İTÜ MATEMATİK BÖLÜMÜ

Enhancing Explainable AI with Robust and Diverse Counterfactual Explanations

Süreyya Akyüz

Bahçeşehir Üniversitesi

Explainable artificial intelligence (XAI) plays a crucial role in high-stakes fields like healthcare, finance, and law by providing transparent insights into model decisions. Counterfactual (CF) explanations—suggesting minimal changes to input features to alter model outcomes—are a vital approach within XAI. However, current CF generation methods face challenges in balancing proximity, diversity, and robustness, which limits their practical use.

One common framework, Diverse Counterfactual Explanations (DiCE), emphasizes diversity but often lacks robustness, making explanations sensitive to minor input variations. To overcome these limitations, we propose DiCE-Extended, an improved CF explanation framework that employs multi-objective optimization to enhance robustness while preserving interpretability. Our method introduces a new robustness metric based on the Dice-Sørensen coefficient to ensure stability against small perturbations, and uses weighted loss components to balance proximity, diversity, and robustness.

We tested DiCE-Extended on several benchmark datasets—such as COMPAS, Lending Club, German Credit, and Adult Income—using different machine learning platforms like Scikit-learn, PyTorch, and TensorFlow. The results show that our approach produces CF explanations that are more valid, stable, and better aligned with decision boundaries compared to standard DiCE explanations.

This work demonstrates the potential of DiCE-Extended in generating more reliable and interpretable counterfactuals for applications with significant real-world impact. Future research will focus on incorporating adaptive optimization techniques and domain-specific constraints to further improve CF generation in complex environments.

***Tarih*: 21 Mayıs 2025 Çarşamba**

***Saat*: 14:30-15:30**

***Yer*: Fen-Edebiyat Fakültesi B1-326**

***İletişim:* sezert22@itu.edu.tr**