[Re] FOCUS: Flexible and Optimizable Counterfactual Explanations for Tree Ensembles

Kyosuke Morita

Heidelberg University

Setup and Motivation

- Counterfactual explanations have been developed to cope with the idea of explaining a machine learning model algorithmically
- A counterfactual, $\overline{x}$, represents a perturbation of the input x within the framework of a tree-based binary classification model f
- The perturbation is designed to yield a divergent prediction such as $f(x) \neq f(\overline{x})$

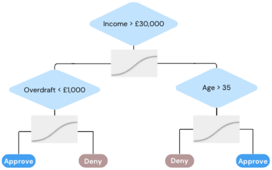
FOCUS

- FOCUS [1] can be applied to non-differentiable models such as tree-based algorithms to generate counterfactual explanations
- This can be done by introducing a probabilistic model approximation $sig(z) = (1 + exp(\sigma \cdot z))^{-1}$, where $\sigma \in \mathbb{R}_{>0}$

Approximated activation $t_j(x)$ with sigmoid function

$$\tilde{t}_j(x) = \begin{cases} 1, & \text{if } j \text{ is the root,} \\ \tilde{t}_{p_j}(x) \cdot sig(\theta_j - x_{f_j}), & \text{if } j \text{ is left child,} \\ \tilde{t}_{p_j}(x) \cdot sig(x_{f_j} - \theta_j), & \text{if } j \text{ is right child,} \end{cases}$$

Decision tree with sigmoid functions approximation



Goal

This paper investigates:

- Whether FOCUS can generate counterfactuals for all instances
- If the mean distance between the original input x and generated counterfactuals $\overline{x}$ is smaller than the existing method
- If FOCUS can perform well with other datasets rather than already tested ones
- How hyperparameters of FOCUS affect its performance

Experimental setup - Data and Evaluation

This paper applied FOCUS on the Decision Tree (DT), Random Forest (RF) and Adaptive Boosting (AB) model on 4 datasets and evaluated them with 4 distance metrics.

Dataset	Sample size	# of features	Positive class ratio
Wine [2]	4,898	9	22%
HELOC [3]	10,459	23	48%
COMPAS [4]	6,172	6	48%
Shopping [5]	12,330	9	15%

Results - Reproducibility

The main findings are:

- FOCUS can find counterfactual explanations for all instances in the datasets
- There were slight deviations from the original paper in terms of the mean distances
- Yet, half of them outperformed the existing method's score

Results - Generality

To examine the generality of FOCUS, this paper applied FOCUS on the German Credit dataset [6].
This paper found:

- FOCUS can find counterfactual explanations for all instances of the DT model
- This study was unable to run one experiment due to the large memory consumption

Results - Hyperparameters

This study found that the quality of model approximation has a significant effect on the performance of FOCUS.

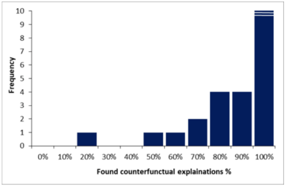


Figure: Found counterfactual explanations % on COMPAS dataset

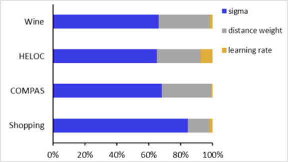


Figure: Hyperparameter importance for the 4 datasets

Conclusion

- FOCUS can find counterfactuals for most instances across the experiments
- The majority of those counterfactuals have smaller distances than the existing method's counterfactual explanations
- The computational cost of FOCUS can be demanding, which leads to a run failure
- Hyperparameters, especially sigma have a significant effect on the performance of FOCUS

References I

- A. Lucic, H. Oosterhuis, H. Haned, and M. de Rijke, "Focus: Flexible optimizable counterfactual explanations for tree ensembles," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp. 5313–5322, 2022.
- P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, "Modeling wine preferences by data mining from physicochemical properties," *Decision support systems*, vol. 47, no. 4, pp. 547–553, 2009.
- FICO2017, "Heloc dataset," 2017.
- D. Ofer, "Compas dataset," *Kaggle: https://www. kaggle. com/danofer/compass*, p. 19, 2017.
- C. O. Sakar, S. O. Polat, M. Katircioglu, and Y. Kastro, "Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and lstm recurrent neural networks," *Neural Computing and Applications*, vol. 31, no. 10, pp. 6893–6908, 2019.
- D. Dua and C. Graff, "UCI machine learning repository," 2017.