Mitigating Bias in Machine Learning: A Comprehensive Review and Novel Approaches

Mahdi Khalili

The Ohio State University, Columbus, Ohio
khalili.17@ osu.edu

Abstract

Machine Learning (ML) algorithms are increasingly used in our daily lives, yet often exhibit discrimination against protected groups. In this talk, I discuss the growing concern of bias in ML and overview existing approaches to address fairness issues. Then, I present three novel approaches developed by my research group. The first leverages generative AI to eliminate biases in training datasets, the second tackles non-convex problems arise in fair learning, and the third introduces a matrix decomposition-based post-processing approach to identify and eliminate unfair model components.

Summary of Talk

As machine learning algorithms are increasingly being used in applications such as education, lending, recruitment, healthcare, criminal justice, etc., there is a growing concern that the algorithms may exhibit discrimination against protected population groups. Various fairness notions have been proposed in the literature to measure and remedy the biases in ML systems; they can be roughly classified into two categories: 1) *individual fairness* focuses on equity at the individual level and it requires similar individuals to be treated similarly (Zuo, Khalili, and Zhang 2023; Zuo et al. 2024); 2) *group fairness* requires certain statistical measures to be (approximately) equalized across different groups distinguished by some sensitive attributes (Khalili et al. 2021; Khalili, Zhang, and Abroshan 2021; Zhang et al. 2022).

Several approaches have been developed to satisfy a given definition of fairness; they fall under three categories: 1) *pre-processing*, by modifying the original dataset, (e.g., (Abroshan, Elliott, and Khalili 2024)); 2) *in-processing*, by imposing fairness constraints or changing objective functions during training, (e.g., (Khalili, Zhang, and Abroshan 2023)); 3) *post-processing*, by adjusting the output of the algorithms based on sensitive attributes, (e.g., (Khalili, Zhang, and Abroshan 2021)).

In this talk, we review pre-processing, in-processing, and post-processing algorithms and discuss their strengths and weaknesses. Then, we will discuss the approaches that have been developed by my research group. In particular, **first**, we discuss how we can take advantage of *generative AI* to eliminate biases that exists in the training dataset. We explain how the current existing generative models can be modified to take into account various fairness notions and create synthetic data that can improve fairness in downstream tasks. **Second**, the talk focuses on the in-processing fair algorithms developed by my group and explain how fair learning problems can be expressed as non-convex problems and how we can find the global optimal solution in such non-convex problems. **Third**, I introduce a novel post-processing approach for learning fair AI models that can take advantage of SVD to identify components leading to an unfair decision. Then, we discuss how these components can be removed from the the model while ensuring high accuracy.

Acknowledgment: This work is supported by the U.S. NSF under award IIS-2301599 and CMMI-2301601, and by grants from the Ohio State University's TDAI and College of Engineering Strategic Research Initiative.

References

Abroshan, M.; Elliott, A.; and Khalili, M. M. 2024. Imposing Fairness Constraints in Synthetic Data Generation. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*.

Khalili, M. M.; Zhang, X.; and Abroshan, M. 2021. Fair sequential selection using supervised learning models. *Advances in Neural Information Processing Systems*, 34: 28144–28155.

Khalili, M. M.; Zhang, X.; and Abroshan, M. 2023. Loss Balancing for Fair Supervised Learning. In *Proceedings of the 40th International Conference on Machine Learning*.

Khalili, M. M.; Zhang, X.; Abroshan, M.; and Sojoudi, S. 2021. Improving fairness and privacy in selection problems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 8092–8100.

Zhang, X.; Khalili, M. M.; Jin, K.; Naghizadeh, P.; and Liu, M. 2022. Fairness Interventions as (Dis) Incentives for Strategic Manipulation. In *International Conference on Machine Learning*, 26239–26264. PMLR.

Zuo, Z.; Khalili, M. M.; and Zhang, X. 2023. Counterfactually Fair Representation. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Zuo, Z.; Xie, T.; Tan, X.; Zhang, X.; and Khalili, M. M. 2024. Lookahead Counterfactual Fairness. *Transactions on Machine Learning Research*.