Counterfactual Situation Testing: Fairness given the Difference

Jose M. Alvarez
University of Pisa & Scuola Normale Superiore
Pisa, Italy
jose.alvarez@sns.it

Salvatore Ruggieri
University of Pisa
Pisa, Italy
salvatore.ruggieri@unipi.it

We present counterfactual situation testing (cfST), a new tool for detecting individual discrimination in datasets that operationalizes the Kohler-Hausmann Critique (KHC) of "fairness given the difference" [2]. In standard situation testing (ST) [5, 6], as with other discrimination analysis tools, the discrimination claim is recreated and thus tested by finding similar individual profiles to the one making the claim, the complainant c, and constructing a control group (*what is*) and a test group (*what would have been if*) of protected and non-protected individuals, respectively. ST builds both groups around c, performing an idealized comparison. Implicitly, from an individual fairness [1] perspective, it assumes that it is fair to compare observably similar protected and non-protected individual profiles to test for discrimination. This approach is wrong under KHC as it fails to account for the pervasive effects of the protected attribute on the relevant attributes used for the decision in question. Under cfST, we extend ST by constructing the control group around the complainant, which is the factual, and the test group around its counterfactual using the abduction, action, and prediction steps for a given structural causal model [4]. Unlike ST, we do not assume the same within-group ordering across protected and non-protected groups. We thus compare groups of not so similar individual profiles: one based on what we observe about c versus one based on a hypothetical, counterfactual representation of c. We compare cfST to existing ST methods along with counterfactual fairness [3] using synthetic data for a loan application process. The results show that cfST detects a higher number of individual discrimination cases than the other methods. The results do not necessarily imply one method is better than the other; we believe they instead highlight the implications of moving away from the dominant idealized comparison model used for detecting individual discrimination.

Acknowledgments and Disclosure of Funding

This work received funding from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie Actions (GA number 860630) for the project "NoBIAS".

References

[1] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. S. Zemel. Fairness through awareness. In *ITCS*, pages 214–226. ACM, 2012.

[2] I. Kohler-Hausmann. Eddie Murphy and the dangers of counterfactual causal thinking about detecting racial discrimination. *Nw. UL Rev.*, 113:1163, 2018.

[3] M. J. Kusner, J. R. Loftus, C. Russell, and R. Silva. Counterfactual fairness. In *NIPS*, pages 4066–4076, 2017.

[4] J. Pearl, M. Glymour, and N. P. Jewell. *Causal Inference in Statistics: A Primer*. John Wiley & Sons, 2016.

[5] B. L. Thanh, S. Ruggieri, and F. Turini. k-NN as an implementation of situation testing for discrimination discovery and prevention. In *KDD*, pages 502–510. ACM, 2011.

[6] L. Zhang, Y. Wu, and X. Wu. Situation testing-based discrimination discovery: A causal inference approach. In *IJCAI*, pages 2718–2724. IJCAI/AAAI Press, 2016.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).