2025 IEEE International Conference on Consumer Electronics - Taiwan (ICCE-Taiwan) | 979-8-3315-8741-3/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICCE-Taiwan66881.2025.11207909

Feature-Driven Counterfactual Explanations: A SHAP-Based Approach to Dimensionality Reduction in XAI

Yu-Lun Chien*, Chia-Yu Lin†, Ted T. Kuo*
* *College of Artificial Intelligence, National Yang Ming Chiao Tung University, Tainan, Taiwan*
† *Department of Computer Science and Information Engineering, National Central University, Taoyuan, Taiwan*
Corresponding Author: Chia-Yu Lin (sallylin0121@ncu.edu.tw)

***Abstract*—Counterfactual Explanation (CE) is a widely used approach in Explainable AI (XAI), aiming to identify minimal changes to input features that would alter a model's prediction, thereby enhancing model interpretability. However, generating high-quality CEs remains challenging due to issues such as high input dimensionality and inefficient optimization. This paper proposes a feature-driven method that integrates SHapley Additive exPlanations (SHAP) with Diverse Counterfactual Explanation (DiCE). The proposed approach incorporates a dynamic learning rate to improve the validity and proximity of CE, and introduces a SHAP-guided sparsity loss to constrain modifications to the most influential features. Experiments on benchmark datasets demonstrate that our method outperforms DiCE in terms of proximity and sparsity, while maintaining or improving validity. These results indicate that the proposed method produces more concise, interpretable, and actionable explanations.**

***Index Terms*—Explainable AI, counterfactual explanation, curse of dimensionality.**

I. Introduction

Explainable AI (XAI) primarily aims to help humans understand the decision-making process of AI models, thereby enhancing trust and ensuring that AI adheres to ethical and legal standards regarding fairness, safety, and controllability.

Counterfactual Explanation [1] (hereinafter referred to as CE) is a method in XAI designed to answer the question: "What input changes are needed to obtain a different prediction?" This approach helps users understand the AI model's decision-making process and provides concrete adjustment suggestions.CE must satisfy several fundamental principles:

- Validity:
  The perturbed input should lead to a change in the model's prediction. Example: The original input results in a loan rejection, but after increasing the salary by $1000, the loan is approved.
- Proximity:
  The change to the input should be as small as possible. Example: If increasing the salary leads to loan approval, then an increase of $1000 is more desirable than an increase of $5000.
- Sparsity:
  The number of features that are changed should be as few as possible.

However, during the process of searching for CE, it may be difficult to find reasonable CE due to the characteristics of the data and the high-dimensionality problem. Existing methods that use gradient descent to search for CE often make excessive changes to the data. However, the closer a CE is to the original input, the more acceptable and trustworthy it is to users. Many current methods modify all features simultaneously to generate multiple CE. However, these methods fail to identify which features are the key reasons for the prediction. Users generally prefer to find CE by changing as few features as possible.

In this paper, our goal is to improve various metrics of CE when using gradient descent methods to search for them, reduce the chance of not finding valid CE, and generate CE that are sparser and closer to the original input.

II. Methods

This study proposes a novel approach that integrates SHapley Additive ExPlanations (SHAP) [2] and Diverse Counterfactual Explanation (DiCE) [3] to achieve efficient and interpretable model prediction and analysis. The proposed method consists of two main components: the data preprocessing stage and the model explanation stage.

In the data preprocessing stage, features in the dataset are divided into continuous and categorical types, each requiring different treatments. Their original values are preserved for continuous features. The common approach for categorical features is to convert them into one-hot encoding. However, one-hot encoding breaks a single categorical feature into multiple sub-features, making it difficult to assess the overall importance of that categorical feature in later analyses. Therefore, we adopt label encoding, transforming categorical feature values into integer values. After label encoding, we further apply normalization to prevent significant numerical disparities.

In the model explanation stage, we build upon DiCE and incorporate two additional enhancement methods. The first method is learning rate enhancement. When using gradient descent to search for CE, we start with a relatively small learning rate. For input instances where a CE cannot be found, we slightly increase the learning rate at the end of each search round until a CE is found or a predefined number of rounds is reached. The advantage of this approach is that, for instance, close to the decision boundary, small perturbations are sufficient, avoiding excessive shifts. For instance, farther from the boundary, larger updates allow exploration of more distant regions, improving the chance of finding valid CE.

The second method is shapley sparsity loss. We first calculate feature importance using SHAP, then apply an L1 norm regularization during the gradient descent-based CE search to focus more on these important features. This encourages sparsity in the generated CEs, helping reduce the dimensionality of the counterfactual search space and improve the efficiency of the explanation process. For identifying the optimal number

of top important features to focus on, we employ genetic algorithms.

We made modifications based on the formula in the DiCE [3] paper, and the overall loss function is:

$$\mathcal{L} = \arg\min_{c_1,\ldots,c_k} \left\{ \frac{1}{k}\sum_{i=1}^{k} \underbrace{\text{yloss}(f(c_i), y)}_{\text{classification loss}} + \frac{\lambda_1}{k}\sum_{i=1}^{k} \underbrace{\text{dist}(c_i, x)}_{\text{proximity loss}} - \lambda_2 \cdot \underbrace{\text{dpp\_diversity}(c_1, \ldots, c_k)}_{\text{diversity loss}} + \frac{\lambda_3}{k}\sum_{i=1}^{k}\sum_{j=1}^{d} \underbrace{M_j \cdot (|c_{ij} - x_j|)}_{\text{sparsity loss (selected features)}} \right\} \tag{1}$$

Here, c_i represents the i-th CE, k is the number of CE to generate, $f()$ denotes the model, yloss measures the difference between $f(c_i)$ and the desired target y. x represents the input instance, dist() calculates the distance between the CE and the input instance, and dpp_diversity measures the diversity among the CEs. d is the number of features, and M_j is the sparsity mask indicating whether a sparsity penalty should be applied to the j-th feature. It is set to 1 if the penalty should be applied, and 0 otherwise. λ_1, λ_2, and λ_3 represent the weights for the different loss components.

By integrating the learning rate enhancement and SHAP-based shapley sparsity loss methods, this study enhances model interpretability, allowing users to better understand its internal mechanisms and increasing the practical applicability of counterfactual explanations.

III. Experiment

We selected two publicly available datasets from Kaggle for experimentation to evaluate our method: the airline [4] and company [5] datasets. The former contains over 100,000 samples and 22 features, including 4 continuous features and 18 categorical features. The latter consists of 6,819 samples and 94 features, with 93 continuous features and 1 categorical feature.

For both datasets, we train a neural network with two hidden layers, and we use 80% of the data for training, while the remaining 20% is used for testing and also serves as input instances for computing CE.

For the evaluation metrics of CE, we refer to the metrics designed in DiCE, using validity, proximity, and sparsity for assessment. Validity is used to evaluate whether the generated CE is indeed classified into the desired target class by the model. Proximity measures the distance between the input instance and the CE. Sparsity evaluates the number of features that have been changed.

It is important to note that, according to the definitions in the DiCE paper, higher values for all these metrics indicate better performance. Furthermore, since features are divided into continuous and categorical types, proximity is also separated into continuous proximity and categorical proximity.

Tables I and II present the experimental results for the airline and company datasets, respectively. Compared to DiCE, our method outperforms DiCE across all metrics, particularly in the Proximity metric. These experimental results demonstrate the effectiveness of our method, indicating that it can significantly improve the proximity and sparsity of the generated counterfactual explanations (CEs) while maintaining similar validity, making the resulting CEs more beneficial for users.

TABLE I
THE EVALUATION METRICS OF CE ON THE AIRLINE DATASET.

Method	Validity	Continuous-Proximity	Categorical-Proximity	Sparsity
DiCE	0.9935	-25.01561	0.72532	0.59971
Our Method	**1.0**	**-3.74987**	**0.87063**	**0.71754**

TABLE II
THE EVALUATION METRICS OF CE ON THE COMPANY DATASET.

Method	Validity	Continuous-Proximity($\times 10^9$)	Categorical-Proximity	Sparsity
DiCE	0.91175	-719.77	0.455	0.06995
Our Method	**0.9405**	**-52.16**	**0.9575**	**0.08905**

IV. Conclusion

We proposed using a learning rate enhancement to guide the data toward the model's decision boundary during the gradient descent process, thereby improving the proximity of CE. In combination with shapley sparsity loss, which increases focus on important features, we further enhanced sparsity. The integration of these two methods resulted in CE that involves minimal and sparse changes, making it easier for users to understand how to make adjustments. Experimental results on the airline and company datasets demonstrated improvements in both proximity and sparsity metrics using our method. Our approach combines two strategies, and the results on public datasets demonstrate that it can produce more accurate and interpretable counterfactual explanations.

Acknowledgements

This work is jointly sponsored by AUO Corporation, AUO · NYCU Joint Research and Development Center, National Central University, and the National Science and Technology Council (NSTC) under the project NSTC 113-2222-E-008-002.

References

[1] Sandra Wachter, Brent Mittelstadt, and Chris Russell, "Counterfactual explanations without opening the black box: Automated decisions and the gdpr," *Harv. JL & Tech.*, vol. 31, pp. 841, 2017.

[2] Scott M Lundberg and Su-In Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[3] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 607–617.

[4] TJ. Klein, "Airline passenger satisfaction, version 1," https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction/version/1, 2021, Retrieved May 30, 2025.

[5] Deron Liang, Chia-Chi Lu, Chih-Fong Tsai, and Guan-An Shih, "Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study," *European Journal of Operational Research*, vol. 252, no. 2, pp. 561–572, 2016.