Progress Report on Counterfactual Explanation Pipeline Using Potential Outcomes and Structural Causal Models

Student Name: Ashutosh Kumar Jha
Degree: Master of Technology (M.Tech)
Department: Computer Science and Engineering
Institution: Tezpur University
Supervisor: Dr.Arindam Karmakar
Academic Year: 2024–2026

Abstract

This progress report presents the current status of the research work titled *"Counterfactual Explanation Pipeline Using Potential Outcomes and Structural Causal Models."* The objective of this research is to develop a causal-aware explainable artificial intelligence framework that integrates counterfactual reasoning with formal causal inference principles. Traditional explainable AI techniques often rely on correlation-based explanations, which may lead to misleading or non-actionable interpretations. To address this limitation, the proposed work leverages the potential outcomes framework and structural causal models to generate causally valid counterfactual explanations.

The work completed so far includes an extensive literature review, dataset preparation, implementation of treatment effect estimation, mediation analysis, and counterfactual prediction. Preliminary results demonstrate the effectiveness of causal modeling in producing interpretable and actionable explanations. The remaining tasks include evaluation on real-world datasets and extension to non-linear models.

1 Introduction

Explainability has become a critical requirement for machine learning systems deployed in real-world, high-stakes environments. Counterfactual explanations are particularly appealing due to their human-centric "what-if" nature. However, many existing counterfactual methods lack causal grounding and may suggest unrealistic or causally invalid changes.

This research aims to address these issues by proposing a counterfactual explanation pipeline rooted in causal inference. By combining Rubin's potential outcomes framework

with structural causal models, the proposed approach ensures that explanations reflect true causal relationships.

2 Objectives of the Research

The key objectives of the proposed research are:

- To study existing counterfactual explanation techniques and their limitations.
- To integrate causal inference principles into counterfactual explanation generation.
- To estimate individual and average causal treatment effects.
- To perform mediation analysis using structural causal models.
- To generate causally valid individual-level counterfactual predictions.

3 Literature Review (Progress)

An extensive review of literature has been conducted covering causal inference, potential outcomes, structural causal models, and counterfactual explanations in explainable AI. Foundational works by Rubin and Pearl provide the theoretical basis for causal reasoning, while recent research in XAI highlights the growing importance of counterfactual explanations for model interpretability.

The literature review also identifies key gaps, particularly the lack of causal guarantees in many counterfactual explanation methods, motivating the need for the proposed causal-aware approach.

4 Methodology Adopted

The proposed methodology consists of the following steps:

1. Data collection and preprocessing.
2. Estimation of potential outcomes under treatment and control.
3. Computation of Individual Treatment Effect (ITE), Average Treatment Effect (ATE), and Average Treatment Effect on the Treated (ATT).
4. Construction of structural causal models for mediator and outcome variables.
5. Mediation analysis to compute direct and indirect causal effects.
6. Generation of counterfactual outcomes through causal intervention.

5 Work Completed So Far

The following tasks have been successfully completed:

- Problem formulation and objective definition for causal-aware counterfactual explanations.
- Literature review on potential outcomes, structural causal models, and counterfactual explanations.
- Synthetic dataset usage and preprocessing for causal analysis.
- Implementation of individual and average treatment effect estimation.
- Development of mediation models using structural causal modeling.
- Generation of individual-level counterfactual predictions through causal intervention.
- Visualization of causal effects and observed versus counterfactual outcomes.

6 Preliminary Results

Initial experiments demonstrate heterogeneous treatment effects across individuals. Mediation analysis reveals both direct and indirect causal pathways between treatment and outcome. Counterfactual predictions illustrate how intervention on the treatment variable leads to changes in predicted outcomes, enhancing interpretability.

7 Work Remaining

The remaining tasks include:

- Evaluation on real-world datasets.
- Extension of the pipeline to non-linear and deep learning models.
- Comparative analysis with existing XAI counterfactual methods.
- Robustness and sensitivity analysis.
- Thesis writing and final documentation.

8 Expected Contributions

The expected contributions of this research are:

- A unified causal-aware counterfactual explanation pipeline.
- Improved interpretability and causal validity of explanations.
- A bridge between causal inference and explainable AI.
- Practical guidelines for deploying causal counterfactual explanations.

9 Conclusion

This progress report summarizes the research conducted to date on developing a causal-based counterfactual explanation framework. The completed work demonstrates promising results, and the remaining tasks aim to further validate and extend the proposed approach. The research is on track to meet the objectives within the stipulated timeline.

References

[1] Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology*.

[2] Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

[3] Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box. *Harvard Journal of Law & Technology*.