Algorithms, fairness, and race: Comparing human recidivism risk assessment with the COMPAS algorithm

Arpita Biswas, Ph.D. student, Department of Computer Science and Automation, Indian Institute of Science

Marta Kołczyńska, Postdoctoral researcher, Institute of Philosophy and Sociology, Polish Academy of Sciences

Saana Rantanen, Postgraduate student, Department of Social Research, University of Turku

Polina Rozenshtein, Ph.D. student in Data Mining Group, Aalto University

Abstract

While in popular perceptions decisions made by computer programs continue to be considered more objective (Sundar and Nass 2001) and accurate (Logg, Minson and Moore 2018) than human decisions, discrimination in algorithmic decision making has become an important topic across different research communities (e.g., social science: Binns et al. 2018, Dressel and Farid 2018; computer science: Liu et al. 2018, Kleinberg et al. 2018; statistics: Johndrow and Lum 2017, Berk et al. 2017). In this paper, we explore the bias in the predictions of the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) algorithm, a criminal risk assessment tool used in sentencing in a number of U.S. states. In doing so we build on and extend prior research by Dressel and Farid (2018), who analyzed the accuracy and fairness of predictions from the COMPAS algorithm matched with a database of 2013-2014 pretrial defendants from Broward County, Florida, and human risk assessment by survey respondents.

Our contribution is twofold. First, we re-analyze the Dressel and Farid (2018) data to (1) explore the consequences of setting risk thresholds, and (2) employ a wider set of fairness metrics than used in prior studies. We find that fairness metrics are highly sensitive to the dichotomization of the COMPAS risk score. We discuss the applicability of the fairness metrics, and the differences between them.

Second, we extend the original study by conducting a new survey with a racial composition that allows for estimating between group differences (an element missing from the study of Dressel and Farid 2018). We explore how the accuracy and fairness of recidivism risk assessment differs between black and white survey respondents, and how it compares to predictions from the COMPAS algorithms. Preliminary results suggest that respondents tend to be more lenient towards the offenders of their own race. We conclude with recommendations for future studies on algorithmic bias.

Secondary data analysis: Re-analysis of existing data

The dataset used by Dressel and Farid (2018) combines data from two sources. The first is a dataset of pre-trial defendants from Broward County, Florida, which are matched to COMPAS predictive risk scores of recidivism (Angwin et al. 2016). The second part comes from a survey conducted by Dressel and Farid (2018) using Mechanical Turk. Respondents were presented descriptions (vignettes) of defendants from the Broward County dataset and asked to assess whether the defendant will commit another crime in the next two years.

We start with a replication of the analyses in Dressel and Farid (2018). Next, we make several extensions and refinements. First, we analyze the consequences of constructing the recidivism risk variable. COMPAS assesses recidivism risk as a score (integer) between 1 and 10, where 1-4 is considered as low, 5-7 as medium, and 8-10 as high risk. Researchers, including Dressel and Farid (2018), typically dichotomize these predictions as 1-4 corresponding to low and 5-10 to high risk. At the same time, in the analysis of the survey results, defendants were considered "high risk" if more than 50% of respondents gave positive answers to predicted recidivism. We show that basic measures of fairness, such as false positive, false negative, false discovery, and false omission rates are highly sensitive to the dichotomization of COMPAS risk score, which raises the issue of comparisons between ranked and binary-labelled data. Next, we perform the analysis considering all three levels of risk and explicit ranking, and also take into account the type of crime. Second, aiming for algorithmic explanation of COMPAS and respondents classification, we perform feature selection and decision rule mining to identify and compare the key features.

Further, we extend the list of fairness metrics. As recent studies have shown, basic statistical metrics used in Dressel and Farid (2018), such as accuracy, false-positive and false-negative rates, can often be misleading. At the same time, reasonable fairness metrics can lead to contradictory results (cf. studies for the same COMPAS dataset by Dieterich et al. 2016, and by Angwin et al. 2016). Other studies also show that it is impossible to simultaneously satisfy a set of common fairness measures (Chouldechova 2016, Corbett-Davies et al. 2017, Kleinberg et al. 2017). Thus, we experiment with a set of novel metrics and methods, which reflect different aspects of bias and discrimination (Verma and Rubin 2018). Among others, we perform benchmark tests, outcome tests (Becker 1957) and recently proposed threshold tests (Simoiu et al. 2017), as well as ranking-based fairness evaluation (Yang and Stoyanovich 2017).

Primary data collection and analysis

The second part of our analysis involves collecting our own data, also via MTurk, and mimicking the design in the Dressel and Farid (2018) study, with some necessary modifications. We decided to use vignettes that mention the defendant's race only (the Dressel and Farid 2018 survey had two variants: with and without race), as this allows us to compare risk assessment by the race of the survey respondent and of the defendant, as well as to investigate in-group bias. Further, we limit the number of vignettes 25 per respondent, and impose a target of 50:50 for the blacks-to-whites ratio in order to test for between-race differences in risk assessment. We have completed the first wave of data collection.

Preliminary analyses show that respondents are more lenient towards the offenders of their own race. However, if we exclude the defendants with medium-risk COMPAS scores and cases with high disagreement among the respondents (majority is supported by less than 3/4 of the respondents of the same race), then race does not play a role and prediction rates agree. This suggests that ambiguous medium-risk cases must be treated with caution and require further analysis. Further analyses and drawing stronger conclusions will be enabled by the second wave of data collection, in the autumn 2018.

References

Angwin, J., Larson, J., Mattu, S. and Kirchner, L., 2016. Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. ProPublica, 23 May 2016; www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Becker, G.S., 1957. The economics of discrimination. University of Chicago Press.

Berk, R., Heidari, H., Jabbari, S., Kearns, M. and Roth, A., 2017. Fairness in criminal justice risk assessments: the state of the art. arXiv preprint arXiv:1703.09207.

Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J. and Shadbolt, N., 2018, April. 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (p. 377). ACM.

Chouldechova, A., 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big data, 5(2), pp.153-163.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S. and Huq, A., 2017, August. Algorithmic decision making and the cost of fairness. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 797-806). ACM.

Dieterich, W., Mendoza, C. and Brennan, T., 2016. COMPAS risk scales: Demonstrating accuracy equity and predictive parity. Northpoint Inc.

Dressel, J. and Farid, H., 2018. The accuracy, fairness, and limits of predicting recidivism. Science advances, 4(1), p.eaao5580.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O. and Zemel, R., 2012, January. Fairness through awareness. In Proceedings of the 3rd innovations in theoretical computer science conference (pp. 214-226). ACM.

Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C. and Venkatasubramanian, S., 2015, August. Certifying and removing disparate impact. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 259-268). ACM.

Hardt, M., Price, E. and Srebro, N., 2016. Equality of opportunity in supervised learning. In Advances in neural information processing systems (pp. 3315-3323).

Johndrow, J.E. and Lum, K., 2017. An algorithm for removing sensitive information: application to race-independent recidivism prediction. arXiv preprint arXiv:1703.04957.

Kamishima, T., Akaho, S., Asoh, H. and Sakuma, J., 2012, September. Fairness-aware classifier with prejudice remover regularizer. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 35-50). Springer, Berlin, Heidelberg.

Kleinberg, J., Ludwig, J., Mullainathan, S. and Rambachan, A, 2018. Algorithmic Fairness. AEA Papers and Proceedings 108:22-27.

Kleinberg, J., Mullainathan, S. and Raghavan, M., 2017. Inherent trade-offs in the fair determination of risk scores. In Innovations in Theoretical Computer Science.

Liu, L.T., Dean, S., Rolf, E., Simchowitz, M. and Hardt, M., 2018. Delayed Impact of Fair Machine Learning. In Thirty-fifth International Conference on Machine Learning (ICML).

Logg, J., Minson, J. and Moore, D.A., 2018. Algorithm Appreciation: People Prefer Algorithmic To Human Judgment. Harvard Business School NOM Unit Working Paper No. 17-086. Available at SSRN: https://ssrn.com/abstract=2941774 or http://dx.doi.org/10.2139/ssrn.2941774.

O'Neil, C., 2016. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. New York, NY: Crown Publishers.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J. and Weinberger, K.Q., 2017. On fairness and calibration. In Advances in Neural Information Processing Systems (pp. 5680-5689).

Simoiu, C., Corbett-Davies, S. and Goel, S., 2017. The problem of infra-marginality in outcome tests for discrimination. The Annals of Applied Statistics, 11(3), pp.1193-1216.

Sundar, S.S. and Nass, C., 2001. Conceptualizing sources in online news. Journal of Communication, 51(1), pp.52-72.

Verma, S. and Rubin, J., 2018. Fairness Definitions Explained. 2018 ACM/IEEE International Workshop on Software Fairness. http://fairware.cs.umass.edu/papers/Verma.pdf

Woodworth, B., Gunasekar, S., Ohannessian, M.I. and Srebro, N., 2017. Learning non-discriminatory predictors. arXiv preprint arXiv:1702.06081.

Yang, K. and Stoyanovich, J., 2017, June. Measuring fairness in ranked outputs. In Proceedings of the 29th International Conference on Scientific and Statistical Database Management (p. 22). ACM.

Zafar, M.B., Valera, I., Rodriguez, M.G. and Gummadi, K.P., 2017. Fairness constraints: Mechanisms for fair classification. In Artificial Intelligence and Statistics (pp. 962-970).

Zemel, R., Wu, Y., Swersky, K., Pitassi, T. and Dwork, C., 2013, February. Learning fair representations. In International Conference on Machine Learning (pp. 325-333).