Inequality-Averse Outcome-Based Matching

John Körtner[†] and Ruben Bach[‡]

[†]IDHEAP – University of Lausanne [‡]MZES – University of Mannheim

Automating Inequality?

Algorithmic decision-making raises concerns about reproducing and reinforcing inequalities (popularly formulated by Eubanks, 2018, in her book on "*Automating Inequality*").

★ How to measure and regulate the impact of algorithmic decision-making on inequality?

We study the specific case of outcome-based matching, introduced by Bansak, Ferwerda, Hainmueller, et al., 2018; Acharya, Bansak, and Hainmueller, 2022, where supervised learning models – predicting potential outcomes under available policy options – are leveraged to optimize the mean impact of the policy.

Outcome-Based Matching

Problem. Individuals $j = 1, ..., n$, each described by a vector of features x_j, are to be matched to policy options $k = 1, ..., m$.

Approach. Past data is used to estimate $\mu_k(x)$ to predict the risk of an outcome for individuals under each policy option where the outcome is not (yet) known.

$$\mu_k(x) = \Pr[Y_j(k) = 1 | X_j = x],$$

Thereafter, the idea of outcome-based matching is to use the estimated conditional potential outcomes to assign individuals to options such that the mean impact of the policy is optimized.

$$\Phi_i = \arg\min_{\Phi \in \Pi} \sum_{j=1}^{n} \hat{\mu}_{j\phi_{ij}},$$

where $\Phi_i \in \Pi$ denotes a specific allocation procedure out of all feasible procedures, and ϕ_{ij} the individual assignment under this allocation procedure i.

Criteria of Interest Besides the Mean

In many policy areas, objectives go beyond mean impacts. Algorithms are deployed in the service of institutions that have to trade-off different political principles: besides efficiency, common principles are non-discrimination and distributional objectives such as service to particular groups, or the reduction of inequality.

Discrimination

Predictive. Prior work defines criteria such as predictive equality, equal opportunity, or predictive parity measuring error rates and require variants of $f(\hat{\mu}_j, Y_j | G_j = 1) = f(\hat{\mu}_j, Y_j | G_j = 0)$ to hold, where f denotes a statistical metric (see Verma and Rubin, 2018, for an overview).

✗ Not directly applicable to outcome-based matching, where several policy options are considered.

✗ Criteria do not take into account the effect of the policy interventions. The difference between $\hat{\mu}$ and Y can very well reflect the impact of the intervention instead of a prediction error.

✗ Perspective ignores the assignment step of outcome-based matching.

PolMeth XLI – University of California, Riverside

Consequential. Bansak and Martén, 2021 propose to assess discrimination in the context of outcome-based matching by measuring the causal impact of the algorithmic procedure,

$$\frac{1}{\sum_{j=1}^{n} \mathbb{1}(G_j = g)} \sum_{j=1}^{n} \left(\hat{\mu}_{j\phi_{0j}} - \hat{\mu}_{j\phi_{ij}} \right) \cdot \mathbb{1}(G_j = g),$$

requiring the causal impact to be similar across groups.

✓ Applicable to outcome-based matching with several policy options. Takes into account the effect and assignment of the policy interventions.

✗ Focus on discriminatory impact, not on inequality in outcomes.

Inequality

The focus of this article are concerns that algorithms amplify inequality in outcomes. We can extend Bansak and Martén, 2021 to a measure the distributional impact.

$$\frac{1}{\sum_{j=1}^{n} \mathbb{1}(Q_j = q^s)} \sum_{j=1}^{n} \left(\hat{\mu}_{j\phi_{0j}} - \hat{\mu}_{j\phi_{ij}} \right) \cdot \mathbb{1}(Q_j = q^s),$$

with $q^s = \inf \left\{ q : \frac{1}{n} \sum_{l=1}^{n} \mathbb{1}(\hat{\mu}_{j\phi_{0j}} \leq q) \geq s \right\}$.

Modifying the Objective Function

Where and how do we include the preference for equality in outcomes into the design of the algorithm?

We suggest to incorporate preferences in the objective function. In order to improve equality, we propose an objective function that satisfies the well-known Pigou-Dalton transfer principle and the principle of diminishing transfer.

$$\Phi_i = \arg\min_{\Phi \in \Pi} \sum_{j=1}^{n} \hat{\mu}_{j\phi_{ij}}^{\frac{1}{1-\epsilon}},$$

with $0 \leq \epsilon < 1$. With increasing ϵ, the welfare function becomes increasingly averse to inequality.

Our function is based on the welfare function underlying the Atkinson measure of inequality Atkinson, 1970. The welfare original function is $\frac{1}{n} \sum_{j=1}^{n} y_j^{1-\epsilon} / (1-\epsilon)$. We adapted and simplified the measure to our scenario where higher values (high risk) are less desirable than lower ones (low risk), and outcomes are in the range between zero and one.

Using this function has several important benefits.

✓ Focus on inequality in outcomes.

✓ Offers a modification to the allocation (and not only a criteria).

✓ Disadvantaged individuals receive greater weight in the allocations depending on their level of disadvantage.

✓ Allocation will not pursue equality if this would render the policy completely ineffective. ϵ makes the preference for equality explicit.

Empirical Illustration

In the paper, we illustrate the approach for the case of allocating unemployment insurance claimants to active labor market programs. By performing retrospective counterfactual impact evaluations (Samii, Paler, and Daly, 2016), we show that inequality in outcomes can be reduced and we quantify the trade-offs to other policy objectives such as efficiency and equal impact across protected groups.

Contribution

★ Discuss inequality considerations in outcome-based matching, and extend Bansak and Martén, 2021 to measure distributional impact.

★ Develop an approach to regulate the distributional impact of outcome-based matching based on Atkinson's inequality-averse welfare function.

★ Connect the literature on outcome-based matching (Bansak, Ferwerda, Hainmueller, et al., 2018; Bansak and Martén, 2021; Acharya, Bansak, and Hainmueller, 2022) to equality considerations in the literature on policy learning (e.g., Manski, 2004; Kitagawa and Tetenov, 2021).

Working Paper

References

Acharya, Avidit, Kirk Bansak, and Jens Hainmueller (2022). "Combining Outcome-Based and Preference-Based Matching: A Constrained Priority Mechanism". In: *Political Analysis* 30.1, pp. 89–112.

Atkinson, Anthony B (1970). "On the measurement of inequality". In: *Journal of Economic Theory* 2.3, pp. 244–263.

Bansak, Kirk, Jeremy Ferwerda, Jens Hainmueller, et al. (2018). "Improving refugee integration through data-driven algorithmic assignment". In: *Science* 359.6373, pp. 325–329.

Bansak, Kirk and Linna Martén (2021). "Algorithmic Decision-Making, Fairness, and the Distribution of Impact: Application to Refugee Matching in Sweden". In: *PolMeth 2021*.

Eubanks, Virginia (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

Kitagawa, Toru and Aleksey Tetenov (2021). "Equality-minded treatment choice". In: *Journal of Business & Economic Statistics* 39.2, pp. 561–574.

Manski, Charles F (2004). "Statistical treatment rules for heterogeneous populations". In: *Econometrica* 72.4, pp. 1221–1246.

Samii, Cyrus, Laura Paler, and Sarah Zukerman Daly (2016). "Retrospective causal inference with machine learning ensembles: An application to anti-recidivism policies in Colombia". In: *Political Analysis* 24.4, pp. 434–456.

Verma, Sahil and Julia Rubin (2018). "Fairness definitions explained". In: *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*. IEEE, pp. 1–7.