Quantifying the Endogenous Domain and Model Shifts Induced by the CLUE Recourse Generator

TU Delft

1 Background

- Algorithmic recourse is the process of generating **counterfactual explanations** (CEs) to classifications made by a black-box machine learning model.
- When algorithmic recourse is applied, the domain and the model can shift.
- We compare the shifts induced in recourse by CLUE [1] to those induced by the baseline Wachter et al. [2] generator (Fig. 1).

Research question: What are the characteristics of shifts induced by the CLUE recourse generator?

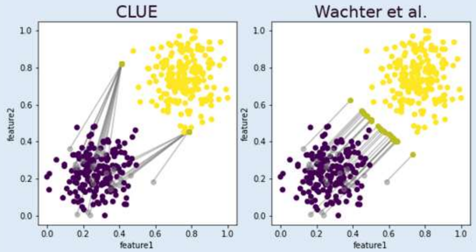

Figure 1. One round of recourse (25 CE) generated using CLUE (left) and Wachter et al. (right)

2 Methods

- Train a classifier model
- Train recourse methods
- Generate CEs using neg. factuals
- Apply CEs to the dataset
- Measure shifts in dynamics

x5

Measurements:

- Model shifts – **Disagreement**, the probability that two classifier models' predictions disagree on an arbitrary point in the domain.
- Domain shifts - **MMD**, a nonparametric statistical measure comparing embeddings of two probability distributions in an RKHS.
- **CE predicted probability**.

3 Results

On synthetic domains, CLUE generates CEs that fall well into the target clusters (Fig. 2) and induces shifts of lower magnitude than Wachter (Tab. 1).

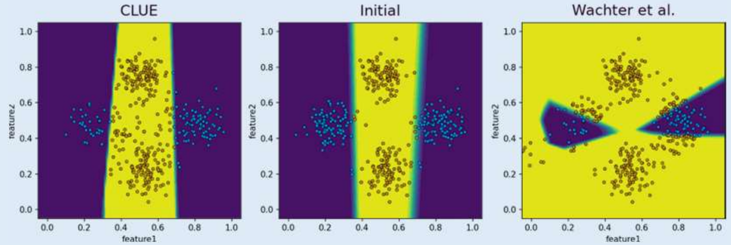

Figure 2. Recourse (10 CEs, 10 rounds) generated by CLUE (left) and Wachter (right) generators on an ANN model with two hidden layers using a synthetic dataset.

For large, real-world datasets, Wachter et al. induces less shifts than CLUE. Both generators perform better on more complex classifier models. CLUE's performance is also influenced by the hyperparameters used to train its VAE.

Dataset		Plus-shaped		GMSC		GC	
Generator		CLUE	Wachter et al.	CLUE	Wachter et al.	CLUE	Wachter et al.
MMD	↓	**0.02**	0.03	0.017	**0.006**	0.0007	**0.0004**
Disagreement	↓	**0.03**	0.07	0.22	**0.18**	0.19	**0.15**
Model MMD	↓	**0.06**	0.09	**0.25**	0.27	0.14	**0.08**
CE Pred. Prob.	↑	**0.88**	0.58	**0.99**	0.87	**0.95**	0.52
Distance	↓	0.29	**0.08**	0.92	**0.03**	2.83	**0.27**
yNN	↑	**0.94**	0.89	**1.00**	0.96	**0.84**	0.61

Table 1. Results for recourse on the plus-shaped dataset (10 CEs, 10 rounds), GMSC (25 CEs, 30 rounds) and GC (15 CEs, 10 rounds).

4 Discussion

- The analysis is limited by the runtimes of the experiments due to long training times of CLUE's VAE and the classifier models.
- The two proposed MMD based model shift metrics pose problems. Model boundary MMD has a high runtime and requires enormous resources, while probability MMD picks up shifts in the dataset.

5 Conclusions

- The proposed metrics and the experimental framework successfully capture and allow analysis of the shifts caused by the recourse process.
- Results show major differences between CEs generated by the two generators stemming from the difference between the objective functions of the generators.
- On all tested domains CLUE's CEs fall better into the target class, while the Wachter et al. generator reduces the distance necessary to employ the explanations.
- It is possible to mitigate shifts to an extent by choosing right classifier models and by providing VAE hyperparameter configurations that are well chosen for the domain under recourse.

6 Future work

- Explore ways of parallelizing the experiments for faster execution times.
- Analyze more combinations of CLUE's VAE in terms of characteristics of induced shifts
- Develop a robust and time efficient model shift metric.

References

[1] J. Antoran, U. Bhatt, T. Adel, A. Weller, and J. M. Hernández-Lobato, "Getting a CLUE: A method for explaining uncertainty estimates," in International Conference on Learning Representations, 2021. DOI: 10.48550/ARXIV.2006.06848.

[2] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the gdpr," Harvard journal of law & technology, vol. 31, pp. 841–887, Apr. 2018. DOI: 10.2139/ssrn.3063289.

Author	Karol Dobiczek k.t.dobiczek@student.tudelft.nl
Supervisor	Patrick Altmeyer
Professor	Cynthia C. S. Liem