Demystifying Algorithmic Fairness in an Uncertain World

Lu Cheng

Computer Science Department, University of Illinois Chicago, Chicago, IL
lucheng@uic.edu

Research Statement

Significant progress in the field of fair machine learning (ML) has been made to counteract algorithmic unfairness against marginalized groups. However, fairness remains an active research area that is far from settled. **One of my research goals** is to address a fundamental gap that exists between fair ML and its execution in the real world, by integrating a critical dimension of "uncertainty", to extend the state-of-the-art fair ML research to a new frontier. In a world that is characterized by *volatility, uncertainty, complexity,* and *ambiguity*, we question whether current fair ML approaches are still effective. For example, recent work has observed a trend of high variance in fairness measures across multiple training runs (Soares et al. 2022). We, therefore, ask: **What fair ML should look like in order to be useful in an uncertain world?**

Uncertainty refers to the extent to which we can confidently predict the future. In neuroscience, researchers found that human brains use uncertain and noisy working memories to make decisions. ML, especially Deep Neural Networks (DNNs), share similarities with the collection of human capacities. Therefore, on one hand, can we similarly harness the power of uncertainty to help ML models make fair and reliable decisions? In the ML ecosystem, predictive uncertainties are commonly decomposed into data (or aleatoric) and model (or epistemic) uncertainties. Data uncertainty refers to the inherent randomness in the outcome of an experiment while model uncertainty can be described as ignorance or a lack of knowledge. On the other hand, a common belief is that bias comes from data and further gets amplified by ML models through different model design choices. So, can we provide a more nuanced understanding of what contributes to algorithmic unfairness from the perspective of predictive uncertainty?

With the goal of *demystifying algorithmic fairness under predictive uncertainties*, in this talk, I will survey three pieces of our work. I will first discuss how to improve algorithmic fairness under data and model uncertainty, respectively. The former regards historical bias reflected in the data and the latter corresponds to the bias perpetuated or amplified during model training due to lack of data or knowledge. In particular, the central hypothesis in the first work (Tahir, Cheng, and Liu 2023) is that aleatoric uncertainty is a key factor for algorithmic fairness, and samples with low aleatoric uncertainty are modeled more accurately and fairly than those with high aleatoric uncertainty. The proposed model is theoretically guaranteed to improve the fairness-utility trade-off. The second work (Wang et al. 2023b) investigates fairness in few-shot learning, where only very few labeled data samples can be collected, leading to potentially inferior fairness performance and difficulty in accurately measuring fairness. To deal with this problem, we devise a novel framework that accumulates fairness-aware knowledge across different meta-training tasks and then generalizes the learned knowledge to meta-test tasks. Lastly, I will introduce coverage-based fairness (Wang et al. 2023a) that ensures different groups enjoy identical treatment and receive equal coverage. To inform the decision maker about the evidence ML can provide while being explicit about the limits of predictive performance, we may want to produce for each sample a predicted interval rather than a point estimate and compare the models' coverage rate across different groups. We propose a new uncertainty-aware fairness – Equal Opportunity of Coverage (EOC) – that aims to achieve equal coverage rates for different groups with similar outcomes and the predetermined coverage rate for the entire population.

Acknowledgements

Cheng is supported by NSF grant #2312862, Cisco Research, and UIC.

References

Soares, I. B.; Wei, D.; Ramamurthy, K. N.; Singh, M.; and Yurochkin, M. 2022. Your Fairness May Vary: Pretrained Language Model Fairness in Toxic Text Classification. In *ACL*.

Tahir, A.; Cheng, L.; and Liu, H. 2023. Fairness through Aleatoric Uncertainty. In *CIKM*.

Wang, F.; Cheng, L.; Guo, R.; Liu, K.; and Yu, P. S. 2023a. Equal opportunity of coverage in fair regression. In *NeurIPS*.

Wang, S.; Ma, J.; Cheng, L.; and Li, J. 2023b. Fair few-shot learning with auxiliary sets. In *ECAI*.