TOP 10 RESEARCH ACHIEVEMENTS

Fair Machine Learning

Department
School of Electrical Engineering

Principal Investigator
Changho Suh

Homepage
http://csuh.kaist.ac.kr

As machine learning (ML) becomes prevalent in our daily lives involving a widening array of applications such as medicine, finance, job hiring and criminal justice, one critical aspect in the design of ML algorithms is to ensure fairness: guaranteeing the irrelevancy of a prediction to sensitive attributes such as gender and race. There has been a proliferation of fair ML algorithms. One challenge that arises in the prior efforts is that they rely upon "proxies" for fairness measures, thereby suffering from achieving the optimal fairness-vs-accuracy tradeoff. In this work, we propose an information-theoretic approach based on mutual information (MI) that can faithfully respect fairness measures, thus obtaining the optimal tradeoff. We also develop a robust version of the method that can well protect against data poisoning. In addition, we develop an upgraded version based on kernel density estimation to address the training instability problem that most MI-based algorithms are faced with. This work was published in the top-tier AI conferences: ICML 2020 and NeurIPS 2020. We expect that our research enables fair ML systems to promote social welfare while embracing equity and inclusion for minority populations.

1. Background

The last decade has witnessed an unprecedented explosion of academic and popular interests in machine learning. Machine learning is now employed to make critical decisions that affect human lives, cultures, and rights, e.g., filtering job applicants, loan screening, and informing bail & parole decision. With a surge of such sensitive applications, there has been a growing concern in the design of trustworthy AI. One critical issue of recent interest is concerning discrimination against minor groups and underrepresented populations. See Fig. 1. Training data often contains biases and this may lead to unfair classifiers that make biased decisions against minority populations. This motivates the need for the design of fair classifiers

Figure1. (Upper) In 2016, ProPublica revealed that COMPAS tends to discriminate black criminals against whites. For instance, a white criminal in the above, named Fugett, was rated a low recidivism score, although he was later rearrested several times.
(lower) U.S. Supreme Court has been employing machine learning software named COMPAS for the purpose of predicting the likelihood of recidivism of criminals. COMPAS outputs a recidivism score, ranging from 0 to 10 (the higher the more likelihood of recidivism).

that can guarantee the irrelevancy of a prediction to sensitive attributes such as gender and race.
There has been a proliferation of fair classifiers. One prominent approach in the prior algorithms is to introduce expressible "proxies" for fairness measures, instead of directly employing fairness measures that are known to be non-straightforward to compute. However, such a proxy-based approach may not ensure the best fairness performance.

2. Contents

During the past years, we have been pioneering this crucial and challenging topic. We made two major contributions. First, we proposed a computationally efficient fair classifier that faithfully respects fairness measures without relying upon their proxies [Paper 1]. The left plot in Fig. 2 demonstrates the accuracy-fairness tradeoff of our proposed algorithm in terms of a popular fairness measure called Disparate Impact (the higher the fairer). We see that ours offers a better tradeoff performance relative to the state of the art that relies upon a proxy of Disparate Impact. Our second contribution is to develope a unified framework that holistically satisfies the two crucial aspects in trustworthy AI: (i) achieving fairness; (ii) ensuring robustness against data poisoning [Paper 2]. As demonstrated in the right plot of Fig. 2, our unified approach can achieve the optimal accuracy-fairness tradeoff, while attaining minimal tradeoff degradation in the presence of data poisoning. In contrast, prior fair algorithms are vulnerable to data poisoning.

3. Expected effect

As machine learning is prevalent in a widening array of sensitive applications that require critical decisions, fair machine learning plays a crucial role to bring about healthy, fruitful, and fair societies that can nurture truly innovative and diverse ideas from a variety of different demographics. Applications are everywhere, including government agencies, financial institutions, companies, public institutions, medical institutions, to name a few. Our fair-robust classifier is a core technology that can equip them with two crucial capabilities: ensuring fairness and protecting against data poisoning, and it is expected to improve the lives of historically underrepresented populations as well as equally support all populations. We believe that our work can promote social welfare while embracing equity and inclusion for minority populations.

[Ref 1]
D. Dua and C. Graff, "UCI machine learning repository," 2017.

[Ref 2]
B. H. Zhang, B. Lemoine and M. Mitchell, "Mitigating unwanted biases with adversarial learning," AIES, 2018.

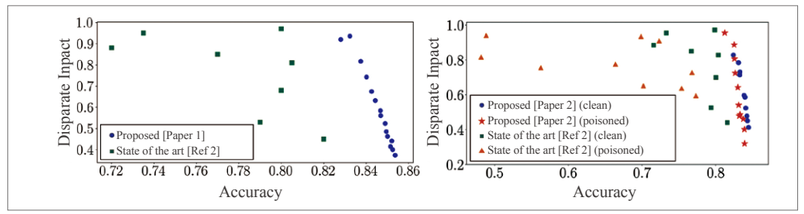

Figure 2. (Left) Accuracy-fairness tradeoff of our proposed algorithm evaluated on AdultCensus [Ref 1]; (Right) Accuracy-fairness tradeoff on clean and poisoned (10% of data) datasets. We observe that our proposed algorithms achieve better tradeoffs compared to Google's recent fair classifier: Adversarial Debiasing [Ref 2].

Research outcomes

[Paper] J. Cho, G. Hwang and C. Suh, "A fair classifier using kernel density estimation," NeurIPS, Dec. 2020 (top-tier AI conference).
Y. Roh, K. Lee, S. E. Whang and C. Suh, "FR-Train: A mutual information-based approach to fair and robust training," ICML, Aug. 2020 (top-tier AI conference).
[Award] A two-year grant from U.S. Air Force Office of Scientific Research (AFOSR)
Naver paper award, Aug. 2020.

Research funding

Air Force Office of Scientific Research (AFOSR) grant funded by the U.S. government (No. FA2386-19-1-4050,"Validating Simulator-based Learning via Interpretation")
Institute for Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2019-0-01396,"Development of framework for analyzing, detecting, mitigating of bias in AI model and training data")