南京郵電大學
Nanjing University of Posts and Telecommunications

Intra-Processing Methods for Debiasing Neural Networks

Yash Savani
Abacus.AI
San Francisco, CA 94103
yash@abacus.ai

Colin White
Abacus.AI
San Francisco, CA 94103
colin@abacus.ai

Naveen Sunda
Govindarajulu RAIR Lab RPI
Troy, NY 12180
naveensundarg@gmail.com


汇报人：王亦恺（研一）

指导老师：李云

Advances in Neural Information Processing Systems 33 (NeurIPS 2020)

目录

CONTENTS

1.研究背景

机器学习在社会各领域广泛应用的同时，引发了人们对机器学习可信性的担忧，其中包括公平性。研究发现，分类器在依据性别、种族、宗教等敏感属性划分的不同群体上预测同一目标属性时，预测的准确率存在明显差异。这种差异在一些特殊的应用场景下可能被进一步放大，从而造成严重的社会问题。

例如，美国政府目前正在使用的某一面部识别系统在预测亚裔或非裔人群时，预测结果的误判率达到了欧裔人群的10倍甚至更高[1]，这可以认为是一种偏见或歧视现象。如果将该工具运用于执法系统，意味着亚裔或非裔人群更有可能被无辜逮捕。

[1]A us government study confirms most face recognition systems are racist

1.研究背景

公平性算法

- **预处理方法(Pre-processing)**

通常认为，数据集的不平衡是偏见的直接来源。预处理方法从数据集出发，通过对原始数据进行去偏操作，得到相对平衡的数据集。

- **处理中方法(In-processing)**

处理中方法通过在模型训练过程中添加公平性约束项，消除源于数据的偏差。

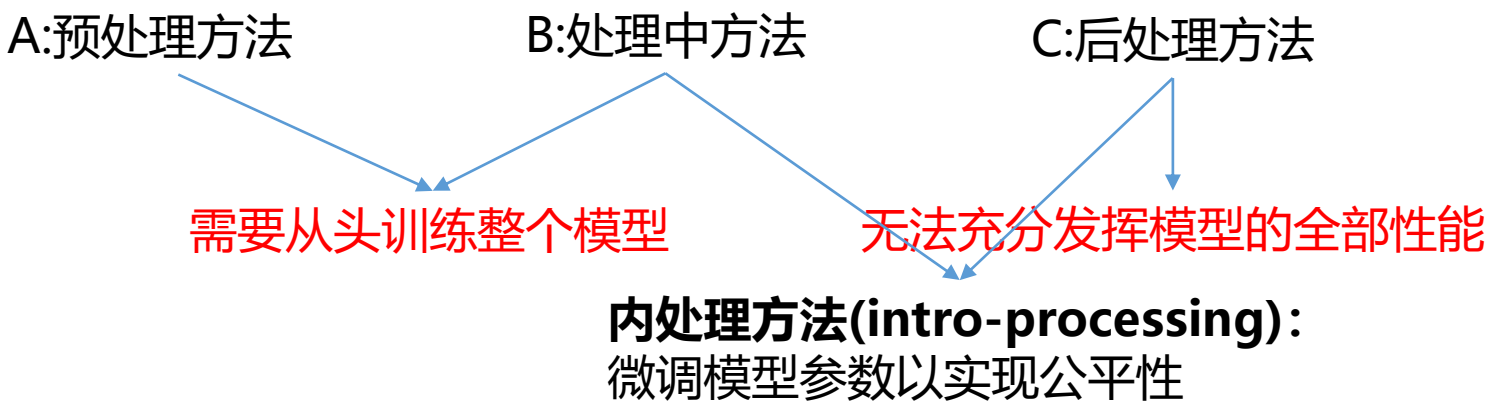
- **后处理方法(Post-processing)**

后处理方法直接对模型的输出进行干预，从而实现公平性。因此适用于整个模型为黑盒的应用场景。

1.研究背景

方法提出

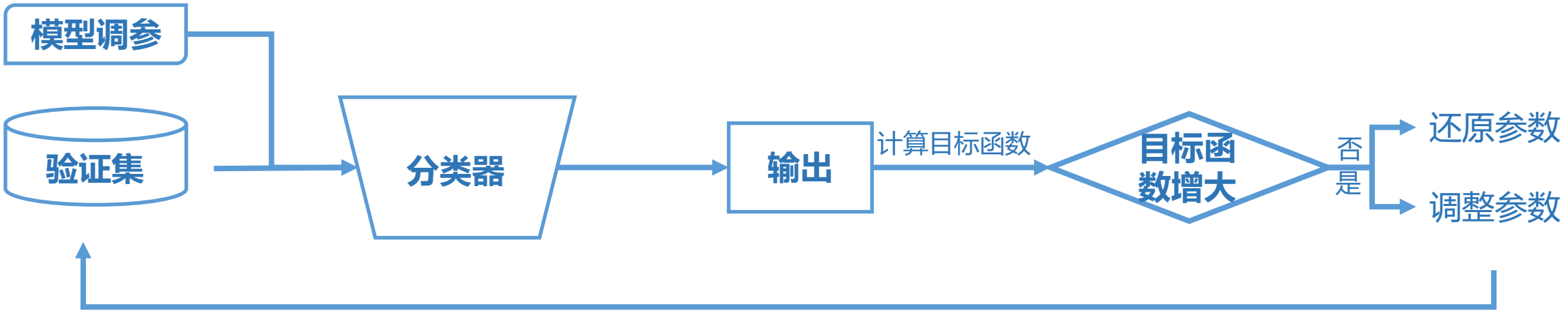
现实许多应用场景下，往往会使用一个预训练好的大模型，再根据具体应用场景对模型微调。选择何种方法合适?



目录

CONTENTS

2.基本思想



最大化目标函数：

$$\max \phi_{\mu,\rho,\epsilon}(\mathcal{D}, \hat{y}, A) = \begin{cases} \rho & \text{if } \mu < \epsilon \\ 0 & otherwise \end{cases}$$

函数输入：
$\mathcal{D}$: 数据集
$\hat{y}$: 预测结果
A: 受保护属性

函数输出：
ρ: 评估模型性能的函数
其他：
μ: 评估模型公平性的函数
ϵ: 公平性指标的阈值

2.基本思想

公平性指标

奇偶校验差(Statistical Parity Difference, SPD):

$$SPD(\mathcal{D},\hat{y},A)=P_{(x_i,Y_i)\in\mathcal{D}}(\hat{Y}_i=1\mid a_i=0)-P_{(x_i,Y_i)\in\mathcal{D}}(\hat{Y}_i=1\mid a_i=1)$$

平等机会差异(Equal opportunity difference, EOD):

$$EOD(\mathcal{D},\hat{y},A)=TPR_{A=0}(\mathcal{D},\hat{y})-TPR_{A=1}(\mathcal{D},\hat{y})$$

平均机会差异(Average Odds Difference, AOD):

$$AOD(\mathcal{D},\hat{y},A)=\frac{\left(FPR_{A=0}(\mathcal{D},\hat{y})-FPR_{A=1}(\mathcal{D},\hat{y})\right)+\left(TPR_{A=0}(\mathcal{D},\hat{y})-TPR_{A=1}(\mathcal{D},\hat{y})\right)}{2}$$

图像数据：$\mu(\mathcal{D},\hat{y},\mathrm{A})=AOD(\mathcal{D},\hat{y},A)$

表格数据：$\mu(\mathcal{D},\hat{y},\mathrm{A})=SPD(\mathcal{D},\hat{y},A)$

2.基本思想

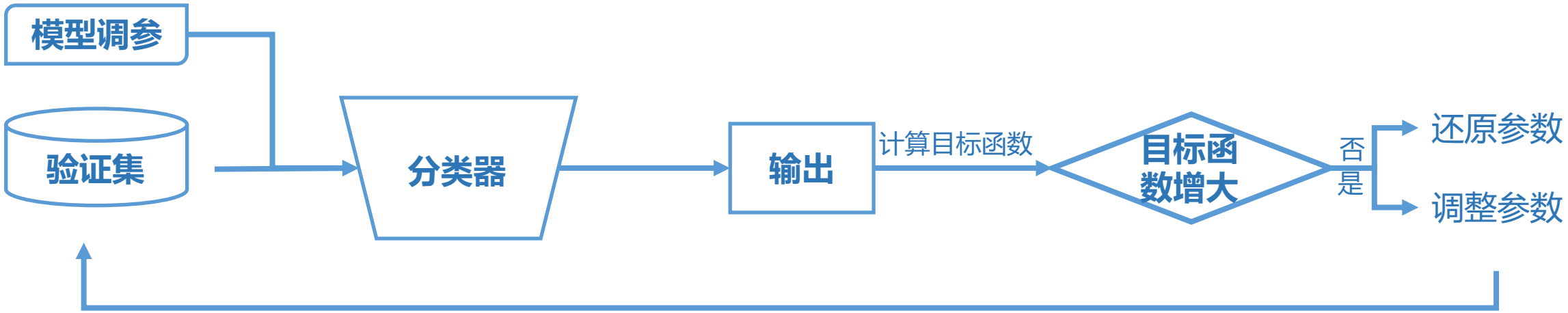
性能评估指标

常见的性能评估指标：准确率（Accuracy）、精确率（Precision）、召回率（Recall）、AUC-ROC曲线

平衡精度(balanced accuracy):

$$\rho(y,\hat{y})=\frac{TPR+TNR}{2}$$

2.基本思想



最大化目标函数：

$$\max \phi_{\mu,\rho,\epsilon}(\mathcal{D}, \hat{y}, A) = \begin{cases} \rho & \text{if } \mu < \epsilon \\ 0 & otherwise \end{cases}$$

目录

CONTENTS

3.1随机扰动算法

从均值为1方差为0.1的高斯分布中随机得到$|\theta|$个随机数，用这些随机数扰动模型参数。

Algorithm 1 Random Perturbation

Input: Trained model f with weights θ, validation dataset $\mathcal{D}_{\text{valid}}$, objective $\phi_{\mu,\rho,\epsilon}$, parameter T
Set $\theta^* = \theta$, $\text{val}^* = -\infty$, and $\tau^* = 0$
for $i = 1$ to T **do**
    Sample $q_j \sim \mathcal{N}(1, 0.1)$ for all $j \in \{1, 2, ..., |\theta|\}$
    $\theta'_j = \theta_j \cdot q_j$
    Select threshold $\tau \in [0, 1]$ which maximizes the objective $\phi_{\mu,\rho,\epsilon}$ on the validation set
    Set $\text{val} = \phi_{\mu,\rho,\epsilon}(\mathcal{D}_{\text{valid}}, \{\mathbb{I}\{f_{\theta'}(\boldsymbol{x}) > \tau\} \mid (\boldsymbol{x}, Y) \in \mathcal{D}_{\text{valid}}\}, A)$
    If $\text{val} > \text{val}^*$, set $\text{val}^* = \text{val}$, $\theta^* = \theta'$, and $\tau^* = \tau$.
end for
Output: θ^*, τ^*

3.2层级优化

核心思想：迭代多棵回归树共同决策最终结果

优点：属于零阶优化器，适用于目标函数不可微的情况，因此又称为“黑盒优化器”

缺点：需要进行多次迭代，因此调参过程较为复杂，训练时间可能较长

3.2层级优化

梯度增强回归树(GBRT, gradient-boosted regression trees)

使用优化器调整每一层的参数，而不是使用随机扰动

Algorithm 2 Layer-wise optimization

Input: Trained model $f = f^{(\ell)} \circ \ldots \circ f^{(1)}$ with weights $\theta_1, \ldots, \theta_\ell$, objective $\phi_{\mu,\rho,\epsilon}$, black-box optimizer $\mathcal{A}$
Set $\theta^* = \emptyset$, $\text{val}^* = -\infty$, and $\tau^* = 0$
for $i = 1$ to ℓ **do**
  Run optimizer $\mathcal{A}$ to optimize weights θ_i to θ_i' with respect to $\phi_{\mu,\rho,\epsilon}$.
  Select threshold $\tau \in [0, 1]$ which maximizes objective $\phi_{\mu,\rho,\epsilon}$
  Set $\text{val} = \phi_{\mu,\rho,\epsilon}(\mathcal{D}_{\text{valid}}, \{\mathbb{I}\{f_{\theta'}(\boldsymbol{x}) > \tau\} \mid (\boldsymbol{x}, Y) \in \mathcal{D}_{\text{valid}}\}, A)$, where $\theta' = (\theta_1, \ldots, \theta_i', \ldots, \theta_\ell)$
  If $\text{val} > \text{val}^*$ set $\text{val}^* = \text{val}$, $\theta^* = \theta'$, and $\tau^* = \tau$.
end for
Output: θ^*, τ^*

3.3对抗优化

批评模型：评判某一批次训练结果中的偏见

训练批评模型

训练预测模型

Algorithm 3 Adversarial Fine-Tuning

Input: Trained model $f = f^{(\ell)} \circ f'$ with weights θ, validation dataset $\mathcal{D}_{\text{valid}}$, parameters λ, ϵ, δ, n, m, m', T.
Set g as the critic model with weights ψ.
for $i = 0$ to n **do**
  for $j = 0$ to m **do**
    Sample a minibatch $(\boldsymbol{X}_k, \boldsymbol{y}_k)$ with replacement from $\mathcal{D}_{\text{valid}}$
    Evaluate the bias in the minibatch, $\bar{\mu} \leftarrow \mu((\boldsymbol{X}_k, \boldsymbol{y}_k), f(\boldsymbol{X}_k))$.
    Update the critic model g by updating its stochastic gradient

$$\nabla_\psi (\bar{\mu} - (g \circ f')(\boldsymbol{X}_k))^2$$

  end for
  for $j = 0$ to m' **do**
    Sample a minibatch $(\boldsymbol{X}_k, \boldsymbol{y}_k)$ with replacement from $\mathcal{D}_{\text{valid}}$
    Update the original model by updating its stochastic gradient

$$\nabla_\theta \left[\max\{1, \lambda \cdot (|(g \circ f')(\boldsymbol{X}_k)| - \epsilon + \delta) + 1\} \cdot \text{BCELoss}(\boldsymbol{y}_k, f(\boldsymbol{X}_k))\right]$$

  end for
  Select threshold $\tau \in [0, 1]$ that minimizes the objective $\phi_{\mu,\rho}$
end for
Output: Debiased model f, threshold τ

3.3对抗训练

$$\max\{1, \lambda \cdot (|(g \circ f')(\boldsymbol{X}_k)| - \epsilon + \delta) + 1\} \cdot \mathrm{BCELoss}\big(\boldsymbol{y}_k, f(\boldsymbol{X}_k)\big)$$

$$\mu(\mathcal{D}, \hat{\mathcal{Y}}, \mathrm{A})$$

对于一批数据，

若符合公平性约束，则损失函数的系数为1；

若不符合公平性约束，则提升惩罚强度。

目录

CONTENTS

4实验结果

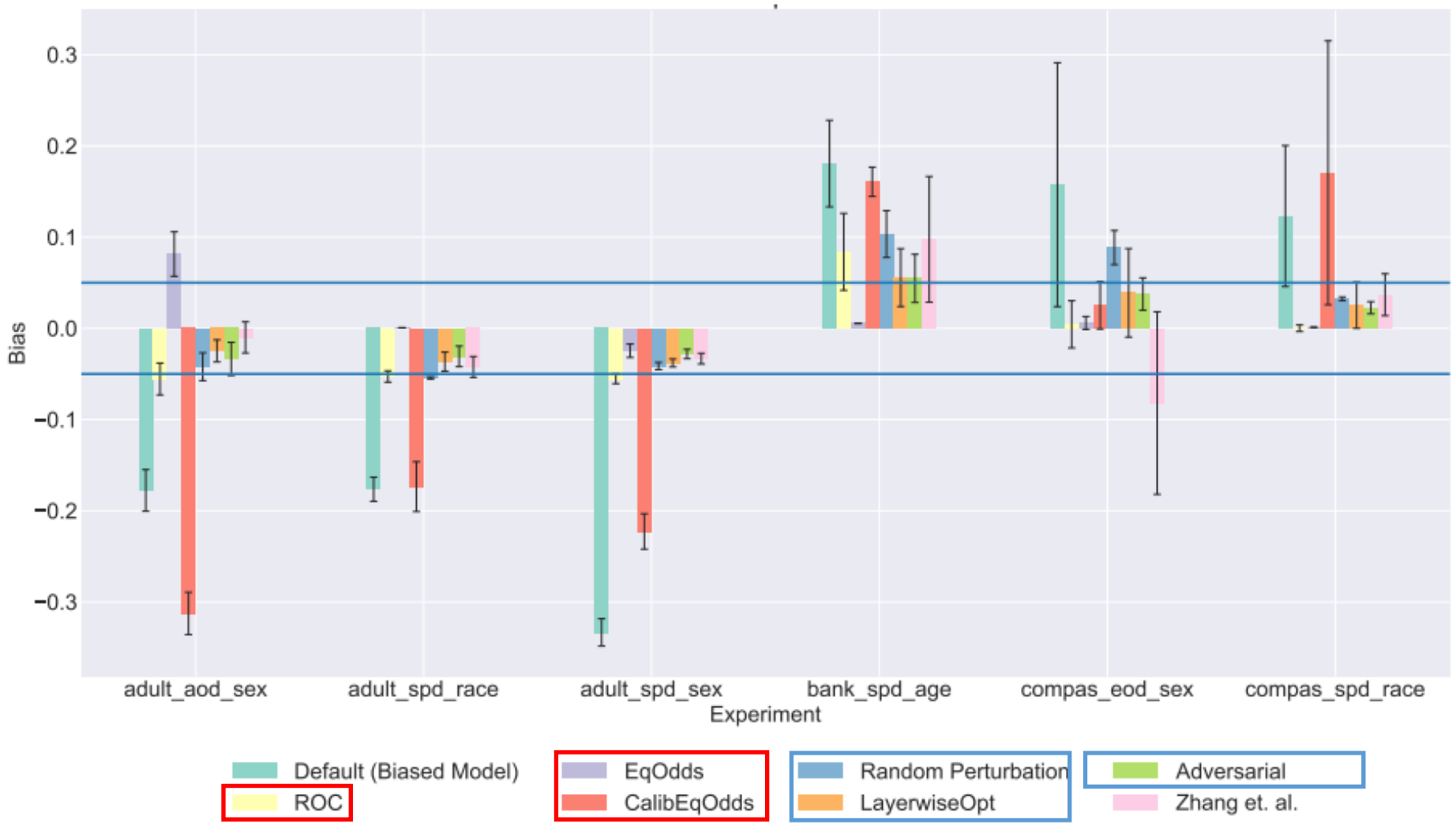
数据集：Adult、Bank、Compas

AOD:平均机会差异

SPD:奇偶校验差

EOD:平等机会差异

结论：在公平性上，本文提出的内处理方法整体上效果优于后处理方法；与处理中方法效果接近。



4实验结果

结论：在模型性能上，本文提出的内处理方法效果优于后处理方法；略低于处理中方法。

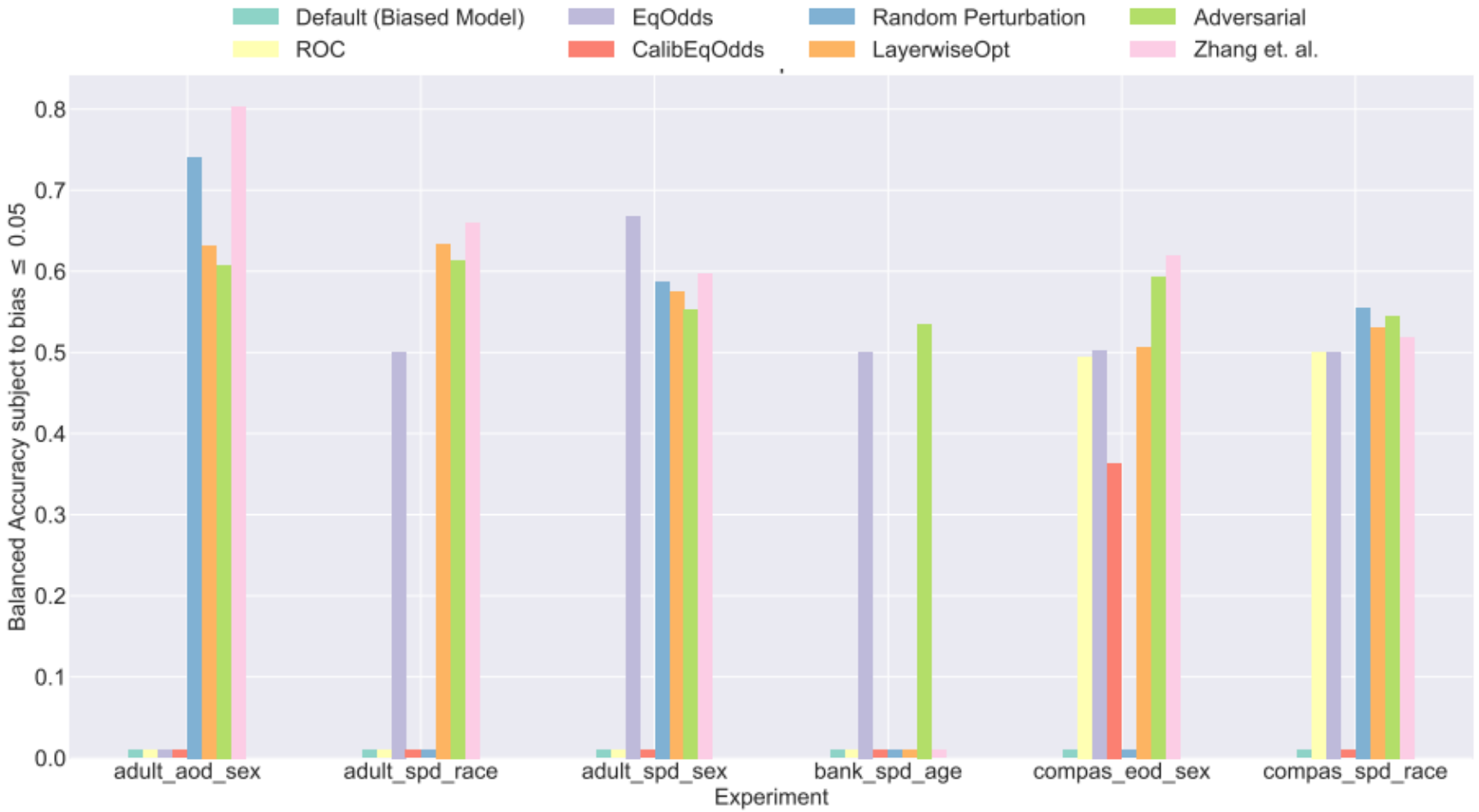


Figure 1: Results for tabular datasets over 5 runs with different seeds. mean Bias with std error bars (Top) and the median of the objective function in Equation 1 (Bottom).

南京邮电大学

4实验结果

数据集：CelebA

Table 1: Results on the CelebA datasets for a pretrained ResNet with three initial random seeds. Results are the balanced accuracy scores after fine-tuning. The crossed out scores are those that did not have biases lower than 0.05.

	Default	ROC	EqOdds	CalibEqOdds	Random	LayerwiseOpt	Adversarial
1	~~0.533~~	~~0.533~~	~~0.983~~	0.519	~~0.567~~	~~0.530~~	0.914
2	~~0.523~~	0.521	~~0.983~~	0.487	0.529	0.508	0.917
3	0.535	0.533	~~0.982~~	0.514	~~0.591~~	0.529	0.905

结论：在实现公平性的模型中（本文的标准是公平性指标小于阈值0.05），Baseline模型、后处理方法得到的模型、随机扰动算法和层级优化方法得到的模型在性能上相似，通过对抗优化微调得到的模型在性能上优于其他方法得到的模型。在更复杂的模型与数据上，更复杂的方法更加有效。

4实验结果

数据集：CelebA

Before	74%	28%	23%	51%	77%	5%	3%	93%	99%	2%	99%	83%
After	92%	35%	29%	61%	84%	11%	2%	92%	99%	2%	99%	84%

Figure 2: Probability of smiling on the CelebA dataset, before and after debiasing w.r.t. race.

结论：Baseline模型在预测目标是否微笑时，预测概率会受其种族因素影响而不平衡。经过对抗优化微调得到的模型有效提升了模型的公平性。

总结

- 提出了一种介于处理中方法与后处理方法之间的黑盒模型公平性方法——“内处理方法”，实验证明该方法能够有效弥补后处理方法导致模型性能下降的缺点，并具有现实应用价值。
- 对内处理方法的性质进行研究，发现公平性问题与模型的初始条件密切相关。
- 通过调整超参数，能够将一些处理中方法转化为内处理方法。

南京郵電大學
Nanjing University of Posts and Telecommunications

敬请各位老师批评指正