Global Counterfactual Explanations: Investigations, Implementations and Improvements

Dan Ley, Saumitra Mishra, Daniele Magazzeni

J.P. Morgan AI Research, London, UK

Abstract

Counterfactual explanations have been widely studied in explainability, with a range of application dependent methods emerging in fairness, recourse and model understanding. However, the major shortcoming associated with these methods is an inability to yield explanations beyond a local level. While some works touch upon the notion of a global explanation, typically suggesting to aggregate masses of local explanations in the hope of ascertaining global properties, few provide frameworks that are reliable or computationally tractable. Meanwhile, practitioners are requesting efficient and interactive explainability tools. We investigate existing global methods, implementing and improving Actionable Recourse Summaries (AReS), the only known global counterfactual explanation framework for recourse.

Investigations: Existing Methods

Counterfactual explanations (CEs) identify input perturbations that result in desired predictions from machine learning (ML) models. A key benefit of these explanations is their ability to offer recourse to affected individuals in certain scenarios (e.g., automated credit decisioning). However, the research efforts so far have largely centred around local analysis, generating explanations for individual inputs. Such analyses can help vet model behaviour at an instance-level, though it is seldom obvious if the insights gained therein would generalise globally.

Implementations: AReS

[1] investigates this problem, proposing Actionable Recourse Summaries (AReS), a framework that constructs global counterfactual explanations (GCEs). AReS adopts an original, interpretable structure of triples of the form If/If/Then conditions, pictured in Figure 1, Left. However, there exist shortcomings that limit its real-world use. Specifically, we find that AReS is a) **computationally expensive** and b) **sensitive to continuous features**. We propose amendments to the algorithm and demonstrate that these lead to significant performance improvements on two benchmarked financial datasets.

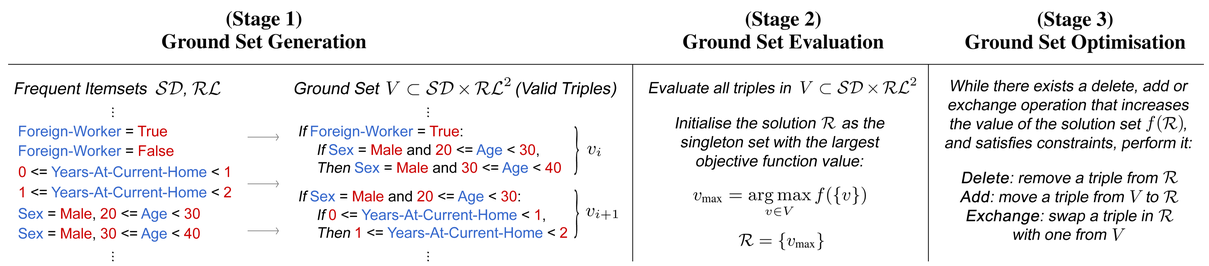

Figure 1: Workflow for our AReS implementation (without improvements). $\mathcal{SD}$ and $\mathcal{RL}$ are assigned to the same set generated by apriori. $\mathcal{SD} \times \mathcal{RL}^2$ is iterated over to compute valid triples (If/If/Then conditions) for the ground set V (Stage 1). Each item in V is evaluated (Stage 2), and the optimisation procedure in [2] is applied (Stage 3), returning the smaller two level recourse set, R.

Improvements: Stage 1

RL-Reduction Iterating over $\mathcal{SD} \times \mathcal{RL}^2$ is wasteful, as many items in $\mathcal{RL}$ will never form valid "If-Then" conditions. We iterate instead over $\mathcal{RL}$ and remove items that contain a feature combination that only occurs once, yielding a reduced $\mathcal{RL}$.

Then-Generation (q) Instead of searching $\mathcal{SD} \times \mathcal{RL}^2$ for triples, we search $\mathcal{SD} \times \mathcal{RL}$ for If conditions, and re-apply apriori, with threshold q, on a filtered dataset to generate Then conditions.

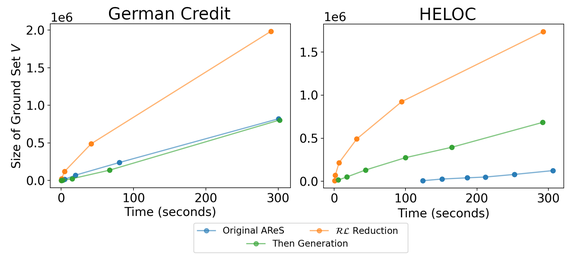

Figure 2: Stage 1 speedups: size of ground set vs time.

Improvements: Stage 2

***V*-Reduction** (r, r') We propose to evaluate a fixed number of triples and form a new ground set by either adding each new triple, or by only adding triples that increase the accuracy of the new set (i.e. vertical blue steps in Figure 3), which we denote r and r' respectively.

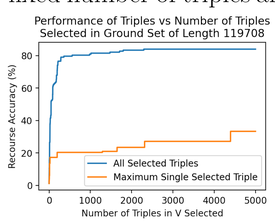

Figure 3: Redundancy in V.

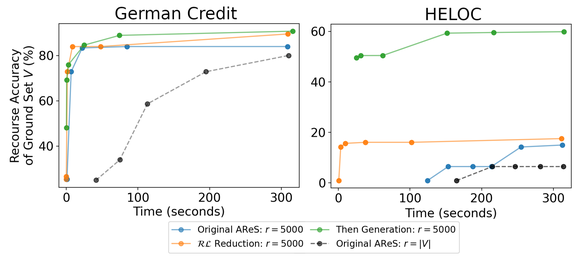

Figure 4: Stage 2 speedups: ground set $acc(V)$ vs time.

Improvements: Stage 3

***V*-Selection** (s) We achieve speedups by further shrinking the ground set pre-optimisation, by sorting the ground set by recourse accuracy (already calculated), and select the s highest-performing triples. If $s = r$ (or $s = r'$) then no sorting occurs.

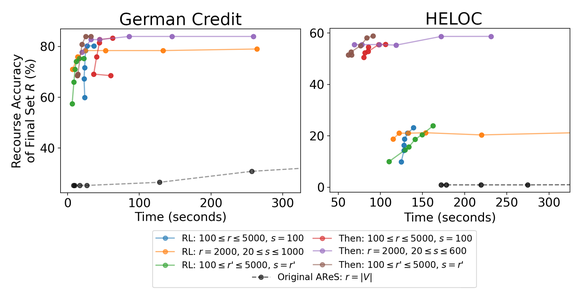

Figure 5: Stage 3 speedups: final set $acc(R)$ vs time.

Experimental Results

In Stage 1, we demonstrate that $\mathcal{RL}$-*Reduction* is capable of generating an equivalent ground set V orders of magnitude faster than the original method, and *Then Generation* also constructs (different) ground sets rapidly (Figure 2). In Stage 2, the saturation of ground set performance after only 1 to 2 percent of full evaluation (Figure 3), causes shrinking ($r = 5000$) to perform significantly better than full evaluation, and *Then Generation* also erases many of the limitations surrounding continuous features (Figure 4). In Stage 3, we finally observe vast speedups, owing to the generation of very small yet high-performing ground sets: r, r' and s restrict the size of V yet retain a near-optimal V (Figure 5).

Conclusion

This work studies the current state of global counterfactual explanations (GCEs), and addresses in detail the scalability issues in the recently proposed AReS framework [1]. We propose improvements to the AReS framework that speed up the generation of GCEs by orders of magnitude, also witnessing significant accuracy improvements on continuous data.

References

[1] Kaivalya Rawal and Himabindu Lakkaraju.
Beyond Individualized Recourse: Interpretable and Interactive Summaries of Actionable Recourses.
In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12187–12198, Virtual-Only, December 2020.

[2] Jon Lee, Vahab Mirrokni, Viswanath Nagarjan, and Maxim Sviridenko.
Non-Monotone Submodular Maximization under Matroid and Knapsack Constraints.
In *Proceedings of the Annual ACM Symposium on Theory of Computing*, Maryland, USA, May 2009.

Acknowledgements

We thank the original authors Kaivalya Rawal and Himabindu Lakkaraju for their helpful discussion of the proposed AReS framework in [1].

Poster: Computational Physics and Biophysics Group, Jacobs University