GLOBE-CE: A Translation Based Approach for Global Counterfactual Explanations

Dan Ley*[1], Saumitra Mishra[2], Daniele Magazzeni[2]

*Work done as an intern at J.P. Morgan [1]Harvard University, US [2]J.P. Morgan AI Research, UK

Abstract

Counterfactual explanations (CEs) have been widely studied in explainability, though the major shortcoming associated with these methods, is their inability to provide explanations beyond the instance-level. While many works touch upon the notion of a global explanation, typically suggesting to aggregate masses of local explanations, few provide frameworks that are both reliable and computationally tractable. We take this opportunity to propose Global & Efficient Counterfactual Explanations (GLOBE-CE), a flexible framework that tackles the reliability and scalability issues associated with current state-of-the-art, particularly on higher dimensional datasets and in the presence of continuous features. Furthermore, we provide a unique mathematical analysis of categorical feature translations, utilising it in our method. Experimental evaluation demonstrates improved performance across multiple metrics (e.g., speed, reliability).

Introduction

Counterfactual explanations (CEs) in machine learning provide valuable insights into model decisions, but their local focus can limit understanding of global model biases. We seek to address this in the context of global counterfactual explanations (GCEs). We define a GCE to be *a global direction* along which a group of inputs may travel to alter their predictions. Our contributions are as follows:

- We propose a framework that permits GCEs to have variable magnitudes while preserving a fixed translation direction, mitigating the commonly accepted trade-off between coverage and cost.
- We prove that arbitrary translations on one-hot encodings can be expressed using If/Then rules. To the best of our knowledge, this is the first work that addresses mathematically the direct addition of translation vectors to one-hot encodings.
- We demonstrate that GLOBE-CE outperforms competing methods in coverage, cost, and runtime (executing orders of magnitude faster across 4 benchmark datasets and 3 model types).

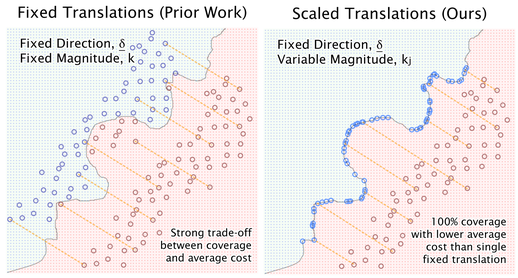

Figure 1: **Left:** Prior work assumes GCEs to be fixed translations/rules. The resulting trade-off between coverage and average cost was discussed in [1]. **Right:** We argue that a fixed direction, variable magnitude set-up can greatly improve performance while retaining interpretability. Each figure demonstrates how one type of GCE works to transfer points from red to blue.

The GLOBE-CE Framework

We propose a novel and interpretable GCE representation: scaled translation vectors, as depicted in Figures 1 and 3. Using too few translations limits the performance of previous methods, yet large numbers of GCEs cannot easily be interpreted. Figure 1 demonstrates conceptually how one can achieve maximum coverage with a single translation at comparably lower average costs to previous methods which do not utilise variable magnitudes (Figure 2 details why this is necessary for bias assessment). The main contribution of the GLOBE-CE framework lies in the notion of *scaling magnitudes*.

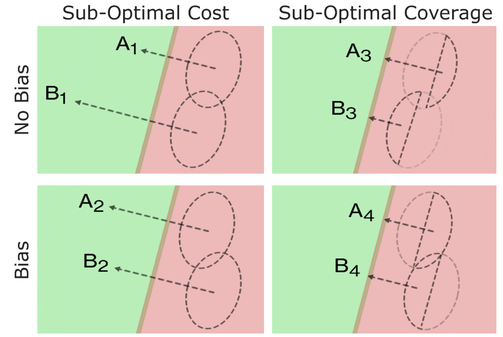

Figure 2: Common pitfalls with *unreliable* recourse bias assessment for GCEs A_i, B_i (ℓ_2 distance represents cost). Conceptual situations of bias/no bias vs sub-optimal cost/coverage. Light dotted lines are all inputs in subgroup A or B; dark dotted lines are inputs in the subgroup for which the respective GCEs apply.

Recourse Bias Assessment

In the absence of minimum cost recourses, biases may be detected where not present (A_1, B_1) or not detected where present (A_2, B_2). Similarly, without sufficient coverage, the same phenomena may occur (A_3, B_3 and A_4, B_4, respectively). The further these metrics stray from optimal, the less likely any potential subgroup comparisons are of being reliable. We argue that maximising reliability thus amounts to maximising coverage while minimising cost. Our experiments evaluate performance along these two dimensions, as well as efficiency (Table 1).

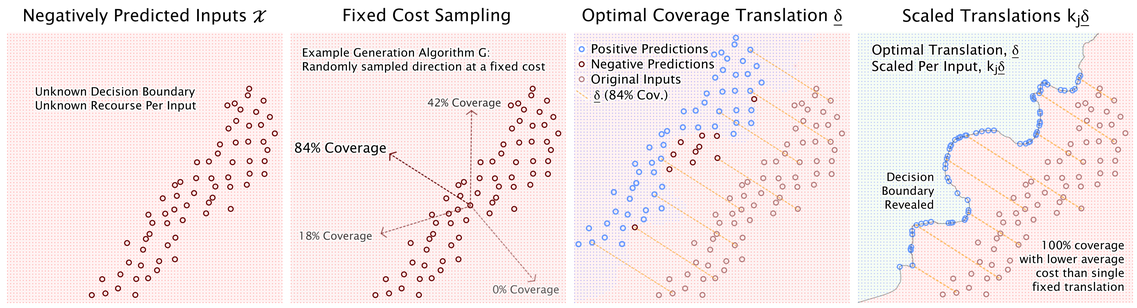

Figure 3: The GLOBE-CE framework for an example generation algorithm G. Cost is ℓ_2 distance. **Left:** Negative predictions, $\mathcal{X}$. **Left Center:** We sample translations at a fixed cost, computing the coverage of each translation. **Right Center:** The translation with highest coverage is selected. **Right:** We scale $\underline{\delta}$ per input, returning the scalar value required for each input.

Table 1: Evaluating the reliability (coverage/cost) and efficiency of GLOBE-CE against AReS [2]. Highlighted in red are GCEs that a) achieve below 10% coverage or b) require computation time in excess of 10,000 seconds (≈3 hours). Best metrics are shown in **bold**. Fast AReS includes our AReS optimisations; dGLOBE-CE uses $d = 3$ GCE directions.

Models	Algorithms	Default Credit			HELOC		
		Cov.	*Cost*	*Time*	*Cov.*	*Cost*	*Time*
DNN	AReS	7.22%	1.0	7984s	5.4%	1.0	9999s
	Fast AReS	99.8%	4.2	37.3s	52%	5.5	109.1s
	GLOBE-CE	98.5%	1.3	**3.6s**	93%	4.3	**4.66s**
	dGLOBE-CE	**100%**	**1.1**	7.86s	**95%**	3.8	5.46s
XGB	AReS	11%	1.0	9999s	1.7%	1.0	9999s
	Fast AReS	93%	2.3	29.97s	28%	**2.1**	93.58s
	GLOBE-CE	96%	1.1	**2.94s**	58%	2.4	**4.7s**
	dGLOBE-CE	**100%**	**0.7**	6.35s	**80%**	2.4	5.6s
LR	AReS	31%	1.2	9999s	4.8%	1.0	9999s
	Fast AReS	99%	2.1	17.82s	92%	1.6	127.3s
	GLOBE-CE	**100%**	**1.0**	**3.42s**	**100%**	**0.5**	**3.11s**
	dGLOBE-CE	**100%**	**1.0**	7.21s	**100%**	**0.5**	3.85s

GLOBE-CE demonstrates consistent improvement across various models and datasets, while operating orders of magnitudes faster than existing methods. AReS can improve coverage over long durations, though this is impractical for relatively simple datasets. Full results, user studies, and translations for categorical features are included in the main text.

Conclusion

We introduce GLOBE-CE, a Global Counterfactual Explanation (GCE) framework that outperforms existing methods by relaxing the objective to permit variable magnitudes per direction. We encourage further research in the underexplored area of GCEs.

References

[1] K. Kanamori et al. "Counterfactual Explanation Trees: Transparent and Consistent Actionable Recourse with Decision Trees". AISTATS 2022.

[2] K. Rawal and H. Lakkaraju. "Beyond Individualized Recourse: Interpretable and Interactive Summaries of Actionable Recourses". NeurIPS 2020.

Poster: Computational Physics and Biophysics Group, Jacobs University