Density-Guided Robust Counterfactual Explanations on Tabular Data under Model Multiplicity

Jun Tan, Qing Guo, Zicheng Xu, Jinglin Li, Qi Fang*, Ning Gui*

School of Computer Science and Engineering, Central South University, Changsha, China

Email: juntan.csu@gmail.com, ninggui@gmail.com

1 Problem & Motivation

Problem: Given an input $\boldsymbol{x}$ $(f(\boldsymbol{x}) = 0)$, we want to find a counterfactual explanation(CE) $\boldsymbol{x}'$ $(f(\boldsymbol{x}') = 1)$,with minimal cost.

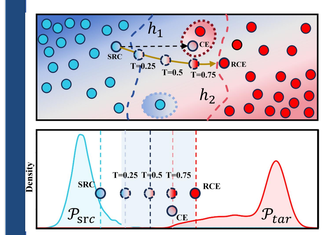
Challenge: Standard CEs often minimize distance to the decision boundary. However, the nearest boundary may lie in low-density regions where different models disagree.

Model multiplicity: multiple plausible classifiers fit the same data but disagree in sparse regions

Goal: How can we generate CEs that are both close to the original input and robust under model multiplicity?

2 Key Insight

Data Density is a Practical Trust Signal — but It is Hard to Obtain



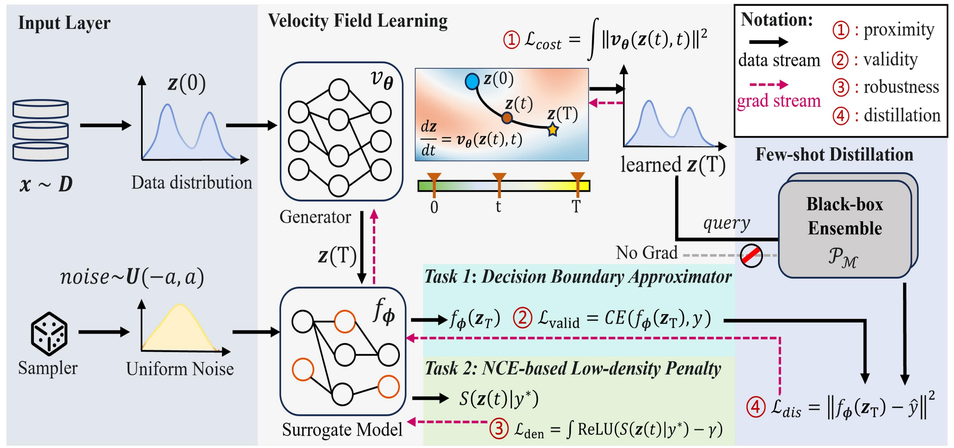
classification in low-density regions are prone to divergence, robust CE should near high-density regions

High-density target support

⬇

Lower model disagreement

3 Framework of Density Flow



Density Flow: Turn density estimation into a Data-vs-Noise discrimination problem and solved using a (K+1) way surrogate classifier

4 Method

1 Surrogate Classifier
(Obtain the density signal)

- Train a $(K+1)$-way classifier *real* *noise*
- Noise class provides reference support

2 Neural ODE
(CE path generation)

- Continuously transport x to x'
- Kinetic energy gives exact path cost

3 Local distillation
(Align with black-box ensemble)

- Use ensemble queries only near the endpoint
- Align the surrogate along visited paths

Final Objective: ***validity*** + ***proximity*** + ***density robustness*** + ***local alignment***

5 Main results

Metrics

$Validity$ (↑):proportion of generated CEs accepted by the black-box ensemble

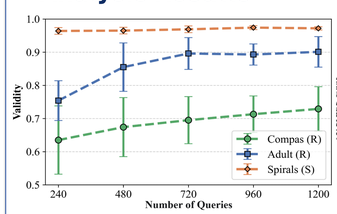
$Cost$(↓) = ℓ_2 distance between instance x and CE x'

Ensemble: [KNN, SVM, RF, MLP, XGB/CatBoost, TabNet]

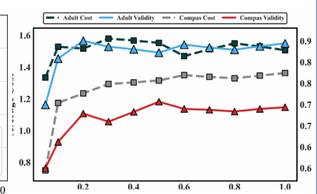
Dataset	Best baseline $V(\uparrow)$ / $C(\downarrow)$	DensityFlow $V(\uparrow)$ / $C(\downarrow)$	ΔV
Adult	0.752 / 2.610	**0.901 / 1.597**	**+14.9**
Compas	0.610 / 1.542	**0.729 / 1.176**	**+11.9**
HELOC	**0.765 / 1.640**	0.757 / 1.812	**−0.8**
Blood	0.509 / 1.548	**0.662 / 1.527**	**+15.3**

Across synthetic and real-world tabular datasets, DensityFlow delivers stronger robustness with low cost.

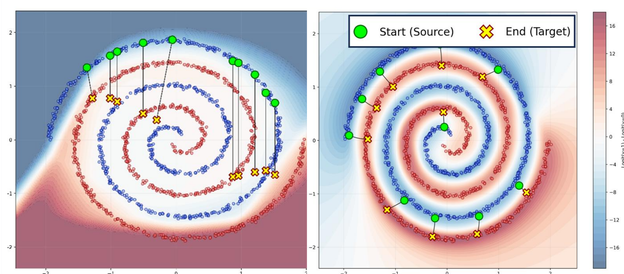
Analysis Results



lower query budget



stable density ratio



Baseline DensityFlow

Visualization: smoother paths on the data manifold

6 Takeaways

Density-guided recourse: valid CEs should also lie on target-class support

NCE support signal: the target-vs-noise logit gap is differentiable and cheap

Black-box efficiency: local distillation queries only the relevant CE trajectory region