Beyond Reasonable Doubt: Improving Fairness in Budget-Constrained Decision Making Using Confidence Thresholds

Michiel Bakker[1,2], Duy Patrick Tu[1,2], Krishna Gummadi[3], Alex Pentland[1,2], Kush Varshney[2,4] and Adrian Weller[5,6]

[1]MIT [2]MIT-IBM Watson AI Lab [3]MPI-SWS [4]IBM Research [5]University of Cambridge [6]The Alan Turing Institute

Contact information: bakker@mit.edu

Motivation

Two equally qualified candidates apply for the same job. However, one has a traditional background while the other has taken a more unconventional path. An algorithmic recruiter will choose certainty and hire the familiar candidate. A fair hiring manager, in contrast, would instead first acquire more information before making an equally confident decision for both candidates.

We argue that every individual should have an equal error rate in expectation which we achieve by additional feature collection at prediction time.

Age	28 yrs	26 yrs
Education	Bachelors	Masters
Nationality	British	American
Work Exp	Startups	Banking
University	Liverpool	Colombia
$\mathbf{P_{hiring}}$	**0.7**	**0.9**

Prediction-time active-feature acquisition

We work in a setting where one starts with no information about an individual and additional features can be acquired at feature-specific cost. For this, we need

- A classifier that can handle partial feature sets.
  We use distribution-based imputation for random forests.
- An acquisition strategy that determines which unselected feature should be selected.
  We maximize the cost-normalized expected utility of unselected features.
- A stopping criterion that determines when to stop selecting additional features.
  We use confidence thresholds to attain individual error parity.

Individual error parity

Given a partial feature sets O_i and probabilistic classifier *h* we define the individual-level expected error rate or *risk*

$$r_{err}(\mathcal{O}_i) = \mathbb{E}_{(\mathbf{x},y)\sim P}\left[|h(\mathcal{O}_i) - y|\right]$$

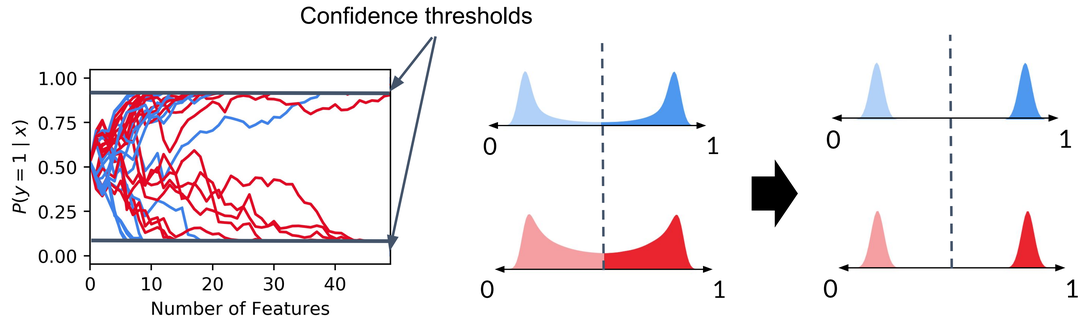
For two individuals *i* and *j* individual error parity requires the individual risk to be equal

$$r_{err}(\mathcal{O}_i) = r_{err}(\mathcal{O}_j)$$

Connection to group fairness

- Perfect individual error parity across a population P necessarily yields **equal accuracy** across groups in P.
- Perfect individual error parity implies **equal false positive and false negative** rates across groups that have equal base rates or across groups with unequal base rate when using group-level calibration.

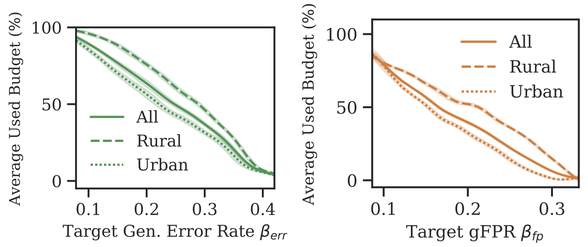
Confidence thresholds



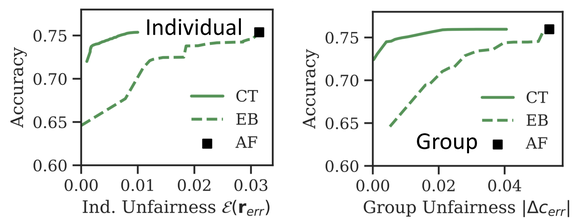
Confidence thresholds distribute feature budget to individuals for which the classifier faces most uncertainty. This yields individual error parity as predictions are equally accurate in expectation.

Results

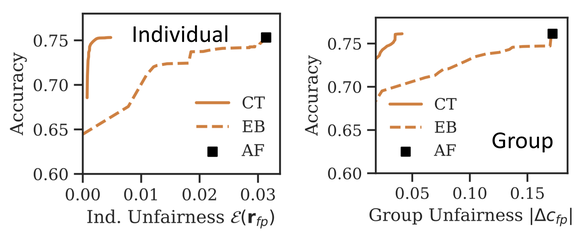
We show how confidence thresholds mitigate both group and individual unfairness using the Mexican Poverty dataset (Noriega-Campero et al. 2019). We benchmark against a baseline where the feature budget is equally distributed across all individuals.



Redistribution of feature budgets across groups.



Residual unfairness when equalizing error rates.



Residual unfairness when equalizing false-positive rates.

Conclusion

We propose individual error parity as an individual fairness notion in an active feature acquisition setting and introduce a method for simultaneously mitigating group and individual unfairness in this setting.

Future work

- Use individual error disparity to guide fairer feature selection at the population level.
- Investigate implications on privacy.
- Effects of miscalibration and mitigation of these effects using individual calibration methods.