The Journal of Open Source Software

mcboost: Multi-Calibration Boosting for R

Florian Pfisterer*[1], Christoph Kern[2], Susanne Dandl[1], Matthew Sun[3], Michael P. Kim[4], and Bernd Bischl[1]

1 Ludwig Maximilian University of Munich **2** University of Mannheim **3** Princeton University **4** UC Berkeley

DOI: 10.21105/joss.03453

Software

- Review
- Repository
- Archive

Editor: Øystein Sørensen

Reviewers:

- @mwt
- @OwenWard

Submitted: 09 June 2021
Published: 27 July 2021

Summary

Given the increasing usage of automated prediction systems in the context of high-stakes decisions, a growing body of research focuses on methods for detecting and mitigating biases in algorithmic decision-making. One important framework to audit for and mitigate biases in predictions is that of Multi-Calibration, introduced by Hebert-Johnson et al. (2018). The underlying fairness notion, Multi-Calibration, promotes the idea of multi-group fairness and requires calibrated predictions not only for marginal populations, but also for subpopulations that may be defined by complex intersections of many attributes. A simpler variant of Multi-Calibration, referred to as Multi-Accuracy, requires unbiased predictions for large collections of subpopulations. Hebert-Johnson et al. (2018) proposed a boosting-style algorithm for learning multi-calibrated predictors. Kim et al. (2019) demonstrated how to turn this algorithm into a post-processing strategy to achieve multi-accuracy, demonstrating empirical effectiveness across various domains. This package provides a stable implementation of the multi-calibration algorithm, called MCBoost. In contrast to other Fair ML approaches, MCBoost does not harm the overall utility of a prediction model, but rather aims at improving calibration and accuracy for large sets of subpopulations post-training. MCBoost comes with strong theoretical guarantees, which have been explored formally in Hebert-Johnson et al. (2018), Kim et al. (2019), Dwork et al. (2019), Dwork et al. (2020) and Kim et al. (2021).

`mcboost` implements Multi-Calibration Boosting for R. `mcboost` is model agnostic and allows the user to post-process any supervised machine learning model. It accepts initial models that fit binary outcomes or continuous outcomes with predictions that are in (or scaled to) the range [0, 1]. For convenience and ease of use, `mcboost` tightly integrates with the **mlr3** (Lang et al., 2019) machine learning eco-system in R by allowing to calibrate regression or classification models fitted either within or outside of mlr3. Post-processing with `mcboost` starts with an initial prediction model that is passed on to an auditing algorithm that runs Multi-Calibration-Boosting on a labeled auditing dataset (Fig. 1). The resulting model can be used for obtaining multi-calibrated predictions. `mcboost` includes two pre-defined learners for auditing (ridge regression and decision trees), and allows to easily adjust the learner and its parameters for Multi-Calibration Boosting. Users may also specify a fixed set of subgroups, instead of a learner, on which predictions should be audited. Furthermore, `mcboost` includes utilities to guard against overfitting to the auditing dataset during post-processing.

Fig 1. Conceptual illustration of Multi-Calibration Boosting with `mcboost`.

Statement of need

Given the ubiquitous use of machine learning models in crucial areas and growing concerns of biased predictions for minority subpopulations, Multi-Calibration Boosting should be widely

*Corresponding author

accessible in form of a free and open-source software package. Prior to the development of `mcboost`, Multi-Calibration Boosting has not been released as a software package for R.

The results in Kim et al. (2019) highlight that MCBoost can improve classification accuracy for subpopulations in various settings, including gender detection with image data, income classification with survey data and disease prediction using biomedical data. Barda, Yona, et al. (2020) show that post-processing for Multi-Calibration can greatly improve calibration metrics of two medical risk assessment models when evaluated in subpopulations defined by intersections of age, sex, ethnicity, socioeconomic status and immigration history. Barda, Riesel, et al. (2020) demonstrate that Multi-Calibration can also be used to adjust an initial classifier for a new task. They re-calibrate a baseline model for predicting the risk of severe respiratory infection with data on COVID-19 fatality rates in subpopulations, resulting in an accurate and calibrated COVID-19 mortality prediction model.

We hope that with `mcboost`, Multi-Calibration Boosting can be utilized by a wide community of developers and data scientists to audit and post-process prediction models and helps to promote fairness in machine learning and statistical estimation applications.

Acknowledgements

We thank Matthew Sun for developing an initial Python implementation of MCBoost. This work has been partially supported by the German Federal Ministry of Education and Research (BMBF) under Grant No. 01IS18036A. The authors of this work take full responsibilities for its content.

References

Barda, N., Riesel, D., Akriv, A., Levy, J., Finkel, U., Yona, G., Greenfeld, D., Sheiba, S., Somer, J., Bachmat, E., Rothblum, G., Shalit, U., Netzer, D., Balicer, R., & Dagan, N. (2020). Developing a COVID-19 mortality risk prediction model when individual-level data are not available. *Nature Communications*, *11*, 4439. https://doi.org/10.1038/s41467-020-18297-9

Barda, N., Yona, G., Rothblum, G. N., Greenland, P., Leibowitz, M., Balicer, R., Bachmat, E., & Dagan, N. (2020). Addressing bias in prediction models by improving subpopulation calibration. *Journal of the American Medical Informatics Association*, *28*(3), 549–558. https://doi.org/10.1093/jamia/ocaa283

Dwork, C., Kim, M. P., Reingold, O., Rothblum, G. N., & Yona, G. (2019). Learning from outcomes: Evidence-based rankings. *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, 106–125. https://doi.org/10.1109/FOCS.2019.00016

Dwork, C., Kim, M. P., Reingold, O., Rothblum, G. N., & Yona, G. (2020). *Outcome indistinguishability*. https://arxiv.org/abs/2011.13426

Hebert-Johnson, U., Kim, M., Reingold, O., & Rothblum, G. (2018). Multicalibration: Calibration for the (Computationally-identifiable) masses. In J. Dy & A. Krause (Eds.), *Proceedings of the 35th international conference on machine learning* (Vol. 80, pp. 1939–1948). PMLR.

Kim, M. P., Ghorbani, A., & Zou, J. (2019). Multiaccuracy: Black-box post-processing for fairness in classification. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 247–254. https://doi.org/10.1145/3306618.3314287

Kim, M. P., Kern, C., Goldwasser, S., Kreuter, F., & Reingold, O. (2021). *Universal generalization versus propensity scoring*. Manuscript submitted for publication.

Lang, M., Binder, M., Richter, J., Schratz, P., Pfisterer, F., Coors, S., Au, Q., Casalicchio, G., Kotthoff, L., & Bischl, B. (2019). mlr3: A modern object-oriented machine learning framework in R. *Journal of Open Source Software*. https://doi.org/10.21105/joss.01903