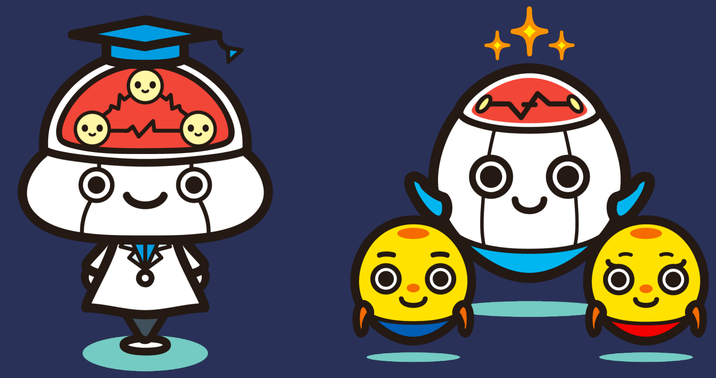
Can I Trust My Fairness Metric? Assessing Fairness with Unlabeled Data and Bayesian Inference

読み会@2021/10/12

楊 明哲



論文情報

著者: Disi Ji[1], Padhraic Smyth[1], Mark Steyvers[2]

所属: University of California, [1] Department of Computer Science

[2] Department of Cognitive Sciences

選んだ理由:

公平性に関する評価とベイズ的アプローチを学ぶため

どんな論文?

目的:

ラベル付きサンプルが少ないが，ラベルなしサンプルはたくさん
ある時に，グループの**公平性を正確に評価したい**

貢献:

ラベル付きだけの方法よりラベルなしデータで補強して、より正
確で分散の少ない推定値を生成できるようになった

背景

公平性配慮型機械学習について

機械学習が意思決定に用いられる

→センシティブ属性に対して偏った出力が問題

公平性配慮型機械学習で主に取り組まれてるもの

1. 機械学習文脈における公平性の定義

2. 公平性を考慮したアルゴリズムの設計

イントロ

対象となる公平性問題

限られたラベル付きサンプルが与えられた中でのモデルの公平性を**正しく評価**

特に二値分類における集団公平性を扱う

集団公平性

グループ(性別,人種…)内での指標(TPR, Accuracy…)が等しいことが公平

イントロ

公平性指標の問題

ラベル付きデータが少量のとき，これらの公平性評価指標の推定値がばらつく

標本分散は$1/n$の速さで小さくなるが，比較的ゆっくり

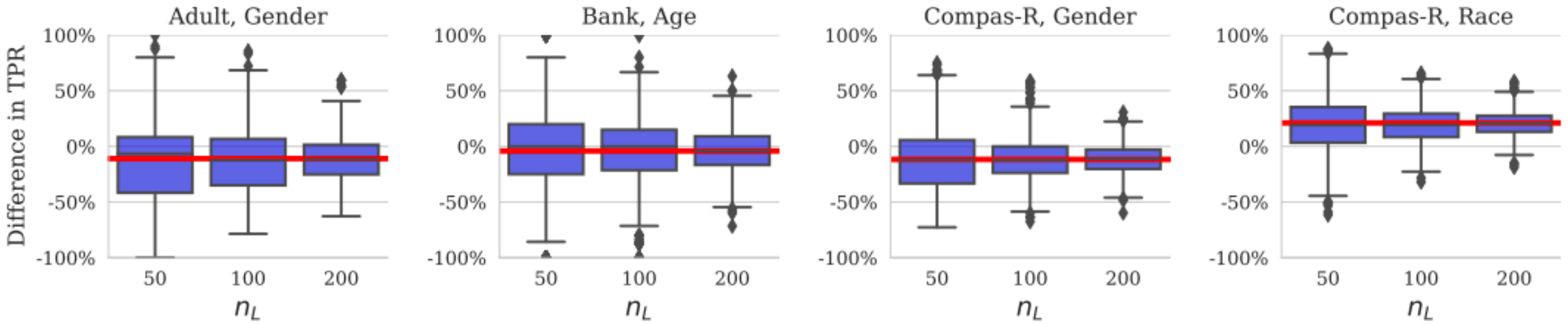


Figure 1: Boxplots of frequency-based estimates of the difference in true positive rate (TPR) for four fairness datasets and binary sensitive attributes, across 1000 randomly sampled sets of labeled test examples of size $n_L = 50, 100, 200$. The horizontal red line is the TPR difference computed on the full test dataset.

イントロ

ラベル分布が不均衡な時

ラベル分布が不均衡な場合，TPRやFPRのグループ差の推定分散は悪化する.

簡単なシミュレーション:

男性:女性=8:2, グループ内の正例が20%，モデルのグループごとの真のTPRが0.95, 0.90の時(公平性指標: 0.95-0.90=0.05).

推定公平性が[0.04,0.06]の中にあることを信頼区間95%にするにはサンプルが少なくとも96,000個必要になる

→ 実世界のデータセットはこれよりも小さいことがほとんどで,
　かつこのようにデータ分布が偏っているのは珍しいことではない

イントロ

現実のデータセット

実世界にあるデータセットだけで公平性評価を
信頼することは困難

さらに公平性が必要な状況(医療や司法)では,
データセットがあるがラベル獲得が困難

イントロ

本研究でやること

提案

大量のラベルなし+少量のラベルありによって，分散が小さい推定を生成

貢献

公平性指標の推定をベイズ的に扱う

ベイズによるキャリブレーションを提案

少量のラベルありでも推定誤差を小さくできることを実証

キャリブレーションとは

モデルの出力を各クラスに属する確率に近づける

モデルの出力値	正解ラベル	Calibrationした値
0.4	1	0.5
0.4	0	0.5
0.9	1	1.0
0.9	1	1.0

モデルの出力=positive になる確率 としていいのか?

キャブレーションによって修正する

準備

表記

M: 学習した訓練モデル, x: 入力, $y \in \{0,1\}$: クラスラベル

モデル生成スコア: $s = P_M(y = 1 \mid x) \in [0,1]$ →モデルの予測確率

$\hat{y}$: モデルの予測ラベル sに応じてラベルが決定する

分類器がキャリブレーションされる → $P(\hat{y} = y \mid s) = s$

スコアsの値の確率で予測が合っていると考えられる

準備

表記 (公平性)

$g \in \{0,1,...,G-1\}$: 対象の集団 (e.g. 人種，性別…)

θ_g: 集団gにおける何かしらの指標 (e.g. Accuracy, TPR, FPR…)

$\Delta = \theta_0 - \theta_1$: 公平性指標, 今回は$g \in \{0,1\}$で考えている

n_L, n_U: それぞれラベルあり，ラベルなしデータセット

$n_L \ll n_U$である状況を考えている

準備

母集団

ラベルなしデータセットのラベルは，スコアsを用いて擬似的に利用

サンプル(x, s, y)は母集団$P(x, y)$もしくは$P(s, y)$からIIDにサンプリングされていると考える.

またn_Uのものは単に$P(x)$や$P(s)$から生成されていると考える

提案手法

概要

提案手法では2つのデータセットを組み合わせる

ラベルあり

→Beta-Binomial Estimation

ラベルなし

→Bayesian Calibration Model

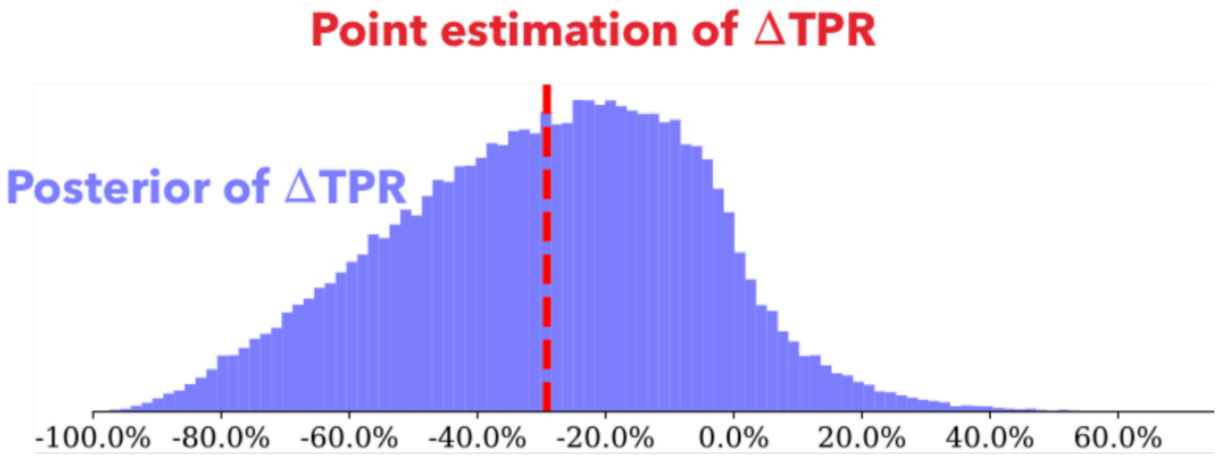
提案手法: 準備

Beta-Binomial Estimation

グループごとの指標 $\theta_g = P(\hat{y} = 1 \mid y = 1, g)$

$$\theta_g \sim Beta(\alpha_g, \beta_g)$$
$\alpha_g = \beta_g = 1$で考える

予測モデルの正誤 $I_i = I((\hat{y}_i = y_i), 1 \leq i \leq n_L$



$$I_i \sim Bernoulli(\theta_g)$$

公平性指標 $\Delta = \theta_1 - \theta_0$としたあとMCMCサンプリングによって事後分布$P(\Delta \mid D_L)$を獲得

一応推定できるが，データ数に精度が依存するのが問題

提案手法: 準備

Leveraging Unlabeled Data with a Bayesian Calibration Model

ラベルなしデータセットn_Uに対してスコア $s_j = P_M(y_j = 1 \mid x_j)$

もしモデルが完全にキャリブレーションされているならスコアを
そのまま予測に用いることができる

$$\hat{\theta}_g = \left(1/n_{U,g}\right) \sum_{j \in g} s_j I\left(s_j \geq 0.5\right) + \left(1 - s_j\right) I\left(s_j < 0.5\right)$$

でグループごとの評価指標を定義できる

提案手法: 準備

モデルスコアをそのまま用いる問題

モデルスコア（キャリブレーションされていない）による予測は
複雑なモデルでは大きく間違えてしまう.

本手法のアプローチとして,
ラベル付きデータを用いてキャリブレーションをして
精度の偏りをなくしていく

提案手法

潜在変数をつくる

$z_j = E[I(\hat{y}_j = y_j)] = P(y_j = \hat{y}_j \mid s_j)$と定義

→ スコアs_jが与えれた時のモデルの精度，これをサンプルごとの潜在変数として利用

提案手法

推定

1. n_Lを用いてグループごとのキャリブレーション関数ϕ_gを推定

2. ϕ_gからz_jの事後分布$P_{\phi_g}(z_j \mid D_L, s_j)$を獲得

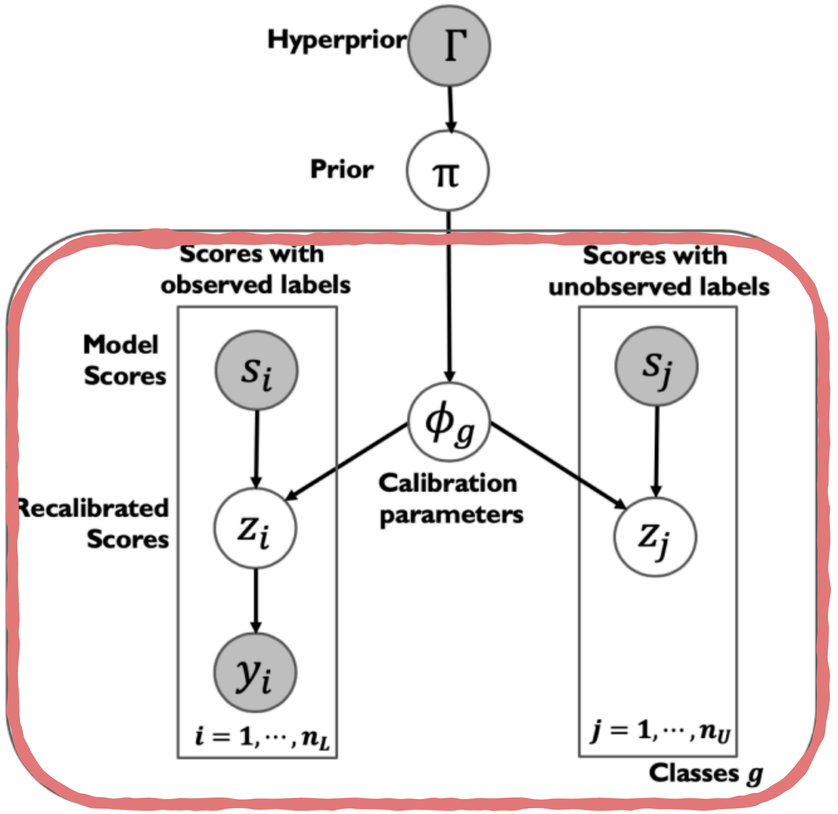
3. z_jとラベル付きデータセットを用いてθ_g, Δを推定

提案手法

グラフィカルモデル

ラベルのないサンプルのスコアとラベル付きのデータセットを組み合わせることで, θ_g^tの推定を行うことができる.

$$\theta_g^t = \frac{1}{n_{L,g} + n_{U,g}} \left(\sum_{i:i \in a} I\left(\hat{y}_i = y_i\right) + \sum_{i \cdot j \in a} z_j^t \right)$$



提案手法

階層ベイズキャリブレーション

キャリブレーション関数ϕ_gの学習を考えていく

モデルスコアのキャリブレーションとして次式を用いる

$$f(s;a,b,c) = \frac{1}{1+e^{-c-a\log s+b\log(1-s)}}$$

各グループに対してキャリブレーションパラメータ

$\phi_g = \{a_g, b_g, c_g\}$を獲得していきたい

提案手法

キャリブレーション関数のモデル化

モデルに適用するために次式で正解ラベルが生成されると仮定

$$y_i \sim \text{Bernoulli}\left(f\left(s_i; a_{g_j}, b_{q_i}, c_{g_i}\right)\right)$$

グループごとのパラメータはそれぞれ共通の分布から生成される

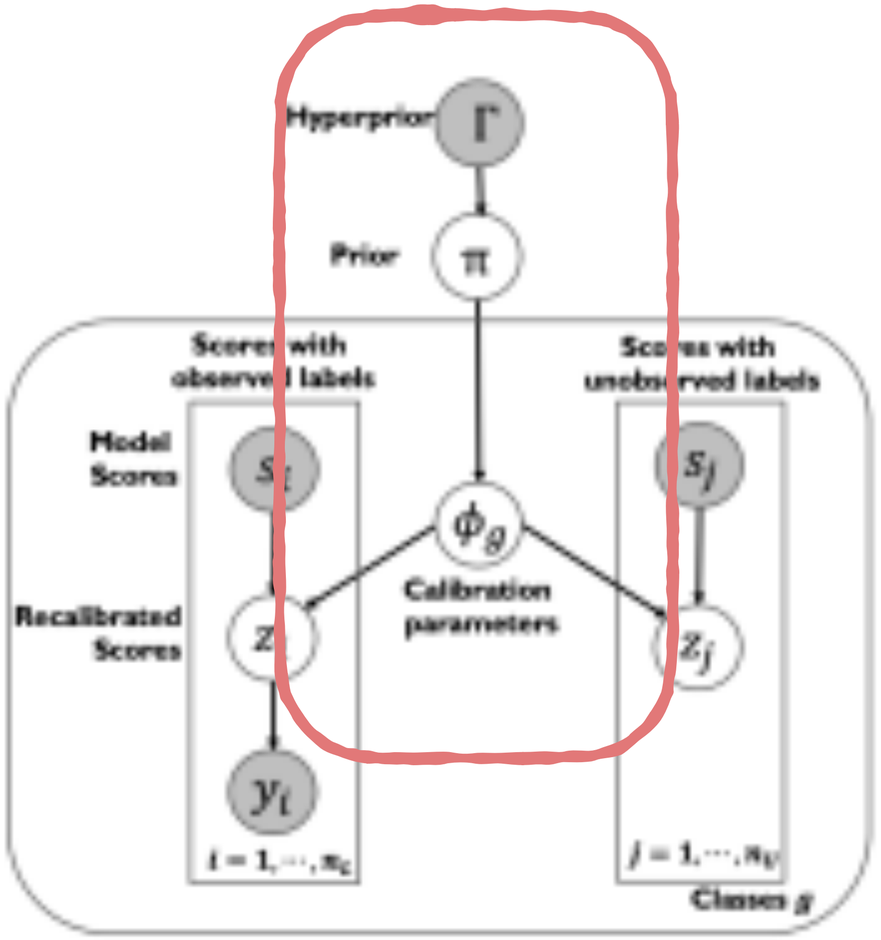
$$loga_g \sim N(\mu_a, \sigma_a), logb_g \sim N(\mu_b, \sigma_b), logc_g \sim N(\mu_c, \sigma_c) \text{where} \pi = \{\mu_{a,b,c}, \sigma_{a,b,c}\}$$

ハイパラπは切断正規分布から生成

提案手法

本研究ではハイパラを設定しているが，これはパラメータとして妥当な値で設定

→すべて同じ設定で実験，提案手法が頑健であることを示す



実験

目的:

限られてたラベル付きデータを用いて，様々な推定の精度を評価すること

データセット:

Adult, German Credit, Ricci, Compas

推定法: logistic回帰, MLP, Random Forest, Gaussian Naive Bayes

実験

推定精度の比較

MLPでグループごとのaccuracy差について評価する

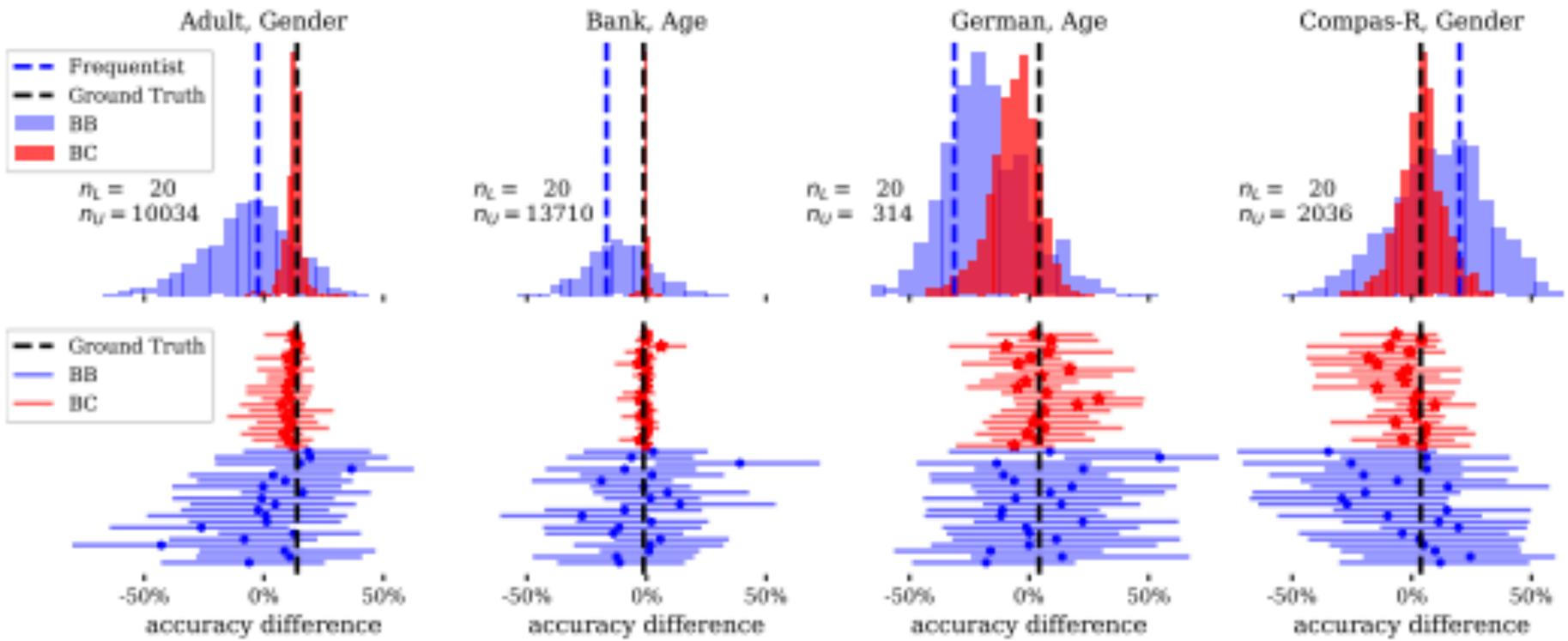
頻度主義: Δの値をテストセットすべてから計算，真の値

ベータ-二項法 (BB): ラベル付きデータセットのみにアクセス可能

ベイズキャリブレーション(BC): ラベルなしにもアクセス可能

実験

BBとBCの推定精度の比較

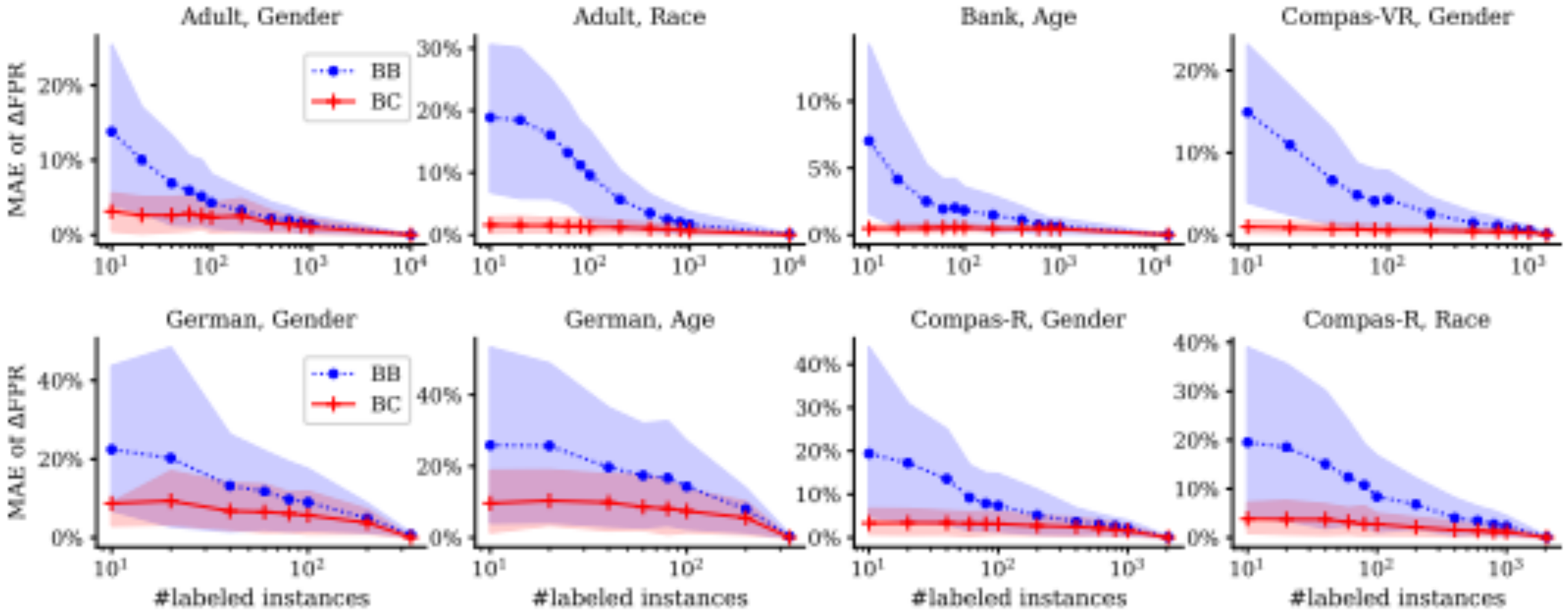


20回実行，95%事後信頼区間と事後平均を示している

実験

データセットサイズによる比較

ラベル付きデータのサイズを変化させたとき



実験

異なるモデルでの推定精度

	Multi-layer Perceptron			Logistic Regression			Random Forest			Gaussian Naive Bayes		
Dataset, Attribute	Freq	BB	BC	Freq	BB	BC	Freq	BB	BC	Freq	BB	BC
Adult, Race	16.5	18.5	**3.9**	16.4	18.7	**2.9**	16.5	18.2	**3.2**	17.6	18.9	**3.6**
Adult, Gender	19.7	17.4	**5.1**	19.1	16.1	**2.2**	17.7	17.4	**4.8**	19.7	16.2	**5.4**
Bank, Age	15.9	13.9	**2.5**	13.9	13.0	**1.4**	11.8	11.1	**1.0**	15.5	13.7	**1.7**
German, Age	34.6	19.8	**5.0**	37.1	21.2	**8.7**	33.6	18.7	**8.2**	36.6	20.4	**11.5**
German, Gender	30.7	21.6	**8.2**	25.6	17.4	**6.3**	27.7	19.3	**8.6**	30.0	20.1	**6.5**
Compas-R, Race	31.5	21.0	**4.2**	31.7	20.4	**4.8**	29.3	20.3	**2.4**	33.5	23.2	**8.4**
Compas-R, Gender	33.7	21.6	**5.0**	34.3	21.9	**3.8**	36.3	23.3	**4.4**	40.5	25.5	**13.7**
Compas-VR, Race	18.7	17.1	**4.0**	18.5	15.6	**4.4**	18.2	15.8	**2.4**	26.6	19.8	**6.5**
Compas-VR, Gender	20.6	16.9	**5.4**	19.9	16.6	**5.3**	22.3	19.0	**6.3**	31.3	21.5	**9.8**
Ricci, Race	23.5	17.7	**14.6**	14.6	14.6	**7.9**	6.3	12.2	**2.1**	8.9	13.1	**1.6**

テストセット全ての評価指標のMAE

$n_L = 200$の時ほとんど同様の結果

まとめ

サンプルサイズが少ないときの公平性評価は不確実だと指摘

ラベルなし，ありを組み合わせて**ベイズキャリブレーション**を用いて推定分散を小さくする手法を提案

今回のフレームワークは，手法に適用するのが容易のため，公平性評価に用いるといいかも