

Highlights:

- We merge reweighing and adversarial approaches for mitigating bias in machine learning models, while keeping the best from both.
- The proposed method can provide interpretable information about fairness of individual instances.
- We provide theoretical analysis of properties of adversarial re-weighting.
- We explore several variants of instance weight estimation including probabilistic ones.
- We evaluate the method on four different real-world datasets, compared to the state-of-the-art techniques, and provide qualitative analysis.