UNIVERSITY OF CALIFORNIA SANTA CRUZ

FAIRNESS WITHOUT HARM: AN INFLUENCE-GUIDED ACTIVE SAMPLING APPROACH

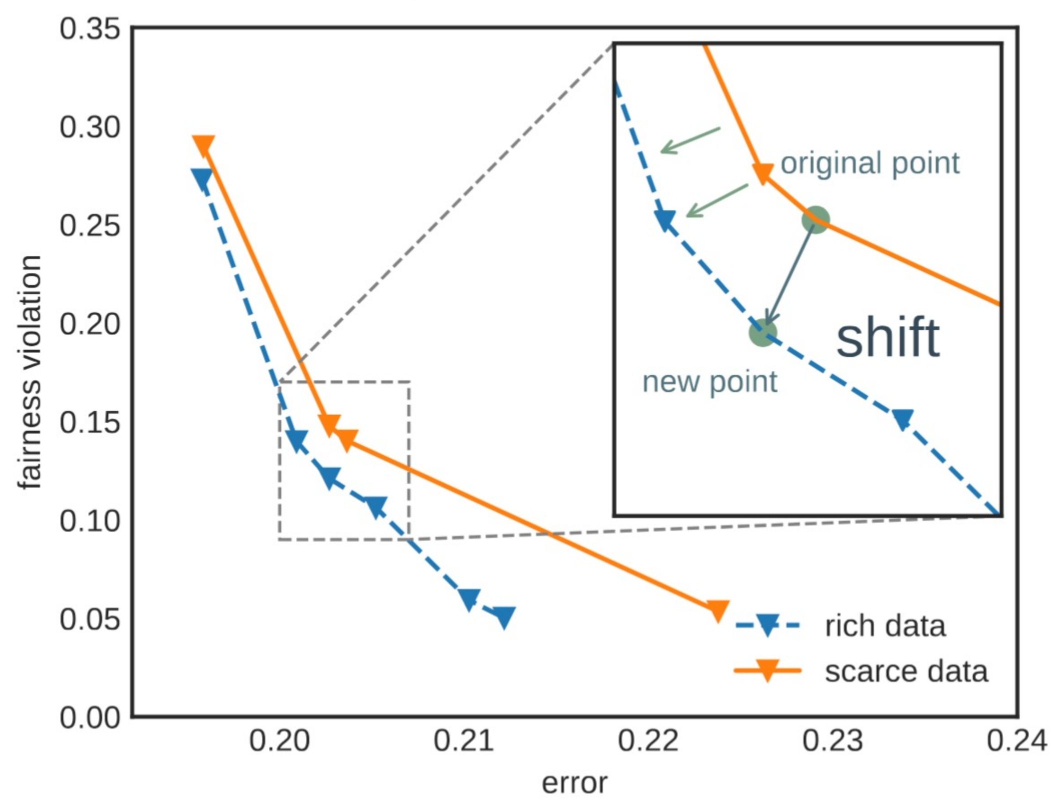
Jinlong Pang†, Jialu Wang†, Zhaowei Zhu‡, Yuanshun Yao*, Chen Qian†, Yang Liu†

†University of California, Santa Cruz ‡Docta.ai *Meta AI

NEURAL INFORMATION PROCESSING SYSTEMS

Problems & Solutions (Overview)

Fairness-accuracy tradeoff phenomenon:
The tradeoff can be explained by the Pareto frontier where given certain resources (e.g., data), reducing the fairness violations often comes at the cost of lowering the model accuracy.



Motivation:
Acquiring more data could help shift to a better Pareto frontier toward low fairness disparity and lower error rates.

One-sentence summary:
We propose a training sensitive attributes-free and tractable active data sampling algorithm solely relying on sensitive attributes on a small validation set.

Solutions: Comparing the gradient direction of the new data with that of the validation set.

Setup

Main goal: Continue training on new active sampling data to find a fair classifier $f \in \mathcal{F}$ using ERM with CE loss:

$$\sum_{n \in \mathbf{P}} \ell(f(x_n; \mathbf{w}), y_n) + \underbrace{\ell(f(x'; \mathbf{w}), y')}_{\text{new inquired examples}}$$

- Original train set: $\mathbf{P} := \{z_n = (x_n, y_n)\}$.
- Unlabeled set: $\mathbf{U} := \{z'_n = (x'_n)\}$ without label.
- Validation set: $\mathbf{Q}_v := \{z^\circ_n = (x^\circ, y^\circ, s^\circ)\}$ with sensitive attributes s°.
- y' is the inquired label for unlabeled examples.
- CE loss $\ell(\cdot,\cdot)$, fairness loss $\phi(\cdot,\cdot)$.

Finding influential examples

SGD process: One step gradient descent on newly acquired example z' is

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \cdot \partial_{\mathbf{w}_t} \ell(\mathbf{w}_t, z')$$

Influence of accuracy/fairness:
The influences of example z' on one validation example z°_n are:

$$\begin{cases} \texttt{Ideal accuracy:} & \mathsf{Infl}_{\mathsf{acc}}(z', z^\circ_n) := \ell(\mathbf{w}_{t+1}, z^\circ_n) - \ell(\mathbf{w}_t, z^\circ_n) \\ \texttt{Ideal fairness:} & \mathsf{Infl}_{\mathsf{fair}}(z', z^\circ_n) := \phi(\mathbf{w}_{t+1}, z^\circ_n) - \phi(\mathbf{w}_t, z^\circ_n) \end{cases}$$

We prove:

Fairness/Accuracy Influence (Lemma 4.1 & Lemma 4.2)

The accuracy/fairness influences of example z' on $\mathbf{Q}_v$ are:

$$\begin{cases} \mathsf{Infl}_{\mathsf{acc}}(z') := \mathsf{MEAN}\Big(\mathsf{Infl}_{\mathsf{acc}}(z', z^\circ_n)\Big) \approx \mathsf{MEAN}\Big(\langle \partial\ell(\mathbf{w}_t, z'), -\eta\partial\ell(\mathbf{w}_t, z^\circ_n)\rangle\Big) \\ \mathsf{Infl}_{\mathsf{fair}}(z') := \mathsf{MEAN}\Big(\mathsf{Infl}_{\mathsf{fair}}(z', z^\circ_n)\Big) \approx \mathsf{MEAN}\Big(\langle \partial\ell(\mathbf{w}_t, z'), -\eta\partial\phi(\mathbf{w}_t, z^\circ_n)\rangle\Big) \end{cases}$$

Intuition: Aligned gradients (reflected by a negative influence score) indicate that this example contributes to improved fairness and accuracy.

Pre-labeling: Utilize lowest-influence labels before querying true labels

$$\hat{y}' = \mathsf{argmin}_{k \in \{1, \cdots, K\}} |\mathsf{Infl}_{\mathsf{acc}}(x', k)|$$

Sampling strategy: Select those samples via influence scores

$$\mathbf{P} \leftarrow \mathbf{P} \cup \{z' \mid \mathsf{Infl}_{\mathsf{acc}}(z') \le 0, \mathsf{Infl}_{\mathsf{fair}}(z') \le 0\}$$

How more data improve fairness without harm?

Fairness definition (Risk disparity)
The model $\mathbf{w}$ would be fair if it achieves the same expected risk on target dataset Q and its group-level subset (Q_k), that is, $\mathcal{R}_{\mathcal{Q}_k}(\mathbf{w}) - \mathcal{R}_{\mathcal{Q}}(\mathbf{w})$.

Proposition (Proposition 3.1 in the paper)
Under appropriate conditions, risk disparity can serve as a lower bound for DP or EOd-based fairness disparities.

Generalization error bound, Theorem 5.1

The generalization error bound of the model trained on P is

$$\mathcal{R}_{\mathcal{Q}}(\mathbf{w}) \le \underbrace{G_P \cdot \mathsf{dist}(\mathcal{P}, \mathcal{Q})}_{\text{distribution shift}} + \sqrt{\frac{\log(4/\delta)}{2|P|}} + \mathcal{R}_P(\mathbf{w}).$$

Upper bound of risk disparity, Theorem 5.2

The upper bound of risk disparity is

$$\mathcal{R}_{\mathcal{Q}_k}(\mathbf{w}) - \mathcal{R}_{\mathcal{Q}}(\mathbf{w}) \le \underbrace{G_k \cdot \mathsf{dist}(\mathcal{P}_k, \mathcal{Q}_k) + G_P \cdot \mathsf{dist}(\mathcal{P}, \mathcal{Q})}_{\text{distribution shift}} + \underbrace{4L^2G^2 \cdot \mathsf{dist}(\mathcal{P}_k, P)^2}_{\text{group gap}} + \Upsilon$$

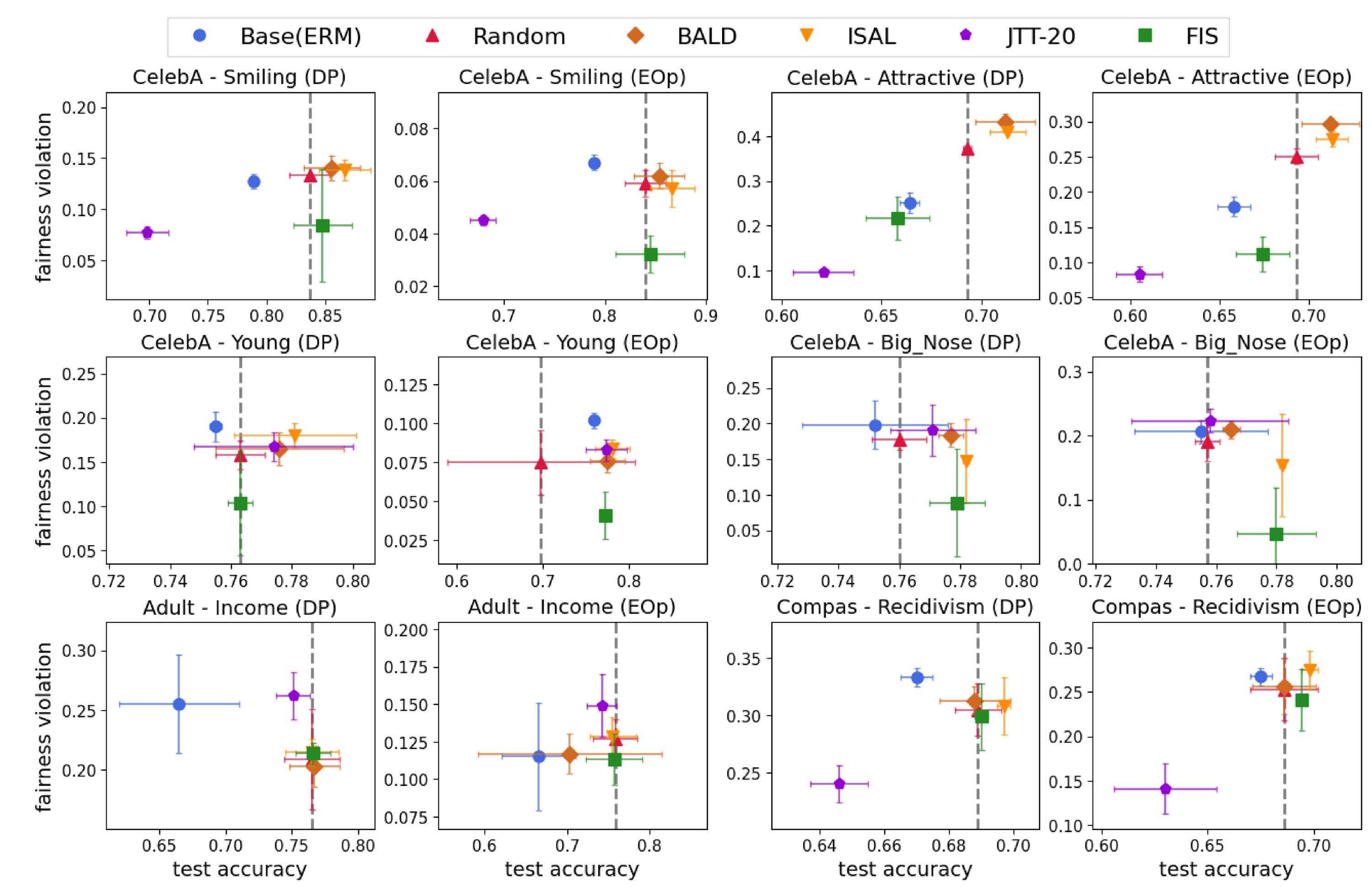
Take-aways:

- Common fair approaches (i.e., reducing group gap) incur additional distribution shifts, leading to an accuracy drop.
- Once the negative impact of distribution shifts can be controlled, it is possible to achieve fairness with harm. (**Our approach**)

Empirical results

Comparison of test accuracy & fairness disparity

- Fairness metrics: DP, EOp, EOd
- Datasets: CelebA, Adult, Compas



Impact of validation set size

Table: Test accuracy & Fairness disparity

	CelebA - Smiling		
	(Test_acc↑, DP↓)	(Test_acc↑, EOp↓)	(Test_acc↑, EOd↓)
$1\times$	(0.848, 0.084)	(0.876, 0.031)	(0.864, 0.030)
$1/2\times$	(0.872, 0.105)	(0.891, 0.042)	(0.880, 0.028)
$1/5\times$	(0.872, 0.117)	(0.863, 0.057)	(0.886, 0.028)

Code