Wen Zhou

Department of Evolutionary Anthropology

Duke University

Overweighing underrepresented groups to combat algorithmic discrimination

While algorithms for decision making based on big data have penetrated all aspects of our daily lives, many of them have been tested positive for discrimination. The best-known cases included that ads about arrest records appeared more with searches of Black-sounding names and that females were less delivered job opportunities in the Science, Technology, Engineering and Math (STEM) fields (Lambrecht & Tucker, 2019; Sweeney, 2013).

The grounds of discrimination may be down to the data that algorithms feed on (Hajian, Bonchi, & Castillo, 2016). The training data are originally derived from our social and cultural lives, which can hardly avoid associations with biases existing in human society. For instance, when images with Amero-centric and Euro-centric representation biases were adopted in model training, the exact same biases manifested themselves in the outcome models (Shankar et al., 2017). That is, a system trained on images from Europe would misclassify a Hyderabadi bridegroom as a chain mail armor. When similar biases happened on grounds of social classifications, discrimination emerges. What also concerns people is that algorithms probably further amplify those social biases (Zhao et al., 2017). The clicks, likes, comments and shares are all a part of the feedback loops which power the algorithmic engine.

Nevertheless, the essence of the issue may imply one solution, that is, to balance the representations of different social groups. Even though the source data are usually inherently biased, a decent amount of research has shown great prospects for the development of equality-aware techniques across different phases of data mining. In the phase of preprocessing, promising ways to remove discrimination included suppressing sensitive attributes and relabeling with respect to different social groups (Kamiran, & Calders, 2012). For instance, a common gender discrimination is that algorithms link gendered pronouns to pro-stereotypical entities (e.g., linking “him” to “the physician”) with higher accuracy rather than anti-stereotypical entities (e.g., linking “her” to “the physician”). The debiasing could be accomplished by adding a simple rule of gender swapping (Zhao et al., 2018). In processing phase, anti-discrimination constraints were suggested to be integrated in the algorithm designs (Hajian & Domingo-Ferrer, 2016). Still taking the example of gender discrimination, 47.5% of the bias magnitude could be eliminated by a corpus-level constraint which made gender indicators appear no more often with other elements of the prediction task (Zhao et al., 2017). In postprocessing phase, algorithmic adjustments can again be adopted to rearrange the weights of different classes, thus avoiding discriminative predictions (Danks & London, 2017; Hajian & Domingo-Ferrer, 2016; Kamiran, F., Calders, T., & Pechenizkiy, 2013). It is worth noting that those anti-discrimination techniques performed with loosing little, if any, accuracy for the underlying recognition tasks.

The main challenges of reweighing social groups lie in the ability to discover biases and point to their sources. While statistical models can be applied to detect potential biases in training data (e.g., Bergsma & Lin, 2006), the classification and labeling require awareness and deeper understandings of explicit and implicit social discrimination, which underlines the necessity of supports from disciplines of law, politics, and social sciences. Although the issue of algorithmic discrimination did not receive much attention until the last decade (Pedreschi et al., 2008), recent years have seen an increasing amount of productive cases of interdisciplinary collaboration (e.g., Awad et al., 2018). After all, algorithms are produced not only to guide machine behaviors, but shape peoples' perceptions of the social environment even without the intention. While it is critical to seek equality at both data and code level, the ultimate question is what ethical principles algorithms should follow.

References

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., ... & Rahwan, I. (2018). The moral machine experiment. *Nature*, *563*(7729), 59-64.

Bergsma, S., & Lin, D. (2006, July). Bootstrapping path-based pronoun resolution. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics* (pp. 33-40). Association for Computational Linguistics.

Danks, D., & London, A. J. (2017, August). Algorithmic Bias in Autonomous Systems. In *IJCAI* (pp. 4691-4697).

Hajian, S., Bonchi, F., & Castillo, C. (2016, August). Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 2125-2126).

Hajian, S., & Domingo-Ferrer, J. (2012). A methodology for direct and indirect discrimination prevention in data mining. *IEEE transactions on knowledge and data engineering*, *25*(7), 1445-1459.

Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, *33*(1), 1-33.

Kamiran, F., Calders, T., & Pechenizkiy, M. (2013). Techniques for discrimination-free predictive models. In *Discrimination and privacy in the information society* (pp. 223-239). Springer, Berlin, Heidelberg.

Lambrecht, A., & Tucker, C. (2019). Algorithmic bias? an empirical study of apparent gender-based discrimination in the display of stem career ads. *Management Science*, *65*(7), 2966-2981.

Shankar, S., Halpern, Y., Breck, E., Atwood, J., Wilson, J., & Sculley, D. (2017). No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *arXiv preprint arXiv:1711.08536*.

Sweeney, L. (2013). Discrimination in online ad delivery. *Queue*, *11*(3), 10-29.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K. W. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K. W. (2018). Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*.