A Machine Learning-based Framework towards Assessment of Labelers' Biases

Wanxue Dong[1], Maria De-Ardearga[1], and Maytal Saar-Tsechansky[1]

[1]The University of Texas at Austin

{wanxue.dong,dearteaga,Maytal.Saar-Tsechansky}@mccombs.utexas.edu

Abstract

Across key domains, human expert assessments and crowd annotations are essential for labeling data to train machine learning models, and constitute a pathway through which societal biases are learned by algorithms. In this research, we propose a machine learning-based framework to produce a relative assessment of the extent of bias contained in labels produced by different sources, when gold standard labels are costly or difficult to acquire and thus available for only a small set of instances. We provide theoretical guarantees, and we then show empirically that our method outperforms the commonly used alternative of relying on statistical parity to assess biases reflected in human assessments. The proposed approach lays the groundwork towards increased transparency in labelers' biases and offers an important building block towards mitigating algorithmic bias stemming from biased labels.

Keywords: Human Labels, Label Bias, Crowdsourcing, Bias Assessment

A Machine Learning-based Framework towards Assessment of Labelers' Biases

Wanxue Dong[1], Maria De-Ardearga[1], and Maytal Saar-Tsechansky[1]

[1]The University of Texas at Austin

{wanxue.dong,dearteaga,Maytal.Saar-Tsechansky}@mccombs.utexas.edu

1 INTRODUCTION

Human-generated assessments are a prevalent source of labels used to train machine learning models in many important domains. This includes judgments made by expert workers, such as recruiters who evaluate job applicants, as well as crowdsourced annotations made by likely inexpert annotators through crowdworking platforms, e.g., Amazon Mechanical Turk to help in the detection of fake news. Societal biases encoded in human assessments have been highlighted as an important source of algorithmic unfairness (e.g. Suresh and Guttag, 2019; Violago and Quevada, 2018). For example, research has established that labels used to train AI recruiting tool showed bias against women (e.g. Dastin, 2018; Kodiyan, 2019); classifiers can have unfair biases toward certain groups of people if the training data made by humans exhibit such biases (Feldman, 2015). Importantly, assessing labelers' biases is key for assessing the usability of human-generated labels for training ML models, and mitigating the risk of encoding decision makers' biases in algorithmic predictions.

Reducing biases in human assessments has also long been a concern in organizational settings, which has in turn motivated quantitative approaches to measuring it. The most prominent metric relies on assessing selection rates, estimating a labeler's bias by the difference of the proportion of positive labels assigned to instances between different protected groups (Mehrabi et al., 2021). While this metric is easy to estimate, because it does not consider the relationship with a ground truth or gold standard, it fails to differentiate between correct and erroneous labels. On the other hand, met-

rics that assess disparities in errors, such as false positive rates (Hardt et al., 2016), rely on the availability of a gold standard to directly estimate whether a decision label matches the true label for a given instance. Yet, the lack of gold standard labels for the same instances for which human assessments are available precludes the use of these metrics to assess human biases.

Given assessing labeling bias is needed for assessing the usability of human-generated labels for training ML models and for mitigating the risk of encoding labeling biases in algorithmic predictions, in this work we aim to develop the groundwork towards reliable assessments of relative biases in human-generated labels with respect to a desired gold-standard. In particular, we aim to recover the correct ranking of labelers by their relative decision (labeling) biases.

In addition to labelers' data, our method utilizes a key source of data that is frequently available: gold-standard labels for a small, and often disjoint, pool of instances. Such limited gold-standard labeled instances are acquired routinely in some contexts, and they can be compiled in a wide variety of contexts to assess experts or other decision makers. In particular, gold-standard labeled instances are often acquired from costly expert panels to constitute a gold-standard when assessing individual experts' labels, and from scarce professional annotators that constitute a gold-standard for assessing crowdsourced labels. The proposed methodology can effectively leverage large amounts of labelers' biased/noisy labels and a small disjoint set of gold standard data to produce a relative bias assessment of human-generated labels. Lastly, in this paper, we consider bias defined as the difference in a given type of error of interest (e.g. false positives) across different sensitive groups, but we also find that our approach effectively applies to different definitions of bias as well.

We first formalize the problem of relative bias assessment when gold-standard labels are only available for a disjoint pool of data. We then propose a machine learning-based solution to this problem and provide theoretical guarantees. Lastly, we evaluate our approach and compare its performance with the benchmark, selection rate (SR), across different settings and find that the proposed approach yields robust performance, and produces relative assessments that are either superior or otherwise comparable to the baseline.

2 RELATED WORK

The risks of learning from noisy or biased labels are a well-known concern in machine learning. In the context of crowdsourcing, the quality of the labels obtained has been subject to doubt (Nowak and Rüger, 2010), and the impact of different aggregation mechanisms when multiple labels are available per instance has been studied (Davani et al., 2021). Separate lines of works develop algorithms for acquiring (Gao and Saar-Tsechansky, 2020) and learning from noisy labels (Li and Bradic, 2018), with a large body of work studying the robustness of such approaches (e.g. Menon et al., 2016). Crucially, these methods typically assume forms of noise that deviate from the scenarios in which multiple labelers share incorrect beliefs, which is particularly plausible when the goal is to assess labelers' bias, as these may be reflective of widely held societal stereotypes. Our research contributes to this body of work by proposing methodology to assess relative biases across labelers without assuming that the majority will be correct nor requiring the modification of the data collection process, which we achieve by leveraging a small disjoint pool of gold-standard labels.

The problem of assessing human bias and decision quality has been a subject of study across disciplines. There are works focusing on evaluating cognitive bias (e.g. De Martino et al., 2006; Cohen, 1993; Aczel et al., 2015; Chapman and Elstein, 2000) serving different goals and using different methodological approaches than ours, including measuring individual differences in cognitive biases, improving rational thinking and mediating decision biases emotionally or psychologically. Relatedly, there exist works evaluating decision makers' biases among individuals with special traits, for example, alcohol dependence (AD) (Miranda Jr et al., 2009). Other related work addresses the problem of either ranking or directly assessing experts' overall decision accuracy with scarce gold standards, e.g., (e.g. Dong et al., 2021; Geva and Saar-Tsechansky, 2021). However, these works do not consider assessing labelers'/decision-makers' biases; furthermore, Geva and Saar-Tsechansky (2021) also do not consider how ground truth data can be brought to bear. In the context of crowdsourcing, researchers have estimated decision reliability based on workers' various behavioural and demographic traits, e.g., (e.g. Kazai et al., 2013). Yet, these works evaluate decision quality by cen-

tering accuracy, while neglecting the risks of biases that may be contained in human-generated labels and potentially shared among the majority of labelers, and which our work aims to assess.

As part of the approach proposed in this paper, we apply algorithmic fairness methodologies developed in the recent years. Bias mitigation strategies broadly fall under three lines of work: casual fairness (e.g. Barabas et al., 2018), individual fairness (e.g. Binns, 2020), and group fairness (e.g. Kamiran and Calders, 2010). Our proposed method leverages the fact that ML models are prone to replicating bias contained in training labels, and we thus also integrate algorithmic fairness methodologies to disentangle the bias introduced during the learning process from the bias coming from the human labels themselves. We do so by implementing a group fairness strategy to mitigate bias with respect to the observed labels via a post-processing approach grounded on Hardt et al. (2016).

3 PROBLEM FORMULATION

We consider a set of K sources of human labels, such as crowd labelers or domain experts, $L = \{L^1, ..., L^K\}$, whose decisions $Y' = \{Y'^1, ..., Y'^K\}$ are encoded in historical data of their decisions. In addition, we consider settings where a small set of gold standard labels, $GS = \{X_l, Y_l\}_{l=1}^m$ is available for instances that may not overlap with any of the labelers' own decision sets. Y is the gold standard label vector, available for the set GS, and likely unavailable for the labelers' instance sets $S = \{S_{L^k}\}_{k=1}^K$, where S_{L^k} indicates labeler L^k's instance set. Figure 1 illustrates our settings, including the labelers' decision sets S (left) and a non-overlapping GS data (right).

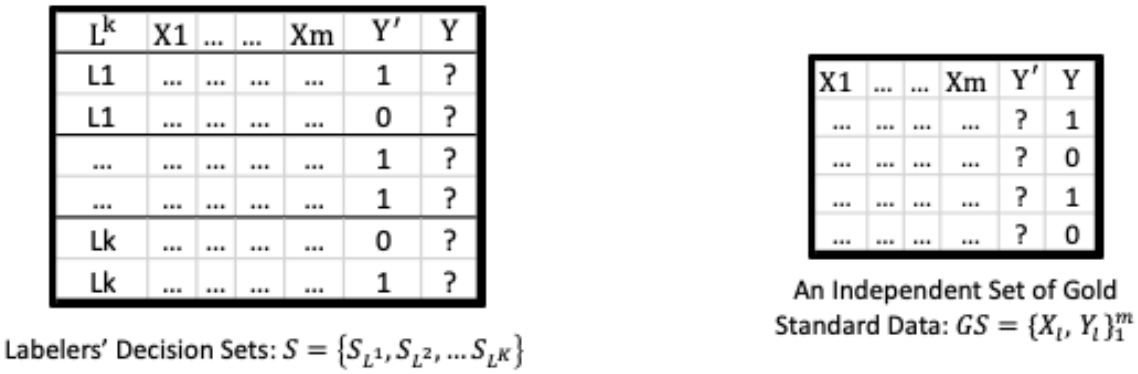


Figure 1. An illustration of labelers' decisions set S (left) and a non-overlapping set with gold-standard labels, GS (right).

For a given labeler, L^k, the labeler's assessment for each instance i, ${Y'}_i^k \in \{0, 1\}$ and its feature vector $X_i^k \sim \mathbb{P}(\mathcal{X})$, are available, such that each labeler has an associated set of instances $S_{L_k} = \{X_i^k, {Y'}_i^k\}_{i=1}^{n_{L^k}}$, where n_{L^k} is the number of instances labeled by labeler L^k. The sets of instances assessed by different labelers need not overlap but should be drawn from the same class distribution. We seek to produce relative assessment of labelers' decision biases, defined as the labelers' relative ranking by their respective biases, where bias can be the difference in true positive rates (TPRs) across groups (GAP) defined by a sensitive attribute A (De-Arteaga et al., 2019), for example, $A = a, \sim a$ (in Eq.1). We consider this measure throughout this paper, but have also found that our approach also applies effectively for different metrics of biases, such as the difference in false positive rates (FPRs) across groups.

$$GAP^{\,k}_{Y'|Y,\,A} = TPR^{\,k}_{Y'|Y,\,a} - TPR^{\,k}_{Y'|Y,\,\sim a} \tag{1}$$

We explore how to leverage scarce and costly gold standard data to assess biases in labelers' decisions. For example, labelers may correspond to a group of crowdworkers or other non-experts, tasked with identifying misinformation in online news stories, which has been proposed as a scalable solution to mitigate misinformation (Allen et al., 2020). In controlled experiments, labelers biases in

this context have been assessed by collecting labels from professional fact-checkers and from crowd-workers for an overlapping pool of cases (Allen et al., 2020). Given the experts limited accessibility, this is both costly and not scalable. In this setting, our work could enable the assessments of relative bias of individual labelers or different sources of labels in newly collected crowdsourced labels (so as to improve learning misinformation detection models from the data) using a previously existing pool of professional assessments.

4 METHOD

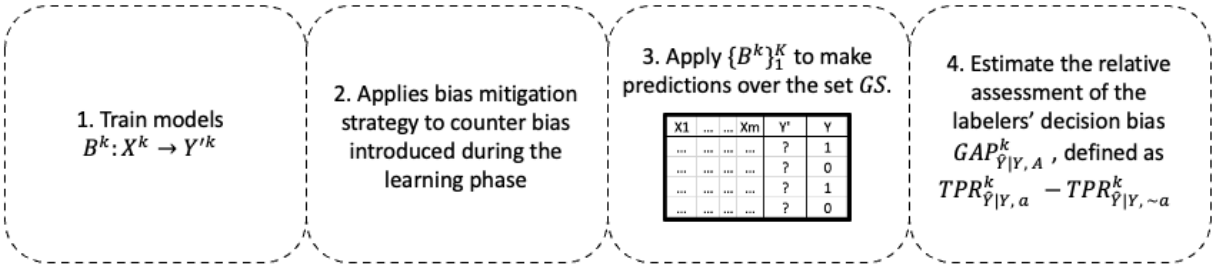
This section first introduces the proposed methodology, then briefly provides theoretical reasoning and guarantees, and finalizes with a detailed description of parameter tuning.

Machine-learning-based labelers' Bias Assessment (MBA)

The proposed **M**achine-learning-based labelers' **B**ias **A**ssessment (MBA) method leverages a typically problematic property of ML models, which are prone to reproducing biases contained in training labels. The proposed approach first trains models to predict each labeler' assessments, yielding a set of models $\{B^k\}_{k=1}^K$, where each model is a mapping $B^k : X^k \mapsto Y'^k$, induced from labeler L^k's data set, S_{L^k}. We ultimate aim to use the models $\{B^k\}_{k=1}^K$ to infer labelers' relative biases. However, biases contained in the models will have multiple sources; in particular, some biases may be introduced during model training and not correspond to (and thereby might compound) the labeler's biases. Thus, the second stage of the proposed algorithm applies a bias mitigation strategy to counter bias introduced during the learning phase, which assesses disparate deviations of a model's prediction $\hat{Y}$ with respect to the label it is trained to predict, Y'. We do so by proposing a recall-versus-precision ratio (RPR) constraint via post-processing, where we consider the group-specific recall and precision, equivalent to true positive rate and positive predictive value of models' predictions $\hat{Y}$ with respect to labelers' decisions Y' given in a protected group, namely $TPR_{\hat{Y}|Y',A}$ and $PPV_{\hat{Y}|Y',A}$. We then apply the set of post-processed models $\{B^k\}_{k=1}^K$ to make predictions over the set GS. Finally, we estimate the relative assessment of the labelers' decision bias, defined

in Equation 1 by assessing biases of $\{B^k\}_{k=1}^K$ with respect to Y, i.e., $GAP_{\hat{Y}|Y,A}$.

Figure 2. Method Key Steps



Algorithm 1: MBA

Algorithm MBA($\{S_{L^k}\}_{k=1}^K$, GS):
  foreach $S_{L^k} \in \{S_{L^k}\}_{k=1}^K$ **do** train base model $B(S_{L^k})$ on S_{L^k} // Step 1
  foreach $B^k \in \{B^k\}_{k=1}^K$ **do**
    $\{\hat{Y}^k\} \leftarrow$ use B^k to classify $\forall\, X_i^k \in S_{L^k}$
    Calculate $\{c'^k_{A=a}, c'^k_{A=\sim a}\}$ based on Eq.16
  $c_{opt.} \leftarrow$ Algorithm 2: Find Optimal C($\{c'^k_{A=a}, c'^k_{A=\sim a}\}_{k=1}^K$)
  $\{\pi^k_{A=a}, \pi^k_{A=\sim a}\}_{k=1}^K \leftarrow$ compute thresholds of $\{B^k\}_{k=1}^K$ with $c_{opt.}$ // Step 2 ends
  foreach $B^k \in \{B^k\}_{k=1}^K$ **do**
    $\{\hat{Y}\}_{l=1}^m \leftarrow$ use B^k with $[\pi^k_{A=a}, \pi^k_{A=\sim a}]$ classify $GS = \{X_l, Y_l\}_{l=1}^m$
    $GAP^{\,k}_{\hat{Y}|Y,A} = TPR^{\,k}_{\hat{Y}|Y,a} - TPR^{\,k}_{\hat{Y}|Y,\sim a}$
  return $\{GAP_{\hat{Y}|Y,A}\}_{k=1}^K$ // Step 3 and 4 end

Figure 2 shows the four key steps in our approach, and the complete procedure is detailed in Algorithm 1 MBA.

Theoretical Analysis

We now show that, given the correct functional form specification of the labelers' models, i.e., functional form of the relationship between the dependent variable and each independent variable, $f : X \mapsto Y'$, our method can recover the correct relative bias assessments of human labelers.

Lemma. *If the correct functional form specification of each labeler' model B, a mapping $f : X \mapsto Y'$ is known, then $\hat{Y} \perp\!\!\!\perp Y|Y'$ and also $Y' \perp\!\!\!\perp Y|\hat{Y}$.*

Algorithm 2: Find Optimal C

Algorithm `Find Optimal C`($\{c'_{j,\,A=a}, c'_{j,\,A=\sim a}\}_{j=1}^{K}$)**:**
  $[c_{min}, c_{max}] \leftarrow$ minimum and maximum of $\{c'_{j,\,A=a}, c'_{j,\,A=\sim a}\}_{j=1}^{K}$
  $c_{step} \leftarrow c_{min}$
  do
    foreach $S_{L_j} = \{X_i^j, Y_i^{\prime j}\}_{i=1}^{n_j} \in \{S_{L_j}\}_{j=1}^{K}$ **do** `/* T-fold cross-validation */`
      Generate T stratified by $A = a, \sim a$ splits: $\{X^j, Y'^j\}_{t=1}^{T}$
      Train a model on the train splits and find corresponding thresholds based on c_{step} and the test split
      $\pi^p_{j,\,A=a}, \pi^p_{j,\,A=\sim a} \leftarrow average(\{\pi^t_{j,\,A=a}, \pi^t_{j,\,A=\sim a}\}_{t=1}^{T})$
    $c_{step} \leftarrow c_{step}+$ step_p
  while $c_{step} \leq c_{max}$
  $steps = \frac{c_{max}-c_{min}}{step\_p}$
  $c_{opt.} \leftarrow c_{step}$ which yields the min($\{std(\{TPR_{j,A=\sim a}\}_{j=1}^{K})_p\}_{p=1}^{steps}$)
  return $c_{opt.}$

Proof. Given the correct functional form for $f_k : X_k \mapsto Y_k'$ then $\hat{Y}_k = {Y_k}' + \epsilon$ where the ϵ is the constant term; and thus, $\hat{Y}_k Y_k | {Y_k}'$. Analogously, if ${Y_k}' = \hat{Y}_k - \epsilon$, then ${Y_k}' Y_k | \hat{Y}_k$. ∎

Theorem. *Given the correct functional form for the labelers models ($f : X \rightarrow Y'$), then there exits a ratio* $\frac{TPR^{l}_{\hat{Y}|Y',\,A}}{PPV^{l}_{\hat{Y}|Y',\,A}} = \frac{TPR^{k}_{\hat{Y}|Y',\,A}}{PPV^{k}_{\hat{Y}|Y',\,A}} = c$*, such that if the biases exhibited in labelers l and k' models are folllowing* $GAP^{\,l}_{\hat{Y}|Y,\,A} > GAP^{\,k}_{\hat{Y}|Y,\,A}$*, then the decision biases of this pair of labelers are also following* $GAP^{\,l}_{Y'|Y,\,A} > GAP^{\,k}_{Y'|Y,\,A}$*, where* $GAP^{\,i}_{\hat{Y}|Y,\,A} = TPR^{\,i}_{\hat{Y}|Y,\,a} - TPR^{\,i}_{\hat{Y}|Y,\,\sim a}$ *and* $GAP^{\,i}_{Y'|Y,\,A} = TPR^{\,i}_{Y'|Y,\,a} - TPR^{\,i}_{Y'|Y,\,\sim a}$.

Proof. Given $GAP^{\,l}_{\hat{Y}|Y,\,A} > GAP^{\,k}_{\hat{Y}|Y,\,A}$, this can be rewritten as:

$$\begin{aligned} &P(\hat{Y}_l = 1|A = 0, Y = 1) - P(\hat{Y}_l = 1|A = 1, Y = 1) > \\ &P(\hat{Y}_k = 1|A = 0, Y = 1) - P(\hat{Y}_k = 1|A = 1, Y = 1) \end{aligned} \tag{2}$$

then

$$\begin{aligned} &P(Y'_l = 1|A = 0, Y = 1) - P(Y'_l = 1|A = 1, Y = 1) > \\ &P(Y'_k = 1|A = 0, Y = 1) - P(Y'_k = 1|A = 1, Y = 1) \end{aligned} \tag{3}$$

It is also true that,

$$P(\hat{Y}_i = 1|A = a, Y = 1) * P(Y_i' = 1|A = a, Y = 1, \hat{Y}_i = 1) = \frac{P(\hat{Y}_i=1,A=a,Y=1)}{P(A=a,Y=1)} * \frac{P(Y_i'=1,A=a,Y=1,\hat{Y}_i=1)}{P(A=a,Y=1,\hat{Y}_i=1)} = \frac{P(Y_i'=1,A=a,Y=1,\hat{Y}_i=1)}{P(A=a,Y=1)} = P(Y_i' = 1, \hat{Y}_i = 1|A = a, Y = 1) \tag{4}$$

By rearranging eq.4, we have

$$P(Y_i' = 1, \hat{Y}_i = 1|A = a, Y = 1) = P(\hat{Y}_i = 1|A = a, Y = 1) * P(Y_i' = 1|A = a, Y = 1, \hat{Y}_i = 1) \tag{5}$$

It is also true that,

$$\frac{P(Y_i'=1,\hat{Y}_i=1|A=a,Y=1)}{P(Y_i'=1|A=a,Y=1)} = \frac{P(Y_i'=1,\hat{Y}_i=1,A=a,Y=1)}{P(A=a,Y=1)} * \frac{P(A=a,Y=1)}{P(Y_i'=1,A=a,Y=1)} = P(\hat{Y}_i = 1|Y_i' = 1, A = a, Y = 1) \tag{6}$$

By rearranging eq.6,

$$P(Y_i' = 1, \hat{Y}_i = 1|A = a, Y = 1) = P(Y_i' = 1|A = a, Y = 1) * P(\hat{Y}_i = 1|Y_i' = 1, A = a, Y = 1) \tag{7}$$

From eq.5 and eq.7,

$$P(\hat{Y}_i = 1|A = a, Y = 1) * P(Y_i' = 1|A = a, Y = 1, \hat{Y}_i = 1) = P(Y_i' = 1|A = a, Y = 1) * P(\hat{Y}_i = 1|Y_i' = 1, A = a, Y = 1) \tag{8}$$

By rearranging eq.8,

$$\frac{P(\hat{Y}_i=1|A=a,Y=1)}{P(Y_i'=1|A=a,Y=1)} = \frac{P(\hat{Y}_i=1|Y_i'=1,A=a,Y=1)}{P(Y_i'=1|\hat{Y}_i=1,A=a,Y=1)} \tag{9}$$

From Lemma 13, it is true that $\hat{Y}Y|Y'$, and $\hat{Y}Y|A, Y'$; therefore,

$$P(\hat{Y}_i = 1|A = a, Y_i' = 1) = P(\hat{Y}_i = 1|Y_i' = 1, A = a, Y = 1) \tag{10}$$

and

$$P(Y_i' = 1|A = a, \hat{Y}_i = 1) = P(Y_i' = 1|\hat{Y}_i = 1, A = a, Y = 1) \tag{11}$$

From eq.9, eq.10, and eq.11, we have

$$\frac{P(\hat{Y}_i=1|A=a,Y=1)}{P(Y_i'=1|A=a,Y=1)} = \frac{P(\hat{Y}_i=1|Y_i'=1,A=a)}{P(Y_i'=1|\hat{Y}_i=1,A=a)} \tag{12}$$

Note that right hand side of eq.12 is the "recall (TPR $_{\hat{Y}|Y',\,A}$) versus precision (PPV $_{\hat{Y}|Y',\,A}$) ratio" and we let the ratio equal to a constant c, so

$$\frac{P(\hat{Y}_i=1|Y_i'=1,A=a)}{P(Y_i'=1|\hat{Y}_i=1,A=a)} = \frac{TPR^{\,i}_{\hat{Y}|Y',\,A}}{PPV^{\,i}_{\hat{Y}|Y',\,A}} = c \tag{13}$$

From eq.13, it is true that

$$\frac{P(\hat{Y}_i=1|A=a,Y=1)}{c} = P(Y_i' = 1|A = a, Y = 1) \tag{14}$$

Given eq.2 above:

$$\begin{aligned} &P(\hat{Y}_l = 1|A = 0, Y = 1) - P(\hat{Y}_l = 1|A = 1, Y = 1) > \\ &P(\hat{Y}_k = 1|A = 0, Y = 1) - P(\hat{Y}_k = 1|A = 1, Y = 1) \end{aligned} \tag{2}$$

dividing both sides by c, we have:

$$\frac{P(\hat{Y}_l = 1|A = 0, Y = 1)}{c} - \frac{P(\hat{Y}_l = 1|A = 1, Y = 1)}{c} > \frac{P(\hat{Y}_k = 1|A = 0, Y = 1)}{c} - \frac{P(\hat{Y}_k = 1|A = 1, Y = 1)}{c} \tag{15}$$

which is equivalent to

$$P(Y'_l = 1|A = 0, Y = 1) - P(Y'_l = 1|A = 1, Y = 1) > P(Y'_k = 1|A = 0, Y = 1) - P(Y'_k = 1|A = 1, Y = 1) \tag{3}$$

Parameter Selection

In this section, we discuss how we derive the ratio c in Theorem to allow recovery of labelers' biases. Note that the ratio c can be simplified as follows:

$$c = \frac{TPR_{\hat{Y}|Y',\,A}}{PPV_{\hat{Y}|Y',\,A}} = \frac{\frac{TP}{TP+FN}}{\frac{TP}{TP+FP}} = \frac{TP+FP}{TP+FN} = \frac{|\hat{Y}=1, A=a, \sim a|}{|Y'=1, A=a, \sim a|} \tag{16}$$

Eq.16 reveals the relationship between a labeler model's positive predictions, $\hat{Y} = 1, A = a, \sim a$, and the actual labeler's positive decisions, $Y' = 1, A = a, \sim a$ for a given group $A = a$ or $A = \sim a$. This relationship implies a corresponding desired probability threshold for classification of instances from each protected group.

There are multiple possible values of c that can satisfy the ratio in Eq.16, each corresponding to a different probability threshold. We use cross validation (cv) to identify a value c. Once c is determined, we adjust the probability threshold of each model to achieve the ratio c. Note that prior to enforcing the desired threshold on all the labelers' models, each model has an initial threshold for each protected group variable value, given by $\{\pi'_{A=a}, \pi'_{A=\sim a}\} = \{0.5, 0.5\}$. The ultimate threshold

pairs, given by $\pi^k_{A=a}$ and $\pi^k_{A=\sim a}$ for labeler L^k, is the averaged across all cv iterations. The procedure for tuning parameter c and identifying the ultimate threshold pair are detailed in Algorithm 2: Find Optimal C.

Once a threshold is identified, each labeler model B^k and the corresponding thresholds pair $\pi^k_{A=a}$ and $\pi^k_{A=\sim a}$, are applied to classify the gold standard instances in GS, based on which the model's prediction biases are computed (8-10 lines in Algorithm 1: MBA), and subsequently ranked.

5 EMPIRICAL EVALUATION

To evaluate our method, we conducted empirical evaluations using simulation studies based on four publicly available datasets: Adult, also known as "Census Income" dataset, Credit dataset from UCI, predicting the default payments of credit card clients [1], Employees Evaluation for Promotion (Employee) dataset from Kaggle[2], and Hospital Readmission Rates dataset from Kaggle [3]. The simulation studies offer controlled settings to allow us to compare the proposed approach with the alternative benchmark, SR, under a variety of settings, including different magnitudes of labelers' decision biases; different class distributions; and different *types* of biases, such as when labelers exhibit correct within-group orderings but have different decision thresholds conditioned on groups, and incorrect within-group orderings driven by the misuse of an interaction variable.

Gold standard labels. We begin by considering a setting where the prevalence of the positive labels is constant across sensitive groups, which yields a scenario where the baseline, SR, may appear to be a sensible choice, given unbiased labels should yield no difference in selection rates across groups. In order to evaluate our method's performance under different class distributions, we consider two scenarios: a positive label prevalence of 20% and 30%, respectively. Note that these two distributions will correspond to settings in which the positive class is smaller, which often arise in practice, e.g., a smaller proportion of candidates would be selected from a large pool of applications.

[1] https://archive-beta.ics.uci.edu/dataset/350/default+of+credit+card+clients
[2] https://www.kaggle.com/muhammadimran112233/employees-evaluation-for-promotion
[3] https://www.kaggle.com/code/iabhishekofficial/prediction-on-hospital-readmission

We then select a pool of 400 instances with synthetic gold-standard labels from each protected group, randomly sampled, as the disjoint set of gold standard data.

Decision Simulation. We run experiments under two types of decision simulations corresponding to two scenarios of interest: "correct within-group ordering" and "incorrect within-group ordering". For the Adult dataset, for example, the "correct within-group ordering" setting means that a labeler infers that women are less likely than others to earn a high income, and thus applies a different threshold for this group, yielding a predefined $TPR_{Y'|Y,\, A=women}$, i.e., true positive rate of labelers' decisions with respect to the gold standard labels within the women group. We assume that labelers correctly assess men, except for random noise that yields an average $TPR_{Y'|Y,\, A=men}$ = 0.95. In the "incorrect within-group ordering" setting, we consider labelers' misuse of an interaction term resulting in biased decisions. Specifically, the interaction *sex* $\times$ *age* reflects how a labeler relates *age* with *sex*; negative deviations from the true coefficient correspond to a higher degree of bias, e.g., assuming that older women are more likely to earn less, for instance.

It is important to note that even though our theoretical analysis provides guarantees when the functional form specification of the labelers' models is correct, our empirical assessment does not make this assumption. The results show that without knowing the correct functional form, i.e., using a different functional form to simulate labelers decisions and for the labelers' models, our approach remains effective under these settings.

Benchmark. We evaluated our proposed approach relative to the "Selection Rate" (SR) benchmark, which is perhaps the most intuitive and widely considered measure Mehrabi et al. (2021) when gold standard labels are unavailable. Specifically, SR estimates a labeler's bias by the difference between the proportion of positive labels the labeler assigns to instances from different groups. For

example, the difference of promotion rates among male and female employees.

$$\widehat{GAP}^{k}_{sr} = \sum_{i=1}^{|S_{L^k}|} I[Y'^k = 1|A = a] - \sum_{i=1}^{|S_{L^k}|} I[Y'^k = 1|A =\sim a] \tag{17}$$

6 RESULTS

In this section, we assess the performance of the proposed approach and compare it with that of the benchmark, SR, under the different settings described in Section 5.

Table 1. Spearman's rank-order ρ for MBA (ours) and the benchmark SR when labelers exhibit correct within-group ordering, ideal setting for SR. The ranks produced by MBA and SR both show significant correlation with true rank.

Dataset	Setting	MBA (ours)	SR
Adult	$P(Y=1\|A)=20\%$	0.947***	0.932**
Credit	$P(Y=1\|A)=20\%$	0.772*	0.936***
Employee	$P(Y=1\|A)=20\%$	0.895***	0.956***
Readmission	$P(Y=1\|A)=20\%$	0.934***	0.979***
Adult	$P(Y=1\|A)=30\%$	0.970***	0.966***
Credit	$P(Y=1\|A)=30\%$	0.860**	0.973***
Employee	$P(Y=1\|A)=30\%$	0.918***	0.983***
Readmission	$P(Y=1\|A)=30\%$	0.979***	0.989***

*: p-value < 0.05, indicating that the correlation coefficient is different from zero and that a linear relationship exists, **: p < 0.01, and ***: p < 0.001.

Table 1 and 3 show Spearman's rank-order correlation and their statistical significance of the proposed method, MBA, and of the benchmark SR, for settings where labelers exhibit either correct or incorrect within-group orderings, respectively. Table 2 and 4 show Pearson correlation coefficients and their statistical significance of the proposed method for the same settings. In each settings and data set, we show results for different class distributions.

Table 3 and 4 show the two methods' performances when labelers exhibit correct within-group orderings, a scenario in which the baseline, SR, is optimal. The results indicate that MBA performs comparably well in this setting. When labelers conditionally misestimate the interaction of the sensitive attribute with a feature (e.g., *sex* $\times$ *age* for the Adult dataset), while appearing to have the same

Table 2. Pearson correlation coefficients r for MBA (ours) and the benchmark SR when labelers exhibit correct within-group ordering, ideal setting for SR. The ranks produced by MBA and SR both show significant correlation with true rank.

Dataset	Setting	MBA (ours)	SR
Adult	$P(Y=1\|A)=20\%$	0.942***	0.943***
Credit	$P(Y=1\|A)=20\%$	0.775*	0.922***
Employee	$P(Y=1\|A)=20\%$	0.901***	0.961***
Readmission	$P(Y=1\|A)=20\%$	0.928***	0.981***
Adult	$P(Y=1\|A)=30\%$	0.964***	0.973***
Credit	$P(Y=1\|A)=30\%$	0.856**	0.976***
Employee	$P(Y=1\|A)=30\%$	0.931***	0.982***
Readmission	$P(Y=1\|A)=30\%$	0.975***	0.992***

*: p-value < 0.05, indicating that the correlation coefficient is different from zero and that a linear relationship exists, **: p < 0.01, and ***: p < 0.001.

Table 3. Spearman's rank-order ρ for MBA (ours) and benchmark SR when labelers exhibit incorrect within-group ordering. The ranks produced by MBA (ours) shows significant correlation with true rank, while the benchmark SR yielded all labelers having the same bias.

Dataset	Setting	MBA (ours)	SR
Adult	$P(Y=1\|A)=P(Y'=1\|A)=20\%$	0.928***	-0.128
Credit	$P(Y=1\|A)=P(Y'=1\|A)=20\%$	0.905**	0.079
Employee	$P(Y=1\|A)=P(Y'=1\|A)=20\%$	0.841*	0.263
Readmission	$P(Y=1\|A)=P(Y'=1\|A)=20\%$	0.975***	-0.337
Adult	$P(Y=1\|A)=P(Y'=1\|A)=30\%$	0.942***	-0.058
Credit	$P(Y=1\|A)=P(Y'=1\|A)=30\%$	0.942***	0.038
Employee	$P(Y=1\|A)=P(Y'=1\|A)=30\%$	0.918***	0.477
Readmission	$P(Y=1\|A)=P(Y'=1\|A)=30\%$	0.977***	0.287

*: p-value < 0.05, indicating that the correlation coefficient is different from zero and that a linear relationship exists, **: p < 0.01, and ***: p < 0.001.

selection rates, Table 3 and 4 show that the SR benchmark exhibits significantly poor performance and thus cannot be relied on in practice. By contrast, MBA produces an accurate rank of labelers' bias ($GAP_{\hat{Y}|Y,A}$) that is significantly correlated to the true rank ($GAP_{Y'|Y,A}$).

Figures 3, 4, 5, 6, show predicted bias, $GAP_{\hat{Y}|Y,A}$, produced by MBA and the SR benchmark, as well as labelers' true bias, $GAP_{Y'|Y,A}$, with 90% confidence bars. Recall that our goal is to recover the correct ranking of labelers' biases; hence, in these plots, we examine whether (and the degree to which) a labeler's bias was correctly positioned relative to others, as shown for the true bi-

Table 4. Pearson correlation coefficients r for MBA (ours) and benchmark SR when labelers exhibit incorrect within-group ordering. The ranks produced by MBA (ours) shows significant correlation with true rank, while the benchmark SR yielded all labelers having the same bias.

Dataset	Setting	MBA (ours)	SR
Adult	$P(Y=1\|A)=P(Y'=1\|A)=20\%$	0.927***	-0.084
Credit	$P(Y=1\|A)=P(Y'=1\|A)=20\%$	0.922***	0.080
Employee	$P(Y=1\|A)=P(Y'=1\|A)=20\%$	0.881**	0.322
Readmission	$P(Y=1\|A)=P(Y'=1\|A)=20\%$	0.964***	-0.329
Adult	$P(Y=1\|A)=P(Y'=1\|A)=30\%$	0.934***	-0.031
Credit	$P(Y=1\|A)=P(Y'=1\|A)=30\%$	0.938***	0.047
Employee	$P(Y=1\|A)=P(Y'=1\|A)=30\%$	0.922***	0.637
Readmission	$P(Y=1\|A)=P(Y'=1\|A)=30\%$	0.972***	0.543

*: p-value < 0.05, indicating that the correlation coefficient is different from zero and that a linear relationship exists, **: p < 0.01, and ***: p < 0.001.

Figure 3. Predicted $GAP_{\hat{Y}|Y,A}$ by MBA (ours) and SR, and true $GAP_{Y'|Y,A}$ when labelers exhibit correct within-group ordering, and for 20% positive rate. Both MBA's and SR's ranking have significant correlation with true rank.

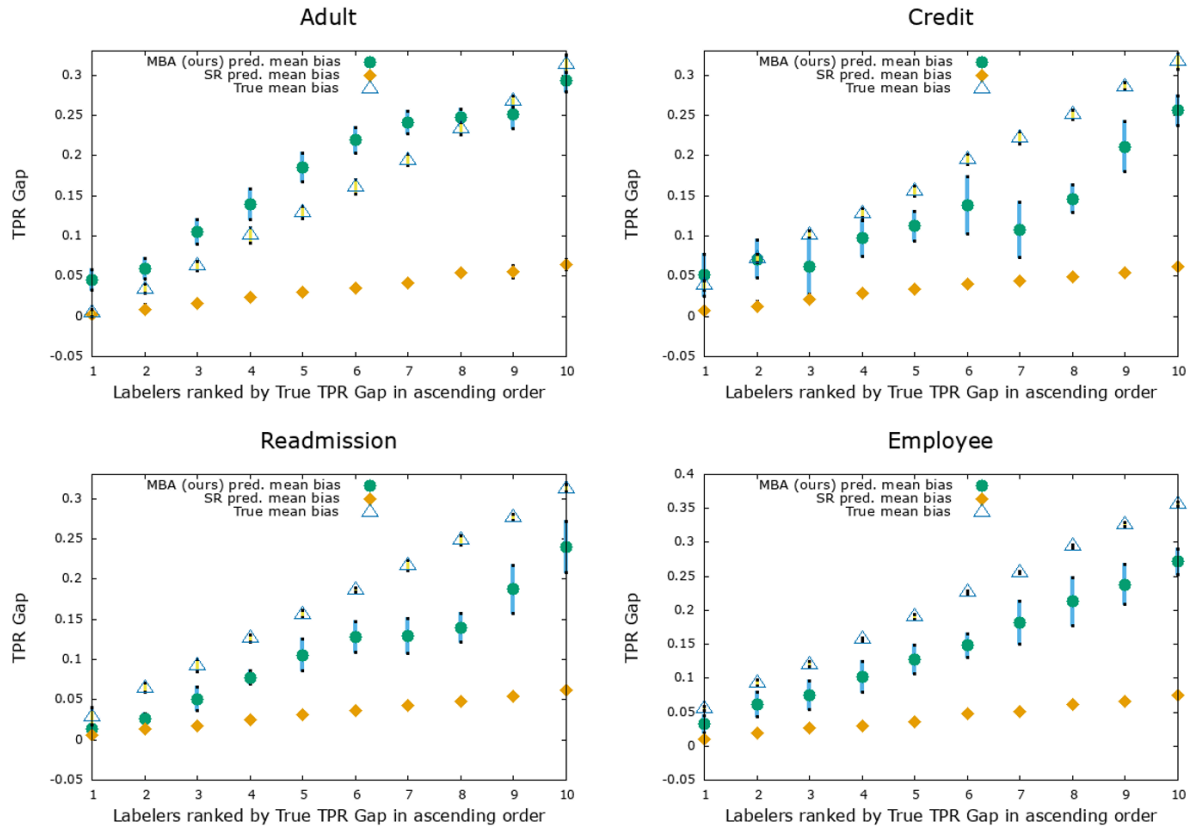
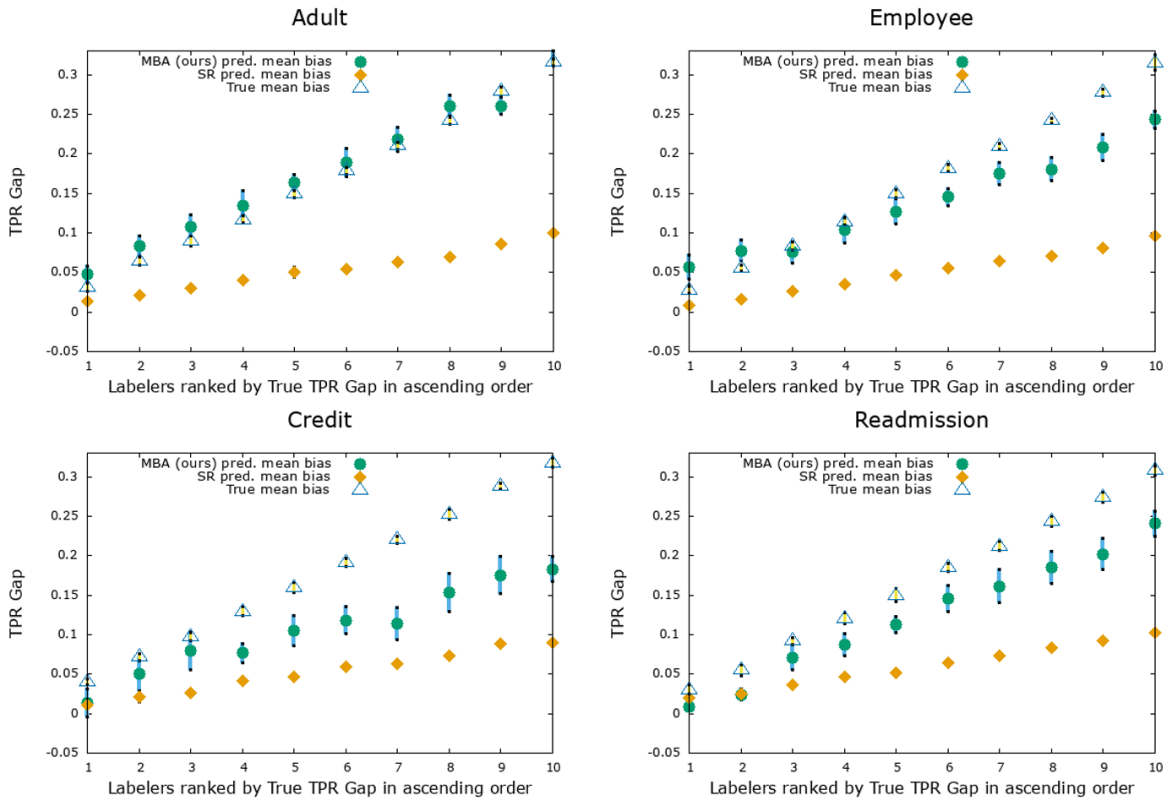


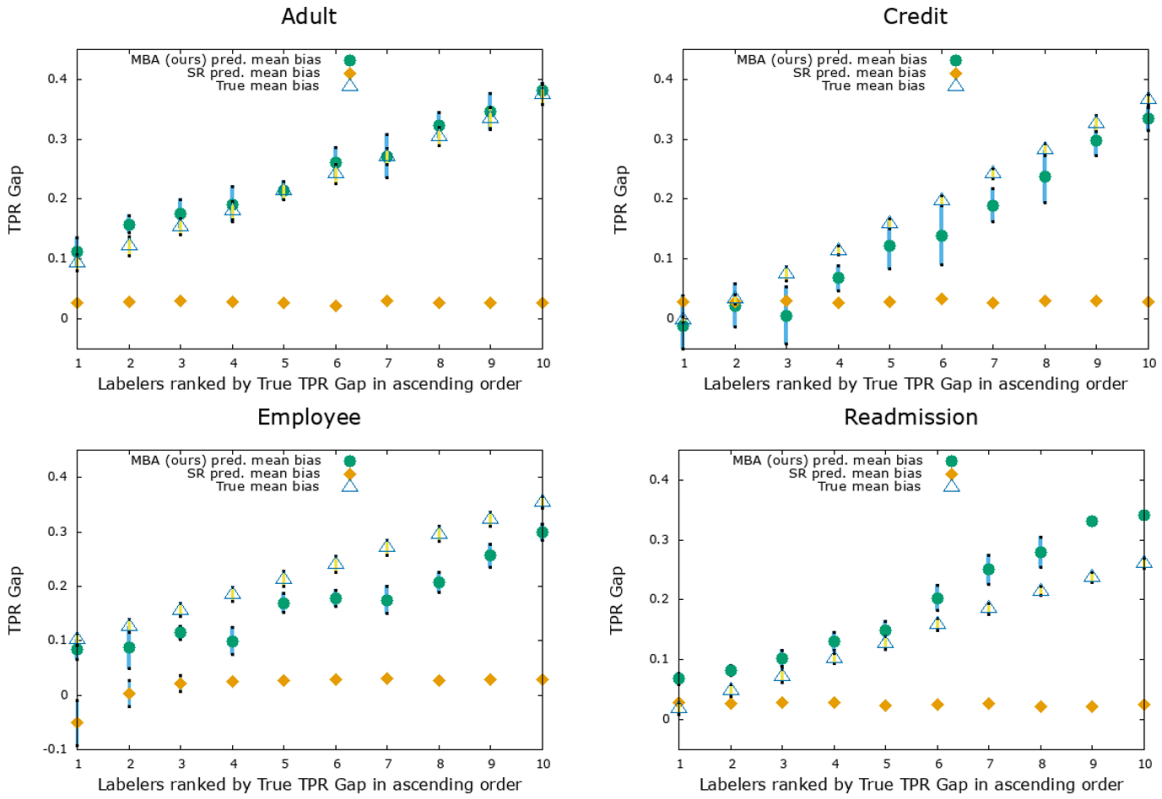
Figure 4. Predicted $GAP_{\hat{Y}|Y,A}$ by MBA (ours) and SR , and the true $GAP_{Y'|Y,A}$ when labelers exhibit correct within-group ordering, and 30% positive rate. MBA estimates follow the true rank better than SR.



ases. Figures 3, 4 show the ranking produced by the two methods for settings where labelers exhibit correct within-group ordering. Interestingly, even though both methods show high correlation with labelers' true rank in Table 1 and 2, the figures reveal how MBA approximates well both the relative ranking as well as the magnitude of the biases in all settings.

Figures 5 and 6 evaluate settings when labelers exhibit incorrect within-group orderings. This assessment visualizes the failure of the benchmark in this setting, which incorrectly yields all labelers as having the same (null) bias. Meanwhile, while MBA tends to underestimate the magnitude of the biases, it effectively recovers the correct rank of labelers' relative biases.

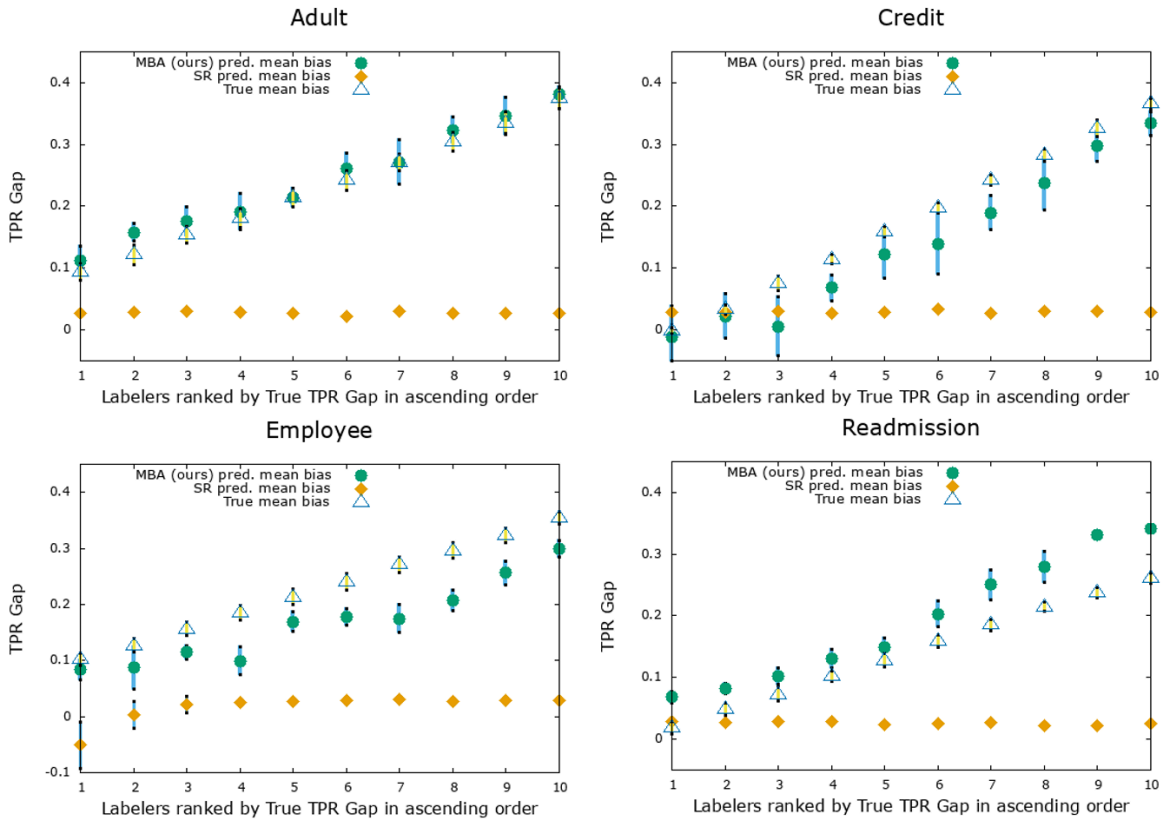
Figure 5. Predicted $GAP_{\hat{Y}|Y,A}$ by (ours) and SR, and the true $GAP_{Y'|Y,A}$ when labelers predict incorrect within-group ordering, and 20% positive rate. MBA yields correct ranking of labelers' biases while SR misestimates the biases to be apprixmately the equivalent.



7 DISCUSSION AND FUTURE WORK

In this paper, we tackle the problem of assessing biases encoded in labelers' decisions. We propose an algorithm that returns an assessment of labelers' relative biases for a set of labelers, without requiring ground truth labels to be available for the instances assessed by the labelers, nor any overlap in the instances assessed across labelers'. The proposed approach estimates biases in terms of gaps in true positive rates, and we illustrate its performance by comparing it to the typically used alternative, selection rates (SR), which has the advantage of not requiring any ground-truth, but, as a result, also cannot account for the correctness of labelers' decisions. After providing theoretical guarantees for the proposed approach, we conduct an empirical assessment in which we consider different scenarios, both favorable and unfavorable for the baseline, SR. We show that our method

Figure 6. Predicted $GAP_{\hat{Y}|Y,A}$ by MBA, SR, and the true $GAP_{Y'|Y,A}$ when labelers predict incorrect within-group ordering, and 30% positive rate. MBA infers the correct ranking of lablers biases, while SR failes to do so.



performs well in what constitutes a best-case-scenario for SR, and then study a scenario in which SR can be misleading, revealing the advantages of the proposed approach in providing consistently good performance in both settings. While assessments of decisions and labeling biases based on selection rates are widespread, our results show how SR may fail to differentiate between labelers exhibiting very different degrees of biases and are prone to being gamed by adversaries. The proposed approach addresses this problem and lays the groundwork towards reliable bias assessment in labeling. In future work we plan on conducting empirical studies using human-generated labels on a variety of tasks, to characterize both when the method succeeds and when the method fails in practice.

Increasing transparency in labelers' biases may have a variety of benefits. We are interested in identifying productive ways to bring the relative bias assessment to bear on related research ques-

tions and downstream tasks, including utilizing the output of our method when training an algorithm on human-generated labels. We are also interested in human-centered interventions that provide this piece of information to labelers as part of strategies meant to counter cognitive biases during labeling or decision-making. Finally, we intend to deepen our study of adversarial settings and modes of failure to better understand how and when different quantitative measures of quality and bias may be misleading and gameable, in order to better characterize its limitations and caution against its misuse as mechanisms for automated assessments.

Bibliography

Aczel, B., Bago, B., Szollosi, A., Foldes, A., and Lukacs, B. 2015. "Measuring individual differences in decision biases: methodological considerations," *Frontiers in psychology* (6), p. 1770.

Allen, J., Arechar, A. A., Pennycook, G., and Rand, D. G. 2020. "Scaling up fact-checking using the wisdom of crowds," *Preprint at https://doi org/1031234/osf io/9qdza* .

Barabas, C., Virza, M., Dinakar, K., Ito, J., and Zittrain, J. 2018. "Interventions over predictions: Reframing the ethical debate for actuarial risk assessment," in *Conference on Fairness, Accountability and Transparency*, , PMLR.

Binns, R. 2020. "On the apparent conflict between individual and group fairness," in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, .

Chapman, G. B., and Elstein, A. S. 2000. "Cognitive processes and biases in medical decision making." *In G B Chapman & F A Sonnenberg (Eds), Decision making in health care: Theory, psychology, and applications (pp 183–210)* .

Cohen, M. S. 1993. "Three paradigms for viewing decision biases," *Decision making in action: Models and methods* (1), pp. 36–50.

Dastin, J. 2018. "Amazon scraps secret AI recruiting tool that showed bias against women," in *Ethics of Data and Analytics*, , Auerbach Publications, pp. 296–299.

Davani, A. M., Díaz, M., and Prabhakaran, V. 2021. "Dealing with disagreements: Looking beyond the majority vote in subjective annotations," *arXiv preprint arXiv:211005719* .

De-Arteaga, M., Romanov, A., Wallach, H., Chayes, J., Borgs, C., Chouldechova, A., Geyik, S., Kenthapadi, K., and Kalai, A. T. 2019. "Bias in bios: A case study of semantic representation bias in a high-stakes setting," in *proceedings of the Conference on Fairness, Accountability, and Transparency*, .

De Martino, B., Kumaran, D., Seymour, B., and Dolan, R. J. 2006. "Frames, biases, and rational decision-making in the human brain," *Science* (313:5787), pp. 684–687.

Dong, W., Saar-Tsechansky, M., and Geva, T. 2021. "A Machine Learning Framework Towards Transparency in Experts' Decision Quality," *arXiv preprint arXiv:211011425* .

Feldman, M. 2015. *Computational fairness: Preventing machine-learned discrimination*, Ph.D. thesis.

Gao, R., and Saar-Tsechansky, M. 2020. "Cost-accuracy aware adaptive labeling for active learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, , vol. 34.

Geva, T., and Saar-Tsechansky, M. 2021. "Who Is a Better Decision Maker? Data-Driven Expert Ranking Under Unobserved Quality," *Production and Operations Management* (30:1), pp. 127–144.

Hardt, M., Price, E., and Srebro, N. 2016. "Equality of opportunity in supervised learning," *Advances in neural information processing systems* (29), pp. 3315–3323.

Kamiran, F., and Calders, T. 2010. "Classification with no discrimination by preferential sampling," in *Proc. 19th Machine Learning Conf. Belgium and The Netherlands*, , Citeseer.

Kazai, G., Kamps, J., and Milic-Frayling, N. 2013. "An analysis of human factors and label accuracy in crowdsourcing relevance judgments," *Information retrieval* (16:2), pp. 138–178.

Kodiyan, A. A. 2019. "An overview of ethical issues in using AI systems in hiring with a case study of Amazon's AI based hiring tool," *Researchgate Preprint* pp. 1–19.

Li, A. H., and Bradic, J. 2018. "Boosting in the presence of outliers: adaptive classification with non-convex loss functions," *Journal of the American Statistical Association* (113:522), pp. 660–674.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. 2021. "A survey on bias and fairness in machine learning," *ACM Computing Surveys (CSUR)* (54:6), pp. 1–35.

Menon, A. K., Van Rooyen, B., and Natarajan, N. 2016. "Learning from binary labels with instance-dependent corruption," *arXiv preprint arXiv:160500751* .

Miranda Jr, R., MacKillop, J., Meyerson, L. A., Justus, A., and Lovallo, W. R. 2009. "Influence of antisocial and psychopathic traits on decision-making biases in alcoholics," *Alcoholism: Clinical and Experimental Research* (33:5), pp. 817–825.

Nowak, S., and Rüger, S. 2010. "How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation," in *Proceedings of the international conference on Multimedia information retrieval*, , ACM.

Suresh, H., and Guttag, J. V. 2019. "A Framework for Understanding Unintended Consequences of Machine Learning," *CoRR* (abs/1901.10002).
URL `http://arxiv.org/abs/1901.10002`

Violago, V., and Quevada, N. 2018. "AI: The Issue of Bias," *Managing Intell Prop* (277), p. 32.