* Flexibility in XAI systems refers to the ability to adapt in many aspects like different user needs and preferences, data types, amount of the dataset, and application areas.
5. *Domain knowledge integration:* Domain knowledge refers to the knowledge and expertise that is specific to a particular application domain and is often held by domain experts such as clinicians, engineers, or financial analysts. Incorporation of expert knowledge from a particular domain into explanation systems is expected to improve their performance, relevance, and interpretability.
6. *Explanation representation:* Explanation representation is a critical component of XAI systems, as it enables users to understand and trust the decisions made by these models. In instance-based explanations, the explanations can be presented as instances, texts, tables, graphics in different forms, or different combinations of them.
7. *Evaluation:* The evaluation of XAI systems can be done in two ways: user evaluations, which measure the effect on users, and quantitative evaluations, which rely on statistical calculations. An ideal XAI system is able to evaluate the generated explanations using both methods and learn from the outcomes.

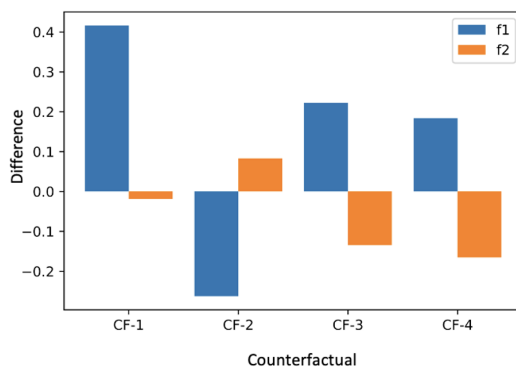
3. Methodology and Progress To Date

In the previous section, the research objectives and points that we aimed to contribute are listed and explained. This section presents the methodology for contributing to some of the listed objectives to date.

In the initial stages of the research, a thorough *literature review* was undertaken, which has been continuously updated and expanded throughout the course of the doctoral studies. The primary contribution lies in the provision of a comprehensive and in-depth analysis of the state-of-the-art approaches in the field. This review encompasses a wide range of perspectives and characteristics, offering valuable insights into the diverse landscape of research in the domain. The final version of the literature review will be published until completion of the

doctoral studies, ensuring its availability to the academic community and further enriching the existing body of knowledge.

The first published work[8] introduces a novel approach to construct an XCBR system that offers counterfactual explanations when required. To generate explanations when it is necessary, an adaptive explanation area is calculated using a sample-centric approach. For each data sample (s), the mean of the k nearest neighbors distance to s is calculated as radius. If there exists at least one counterfactual within the circle with the calculated radius, the circle is marked as an explanation area, and the identified explanation pairs are added to the case-base. When a new query falls within an explanation area, at least one explanation case is activated from the constructed CBR system, resulting in the creation of a two-phase explanation using a text template and a bi-directional bar graph (i.e. Figure 2). The proposed system is a flexible system and has contributions about *explanation representation* by providing multiple explanation pairs similar to the query and combining textual explanations which are powerful to convey statistical data and visual explanations which are powerful to convey comparative data. The proposed XCBR system is notable for its *flexibility* with the amount of data and application area, allowing for multiple explanation pairs that resemble the query, and its contributions to *explanation representation* by combining textual and visual explanations.



” The prediction result is the same with 4 out of 4 closest samples. However, in similar cases, with 0.14 increase in **f1** feature and 0.05 decrease in **f2** model’s decisions can change. ”

Figure 2: An explanation example from a dataset with two features.

In another sub-project, we proposed a novel approach for generating twin XAI systems that utilize CBR to explain black-box models in multi-class classification tasks. The preliminary work has been presented at the XCBR challenge during the 2022 International Conference on Case-Based Reasoning (ICCBR-2022), and details of the preliminary experiments can be found in the proceedings [9]. However, the extended version is under review.

The twin XAI system consists of a multi-agent CBR system (MA-CBR system), where each agent is developed for a specific class and is modeled separately. The system incorporates feature attributions and data distribution to project the different characteristics of classes through separately calculated SHAP values for each class over the black-box model. To develop the local similarity functions, a data-driven similarity measure development method is employed

[10]. In this approach, Verma et al. proposed an Inter Quartile Range-based polynomial modeling. For both global and local similarity function developments, expert knowledge (if available) may be incorporated.

One of the key contributions of this work is the facilitation of *expert knowledge incorporation* into the XAI system, which enhances the reliability and trustworthiness of the explanations provided. Additionally, the multi-agent structure enables the generation of instance-based explanations that incorporate both local and global features, providing a comprehensive understanding of model outputs. An *evaluation metric* is also introduced called "rigidity" which measures the adaptability and *flexibility* of the black-box model's performance through the proposed explanation system. This metric helps assess the quality and reliability of the system's explanations. The proposed system was tested on diverse datasets with varying characteristics, different performance levels of black box models, and varying degrees of expert knowledge. It was observed that the system exhibited a high degree of flexibility. Furthermore, reproducible benchmarking experiments and open-source implementation of the approach and evaluation metric are provided ¹, promoting transparency and further research in the field.

We are currently engaged in ongoing research focused on the development of a *perturbation-based counterfactual generation method*, *PertCF*, that leverages feature attributions generated by SHAP values. This approach combines the strengths of perturbation-based counterfactual generation and feature attribution to produce counterfactuals that are of high quality, stable, and interpretable. Unlike conventional approaches that employ predefined distance metrics such as Euclidean distance, *PertCF* adopts specialized metrics tailored to the specific problem at hand. This specialization of distance metrics offers two distinct advantages. Firstly, it utilizes SHAP values calculated for each class individually, enabling the projection of different class characteristics through feature attribution, in a similar way to the previous work. Secondly, it facilitates the seamless incorporation of domain or expert knowledge, thereby effectively representing the semantics of the data. Preliminary evaluations indicate that *PertCF* demonstrates measurable advancements over state-of-the-art methods. However, further development and refinement of the results are necessary to enhance its performance and efficacy.

4. Conclusion and Future Work

Currently, there are several ongoing tasks and next steps in the research:

- Further development and enhancement of the local-global attribution method to improve its effectiveness. Specifically, exploring data-driven techniques that are independent of the model to strengthen its capabilities.
- Refinement and improvement of the counterfactual generation method that is being worked on, followed by publication to share the findings with the research community.
- Completion of the survey, which aims to provide a comprehensive overview of the approaches developed in recent years. This involves evaluating and comparing these approaches based on various characteristics and presenting them from different perspectives.

¹https://github.com/b-bayrak/Twin_XAI

- Expanding the applicability of the research, such as exploring the integration of multi-modal data to enhance the capabilities and scope of the proposed methods.
- Addressing the challenges related to interactable and customized explanations, seeking solutions that can effectively handle complex scenarios and accommodate specific user requirements.

These tasks collectively aim to advance the understanding and application of explainable artificial intelligence, improving the interpretability and trustworthiness of machine learning models.

References

- [1] S. Tsirtsis, M. Gomez Rodriguez, Decisions, counterfactual explanations and strategic behavior, *Advances in Neural Information Processing Systems* 33 (2020) 16749–16760.
- [2] R. Shang, K. K. Feng, C. Shah, Why am I not seeing it? understanding users' needs for counterfactual explanations in everyday recommendations, in: *2022 ACM Conference on Fairness, Accountability, and Transparency, 2022*, pp. 1330–1340.
- [3] X. Dai, M. T. Keane, L. Shalloo, E. Ruelle, R. M. Byrne, Counterfactual explanations for prediction and diagnosis in xai, in: *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, 2022*, pp. 215–226.
- [4] I. Nunes, D. Jannach, A systematic review and taxonomy of explanations in decision support and recommender systems, *User Modeling and User-Adapted Interaction* 27 (2017) 393–444.
- [5] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, *Information fusion* 58 (2020) 82–115.
- [6] A. Aamodt, E. Plaza, Case-based reasoning: Foundational issues, methodological variations, and system approaches, *AI communications* 7 (1994) 39–59.
- [7] 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*.
- [8] B. Bayrak, K. Bach, When to Explain? Model Agnostic Explanation Using a Case-based Approach and Counterfactuals, in: A. Rutle (Ed.), *Norsk IKT-konferanse for forskning og utdanning, 1, 2022*.
- [9] B. Bayrak, P. Marin Veites, K. Bach, Explaining your neighbourhood: A CBR approach for explaining black-box models (2022) 251–255. URL: <http://ceur-ws.org/Vol-994>.
- [10] D. Verma, K. Bach, P. J. Mork, Modelling similarity for comparing physical activity profiles—a data-driven approach, in: *International Conference on Case-Based Reasoning, Springer, 2018*, pp. 415–430.