Quantifying the Endogenous Domain and Model Shifts Induced by the DiCE Generator

Aleksander Buszydlik

Cynthia C. S. Liem
Responsible Professor

Patrick Altmeyer
Supervisor

Background

Counterfactual explanations (CEs) for black box model decisions in the form of actionable changes are referred to as **algorithmic recourse**.

When recourse is applied, it may lead to **shifts in the domain and model**, we analyze such dynamics for Wachter *et al.* [1] and DiCE [2] generators.

Our research question: ***what are the differences in the characteristics of the domain and model shifts induced by the DiCE and Wachter et al. generators?***

Methods

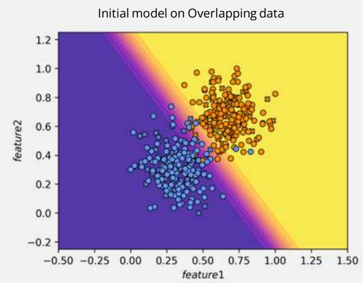
Main metrics for the assessment of shifts:

- **Maximum Mean Discrepancy**, a measure of the distance between the kernel mean embeddings of probability distributions p, q in a Reproducing Kernel Hilbert Space $\mathcal{H}$. It is applied both on the features (MMD) and probabilities predicted by the classifier (PP MMD).
  $MMD[\mathcal{F}, p, q] \coloneqq sup_{f \in \mathcal{F}}(E_x[f(x)] - E_y[f(y)])$.
- **Disagreement Pseudo-distance**, a measure of the overlap between two hypothesis functions.
  $Disagreement(h, h') \coloneqq Pr_{X \sim D}[h(X) \neq h'(X)]$.

Our experimental procedure:

1. Train the classifier
2. Select negative instances
3. Generate recourse
4. Measure observed shifts
5. Update the dataset

Results



On 6 synthetic datasets Wachter *et al.* induces larger domain and model shifts. Its explanations are close to the decision boundary which makes them highly feasible. DICE can generate multiple CEs for each factual instance. They are further from the original points and more dispersed. An example is presented in Figure 1.

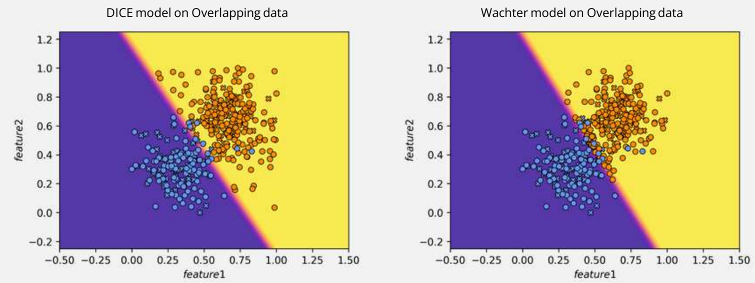
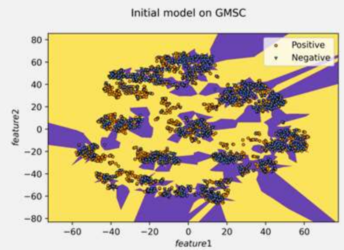


Figure 1. Recourse generated over 10 rounds with 5 counterfactuals per round on Overlapping data.

On the real-world datasets (one of these is shown in Figure 2) DiCE performs much worse than the baseline. It fails to preserve the data manifold. CEs of Wachter *et al.* are clustered with positive factual instances; DiCE generates clusters of counterfactuals.



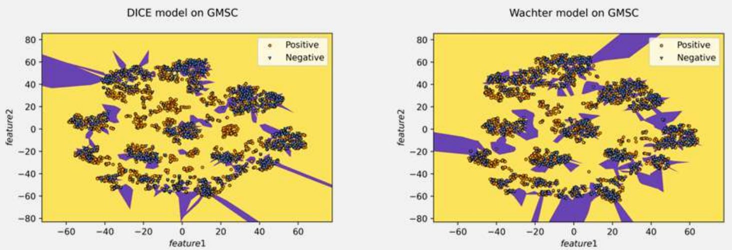


Figure 2. Recourse generated over 15 rounds with 25 counterfactuals per round on GMSC data.

Discussion

Model and Generator	*MMD*	*PP MMD*	*Disagreement*
Synthetic dataset: *Overlapping data*			
(C1) DiCE	**0.0275**	0.2670	**0.0260**
(C1) Wachter *et al.*	0.0854	**0.2492**	0.1535
(C2) DiCE	**0.0401**	**0.1289**	**0.0195**
(C2)Wachter *et al.*	0.0919	0.1677	0.1190
Real-world dataset: *Give Me Some Credit*			
(C1) DiCE	0.1544	0.4138	**0.1737**
(C1) Wachter *et al.*	**0.0567**	**0.3724**	0.2186
(C2) DiCE	0.1619	**0.3422**	**0.0798**
(C2)Wachter *et al.*	**0.0601**	0.3444	0.0955

Table 1. Comparison of the dynamics induced by the two generators.
(C1) is Logistic Regression, (C2) is an ANN with 5 hidden neurons

Wachter *et al.* generates feasible CEs that do not work well on linearly-separable domains.

DiCE induces much larger shifts on the real-world datasets due to its dispersed counterfactuals.

Conclusions

Main findings:

- Both generators typically induce statistically significant domain and model shifts.
- Type of the underlying model and the data distribution influence the magnitude of shifts..

Future work:

- Large-scale comparison of recourse generators.
- Assessment in multi-class scenarios.
- More robust metrics for model shifts.

References

[1] 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, no. 2, 2018.*

[2] R. K. Mothilal, A. Sharma, and C. Tan. "Explaining machine learning classifiers through diverse counterfactual explanations". In: *Proceedings of the 2020 Conferenceon Fairness, Accountability, and Transparency*. ACM, Jan. 2020.

linkedin.com/in/abuszydlik

CSE3000 Research Project

a.j.buszydlik@student.tudelft.nl