Crowdsourcing with Fairness, Diversity and Budget Constraints

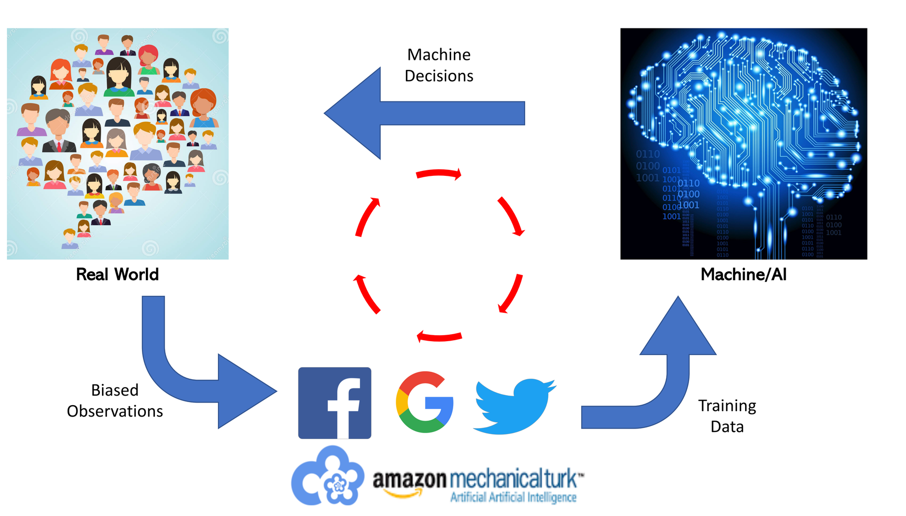
Naman Goel, Boi Faltings

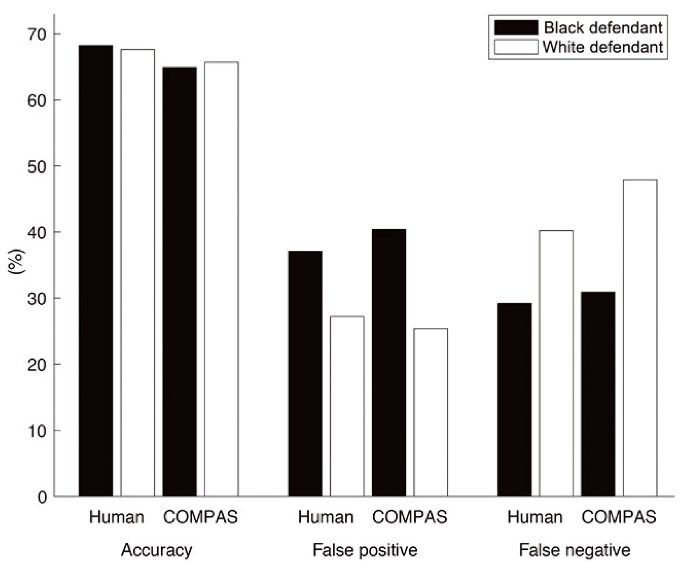
Artificial Intelligence Laboratory, École Polytechnique Fédérale de Lausanne, Switzerland

{naman.goel, boi.faltings}@epfl.ch

Introduction

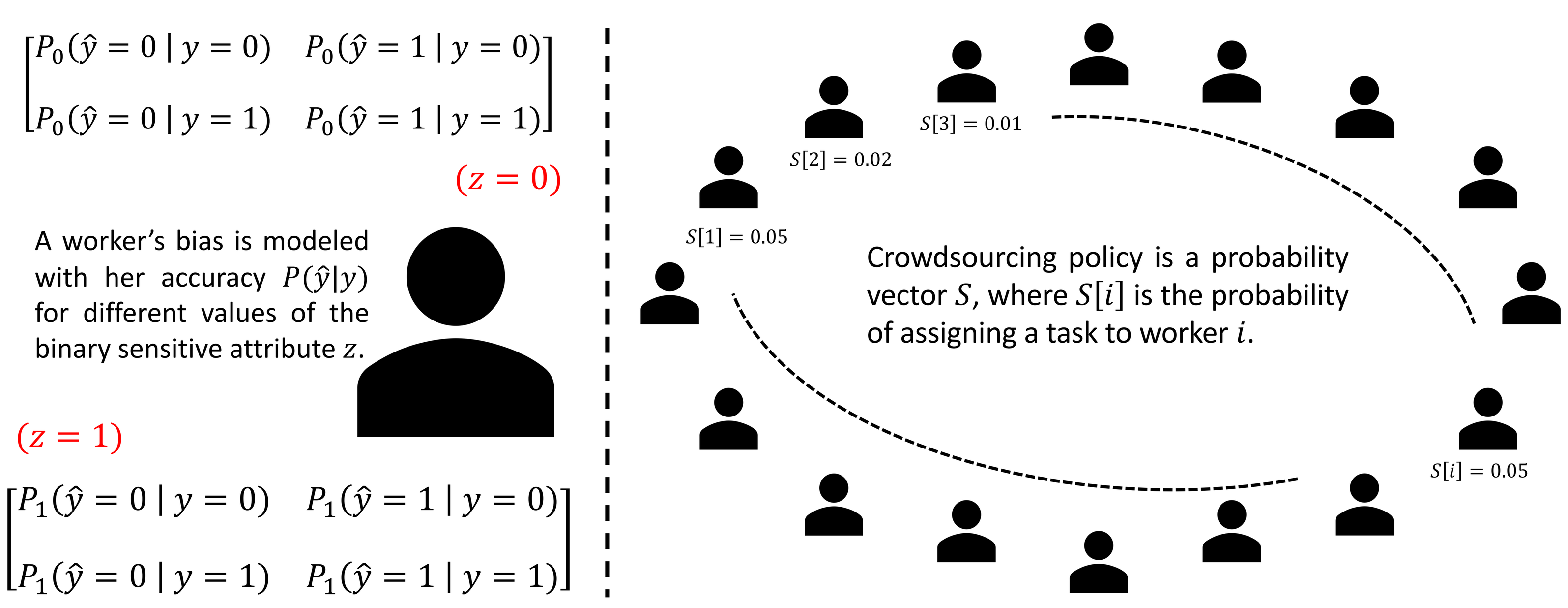
One of the main reasons for bias in algorithms is the bias in the data used for training the algorithms. In this paper, we address the issue of data fairness. More specifically, we consider the problem of collecting fair data from a crowd. Crowdworkers (humans) often have different biases, which are then reflected in the labels collected from them. A recent study conducted by Dressel and Farid [1], in which workers on Amazon Mechanical Turk were asked to submit binary labels to predict recidivism, demonstrated this phenomenon.





- We propose a novel task assignment framework which maximizes the accuracy of the collected data from a biased crowd, while ensuring that 1) the errors satisfy *desired* notions of fairness; 2) the data is collected within the expected budget constraints of the requester; and 3) the task assignment policy is diverse.
- Unlike prior work, our method doesn't require assumptions about the availability of information such as the sensitive attribute and the risk scores (label probabilities) of the tasks. It is also suitable for online crowdsourcing settings (when the requester doesn't know the details or even the number of tasks in advance).

Model



Finding Optimal Crowdsourcing Policy

Expected Accuracy of the Crowdsourcing Policy

$$A_z = \sum_{i=1}^{n} S[i] \cdot A_{iz}$$

Fairness

FPR Parity: $A_0[0,1] = A_1[0,1]$

FNR Parity: $A_0[1,0] = A_1[1,0]$

Error Rate Parity: $A_0 = A_1$

$$\arg\min_S \; -\sum_{z\in\{0,1\}} P(Z=z) \sum_{y\in\{0,1\}} P_z(Y=y) \sum_{i=1}^{n} S[i]\mathcal{A}_{iz}[y,y]$$

subject to

$$\sum_{i=1}^{n} S[i] = 1$$

$$S[i] \geq 0 \quad ,\forall i \in [n]$$

$$S[i] \leq \beta \quad ,\forall i \in [n] \qquad \text{Diversity Constraint}$$

$$\left.\begin{aligned} &\mathcal{A}_0[0,1] - \mathcal{A}_1[0,1] \leq \alpha \\ &-(\mathcal{A}_0[0,1] - \mathcal{A}_1[0,1]) \leq \alpha \\ &\mathcal{A}_0[1,0] - \mathcal{A}_1[1,0] \leq \alpha \\ &-(\mathcal{A}_0[1,0] - \mathcal{A}_1[1,0]) \leq \alpha \end{aligned}\right\} \text{Fairness Constrain}$$

$$\sum_{i=1}^{n} S[i] \cdot c_i \leq C \qquad \text{Budget Constraint}$$

Unknown Accuracy Matrices

Finding the optimal crowdsourcing policy requires knowledge about workers' accuracy matrices. In practice, these are unknown. One way to estimate them is to use a limited number of gold standard tasks. Please see the paper [2] for more discussion and theoretical analysis.

Datasets for Experimental Evaluation

- **Broward County Dataset:** Information about 7214 defendants (3696 black and 2454 white) arrested in Broward County, Florida between 2013 and 2014. The information includes race of defendants among other non-sensitive attributes such as age, prior charges etc. The dataset also contains ground-truth whether the defendants recidivated within 2 years or not.
- **Crowd Judgment Dataset:** Dressel and Farid [1] randomly selected a subset of 1000 defendants from the Broward County dataset and asked 20 random workers on Amazon Mechanical Turk to predict recidivism for each individual. In total, 400 workers participated in their study and each worker submitted answers for 50 different defendants.

Synthetic answers from all 400 crowdworkers were generated for the entire Broward County dataset by 'extrapolating' the Crowd Judgment dataset. Please see the paper for more details about the requirement and process of synthetic answer generation.

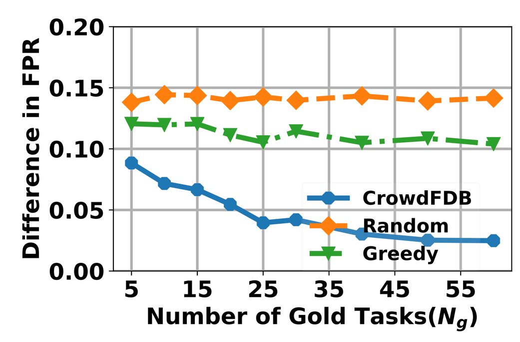
Baselines for Experimental Evaluation

- **Random Policy:** In the random policy, all workers are equally likely to be selected (probability $\frac{1}{400}$) for any task.
- **Greedy Optimization [3]:** Workers are sorted in decreasing order of their "density"(the ratio of the expected accuracy of a worker and her cost). As many as possible tasks are assigned to the highest density worker available (respecting the diversity constraint).

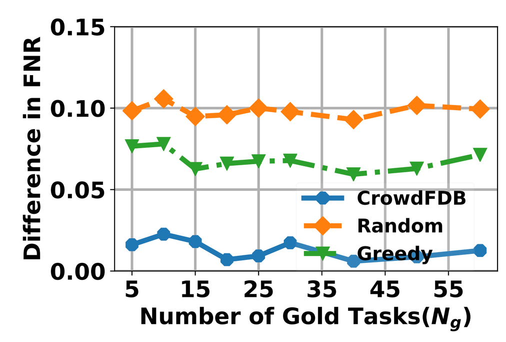
Experiments

- **Uniform cost model** : Every worker has a cost of $\$1$.

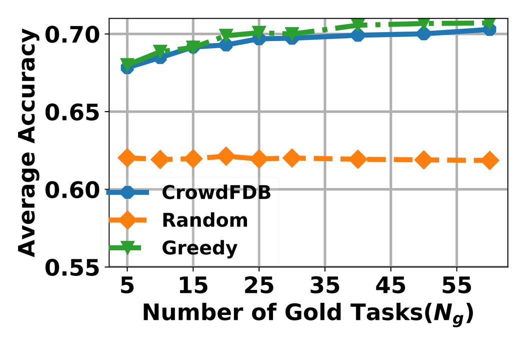
1. Varying N_g (Number of gold tasks), Settings : $\beta = 0.01, \alpha = 0.01$



Absolute Difference in FPR

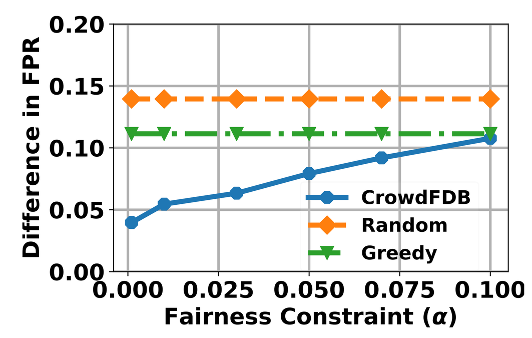


Absolute Difference in FNR

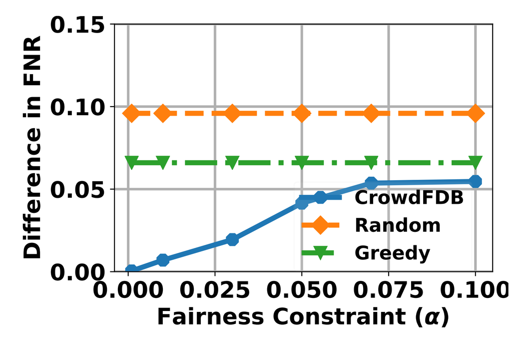


Average Accuracy

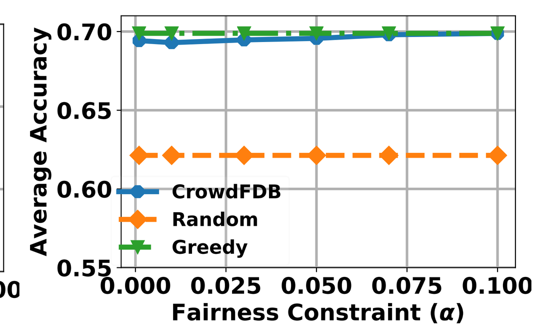
2. Varying α (Fairness Constraint), Settings : $\beta = 0.01, N_g = 20$



Absolute Difference in FPR



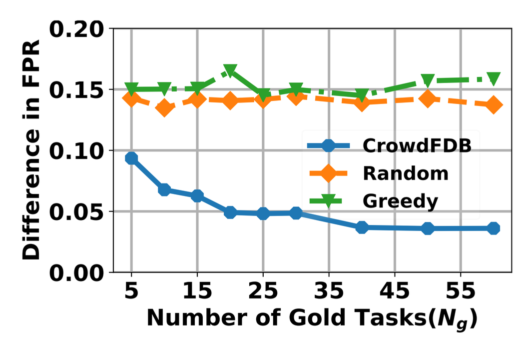
Absolute Difference in FNR



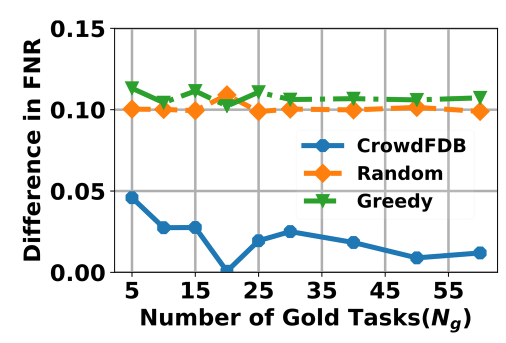
Average Accuracy

- **Non-uniform cost model** : The probability of a worker's cost being $\$3$ is equal to her average accuracy and is $\$1$ otherwise.

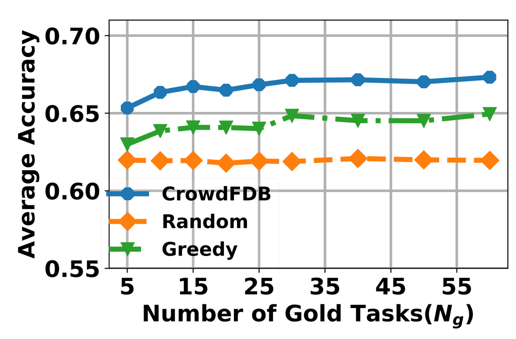
1. Varying N_g (Number of gold tasks), Settings : $\beta = 0.01, \alpha = 0.01, C = 1.5$



Absolute Difference in FPR

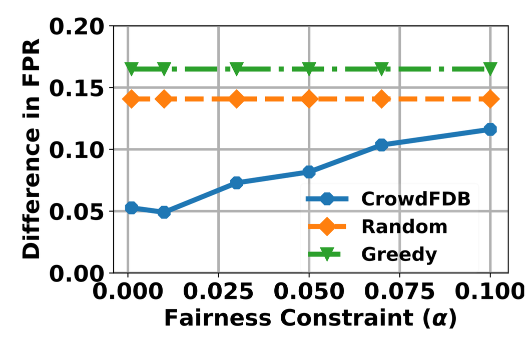


Absolute Difference in FNR

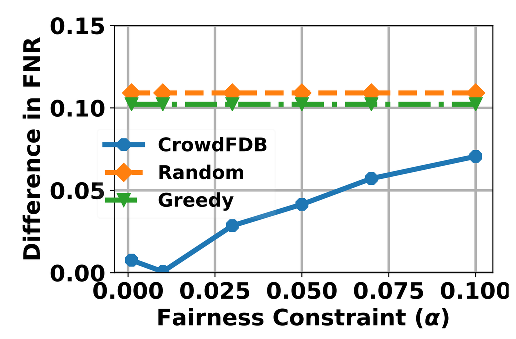


Average Accuracy

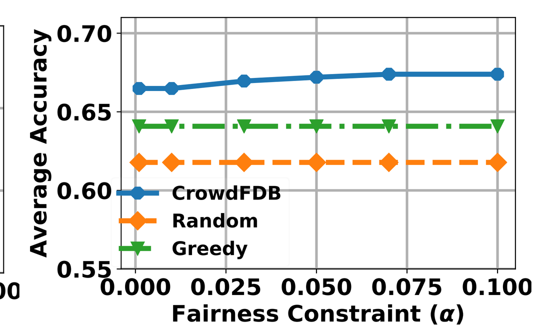
2. Varying α (Fairness Constraint), Settings : $\beta = 0.01, N_g = 20, C = 1.5$



Absolute Difference in FPR



Absolute Difference in FNR



Average Accuracy

References

[1] Dressel J and Farid H. The accuracy, fairness, and limits of predicting recidivism. Science advances. 2018 Jan 1;4(1):eaao5580.

[2] Goel N and Faltings B. Crowdsourcing with Fairness, Diversity and Budget Constraints. Proceedings of the AAAI/ACM Conference on AI, Ethics and Society, 2019.

[3] Tran-Thanh, L., Stein, S., Rogers, A. and Jennings, N. R. Efficient crowdsourcing of unknown experts using bounded multi-armed bandits. Artificial Intelligence 2014.