Beijing Forest Studio

北京理工大学信息系统及安全对抗实验中心

人工智能模型的公平性测试
——既要公平，也要正确

博士研究生 刘洧光

2025年03月30日

问题回溯

- 总结反思
  - 实验结果部分介绍不够详细，缺乏说明
  - 对公平性的定义没有解释清楚

- 相关内容
  - 2024.09.29 刘洧光《人工智能模型的公平性测试》
  - 2024.01.26 刘洧光《FNN模型正确性测试及测试样本生成》
  - 2022.08.23 王若辉《AI测试：历史与发展》
  - 2022.03.12 侯钰斌《神经网络模型的覆盖测试》

内容提要

- 预期收获
- 题目内涵解析
- 研究背景与意义
- 研究历史与现状
- 知识基础
- 算法原理
  - An Empirical Study on Correlations Between Deep Neural Network Fairness and Neuron Coverage Criteria
  - MirrorFair
- 特点总结与工作展望
- 参考文献

背景简介

- 预期收获
  - 掌握人工智能模型的公平性和正确性测试指标
  - 了解公平性和覆盖率之间的关系
  - 了解一种基于正确性和公平性联合决策的公平性提升方法

内涵解析与研究目标

- 题目内涵解析（人工智能模型的公平性/正确性测试）
  - 人工智能模型：包括机器学习和深度学习模型等
  - 公平性：模型的伦理安全属性，输出是否公正，是模型重要的非功能性需求之一
  - 正确性：模型的内生安全属性，输出是否准确和可靠，是模型最主要的功能性需求

- 研究目标
  - 面向人工智能领域的机器学习/深度学习模型公平性测试
  - 研究数据集公平、算法公平、模型公平等关键问题
  - 迁移正确性测试、鲁棒性测试等理论技术
  - 在提高模型公平性的同时不影响模型的正确性

研究背景和意义

- 研究背景
  - 人工智能在决策系统等领域应用广泛，而模型会使用敏感属性做出偏见决策，产生严重的舆论影响和社会问题
  - 研究表明：许多现有的偏差缓解方法经常导致正确性大幅下降；用于提升公平性的现有方法很难在不同任务或算法之间达成与正确性一致的权衡

- 研究意义
  - 发现模型缺陷
    - 测试模型的公平性，及时发现模型的歧视行为
    - 测试模型的正确性，及时发现模型的错误行为
  - 模型公平性和正确性的增强
    - 实现模型正确性和公平性的双赢（win-win）

研究历史

2017 Galhotra等人首次定义了软件公平和歧视，并开发了一种基于测试的方法来衡量软件是否歧视以及歧视的程度，重点关注歧视行为中的因果关系

2022 Chen等人提出了一种公平性-正确性集成方法MAAT。分别针对公平性和正确性训练优化后的子模型，然后综合两个子模型的结果加权平均后做出最终决策

2023 Zheng等人对DNN公平性和覆盖标准之间的相关性进行了研究。覆盖标准和公平性之间的相关性很有限，提升覆盖率甚至可能对公平性产生负面影响

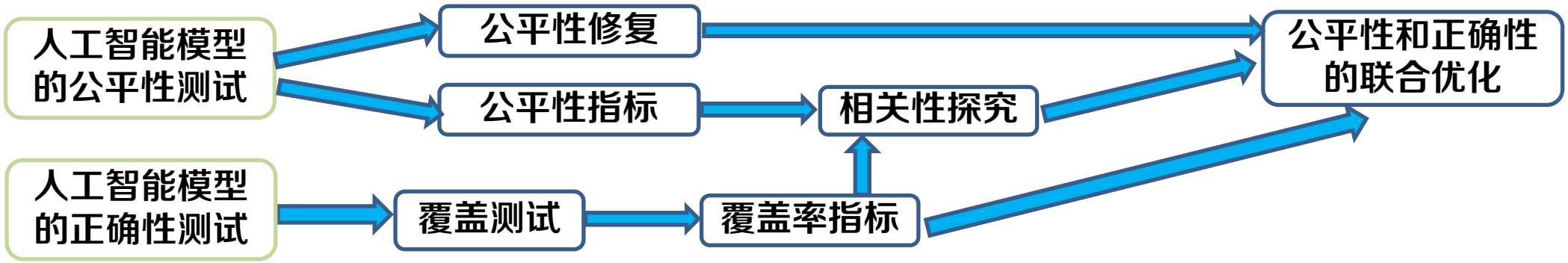
2024 Xiao等人改进了MAAT方法，并提出了MirrorFair。根据原始数据集构建一个反事实数据集，并训练镜像模型。MirrorFair自适应地组合两个模型的预测结果以生成更公平的决策

2017 Pei等人首次提出一种针对深度学习系统的白盒测试方法，借鉴了传统软件测试中的覆盖测试和差分测试思想，并首次提出神经元覆盖率这个指标

2019 Ma等人在Pei等人提出的神经元覆盖率基础上进行了细化，提出了多种细粒度的神经元覆盖率变体，包括KMNC、TKNC、SNAC和NBC

2020 Fabrice-Harel等人对目前的神经元覆盖准则提出了质疑，并通过实验证明神经元覆盖率与模型鲁棒性之间关系性较弱，甚至提升神经元覆盖率的同时会降低鲁棒性

2022 Xie等人为解决传统神经元覆盖率可解释性不足的问题，从DNN模型中提取决策流图，并提出了基于结构的神经元路径覆盖率和基于激活的神经元路径覆盖率



知识基础　模型性能指标（正确性）

- 精确度（ $Precision@c$ ）

$$Precision@c = Pr\left[Y = c \middle| \hat{Y} = c\right] = \frac{TP}{TP + FP}$$

衡量分类器预测为某类别的样本中实际正确的比例（查准能力），反映预测结果的可靠性

- 召回率（ $Recall@c$ ）

$$Recall@c = Pr\left[\left|\hat{Y} = c\right| Y = c\right] = \frac{TP}{TP + FN}$$

衡量分类器对某类别样本的全面识别能力（查全能力），反映实际存在样本被正确找出的比例

- F1值（ $F1@c$ ）

$$F1@c = \frac{2 \times Precision@c \times Recall@c}{Precision@c + Recall@c}$$

通过精确率和召回率的调和平均，综合评估模型在特定类别上的整体表现平衡性

- 准确率（ Acc ）

$$Acc = Pr\left[\left|\hat{Y} = Y\right.\right] = \frac{TP + TN}{TP + FP + TN + FN}$$

衡量模型总体预测正确率，但易受类别分布不平衡影响，对偏斜数据集评估可靠性较低

- Matthews相关系数（Mcc）

$$Mcc = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

  - 真阳性TP、真阴性TN、假阳性FP和假阴性FN
  - 准确率的局限性：在不平衡的数据集中，即正例和负例的数量差异较大时，准确度可能不是一个很好的评估指标。因为如果一个模型倾向于预测较多的样本为多数类别，那么它可能仍然具有高准确度，但在实际应用中可能并不理想
  - 范围： MCC 的值范围为[-1, 1]，其中 1 表示完美预测，0 表示等同于随机预测，-1 表示完全相反的预测
  - 适用性： MCC 尤其适用于样本不平衡的二分类问题，因为它同时考虑了各个类别的预测性能

知识基础　覆盖率指标（正确性）

- 传统神经元覆盖率（Neuron Coverage，NC）

$$NC(T,x)=\frac{|(n|\forall x\in T,out(n,x)>t)|}{|N|}$$

  - 公式说明
    - 其中，$|N|$代表模型的神经元总数，$out(n,x)$是神经元n对于某一测试样本x的输出值，t是预设的神经元激活阈值，T是测试样本全集
- k节神经元覆盖率（KMNC）

$$KMNC(k)=\frac{|\{S_i^n|\exists x\in T:\phi(x,n)\in S_i^n\}|}{k}$$

  - 公式说明
    - 神经元的输出位于区间$[low,high]$，将区间$[low,high]$分成相等的k个区段
    - 当深度学习系统运行测试集T中的一个测试样本x时，神经元的输出$\phi(x,n)$位于一个区段S_i中，则区段S_i被覆盖

知识基础　覆盖率指标（正确性）

- 神经元边界覆盖率（NBC）
  - 计算公式

$$NBC(k) = \frac{\left|\sum_{i=1}^{N}(U_i + L_i)\right|}{2N}$$

  - 公式说明
    - 不同于$KMNC(k)$，$NBC(k)$的目标是覆盖边界，即$(low - k\sigma, low]$和$(high, high + k\sigma]$
    - 其中σ是在训练过程中的神经元输出的标准偏差，k是用户定义的参数。设共有N个测试样本，U_i和L_i分别表示样本i的输出落在上边界和下边界

知识基础　覆盖率指标（正确性）

- 强神经元激活覆盖率（SNAC）
  - 计算公式

$$SNAC(k) = \frac{\left|\sum_{i=1}^{N} U_i\right|}{N}$$

  - 公式说明
    - $SNAC(k)$是$NBC(k)$的一个特例，只考虑了上边界的情况
    - 其中σ是在训练过程中的神经元输出的标准偏差，k是用户定义的参数。设共有N个测试样本，U_i和L_i分别表示样本i的输出落在上边界和下边界

知识基础　覆盖率指标（正确性）

- $Top-k$神经元覆盖率（TKNC）
  - 计算公式

$$TKNC(k)=\frac{\left|U_{x\in T}\left(U_{1\le i\le l}top_k(x,i)\right)\right|}{N}$$

  - 公式说明
    - 对于给定的测试输入x和同一层上的神经元n_1和n_2，如果神经元输出值$\phi(x,n_1)>\phi(x,n_2)$，则表示n_1比n_2更活跃。$top_k(x,i)$表示给定一个x,第i层中值最大的k个神经元。
    - $TKNC(k)$测量了每一层中曾经最活跃的k个神经元的数量，定义为每层top_k神经元总数与模型中神经元总数之比

- 统计奇偶差异（ SPD ）
  - 计算公式

$$SPD = Pr\left[\hat{Y} = 1 | A = 0\right] - Pr\left[\hat{Y} = 1 | A = 1\right]$$

  - 公式说明
    - $\hat{Y}$为模型预测结果， A为敏感属性，对于特权组， A设置为0
    - 用于衡量模型对特权群体和非特权群体的结果平等性
  - 应用场景示例
    - 若男性贷款获批率80%，女性仅60%，则$SPD = 0.2$，说明模型存在性别歧视
  - 局限性
    - 仅测量结果分布的群体差异，但无法区分差异来源
    - 合理差异：岗位对体力要求高→男性申请者通过率自然高
    - 不合理歧视：相同资历下女性简历被过滤更多
    - 数学本质： SPD = 系统偏差 + 数据偏差 + 合理差异

模型公平性指标（群体公平性）

- 平均赔率差异（AOD）
  - 计算公式

$$AOD = \frac{1}{2}\left(\left|Pr\left[\hat{Y} = 1 | A = 0, Y = 0\right] - Pr\left[\hat{Y} = 1 | A = 1, Y = 0\right]\right| + \left|Pr\left[\hat{Y} = 1 | A = 0, Y = 1\right] - Pr\left[\hat{Y} = 1 | A = 1, Y = 1\right]\right|\right)$$

  - 公式说明
    - $\hat{Y}$为模型预测结果，A为敏感属性，Y为真实标签
    - 公式本质为$AOD = \frac{1}{2}[|$假阳性率差异$| + |$真阳性率差异$|]$
    - 用于衡量模型对两个群体的分类性能一致性
  - 应用场景示例
    - 若白人群体的疾病漏诊率比黑人低5%，误诊率高3%，则$AOD = 4\%$

知识基础　模型公平性指标（群体公平性）

- 平等机会差异（EOD）
  - 计算公式

$$EOD = Pr\left[\hat{Y} = 1 | A = 0, Y = 1\right] - Pr\left[\hat{Y} = 1 | A = 1, Y = 1\right]$$

  - 公式说明
    - $\hat{Y}$为模型预测结果，A为敏感属性，Y为真实标签
    - 比较两个群体中真正符合条件者被公平对待的比例
  - 应用场景示例
    - 若高学历和低学历的"实际优秀员工"中，被模型正确晋升的比例相差15%，则$EOD = 0.15$，说明存在学历偏见

- 差别影响（DI）
  - 计算公式

$$DI = min\left\{\frac{Pr\left[\hat{Y}=1|A=0\right]}{Pr\left[\hat{Y}=1|A=1\right]}, \frac{Pr\left[\hat{Y}=1|A=1\right]}{Pr\left[\hat{Y}=1|A=0\right]}\right\}$$

  - 公式说明
    - $\hat{Y}$为模型预测结果，A为敏感属性
    - 计算方式：取两个群体录取率的较小值除以较大值
    - 取值范围：强制限定在[0，1]区间（1表示完全公平，0表示极端歧视）
    - 法律渊源：源自美国EEOC《统一准则》的80%规则（若DI<0.8，视为存在非法歧视）
  - 应用场景示例
    - 合规性检查中特权群体（有房者）获批率85%，未特权群体（无房者）获批率68%，DI = 68/85≈0.8（踩法律红线）

知识基础　模型群体公平性指标对比

- 指标对比表

指标	核心差异维度	敏感场景	公平性理想值
SPD	结果分布的群体平等性（结果公平）	资源分配（贷款/录取）	0
AOD	分类性能的群体一致性（过程公平）	风险评估（医疗/司法）	0
EOD	机会获取的条件平等性（机会公平）	资格认定（晋升/奖学金）	0

- DI与SPD对比分析

指标	计算方式	敏感方向	场景侧重
DI	比率（相对差异）	关注弱势群体利益	法律合规审查
SPD	差值（绝对差异）	双向平等性	学术研究分析

- 泰尔指数（ TI ）——个体与群体公平的熵度量
  - 计算公式

$$\varepsilon(\alpha)=\begin{cases}\dfrac{1}{n\alpha(\alpha-1)}\sum_{i=1}^{n}\left[(\dfrac{b_i}{\mu})^{a}-1\right], \alpha\neq 0,1,\\ \dfrac{1}{n}\sum_{i=1}^{n}\dfrac{b_i}{\mu}ln\dfrac{b_i}{\mu}, a=1,\\ -\dfrac{1}{n}\sum_{i=1}^{n}ln\dfrac{b_i}{\mu}, a=0.\end{cases}$$

  - 公式说明
    - 本质是对个体b_i求期望
    - 测量对象：资源分配或决策结果的分布不平等性。①群体公平：不同敏感属性群体间的收益差异；②个体公平：相似个体（相同特征）的收益差异
    - 敏感方向：对高收益群体的过度分配更敏感（右偏分布惩罚）
  - 应用场景示例
    - 原始分配：100个申请人，90人获贷10k，10人获贷100k；μ= (90×10k + 10×100k)/100 = 19k； $TI\approx 0.63$
    - 优化后：调整为80人15k，20人50k， $TI\approx 0.37$（公平性提升41%）

- 一致性（CNT）——局部相似个体的预测稳定性
  - 计算公式

$$CNT = 1 - \frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i - \frac{1}{n\_neighbors}\sum_{j\in N_{n_{neighbors}}(x_i)}\hat{y}_j\right|$$

  - 公式说明
    - 测量对象：特征相似个体的预测结果一致性
    - 公平准则："Similar individuals should receive similar outcomes“
  - 应用场景示例
    - 两位候选人（年龄28，硕士，5年经验）：候选人A（男性）→ 预测通过（y=1），候选人B（女性）→ 预测拒绝（y=0），在k=5邻域内，发现3次类似不一致 → CNT下降0.15

研究现状

- 正确性与覆盖指标
  - 覆盖指标在领域初期取得了巨大成果，但目前逐步被人证明存在不合理性或局限性，转而研究更具有可解释性的路径覆盖、因果覆盖等指标
  - 覆盖指标无法指导模型公平性的测试
- 公平性与正确性的联合优化
  - 经研究表明，现有大部分公平性提升方法只能单边提升某类公平性指标，且会对正确性造成损害
  - 现有方法考虑到公平性和正确性的联合优化，采用集成学习模式，或者设计多个优化目标

Faire

【 TSE 】

An Empirical Study on Correlations Between Deep Neural Network Fairness and Neuron Coverage Criteria

论文A TIPO

T	目标	系统性地探究DNN公平性和覆盖标准的相关性
I	输入	测试模型*N个，测试指标*N个
P	处理	RQ1：数据集的大小如何影响DNN覆盖率和DNN公平性的评估? RQ2：覆盖标准和DNN公平性之间有何相关性? RQ3：采用不同的偏差缓解方法后，覆盖率是否有显著变化? RQ4：在偏差缓解后，覆盖标准与 DNN 公平性之间是否存在相关性? RQ5：不同的覆盖标准以及不同的公平性指标本身之间是否存在内在相关性?
O	输出	实验结果与分析

P	问题	已知覆盖指标与鲁棒性的相关性较小，欲探究覆盖指标与公平性的关系
C	条件	白盒模型、能访问训练数据
D	难点	如何分析不同指标，不同修复技术之间的关系
L	水平	TSE 2023（CCF A）

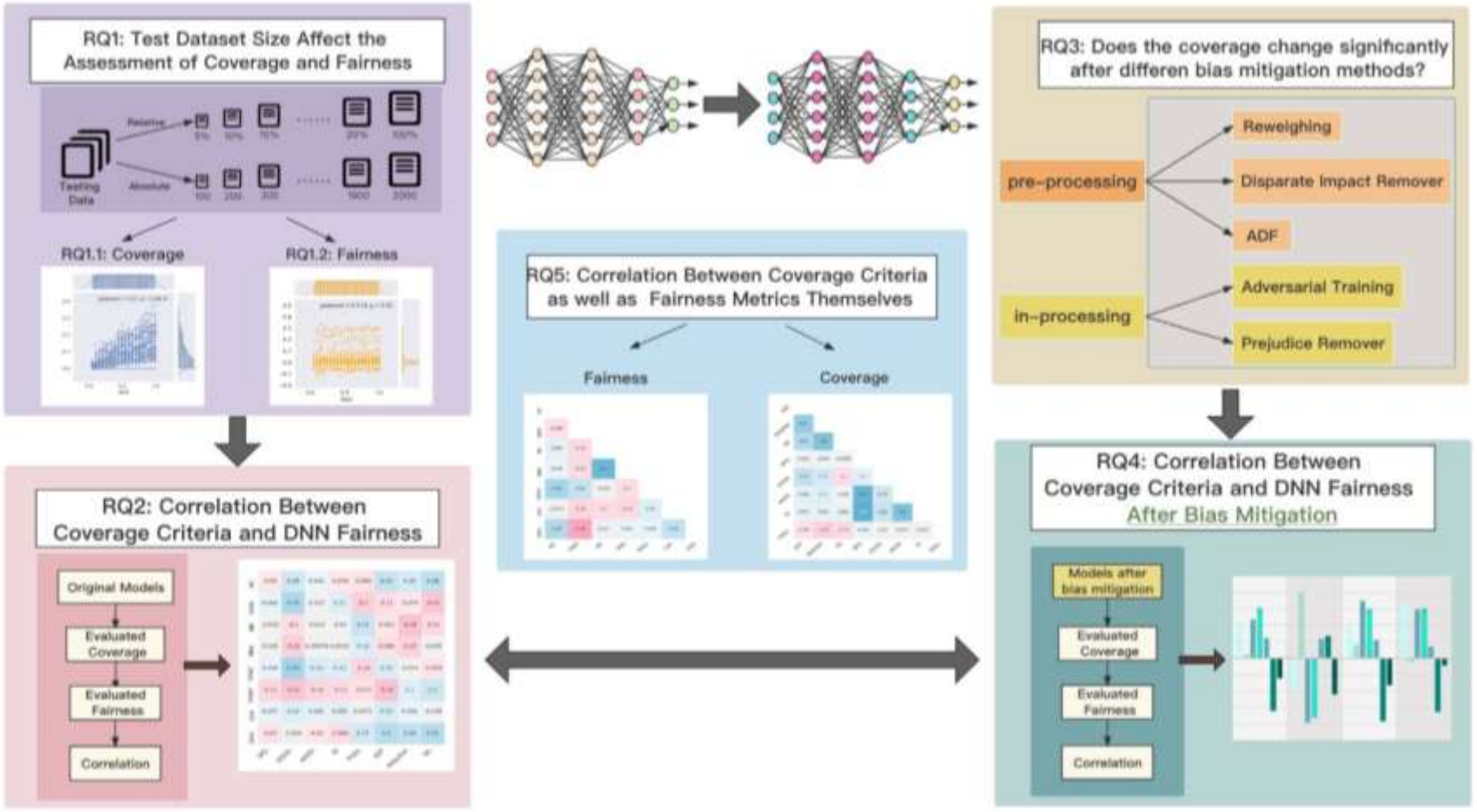
论文A 实验目标图

- 实验目标图

RQ5：不同的覆盖标准以及不同的公平性指标本身之间是否存在内在相关性?

RQ1：数据集的大小如何影响DNN覆盖率和DNN公平性的评估?

RQ2：覆盖标准和DNN公平性之间有何相关性?



RQ3：采用不同的偏差缓解方法后，覆盖率是否有显著变化?

RQ4：在偏差缓解后，覆盖标准与 DNN 公平性之间是否存在相关性?

实验设计　数据资源

- 数据与模型资源
  - 2至6层全连接DNN
    - [4]
    - [8，4]
    - [16，8，4]
    - [32，16，8，4]
    - [64，32，16，8，4]
- 对比方法
  - RW（重新加权：预处理）
  - DIR（不同影响消除：预处理）
  - ADF（歧视样本生成：再训练）
  - 对抗性训练偏差缓解（处理中）
  - PR（偏见消除：处理中）

Name	Abbr.	#Features	Protected Attributes	Size
Adult Income [47]	adult	14	sex,race	45,222
Bank Marketing [48]	bank	20	age	30,488
COMPAS Score [49]	compas	10	sex,race	6,167
Default Credit [50]	default	23	sex	30,000
German Credit [51]	German	20	sex	1,000
Heart Health [52]	heart	14	age	297
Medial Survey 2015 [53]	meps15	41	race	15,830
Medial Survey 2016 [53]	meps16	41	race	34,655
Student Performance [54]	student	33	sex	1,044

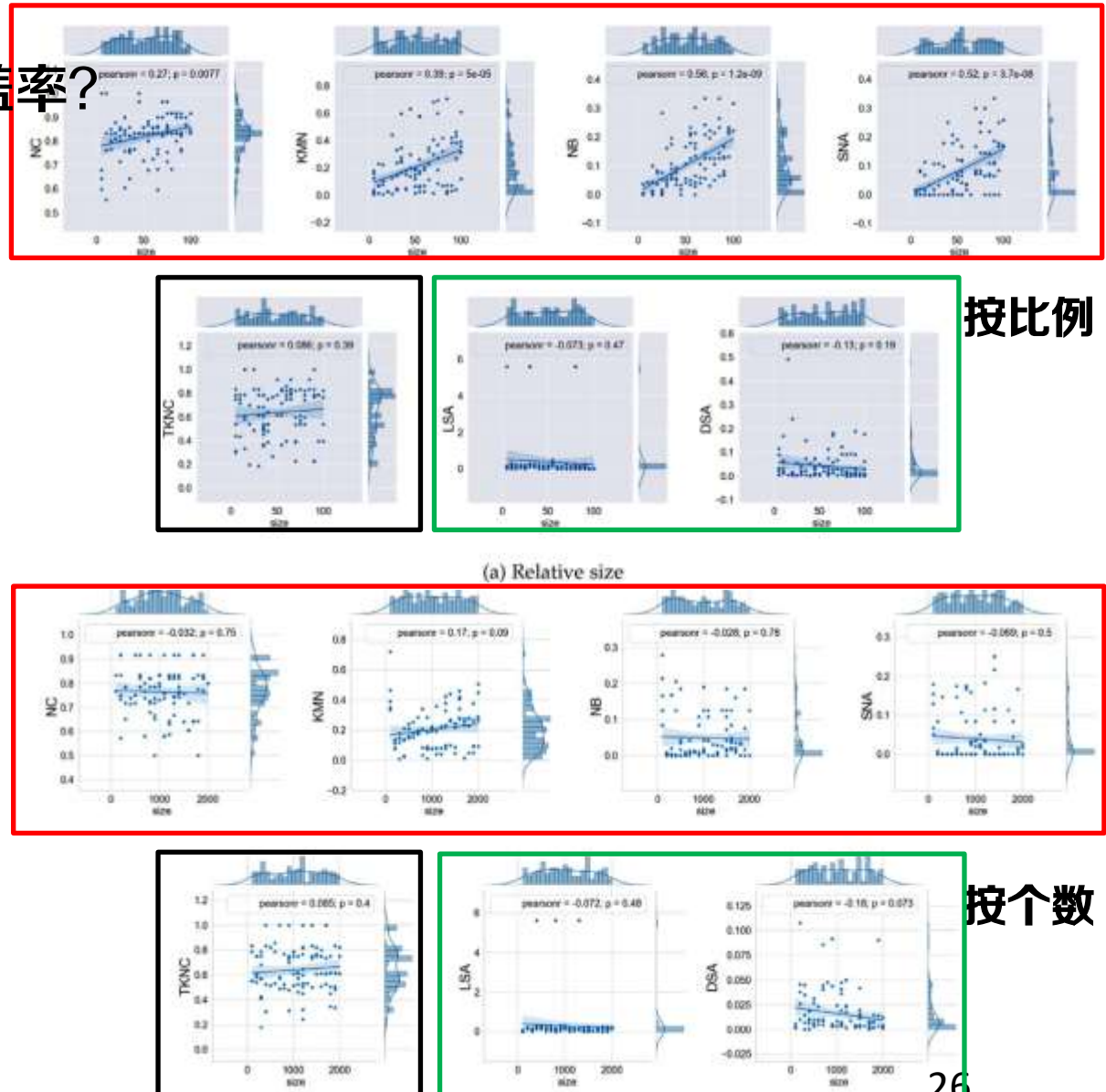
- 评价指标
  - SPD：统计奇偶差异
  - AOD：平均赔率差异
  - EOD：平等机会差异
  - DI：差别影响

  群体公平性
  - TI：泰尔指数
  - CNT：一致性

  个体公平性

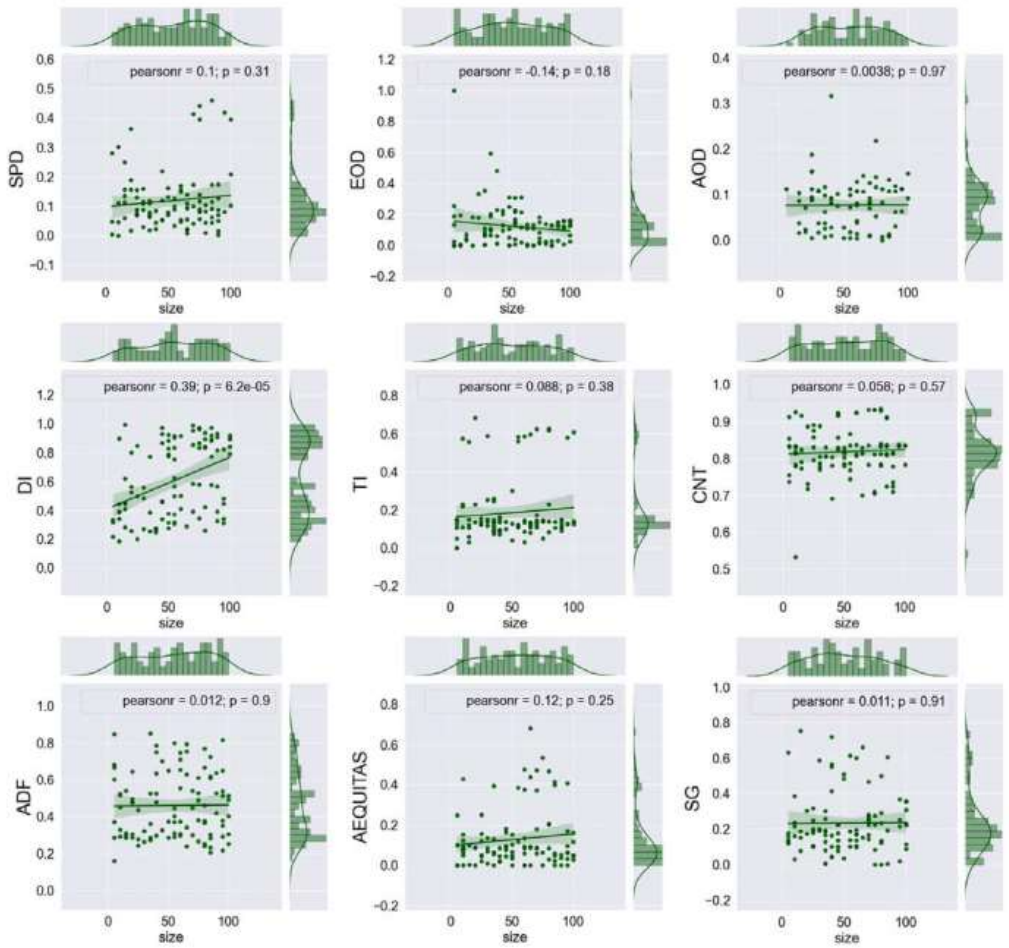
实验设计　实验一

- **RQ1.1：　测试数据集的大小如何影响DNN的覆盖率?**
- **实验分析**
  - **随着测试数据的相对增加，NC、KMN、NB和SNA呈现上升趋势，但绝对增加的变化趋势相对温和**
  - **TKNC在相对和绝对增长方面都显示出一致和适度的变化**
  - **LSA和DSA在相对和绝对增加方面都表现出下降趋势**
- **实验结论**
  - **测试数据集的大小对覆盖率评估的影响是显著的。无论测试数据集的大小是相对增加还是绝对增加，这种显著性都存在**



实验设计　实验一

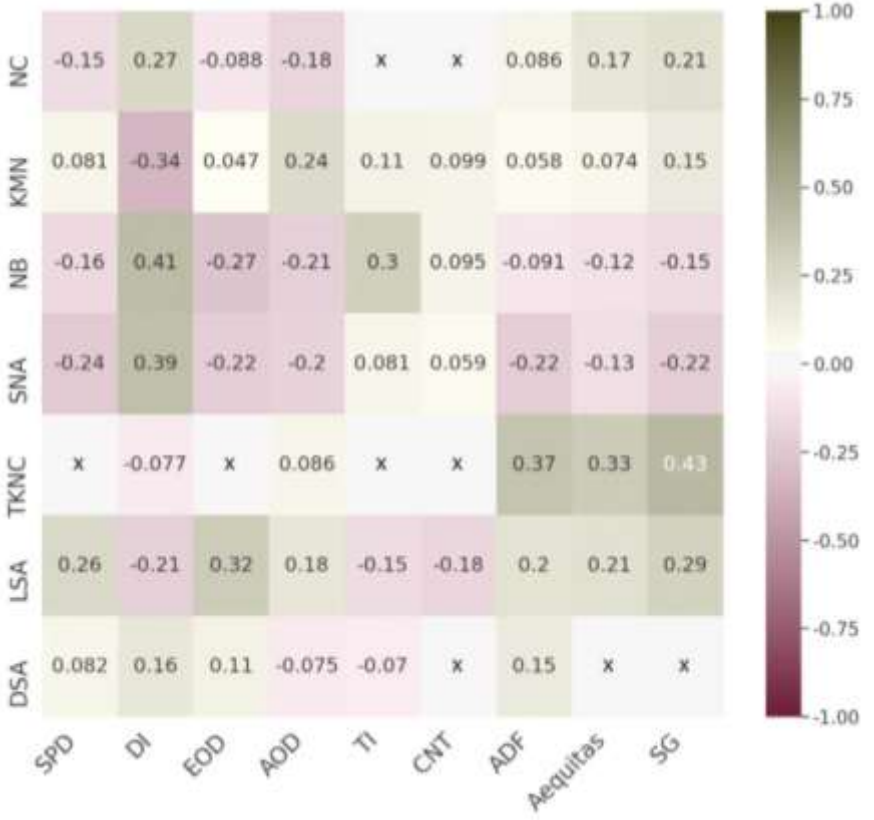
- RQ1.2：测试数据集的大小如何影响DNN公平性的评估?
- 实验分析
  - 令人惊讶的是，对于大多数公平性度量，随着测试数据集大小的增加，变化变得可以忽略不计
- 实验结论
  - DNN的公平性测试技术受测试数据集大小变化的影响较小



(a) Relative size

实验设计　实验二

- RQ2：覆盖标准和DNN公平性之间有什么相关性吗?
- 实验分析
  - 利用τ秩相关系数来分析覆盖率和公平性之间的相关性，取值为[−1，1]
  - 绿色为正相关，粉色为负相关，相关性绝对值越大颜色越深
  - 绝对值<0.4为低相关， [0.4，0.7]为中等相关， [0.7，0.9]为高相关，>0.9为超高相关
- 实验结论
  - 两者之间不存在显著的强正相关。事实上，分析甚至表明了微弱的负相关性。这意味着覆盖标准的更高覆盖实际上可能对DNNs的公平性产生负面影响



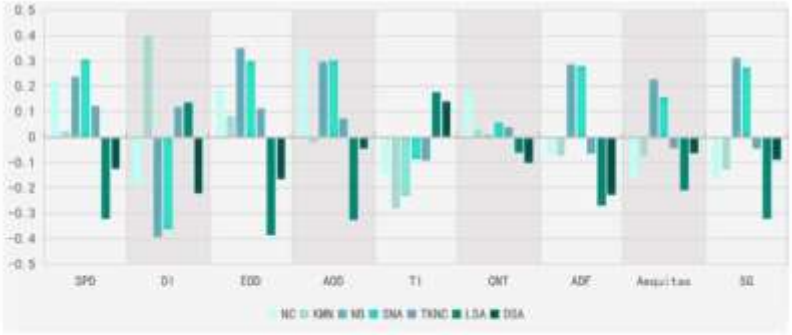
实验设计　实验三

- RQ3：采用不同的偏差缓解方法后，覆盖率是否有显著变化?
- 实验结论
  - 当应用各种偏差缓解技术时，大多数覆盖标准表现出显著变化，但这些变化的方向各不相同。例如，NC、NB、SNA、LSA和DSA表现出显著增加，而KMN表现出显著减少，并且TKNC表现出振荡行为

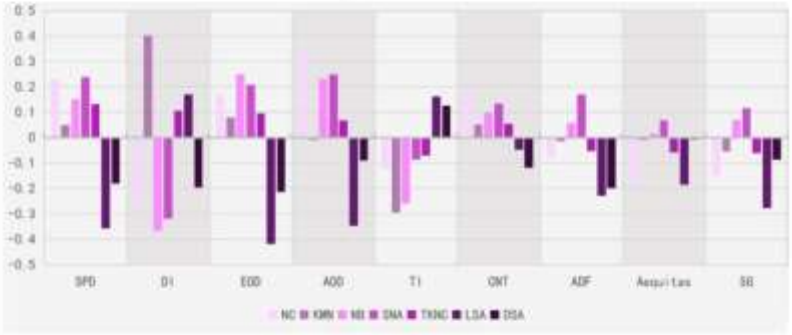
Dataname	Method	NC		KMN		NB		SNA		TKNC		LSA		DSA	
		T	P	T	P	T	P	T	P	T	P	T	P	T	P
adult-race	RW	-18.491	**	4.393	**	-27.970	**	-27.850	**	1.729	>0.05	4.851	**	-7.412	**
	DIR	-9.889	**	12.063	**	-42.416	**	-56.947	**	-6.507	**	15.728	**	-30.825	**
	ADF	-15.075	**	7.972	**	-32.031	**	-26.394	**	-0.176	>0.05	4.933	**	-10.477	**
	Adebias	-16.908	**	6.101	**	-41.738	**	-28.083	**	-1.373	>0.05	7.714	**	-9.159	**
	PR	-20.335	**	5.921	**	-32.375	**	-31.551	**	-1.574	>0.05	2.213	*	-10.691	**
adult-sex	RW	-15.854	**	5.566	**	-30.744	**	-27.596	**	-0.405	>0.05	6.141	**	-8.478	**
	DIR	-35.066	**	16.015	**	-65.593	**	-63.604	**	-0.416	>0.05	15.184	**	-30.836	**
	ADF	-15.153	**	6.955	**	-28.153	**	-25.304	**	-1.198	>0.05	6.202	**	-8.941	**
	Adebias	-17.501	**	4.482	**	-40.659	**	-36.669	**	-1.228	>0.05	6.702	**	-11.384	**
	PR	-21.360	**	5.230	**	-42.193	**	-36.822	**	-1.563	>0.05	5.482	**	-9.932	**
bank-age	RW	-5.961	**	4.780	**	-55.510	**	-41.614	**	5.806	**	8.067	**	-10.231	**
	DIR	-32.641	**	15.927	**	-64.656	**	-62.522	**	1.371	>0.05	15.848	**	-30.891	**
	ADF	-7.140	**	3.889	**	-42.952	**	-16.659	**	5.108	**	13.599	**	-8.316	**
	Adebias	-6.785	**	3.185	**	-37.360	**	-23.556	**	5.430	**	12.968	**	-12.486	**
	PR	-7.092	**	-0.924	>0.05	-46.964	**	-33.107	**	4.860	**	13.202	**	-10.157	**
compas-race	RW	-19.046	**	6.751	**	-13.346	**	-23.894	**	-2.865	**	17.887	**	-14.221	**
	DIR	-34.961	**	16.667	**	-60.903	**	-63.017	**	-0.918	>0.05	15.33	**	-30.651	**
	ADF	-16.734	**	3.305	**	-14.207	**	-26.050	**	-4.303	**	16.490	**	-12.078	**
	Adebias	-17.854	**	7.444	**	-21.761	**	-18.368	**	-2.048	*	21.305	**	-14.642	**
	PR	-17.866	**	7.570	**	-20.722	**	-19.530	**	-2.525	*	23.307	**	-15.467	**
compas-sex	RW	-21.698	**	8.063	**	-19.145	**	-25.219	**	-3.091	**	-11.032	**	-10.364	**
	DIR	-35.178	**	16.611	**	-62.651	**	-63.448	**	-0.868	>0.05	-15.233	**	-30.719	**
	ADF	-17.004	**	7.010	**	-24.352	**	-30.300	**	-3.956	**	-17.350	**	-11.740	**
	Adebias	-17.298	**	6.310	**	-14.387	**	-25.695	**	-3.615	**	-15.223	**	-12.285	**
	PR	-19.309	**	9.002	**	-22.297	**	-23.297	**	-2.055	*	-15.888	**	-13.260	**
default-sex	RW	-7.810	**	18.753	**	-10.275	**	-9.589	**	-9.164	**	-8.219	**	-8.954	**
	DIR	-32.295	**	17.674	**	-57.258	**	-57.489	**	-3.193	**	-16.23	**	-30.53	**
	ADF	-6.280	**	19.576	**	-9.642	**	-10.684	**	-9.204	**	-8.536	**	-12.309	**
	Adebias	-5.229	**	13.678	**	-7.928	**	-11.133	**	-9.759	**	-9.454	**	-11.129	**
	PR	-7.044	**	17.936	**	-8.467	**	-6.236	**	-10.214	**	-7.853	**	-10.201	**
German-sex	RW	-2.844	**	20.996	**	-29.836	**	-27.510	**	8.813	**	-3.324	**	-12.617	**
	DIR	-31.129	**	16.340	**	-63.001	**	-61.831	**	2.735	**	-14.785	**	-31.073	**
	ADF	-4.036	**	17.753	**	-25.962	**	-17.489	**	9.623	**	-3.789	**	-14.541	**
	Adebias	-2.281	*	24.735	**	-18.729	**	-18.074	**	9.789	**	-3.897	**	-14.225	**
	PR	-2.869	**	26.597	**	-19.230	**	-16.355	**	9.549	**	-2.764	**	-11.902	**
heart-age	RW	2.591	*	40.594	**	-37.316	**	-19.047	**	7.144	**	-3.164	**	-4.500	**
	DIR	-28.924	**	15.933	**	-65.419	**	-63.207	**	1.267	>0.05	-16.711	**	-32.370	**
	ADF	1.649	>0.05	23.873	**	-26.579	**	-18.226	**	6.589	**	-13.567	**	-9.817	**
	Adebias	1.333	>0.05	35.936	**	-34.652	**	-25.869	**	7.160	**	-5.838	**	-6.334	**
	PR	2.368	**	27.223	**	-26.674	**	-14.101	**	7.349	**	-16.068	**	-14.136	**

实验设计　实验四

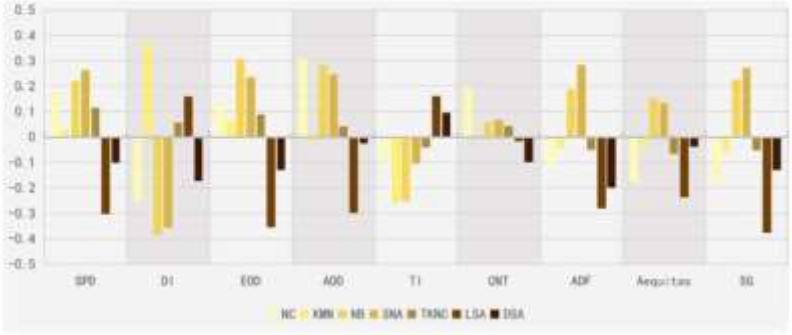
- RQ4：偏差缓解后，覆盖标准和DNN公平性之间有相关性吗?
- 实验分析
  - 每种偏差缓解方法都显著降低了覆盖标准和公平性度量之间的相关性的绝对值
  - 与RQ2的热图结合分析，当原始数据的相关值为正时，应用偏差缓解方法后相关值变化的趋势通常趋于降低。 相反，当原始值为负值时，趋势通常倾向于增加
- 实验结论
  - 在应用偏差缓解方法之后，覆盖标准和公平性度量之间的相关性趋于降低



(a) Adebias



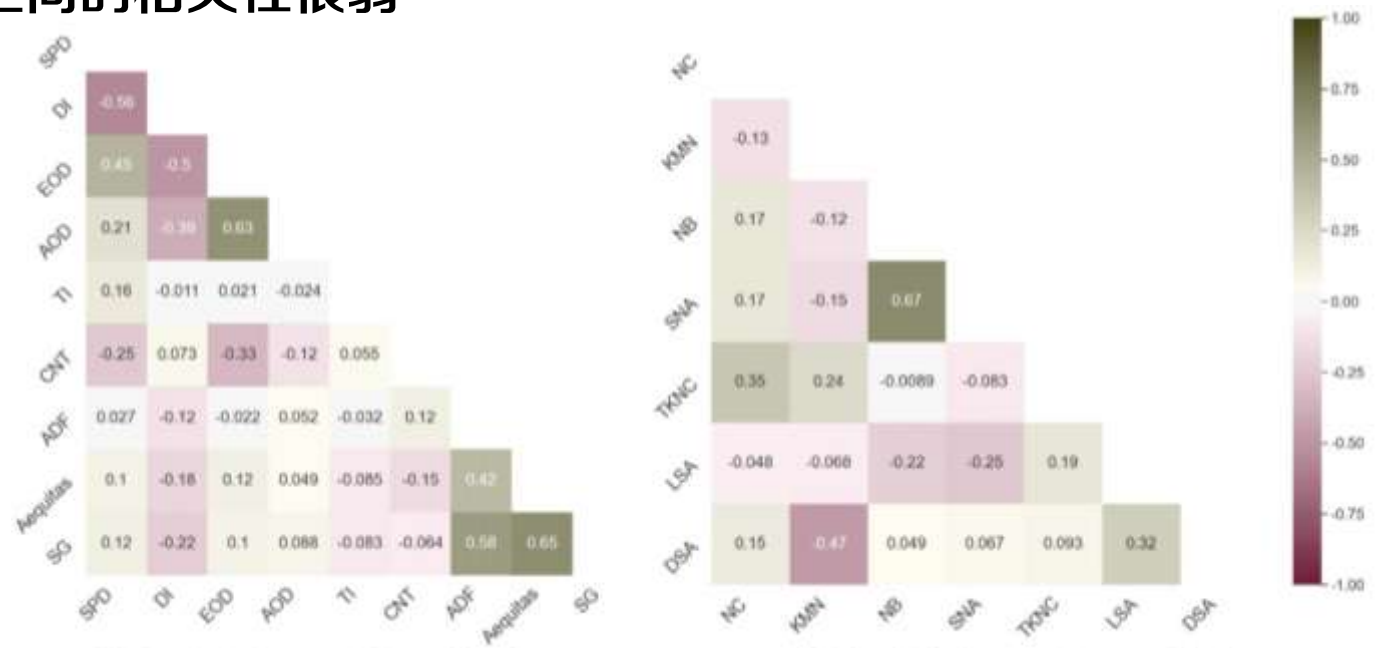
(b) ADF



(c) DIR

实验设计　实验五

- RQ5：不同的覆盖标准以及不同的公平性指标本身之间是否存在内在相关性?
- 实验结论
  - 个体公平性测试技术之间以及群体公平性度量之间存在适度的相关性；然而，个人和群体公平性指标之间的相关性很弱
  - 关于覆盖标准，结果显示NB和SNA之间以及DSA和KMN之间的中度相关性；然而，其余覆盖标准之间的相关性很弱



(a) Correlation between Fairness Metrics　　(b) Correlation between Coverage Criteria

Faire

【 FSE 】

MirrorFair: Fixing Fairness Bugs in Machine Learning Software via Counterfactual Predictions

BET TIPO

T	目标	提升模型公平性的同时不损失模型的正确性
I	输入	训练集*1，原模型*1
P	处理	1.通过突变敏感属性来构建镜像数据集 2.分别从原始训练集和镜像训练集训练两个模型 3. 自适应地集成两个模型的预测来做出决策
O	输出	基于正确性模型和公平性镜像模型的综合决策

P	问题	许多现有的公平性缓解方法经常导致性能大幅下降； 现有方法根据任务、数据集、模型和敏感属性的不同，效果差异很大
C	条件	可访问的训练集
D	难点	1.如何做好正确性和公平性的权衡 2.如何综合两个模型做出最优解
L	水平	FSE 2024（CCF-A）

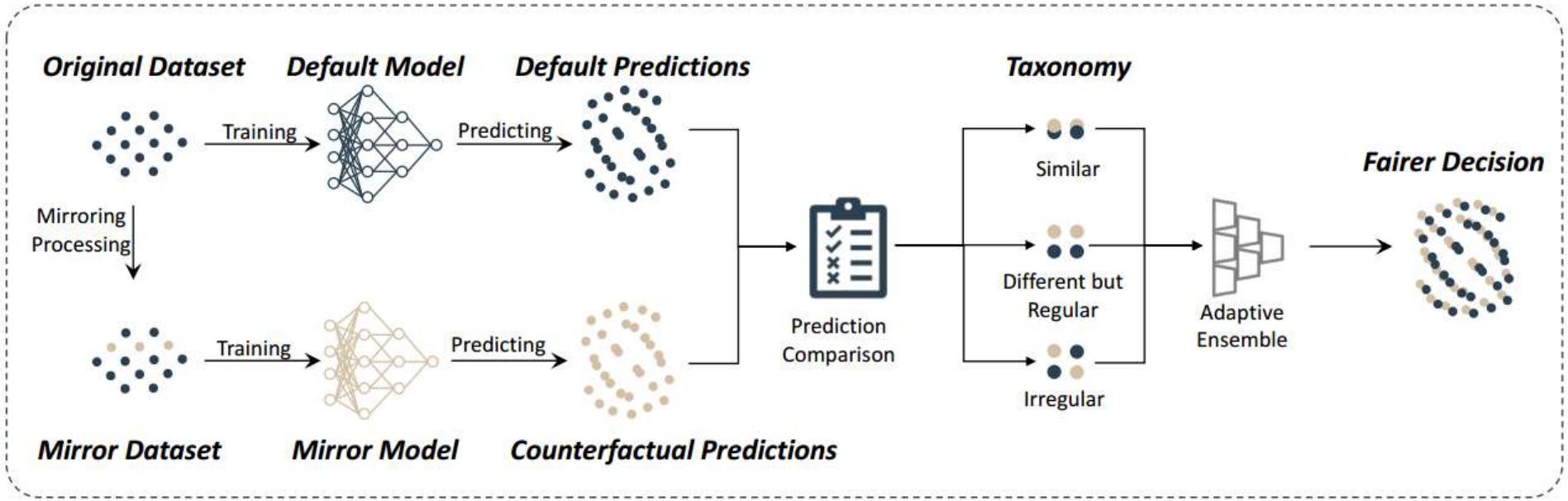
BET 算法原理图

- 算法原理图

步骤1：通过突变敏感属性来构建镜像数据集

步骤2：分别从原始训练集和镜像训练集训练两个模型

步骤3：自适应地集成两个模型的预测来做出决策



创新分析　反事实推理与镜像处理

- 反事实推理
  - 反事实是为已经发生的事件或情况提出的假设情景（与事实相反）
  - 反事实推理是一种推理方法，涉及基于假设场景推断可能性
  - 反事实公平是一种反事实推理的实践，它要求在一个决策任务中，敏感属性值的变化不应改变原始决策
- 镜像处理
  - 当处理多值敏感属性时，确定哪个值作为修改的目标是一个挑战
    - 如“种族”类别中包括像“白人”、“亚洲人”和“爱斯基摩人”等
  - 将敏感属性简化为二进制形式（如，“白人”和“非白人”）
    - 这种简化可能会模糊子组之间的区别，但它大大简化了建模公平性问题的复杂性，并促进了反事实数据集的构建

- 决策情景的分类
  - 基于决策边界附近的特征，将决策任务分为镜像不敏感场景、镜像规则场景和镜像不规则场景三类：

$$S_{type} = \begin{cases} mirror-insensitive & if \forall d_{A=a} \in D_{test}, \left|DIF_{d_{A=a}}\right| \in N^{\delta}(0) \\ mirror-regular & if \forall d_{A=a} \in D_{test}, \left|DIF_{d_{A=a}}\right| \in N^{\delta}(c), c \neq 0 \\ mirror-irregular & if \exists d_{A=a} \in D_{test}, \left|DIF_{d_{A=a}}\right| \in N^{\delta}(c) \end{cases}$$

- S_{type}表示决策场景类型； D_{test}表示测试数据集； $d_{A=a}$表示测试实例； $N^{\delta}(0)$和 $N^{\delta}(c)$ 表示0和c的邻域， c为常数； $DIF_{d_{A=a}}$的计算公式如下：

$$DIF_{d_{A=a}} = P_{def}(\hat{Y} = y | d_{A=a}) - P_{mir}(\hat{Y} = y | d_{A=a})$$

- $\hat{Y}$表示预测输出标签（例如，收入）； y表示标签值（例如，高收入或低收入）； A表示敏感属性（如性别）； a表示敏感属性值（例如，女性或男性）； $P_{def}(\hat{Y} = y | d_{A=a})$表示原模型的概率； $P_{mir}(\hat{Y} = y | d_{A=a})$表示镜像模型的概率

- 决策情景的分类
  - 镜像不敏感场景：原模型与镜像模型的预测几乎相同，即镜像处理对这种场景的预测影响很小
  - 镜像规则场景：原模型与镜像模型的预测不同，但两个模型之间的绝对概率差异是规则的并接近常数值的情况，即镜像处理对这种场景的每个预测具有显著和规则的影响
  - 镜像不规则场景：原模型与镜像模型的预测不同，并且对于不同的测试实例，两个模型之间的绝对概率差异是不规则的，镜像处理对这种场景的预测具有显著的、不规则的和不确定的影响

- 自适应集成策略
  - 为了自适应地处理不同的决策场景，提出了E-Mean和E-Max两种策略来集成原模型和镜像模型的预测
    - 关于镜像规则场景，其中镜像处理对模型预测做出显著和规则的贡献，通过E-Mean策略加权平均来集成两个模型的预测结果，其输出概率向量如下：

$$P_{final} = \left[\frac{P_{def}(Y=0) + P_{mir}(Y=0)}{2}, \frac{P_{def}(Y=1) + P_{mir}(Y=1)}{2}\right]$$

    - 在其他决策场景方面，通过E-Max最大化决策边界附近（$0.5 - c < P_{def}(Y = 1) < 0.5 + c$ ）无特权组的有利标签概率来修复机器学习软件中的公平性问题（边界外的预测仍然遵循E-Mean策略），其输出概率向量如下：

$$P_{final} = \left[min(P_{def}(Y=0), P_{mir}(Y=0), max(P_{def}(Y=1), P_{mir}(Y=1)\right]$$

创新分析　多敏感属性保护

- 多敏感属性保护策略
  - 在机器学习预测任务中，数据集可能包含多个需要保护的敏感特征。例如，在Adult和Compas数据集中都存在性别和种族敏感特征
  - 覆盖多个敏感特征保护是评估偏差缓解方法的重要评估维度，MirrorFair采用如下两种策略：
    - MirrorMulti-S1：通过同时反转原始训练数据集中的性别和种族值来创建性别-种族镜像训练数据集
    - MirrorMulti-S2：分别单独生成性别镜像和种族镜像训练数据集。然后，组合来自原模型、性别镜像模型和种族镜像模型的预测以生成预测。该策略在控制预测中每个敏感特征的权重方面提供了更大的灵活性，并且允许每个特征的不同权重设置以实现更细粒度的公平性水平

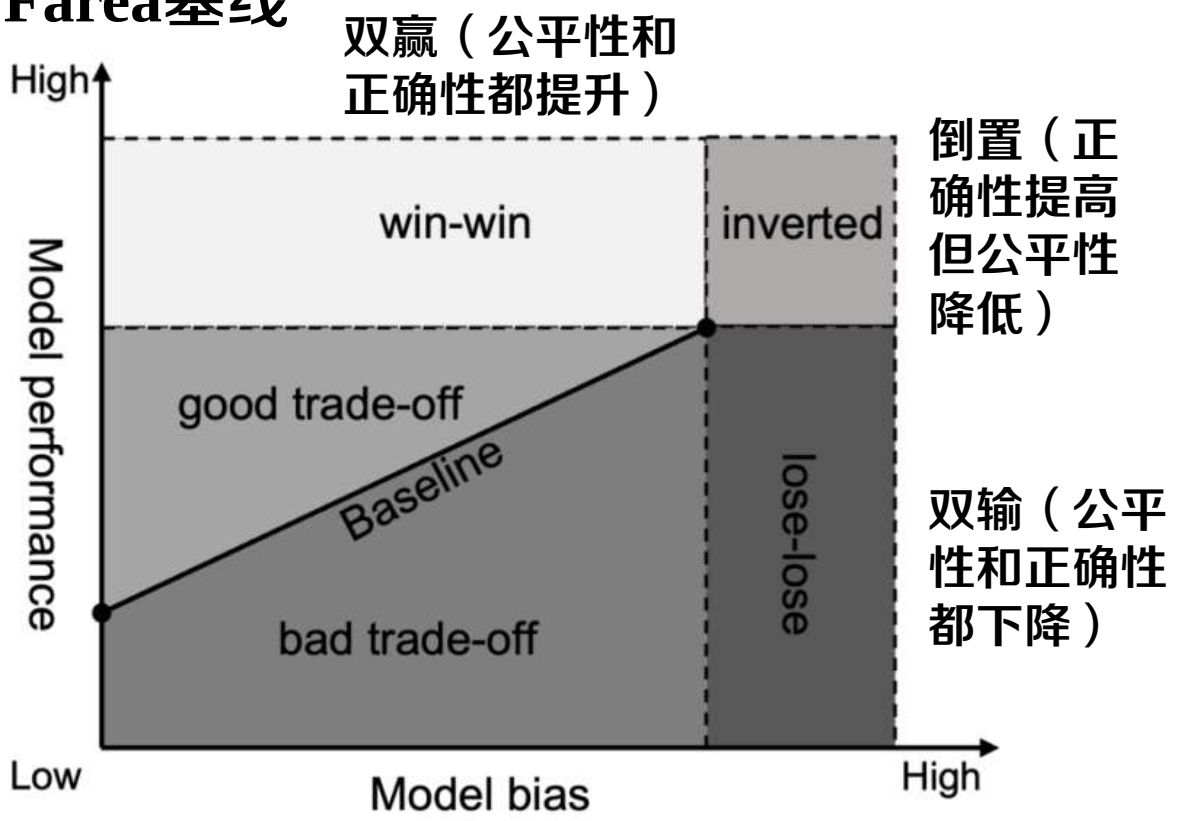
- 评价指标
  - Acc：准确率（+）
  - $Recall$：召回率（+）
  - $Precision$：精确率（+）
  - $F1-score$：宏F1值（+）

  } 正确性

  - SPD：统计奇偶差异（-）
  - AOD：平均赔率差异（-）
  - EOD：平等机会差异（-）

  } 公平性

Farea基线



实验设计　数据资源

- 数据与模型资源
  - 逻辑回归（LR）
  - 随机森林（RF）
  - 支持向量机（SVM）
  - 深度神经网络（DNN）
- 对比方法
  - Fairea基线（2023）

Task	Protected attribute(s)	Dataset	Size	Favourable label	Majority label
1. Adult-sex	Sex	Adult	45,222	1 (income > 50k)	0 (75.2%)
2. Adult-race	Race	Adult	45,222	1 (income > 50k)	0 (75.2%)
3. Compas-sex	Sex	Compas	6,167	0 (no recidivism)	0 (54.5%)
4. Compas-race	Race	Compas	6,167	0 (no recidivism)	0 (54.5%)
5. German-sex	Sex	German	1,000	1 (good credit)	1 (70.0%)
6. German-age	Age	German	1,000	1 (good credit)	1 (70.0%)
7. Bank-age	Age	Bank	30,488	1 (subscriber)	0 (87.3%)
8. Mep-race	Race	Mep	15,830	1 (utilizer)	0 (82.8%)
9. Adult-sex-race	Sex, Race	Adult	45,222	1 (income > 50k)	0 (75.2%)
10. Compas-sex-race	Sex, Race	Compas	6,167	0 (no recidivism)	0 (54.5%)
11. German-sex-age	Sex, Age	German	1,000	1 (good credit)	1 (70.0%)

Method	Type	Venue/Journal	Description
Optimized Pre-processing (OP) [14]	Pre-processing	NeurIPS	Modify data features and labels.
Learning Fair Representation (LFR) [57]	Pre-processing	ICML	Obfuscating information about sensitive attributes
Reweighting (RW) [35]	Pre-processing	KAIS	Set different weights for samples in different groups.
Disparate Impact Remover (DIR) [26]	Pre-processing	SIGKDD	Modify data feature values.
Fairway [17]	Pre-processing	ESEC/FSE	Remove ambiguous data points.
Fair-SMOTE [16]	Pre-processing	ESEC/FSE	Remove ambiguous data points and synthesize new data points.
FairMask [48]	Pre-processing	TSE	Replace the sensitive attribute vector of testing data.
MAAT [19]	Pre-Post-processing	ESEC/FSE	Ensemble prediction of fairness model and performance model.
Prejudice Remover (PR) [37]	In-processing	ECML-PKDD	Add a fair regularization term to the learning objective.
Adversarial Debiasing (AD) [58]	In-processing	AAAI	Reduce the contribution of protected attributes to prediction.
Meta Fair Classifier (MFC) [15]	In-processing	FAT	Optimize classifier with fairness metrics.
CARE [50, 55]	Post-processing	ICSE	Using causality analysis to modify neurons weights.
Reject Option Classification (ROC) [36]	Post-processing	ICDM	Modify prediction near the threshold.
Equalized Odds Post-processing (EOP) [31]	Post-processing	NeurIPS	Modify predictions to make the Odds Difference equal.
Calibrated Equalized Odds Post-processing (CEO) [51]	Post-processing	NeurIPS	Modify predictions with calibrated probability.

实验设计　有效性验证实验

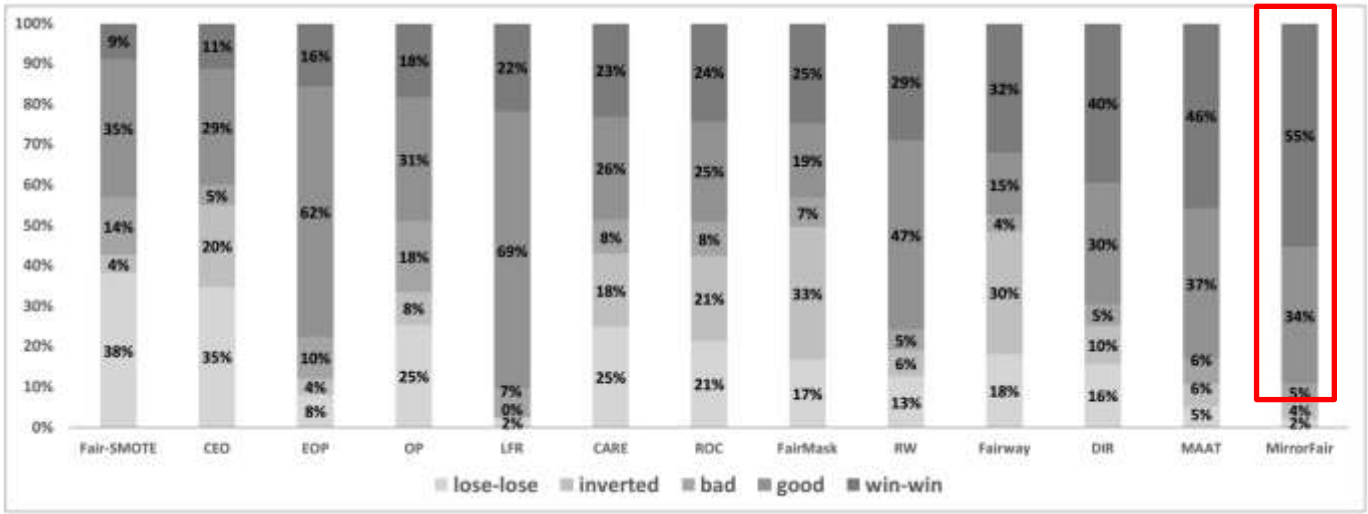
- RQ1：MirrorFair的有效性
- 实验分析
  - 大多数情况下，应用MirrorFair后，整体预测性能保持不变，偏差明显减少
  - 在某些场景中，MirrorFair改进了所有四个性能指标，并减少了所有三个偏差指标
  - 增强公平性但降低性能的情况很少

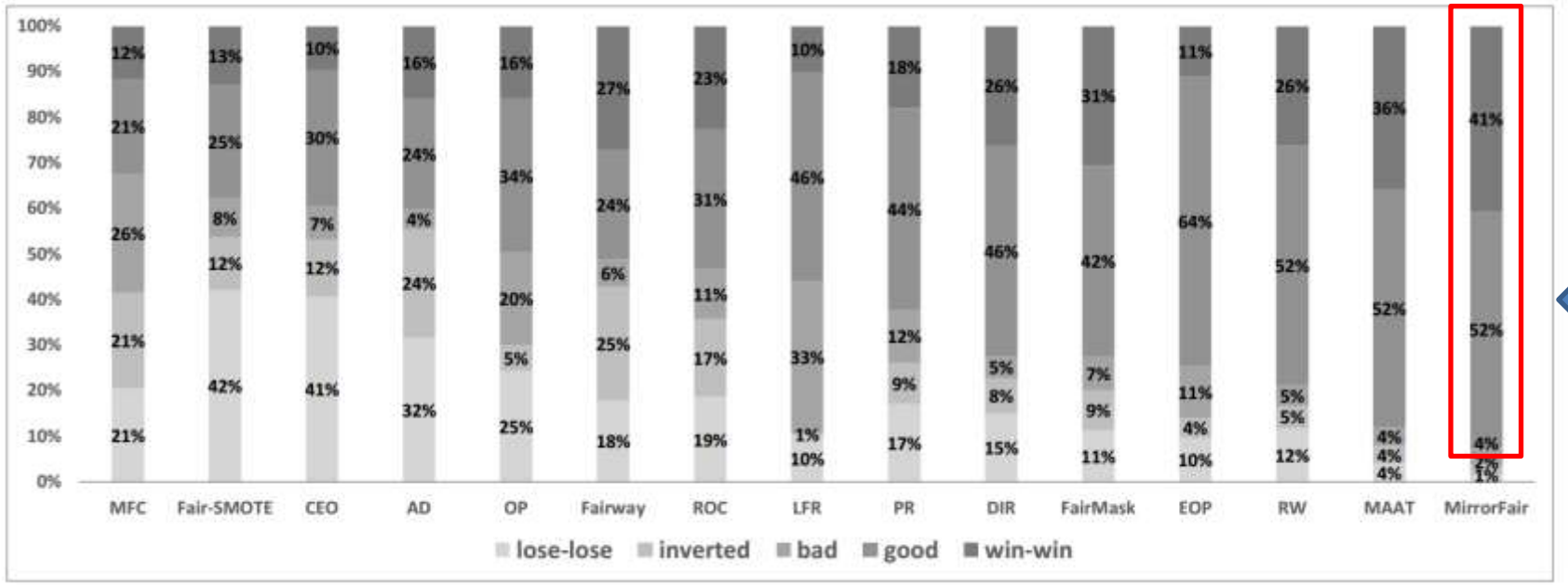
Task	Method	LR Accuracy (+)	LR Recall (+)	LR Precision (+)	LR F1-Score (+)	LR SPD (-)	LR AOD (-)	LR EOD (-)	SVM Accuracy (+)	SVM Recall (+)	SVM Precision (+)	SVM F1-Score (+)	SVM SPD (-)	SVM AOD (-)	SVM EOD (-)
Adult-Sex	Default	0.85	0.76	0.80	0.78	0.19	0.10	0.12	0.85	0.76	0.81	0.78	0.18	0.08	0.09
	MirrorFair	0.84	0.74	0.81	0.76	**0.11**	**0.04**	**0.05**	0.84	0.74	0.81	0.76	**0.11**	**0.05**	**0.07**
Adult-Race	Default	0.85	0.76	0.80	0.78	0.10	0.06	0.09	0.85	0.76	0.81	0.78	0.10	0.05	0.07
	MirrorFair	0.85	0.76	0.80	0.78	**0.07**	**0.02**	**0.02**	0.85	0.76	0.81	0.78	**0.07**	**0.02**	**0.02**
Compas-Sex	Default	0.67	0.66	0.67	0.66	0.28	0.25	0.20	0.66	0.66	0.66	0.66	0.26	0.24	0.18
	MirrorFair	0.67	0.65	0.67	0.65	**0.12**	**0.10**	**0.06**	0.66	0.65	0.67	0.64	**0.11**	**0.08**	**0.04**
Compas-Race	Default	0.67	0.66	0.67	0.66	0.18	0.16	0.11	0.66	0.66	0.66	0.66	0.17	0.15	0.10
	MirrorFair	0.66	0.65	0.67	0.65	**0.06**	**0.05**	**0.02**	0.66	0.64	0.67	0.64	**0.05**	**0.04**	**0.02**
German-Sex	Default	0.75	0.67	0.70	0.68	0.11	0.10	0.07	0.75	0.67	0.70	0.68	0.11	0.10	0.07
	MirrorFair	0.74	0.65	0.69	0.66	**0.05**	**0.08**	**0.04**	0.74	0.63	0.70	0.64	**0.05**	**0.08**	**0.04**
German-Age	Default	0.75	0.67	0.70	0.68	0.21	0.17	0.16	0.75	0.67	0.70	0.68	0.20	0.17	0.16
	MirrorFair	0.74	0.67	0.70	0.68	**0.05**	**0.07**	**0.05**	0.75	0.64	0.71	0.65	**0.05**	**0.10**	**0.05**
Bank-Age	Default	0.90	0.68	0.79	0.72	0.09	0.08	0.13	0.90	0.67	0.79	0.71	0.07	0.05	0.08
	MirrorFair	0.90	0.69	0.79	0.73	**0.05**	**0.03**	**0.04**	0.90	0.70	0.79	0.73	**0.05**	**0.03**	**0.04**
Mep-Race	Default	0.86	0.68	0.78	0.71	0.12	0.11	0.18	0.86	0.67	0.78	0.70	0.10	0.08	0.12
	MirrorFair	0.86	0.67	0.78	0.71	**0.08**	**0.05**	**0.07**	0.86	0.67	0.78	0.70	**0.07**	**0.03**	**0.05**

Task	Method	RF Accuracy (+)	RF Recall (+)	RF Precision (+)	RF F1-Score (+)	RF SPD (-)	RF AOD (-)	RF EOD (-)	DNN Accuracy (+)	DNN Recall (+)	DNN Precision (+)	DNN F1-Score (+)	DNN SPD (-)	DNN AOD (-)	DNN EOD (-)
Adult-Sex	Default	0.84	0.77	0.79	0.78	0.19	0.08	0.08	0.85	0.77	0.80	0.78	0.18	0.08	0.08
	MirrorFair	0.84	0.77	0.79	0.78	**0.16**	**0.04**	**0.02**	0.85	0.75	0.81	0.77	**0.13**	**0.04**	**0.04**
Adult-Race	Default	0.84	0.77	0.79	0.78	0.10	0.05	0.04	0.85	0.77	0.80	0.78	0.09	0.04	0.05
	MirrorFair	0.85	0.77	0.80	0.79	**0.06**	**0.03**	**0.04**	0.85	0.77	0.80	0.78	**0.07**	**0.02**	**0.02**
Compas-Sex	Default	0.65	0.64	0.64	0.64	0.17	0.14	0.12	0.65	0.65	0.65	0.65	0.19	0.16	0.13
	MirrorFair	0.66	0.64	0.66	0.64	**0.03**	**0.03**	**0.03**	0.66	0.65	0.65	0.65	**0.14**	**0.11**	**0.09**
Compas-Race	Default	0.65	0.64	0.64	0.64	0.14	0.12	0.09	0.65	0.65	0.65	0.65	0.16	0.14	0.10
	MirrorFair	0.65	0.64	0.65	0.64	**0.04**	**0.02**	**0.02**	0.65	0.64	0.65	0.64	**0.06**	**0.04**	**0.03**
German-Sex	Default	0.76	0.66	0.73	0.67	0.07	0.07	0.04	0.73	0.66	0.68	0.67	0.09	0.09	0.06
	MirrorFair	0.76	0.64	0.73	0.66	**0.04**	**0.05**	**0.03**	0.74	0.65	0.70	0.66	**0.07**	**0.07**	**0.05**
German-Age	Default	0.76	0.66	0.73	0.67	0.13	0.11	0.07	0.73	0.66	0.68	0.67	0.19	0.16	0.15
	MirrorFair	0.76	0.65	0.73	0.66	**0.05**	**0.07**	**0.04**	0.74	0.65	0.69	0.66	**0.08**	**0.08**	**0.06**
Bank-Age	Default	0.90	0.72	0.79	0.75	0.08	0.05	0.06	0.90	0.75	0.77	0.76	0.09	0.06	0.07
	MirrorFair	0.90	0.73	0.79	0.75	**0.06**	**0.04**	**0.05**	0.90	0.77	0.78	0.77	0.09	**0.05**	**0.06**
Mep-Race	Default	0.86	0.67	0.76	0.70	0.09	0.07	0.09	0.85	0.67	0.74	0.69	0.11	0.09	0.13
	MirrorFair	0.86	0.68	0.75	0.71	**0.06**	**0.02**	**0.02**	0.86	0.68	0.75	0.70	**0.08**	**0.05**	**0.06**

实验设计　对比实验

- RQ1：MirrorFair的有效性
- 实验结论
  - MirrorFair实现了最高的“双赢”比例，在41%的测试用例中同时增强了模型性能和公平性





DNN模型上的效果

实验设计　对比实验

- **RQ1：MirrorFair的有效性**
- **实验结论**
  - **MirrorFair以最小的性能影响将总体偏差减少了50%，而最先进的方法仅将其减少了45%（EOP）和41%（MAAT）**

Method	Accuracy	Recall	Precision	F1-Score	Overall Performance	SPD	AOD	EOD	Overall Bias
FairMask	-0.23%	-0.94%	-1.10%	-1.51%	-1.48%	-15.53%	-19.18%	-20.24%	-18.32%
DIR	-0.39%	-0.75%	-0.37%	-0.94%	-0.61%	-18.94%	-17.54%	-18.95%	-18.47%
RW	-0.32%	-0.78%	-0.37%	-0.71%	-0.54%	-49.24%	-37.32%	-26.25%	-37.60%
MAAT	0.03%	-1.33%	0.83%	-1.01%	-0.37%	-37.02%	-43.44%	-42.51%	-40.99%
EOP	-2.39%	-3.00%	-3.55%	-3.61%	-3.14%	-44.54%	-47.42%	-44.33%	-45.43%
MirrorFair	0.03%	-0.46%	0.47%	-0.54%	**-0.12%**	-44.69%	-51.93%	-55.43%	**-50.68%**

Method	VS Default	VS FairMask	VS DIR	VS EOP	VS RW	VS MAAT	VS MirrorFair
Default	0%	9%	13%	4%	7%	0%	0%
FairMask	61%	0%	24%	20%	18%	7%	0%
DIR	71%	55%	0%	22%	27%	20%	10%
EOP	77%	65%	54%	0%	29%	44%	42%
RW	80%	54%	44%	21%	0%	40%	34%
MAAT	92%	65%	51%	31%	38%	0%	16%
MirrorFair	99%	92%	67%	34%	45%	44%	0%

实验设计　对比实验

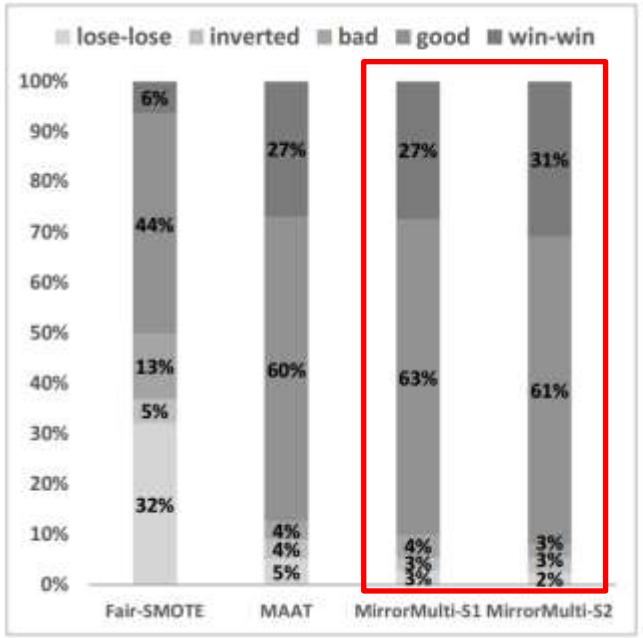
- **RQ2：MirrorFair的普适性**
- **实验结论**
  - **最先进的方法相比，MirrorFair展示了更高的效能和更窄的效能波动范围，展示了其在不同算法和任务中的普适性**

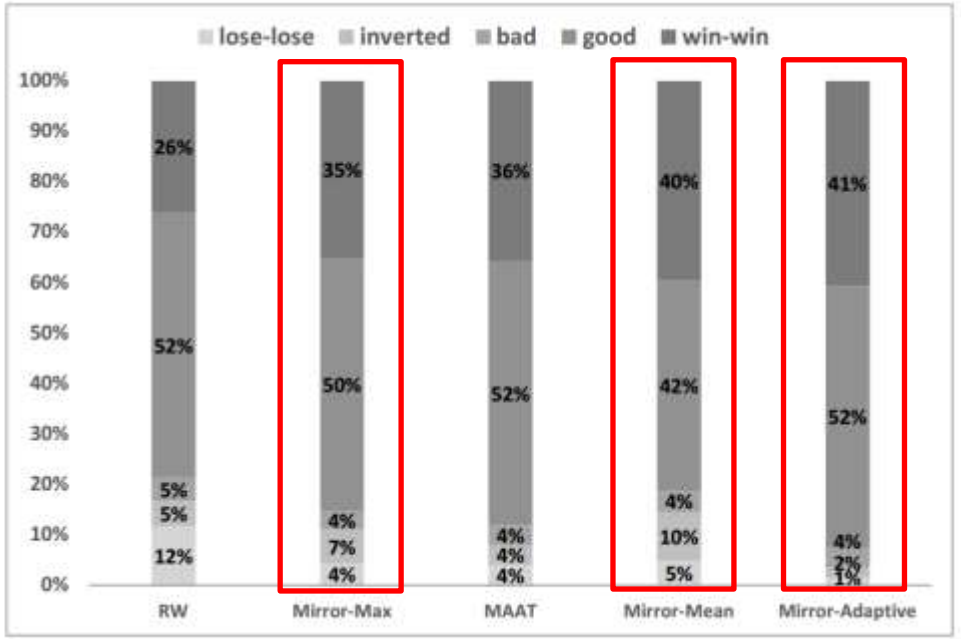
在不同的算法和任务中超过基线的公平性比例

Method	Algoritm				Task							
	LR	RF	SVM	DNN	Adult-Sex	Adult-Race	Compas-Sex	Compas-Race	German-Sex	German-Age	Bank-Age	Mep-Race
FairMask	76.52%	65.08%	75.97%	43.27%	63.80%	66.73%	78.53%	42.93%	70.33%	66.97%	57.77%	74.60%
DIR	77.02%	70.10%	72.58%	69.70%	54.23%	89.10%	**100.00%**	97.20%	64.87%	76.23%	7.43%	89.73%
RW	89.52%	65.80%	83.22%	75.95%	51.40%	73.17%	96.20%	90.37%	72.30%	77.27%	78.17%	90.10%
EOP	84.15%	53.05%	83.45%	77.70%	79.13%	72.50%	95.20%	90.57%	68.47%	52.50%	61.90%	76.43%
MAAT	93.35%	85.53%	90.23%	82.97%	90.53%	93.40%	98.77%	97.00%	71.70%	73.07%	82.63%	**97.07%**
MirrorFair	**95.38%**	**89.23%**	**95.10%**	**90.35%**	**96.60%**	**96.43%**	98.13%	**99.93%**	**83.03%**	**84.03%**	**85.90%**	96.07%

实验设计　对比/消融实验

- RQ3：MirrorFair面对多敏感属性时的有效性
- 实验结论
  - MirrorFair的两个变体（MirrorMulti-S1和MirrorMulti-S2）在保护多个敏感属性方面都比现有方法更有效
  - 自适应的集成策略更加有效





实验设计　消融实验

- RQ4：MirrorFair反事实数据集和自适应策略的有效性
- 实验分析
  - 镜像处理对随机森林（RF）和DNN有不规则的影响，可能是因为随机森林和DNN是更复杂的模型，能够捕捉特征和目标变量之间的非线性关系
  - 相反，逻辑回归（LR）和SVM在如何使用特征进行预测方面更直接，通常依赖于线性边界，因此敏感特征的变化具有更可预测的结果

Adult-Sex-LR			Adult-Sex-SVM			Adult-Sex-RF			Adult-Sex-DNN		
$P_{def}(1,0)$	$P_{mir}(1,0)$	DIF	$P_{def}(1,0)$	$P_{mir}(1,0)$	DIF	$P_{def}(1,0)$	$P_{mir}(1,0)$	DIF	$P_{def}(1,0)$	$P_{mir}(1,0)$	DIF
0.34	0.50	-0.17	0.36	0.50	-0.14	0.60	0.42	0.17	0.30	0.52	-0.22
0.44	0.62	-0.17	0.48	0.68	-0.20	0.16	0.72	-0.56	0.31	0.63	-0.31
0.36	0.53	-0.17	0.46	0.66	-0.20	0.52	0.43	0.09	0.46	0.58	-0.12
0.45	0.60	-0.16	0.48	0.68	-0.20	0.10	0.67	-0.57	0.41	0.75	-0.35
0.39	0.56	-0.17	0.43	0.61	-0.18	0.28	0.76	-0.49	0.47	0.79	-0.32
$P_{def}(1,1)$	$P_{mir}(1,1)$	DIF	$P_{def}(1,1)$	$P_{mir}(1,1)$	DIF	$P_{def}(1,1)$	$P_{mir}(1,1)$	DIF	$P_{def}(1,1)$	$P_{mir}(1,1)$	DIF
0.59	0.42	0.17	0.62	0.44	0.18	0.38	0.50	-0.12	0.52	0.43	0.09
0.64	0.47	0.17	0.56	0.35	0.21	0.55	0.46	0.10	0.54	0.20	0.35
0.55	0.38	0.17	0.52	0.30	0.22	0.53	0.38	0.15	0.72	0.35	0.37
0.55	0.38	0.17	0.52	0.30	0.22	0.47	0.53	-0.06	0.70	0.43	0.27
0.59	0.41	0.17	0.61	0.43	0.18	0.56	0.47	0.09	0.56	0.24	0.31
Compas-Race-LR			**Compas-Race-SVM**			**Compas-Race-RF**			**Compas-Race-DNN**		
$P_{def}(1,0)$	$P_{mir}(1,0)$	DIF	$P_{def}(1,0)$	$P_{mir}(1,0)$	DIF	$P_{def}(1,0)$	$P_{mir}(1,0)$	DIF	$P_{def}(1,0)$	$P_{mir}(1,0)$	DIF
0.50	0.50	0.00	0.55	0.54	0.01	0.29	0.52	-0.23	0.62	0.42	0.20
0.50	0.50	0.00	0.50	0.50	0.00	0.55	0.45	0.10	0.69	0.32	0.37
0.50	0.50	0.00	0.48	0.47	0.01	0.97	0.16	0.81	0.43	0.53	-0.10
0.50	0.50	0.00	0.47	0.45	0.01	0.42	0.65	-0.23	0.38	0.52	-0.14
0.50	0.50	0.00	0.49	0.48	0.01	0.27	0.87	-0.60	0.33	0.64	-0.32
$P_{def}(1,1)$	$P_{mir}(1,1)$	DIF	$P_{def}(1,1)$	$P_{mir}(1,1)$	DIF	$P_{def}(1,1)$	$P_{mir}(1,1)$	DIF	$P_{def}(1,1)$	$P_{mir}(1,1)$	DIF
0.45	0.45	0.00	0.47	0.47	0.01	0.50	0.98	-0.48	0.50	0.61	-0.11
0.51	0.51	0.00	0.51	0.52	0.00	0.45	0.99	-0.54	0.28	0.54	-0.26
0.49	0.49	0.00	0.49	0.49	0.00	0.48	0.61	-0.13	0.67	0.42	0.24
0.53	0.52	0.00	0.52	0.53	0.00	0.45	0.58	-0.13	0.37	0.61	-0.24
0.50	0.50	0.00	0.51	0.51	0.00	0.32	0.69	-0.37	0.46	0.66	-0.19

Decision Task	LR			RF			SVM			DNN		
	$DIF \in N^{\delta}(c)$	$Mean_{DIF}$	Type	$DIF \in N^{\delta}(c)$	$Mean_{DIF}$	Type	$DIF \in N^{\delta}(c)$	$Mean_{DIF}$	Type	$DIF \in N^{\delta}(c)$	$Mean_{DIF}$	Type
Adult-Sex	✓	0.17	Regular	×	-	Irregular	✓	0.18	Regular	×	-	Irregular
Adult-Race	✓	0.03	Regular	×	-	Irregular	✓	0.04	Regular	×	-	Irregular
Compas-Sex	✓	0.08	Regular	×	-	Irregular	✓	0.06	Regular	×	-	Irregular
Compas-Race	✓	0.00	Insensitive	×	-	Irregular	✓	0.01	Insensitive	×	-	Irregular
German-Sex	✓	0.02	Regular	×	-	Irregular	✓	0.02	Regular	×	-	Irregular
German-Age	✓	0.02	Regular	×	-	Irregular	✓	0.01	Insensitive	×	-	Irregular
Bank-Age	✓	0.08	Regular	×	-	Irregular	✓	0.13	Regular	×	-	Irregular
Mep-Race	✓	0.11	Regular	×	-	Irregular	✓	0.09	Regular	×	-	Irregular

特点总结与未来展望

算法总结 MirrorFair

- 算法创新
  - 采用集成学习的思想
- 算法优势
  - 允许多个模型相互补充，获得更好的结果
  - 自适应策略更好地减轻模型偏差
- 算法不足
  - MirrorFair将敏感属性简单转换为二进制类别（例如，“白人”和“非白人”），可能会模糊二元类别亚组内的区别（例如，“非白人”中的“亚洲人”和“爱斯基摩人”），从而引入新的偏见
- 未来改进
  - 引入新的公平性度量，计算各组公平性得分的总体标准差
  - 允许测量具有非二元敏感属性的更多粒度偏差

参考文献

[1]Xiao Y, Zhang J M, Liu Y, et al. MirrorFair: Fixing fairness bugs in machine learning software via counterfactual predictions[J]. Proceedings of the ACM on Software Engineering, 2024, 1(FSE): 2121-2143.

[2] Zheng W, Lin L, Wu X, et al. An empirical study on correlations between deep neural network fairness and neuron coverage criteria[J]. IEEE Transactions on Software Engineering, 2024, 50(3): 391-412.

知人者智，自知者明。胜人者有力，自胜者强。知足者富。强行者有志。不失其所者久。死而不亡者，寿。

谢谢！