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

博士研究生 刘洧光

2024年09月28日

问题回溯

- 总结反思
  - 标题起名不太合理
  - 讲述的时候按顺序讲，少留伏笔

- 相关内容
  - 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》

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
  - FairNeuron
  - Faire
- 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测试相关知识
  - 了解两种深度学习模型的公平性修复方法

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测试）
  - 人工智能模型：泛指人工智能领域中的模型，包括机器学习和深度学习模型
  - 公平性：模型的属性，是模型重要的非功能性需求之一

- 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
  - 发现模型的公平性问题、生成歧视样本并指导模型的修复与再训练

研究背景

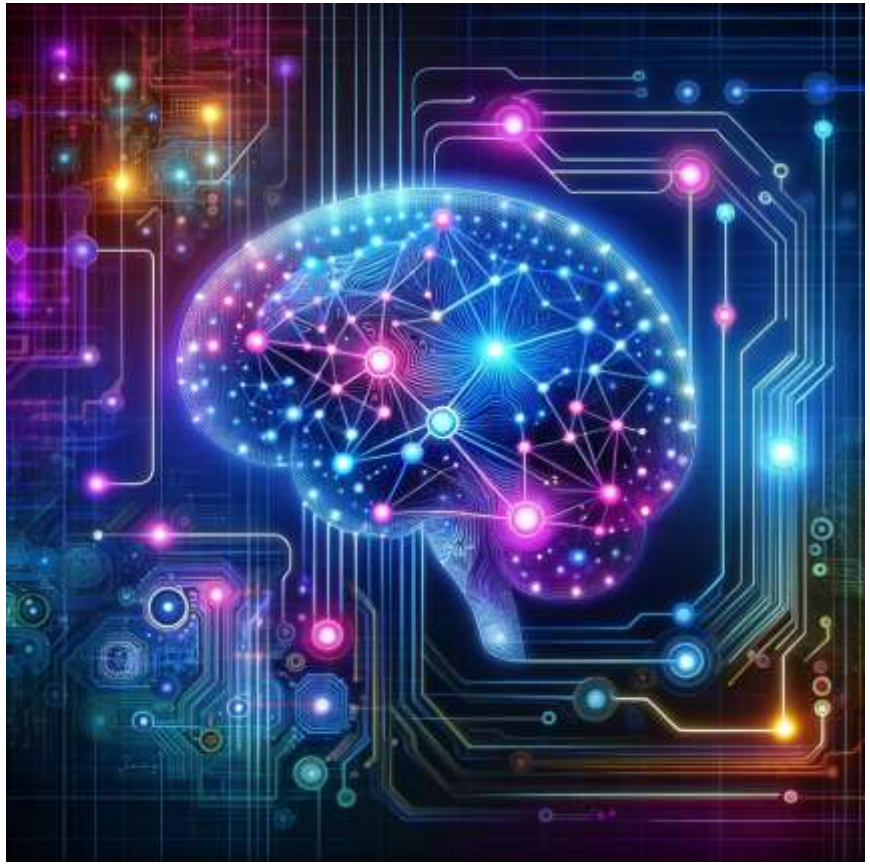
- 研究背景
  - 人工智能技术发展迅速，不仅在图像领域，在决策系统等领域也发挥了重要作用
  - 用于模型训练的数据集中含有显示或者隐式的敏感属性（性别、种族）
  - 人工智能模型往往会利用敏感属性的特征做出决策
  - 这将导致人工智能模型在公平性方面出现偏差，产生严重的舆论影响和社会问题

2020年，ExamSoft 远程考试的人脸识别系统被发现对有色人种识别成功率更低

2016年一项研究发现，词嵌入时“he”会与“genius”这类褒义词更加接近，而“ she”与“sassy”更接近

研究意义

- 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
  - 自动化生成歧视样本、进行样本优先级排序
    - 高效且自动化生成歧视样本，避免手工标记或检查，提升模型的公平性修复效率
    - 发现导致模型歧视行为的样本分布
  - 模型公平性增强
    - 对有偏模型进行再训练，以提高模型公平性



发现模型缺陷，自动化生成歧视样本，提升模型的公平性

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

2018 Udeshi等人针对机器学习模型，提出一种会自动发现突出违反公平性的歧视性输入的方法Aeqitas，其核心是三种新颖的策略，目的是发现违反公平性的行为

2020 Zhang等人提出了一种可扩展的方法来搜索DNN的个体歧视性实例，只采用了梯度计算和聚类等轻量级过程，这使其具有更大的可扩展性

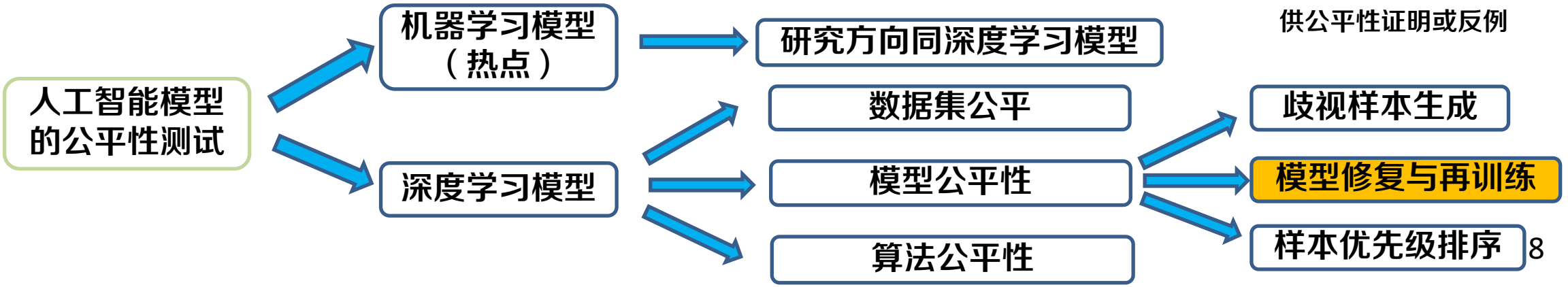
2021 Zhang等人提出了一个有效发现个人公平违规的框架EIDIG。结合了快速生成一组多样化的判别种子的全局生成阶段和梯度指导下在这些种子周围生成尽可能多的个体判别实例的局部生成阶段

2022 Tao等人提出了一种新的模型修复技术 RULER，即在生成用于模型修复的测试用例时区分敏感和非敏感属性。将生成的样本用于训练，以提高 DNN 的公平性

2022 Gao等人提出了FairNeuron，一个DNN模型自动修复工具，以减轻公平性问题，平衡准确性和公平性之间的权衡，而无需引入另一个模型

2022 Zhang等人提出了一种基于因果关系分析的自适应选择公平性改进方法。根据负责不公平的神经元和属性如何在输入属性和隐藏神经元之间分布来选择方法

2023 Li等人提出Fairify ，利用对生产中的模型的白盒访问，然后应用基于修剪的形式分析。对每个分区的神经网络进行修剪，以提供公平性证明或反例



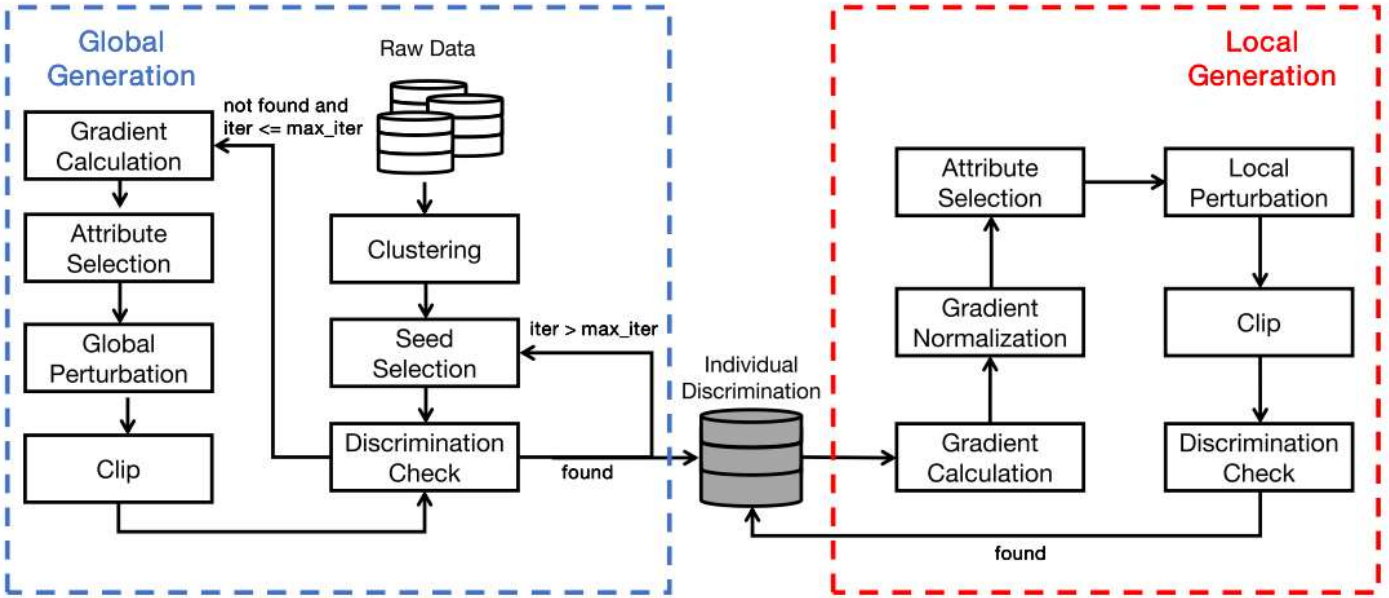
知识基础　公平性的定义

- 个体公平（Individual fairness）
  - 无意识公平：系统可以通过避免在决策过程中明确使用敏感属性来实现公平结果
  - 有意识公平：系统需要为相似的个体产生相似的结果
  - 反事实公平：个体的预测在现实世界中应该保持不变
  - 因果公平：捕获敏感属性与预测结果之间的因果关系
- 群体公平（Group fairness）
  - 人口平等：要求在不同人口群体中出现有利结果的概率相同
  - 均等几率：要求特权群体和非特权群体具有相等的真阳性率和假阳性率
  - 平等机会：要求特权群体和非特权群体的真阳性率相等

个体公平要求为相似的个人产生相似的预测结果，
群体公平以相似的方式对待不同的人口统计群体

通用的歧视样本生成方法

- 阶段一：全局搜索（追求多样性）
  - 基于梯度/保护特征翻转等方式在全数据集范围生成歧视样本
    - 生成标准：仅改变保护属性，模型的预测结果发生改变
- 阶段二：局部生成（追求数量）
  - 采用KNN等算法在阶段一所生成的歧视样本周围大量寻找相似歧视样本
    - 生成标准：在给定范围内相较原歧视样本不改变模型的预测结果



研究现状

- 歧视样本生成方法
  - 主流方法都使用“两阶段”生成法，新的方法跳不出这个框架，只能修改两阶段中的生成细节，且这样生成费时费力，存在大量相似的歧视样本，对模型的公平性修复意义不大
- 模型修复方法
  - 通过还原公平数据集对模型进行重新训练，避免使用生成的歧视样本进行修复
  - 通过对模型添加隐藏层/Dropout层的方法进行再训练
  - 普遍只适用于表格数据集，对于图像数据集效果欠佳

力求轻量化修复，不过多依靠生成的歧视样本

【 ICSE 】

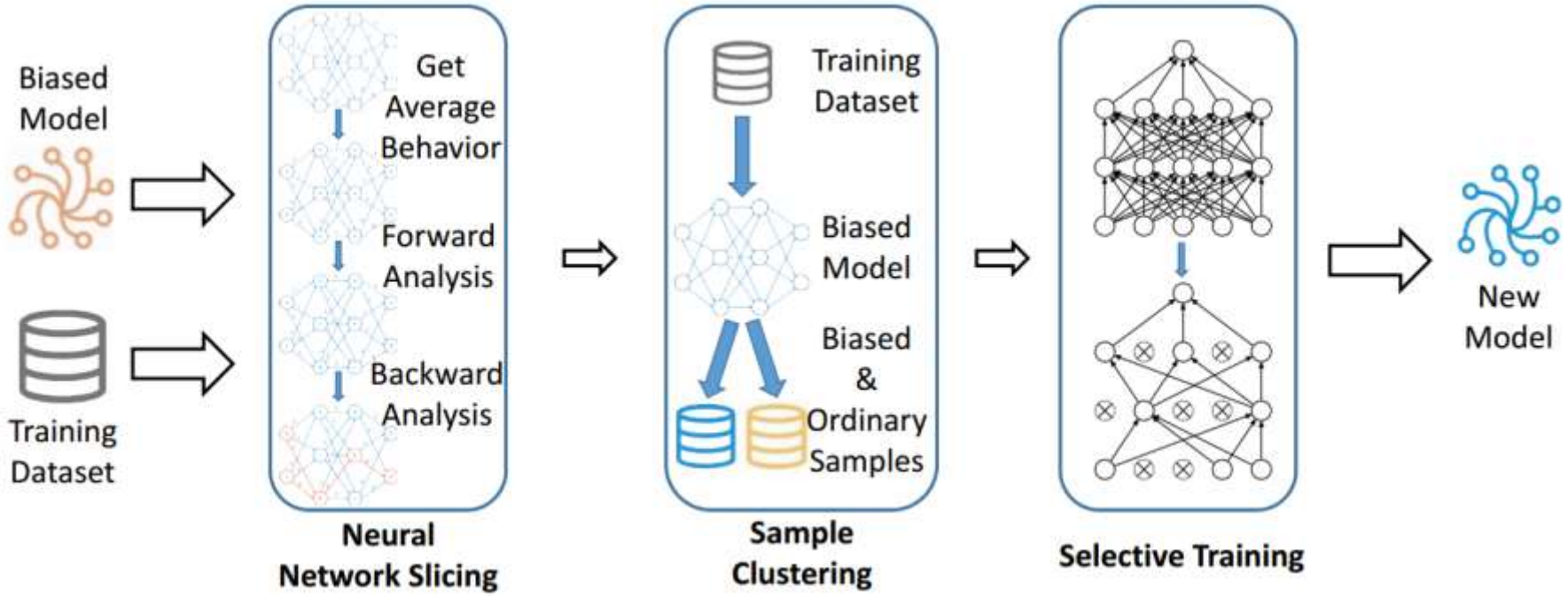
FairNeuron: improving deep neural network fairness with adversary games on selective neurons

FairNeuron TIPO

T	目标	不额外生成歧视样本并提高模型的公平性
I	输入	有偏模型*1个，训练集数据全集
P	处理	1. 对模型进行神经网络切片 2. 样本聚类，得到有偏数据样本和良性样本 3. 对模型进行再训练，对不同样本执行不同训练策略
O	输出	修复后提升公平性的新模型

P	问题	现有训练对抗网络的方法困难且难以收敛
C	条件	白盒模型、能访问训练数据
D	难点	1. 如何以轻量级的方法修复模型 2. 如何找到正确性和公平性之间的平衡
L	水平	ICSE 2022（CCF A）

- 算法原理图
  - 步骤1：神经网络切片（识别冲突路径，即包含导致偏差预测的特征的路径）
  - 步骤2：样本聚类（分离良性样本和有偏样本）
  - 步骤3：选择性训练（对正常样本进行普通训练，对有偏样本进行dropout训练）

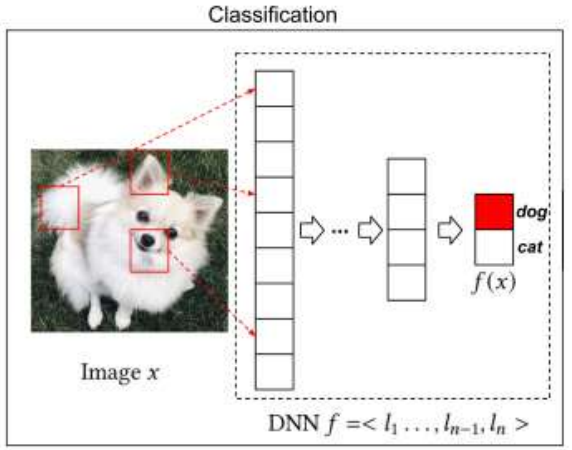


- 逐层相关性传播（Layer-Wise Relevance Propagation，LRP）
  - 算法思想：一种通过计算输入的关键特征来解释决策的有效方法
  - 算法原理
    - 正向计算输出

输入样本

置信度矩阵

$$f(x) = argmax_y O_n$$



    - 反向计算相关性

输入的维度

分类器f对于输入样本x在类y上的输出

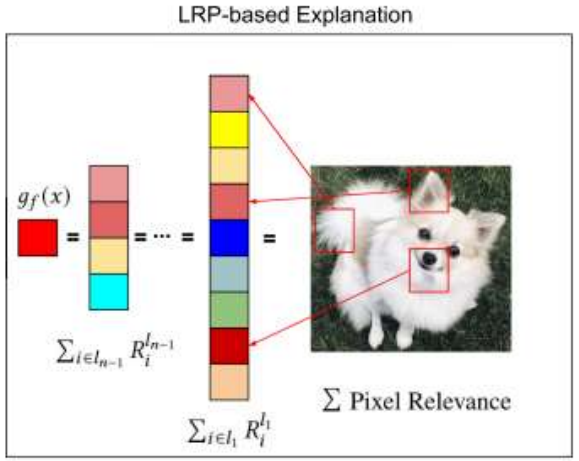
输入层l_1(如图像中的像素)中维度i的相关性

神经元i在第l_{n-1}层的相关性

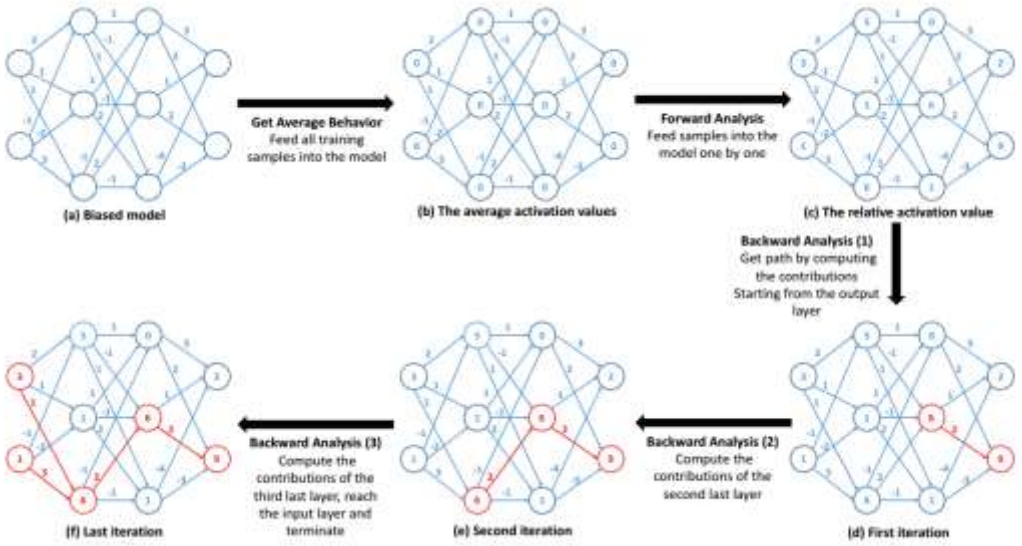
$$g_{f(x)} = \sum_{i=1}^{d} R_i^{l_1}$$

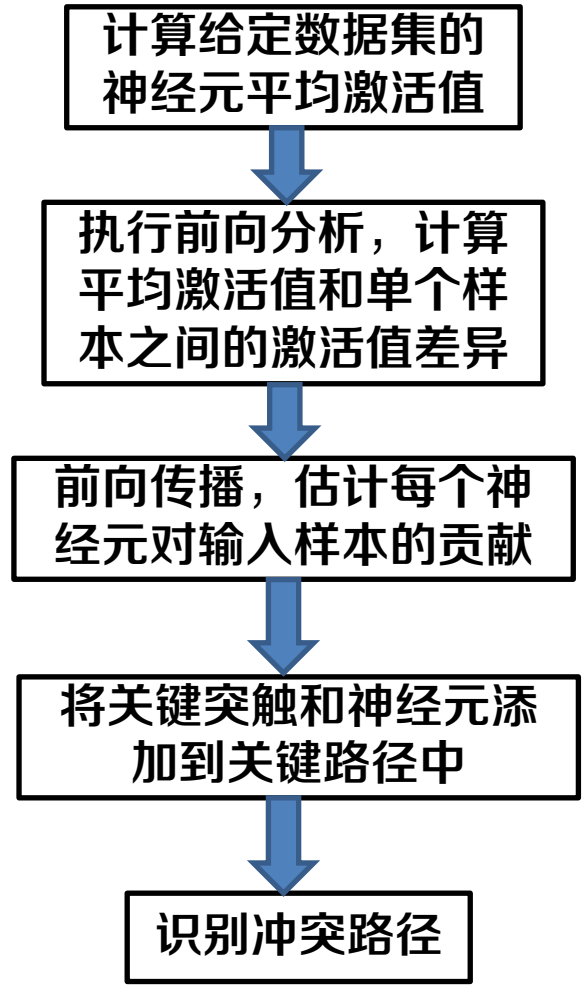
$$g_{f(x)} = R_y^{(l_n)} = \sum_{i \in (l_{n-1})} R_i^{l_{n-1}}$$

$$R_j^{(l_{m+1})} = \sum_{j \in (l_{m+1})} R_{i \leftarrow j}^{(l_m, l_{m+1})}$$



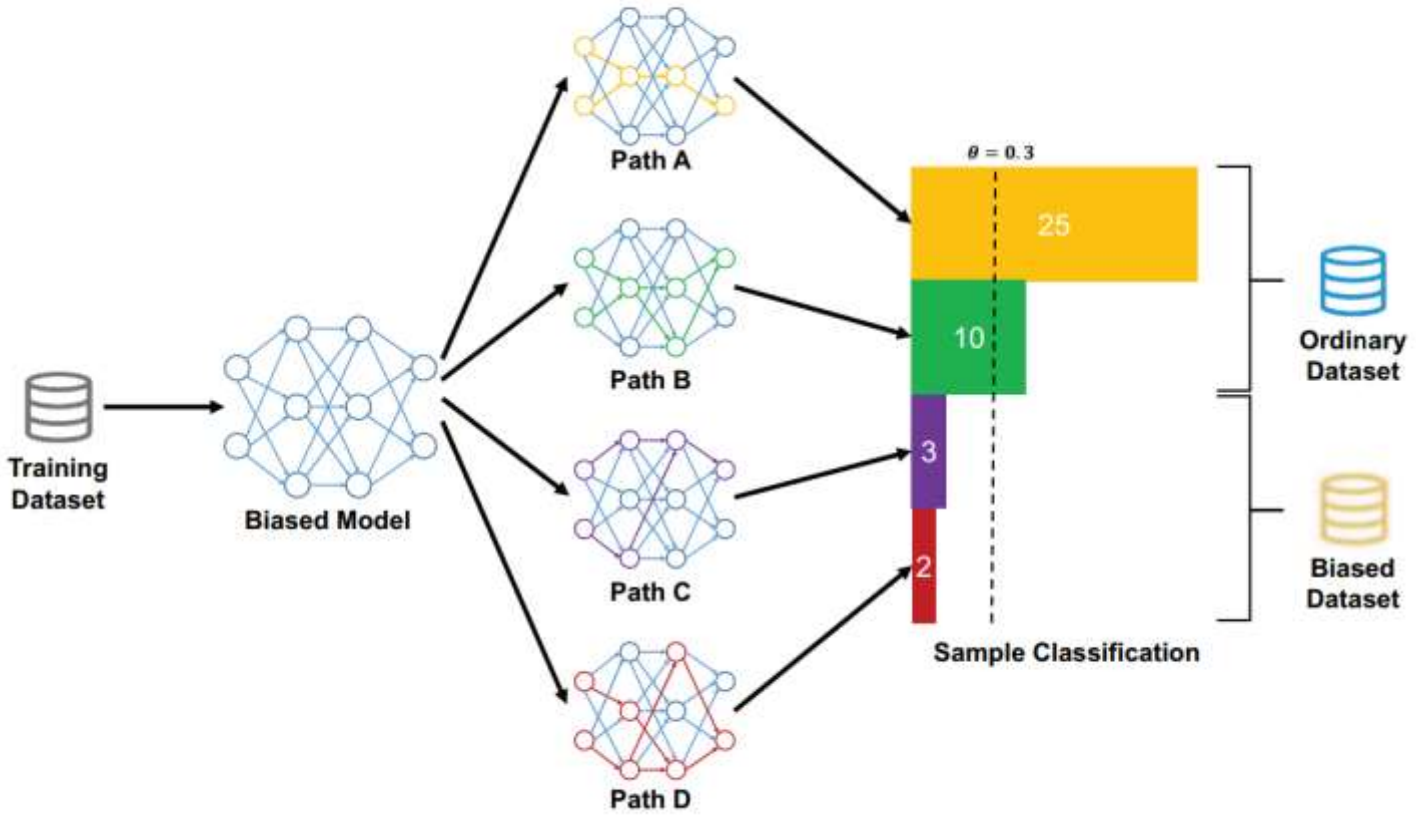
- 几个直观经验
  - 模型使用良性特征集对良性样本进行预测，使用有偏特征集对有偏样本进行预测
  - 有偏路径/神经元只占整个神经网络的一小部分（否则，网络会对大量有偏差的特征进行预测，导致准确率较低）
  - 如果某条路径的激活频率小于某一标准的百分比，则可以认为是一条有偏路径

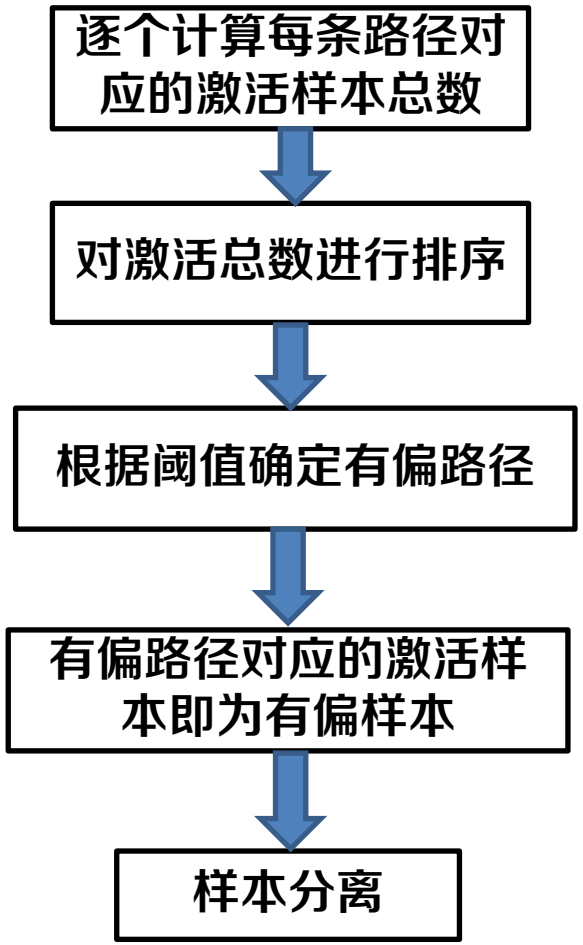




创新分析　样本聚类

- 样本聚类的目的
  - 衡量输入样本对公平性的影响，从而分离样本





样本聚类并非使用了聚类算法

实验设计　数据资源

- 数据与模型资源

数据集名称	UCI Adult Census	COMPAS	German Credit
数据集用途	人口普查	再犯罪风险预测	评估个人信用
敏感属性	性别	种族	性别
样本数量	32561	10000+	600
特征总数	9	9	20
模型准确率	83.9%	73.4%	62.1%

- 对比方法：
  - FAD [in-processing] (2019)
  - Ethical Adversaries [in-processing] (2021)
  - Reweighing [pre-processing] (2012)
  - ROC [post-processing] (2012)

实验设计　评价指标

- 评价指标：
  - Demographic parity (DP)【人口平等 群体公平】

$$DP = \left| P\left(\hat{Y} = 1 \mid S = 0\right) - P\left(\hat{Y} = 1 \mid S = 1\right) \right|$$

S：敏感属性，
$S = 0$：弱势群体

$\hat{Y}$：真实标签，
$\hat{Y} = 1$：阳性结果

  - Demographic parity ratio (DPR)【人口平等比例 群体公平】

$$DPR = \frac{P\left(\hat{Y} = 1 \mid S = 1\right)}{P\left(\hat{Y} = 1 \mid S = 0\right)}$$

  - Equal opportunity (EO)【平等机会 群体公平】

$$EO = \left| P\left(\hat{Y} = 1 \mid S = 0, Y = 1\right) - P\left(\hat{Y} = 1 \mid S = 1, Y = 1\right) \right|$$

假阳性率FPR　　真假阳性率TPR

理想情况下：DP=0，EO=0，DPR=1，模型公平

实验设计　对比实验

- RQ1： FairNeuron的有效性
- 实验结论
  - FairNeuron可以有效地修正在不同数据集上训练的所有模型的公平性偏差
  - FairNeuron在Credit上的EO和DP结果并不令人满意，是因为神经元切片功能不全
  - FairNeuron在成功修复公平性问题后对模型正确性的影响很小，甚至具有通过修复公平性问题来提高准确率的优势

Dataset	Model	Acc	DP	EO	DPR
Census	Naive model	0.839	0.079	0.102	0.609
	ROC	0.597	0.044	0.051	0.773
	Reweighing	0.719	0.059	0.0141	1.497
	FAD	0.612	0.059	0.061	0.518
	Ethical Adversaries	0.814	0.031	0.179	0.784
	FairNeuron	**0.832**	**0.020**	**0.031**	**0.869**
Credit	Naive model	0.734	0.048	0.142	0.407
	ROC	0.646	0.041	0.073	1.273
	Reweighing	0.632	0.067	0.066	0.828
	FAD	0.710	**0.000**	**0.000**	inf
	Ethical Adversaries	0.715	0.041	0.031	2.442
	FairNeuron	**0.744**	0.047	0.112	**0.834**
COMPAS	Naive model	0.621	0.341	0.095	1.860
	ROC	0.618	0.083	0.069	0.890
	Reweighing	0.671	0.193	0.176	1.406
	FAD	0.567	0.057	0.114	0.926
	Ethical Adversaries	0.759	0.095	0.095	1.203
	FairNeuron	**0.799**	**0.013**	**0.058**	**1.021**

实验设计　效率实验

- RQ2：FairNeuron的效率

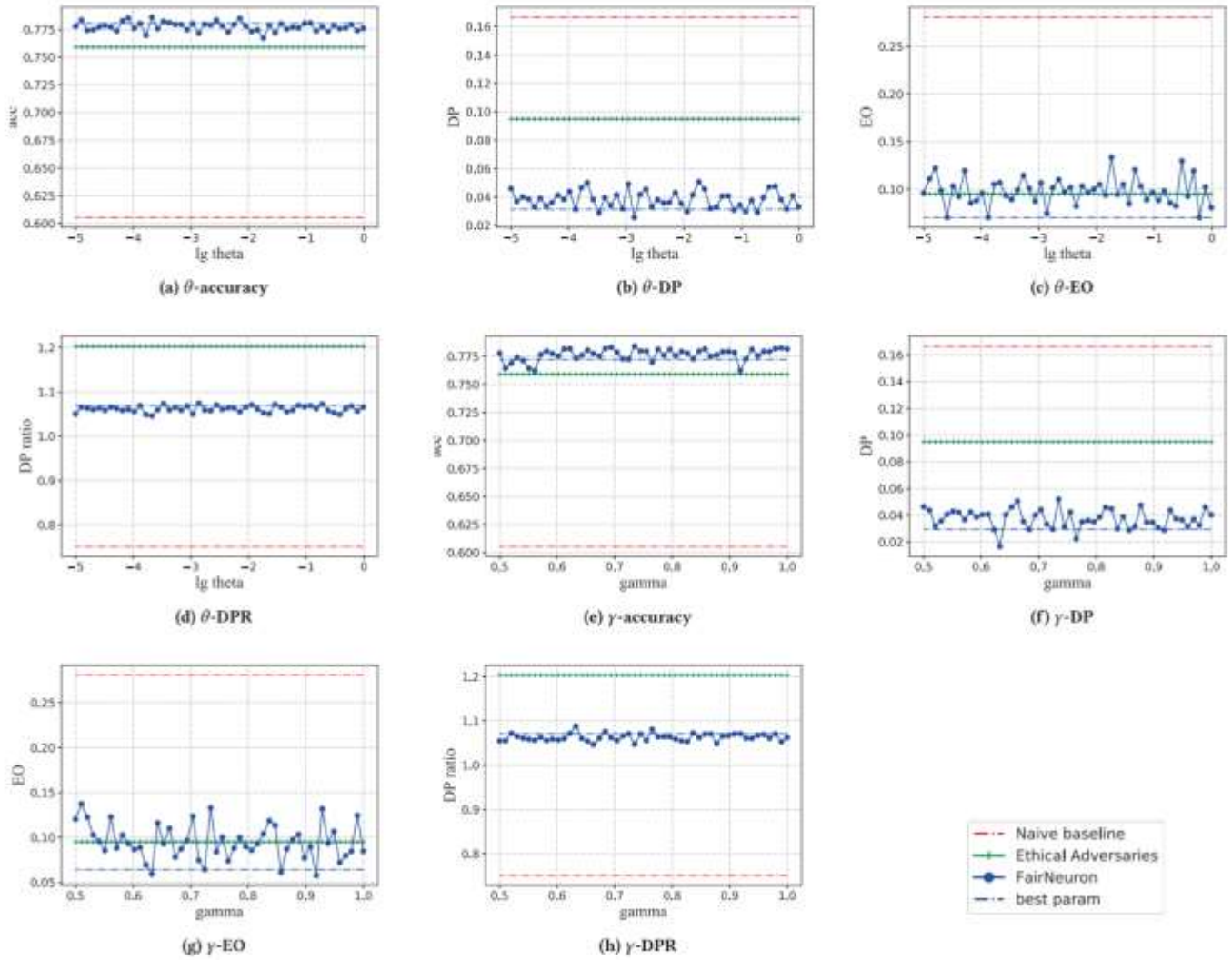
Dataset	Naive	EA (/iteration)	FairNeuron (/trial)
Census	115.74s	1439.96s	254.41s
Credit	3.07s	33.24s	31.49s
COMPAS	11.92s	81.93s	44.31s

Dataset	Para selection	Slicing	Clustering	Training
Census	115.41s	25.37s	43.70s	74.37s
Credit	30.98s	0.20s	6.73e-4s	0.30s
COMPAS	40.76s	2.09s	0.06s	1.40s

- 实验结论
  - 对于普通训练，运行时开销全部来自训练过程，但对于FairNeuron，超参数调优占总时间使用的比例更大
  - 如果FairNeuron尝试更多的次数，平均时间将会减少，因为超参数调整只进行一次

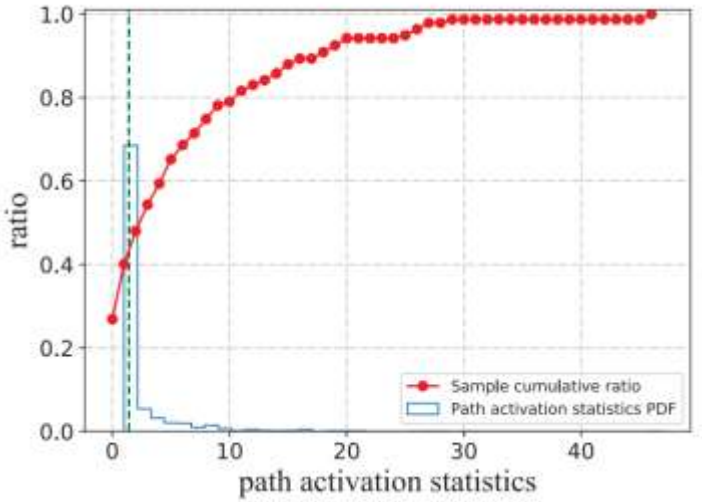
实验设计　超参数实验

- **RQ3：可配置超参数的影响**
- γ：表示神经元激活的阈值，随着其值的减小，路径中包含更多的神经元和突触，导致路径更加复杂
- θ：表示神经元网络切片的阈值。参数越低，有偏路径越少
- **实验结论**
  - **FairNeuron对超参数不敏感（除了EO）**



实验设计　消融实验

- 神经元网络切片的有效性
  - 大多数非零路径都集中在1附近，但是它们对应的样本比例并不高。这些路径是FairNeuron检测到的异常路径



- 样本聚类的有效性
  - 对比随机聚类方法，准确率提高了6.68%，公平性性能也有很大提高

Method	Acc	DP	DPR	EO
Random	0.749	0.325	1.89	0.159
Ours	0.799	0.013	1.02	0.058

- 选择性训练的有效性
  - FairNeuron的选择性训练可以达到较高的准确率和公平性

Training approach	Acc	DP	DPR	EO
Ordinary	0.575	0.733	0.183	0.683
Dropout	0.621	0.341	1.860	0.095
Selective	0.799	0.013	1.021	0.058

算法总结 FairNeuron

- 算法流程
  - 对模型进行神经网络切片，使用原数据集样本，通过LRP逐层相关性前向传播识别冲突路径关键
  - 通过样本聚类，分离正常样本和有偏样本
  - 使用选择性训练对模型进行修复与再训练
- 算法优势
  - 轻量级、高效（不引入对手模型）
  - 无需额外生成歧视样本进行再训练
- 算法不足
  - 对于CNN模型， FairNeuron只能在最后一个全连接层上执行，且性能并不理想
  - 如实验结果所示，对于小数据集无法很好地进行神经网络切片

Faire

【 TOSEM 】

Faire: Repairing Fairness of Neural Networks via Neuron Condition Synthesis

BET TIPO

T	目标	在不显式删除受保护属性的同时减少其对公平性的影响
I	输入	有偏模型*1个，训练集全集
P	处理	1.基于原模型训练保护特征分类器 2.对数据集的特征进行神经元层面的分析 3.通过添加隐藏层继续训练以提高公平性
O	输出	修复后提升公平性的新模型

P	问题	基于生成歧视样本并再训练的方法开销大且无法确定样本的测试预言；直接移除受保护的属性并不能解决不公平问题，因为受保护的属性和未受保护的属性之间通常存在很强的相关性
C	条件	白盒模型，可访问训练集
D	难点	1.如何不使用过多额外的歧视样本完成对模型的公平性修复 2.如何轻量级地修复模型
L	水平	TOSEM 2023（CCF-A）

BET 算法原理图

- 算法原理图

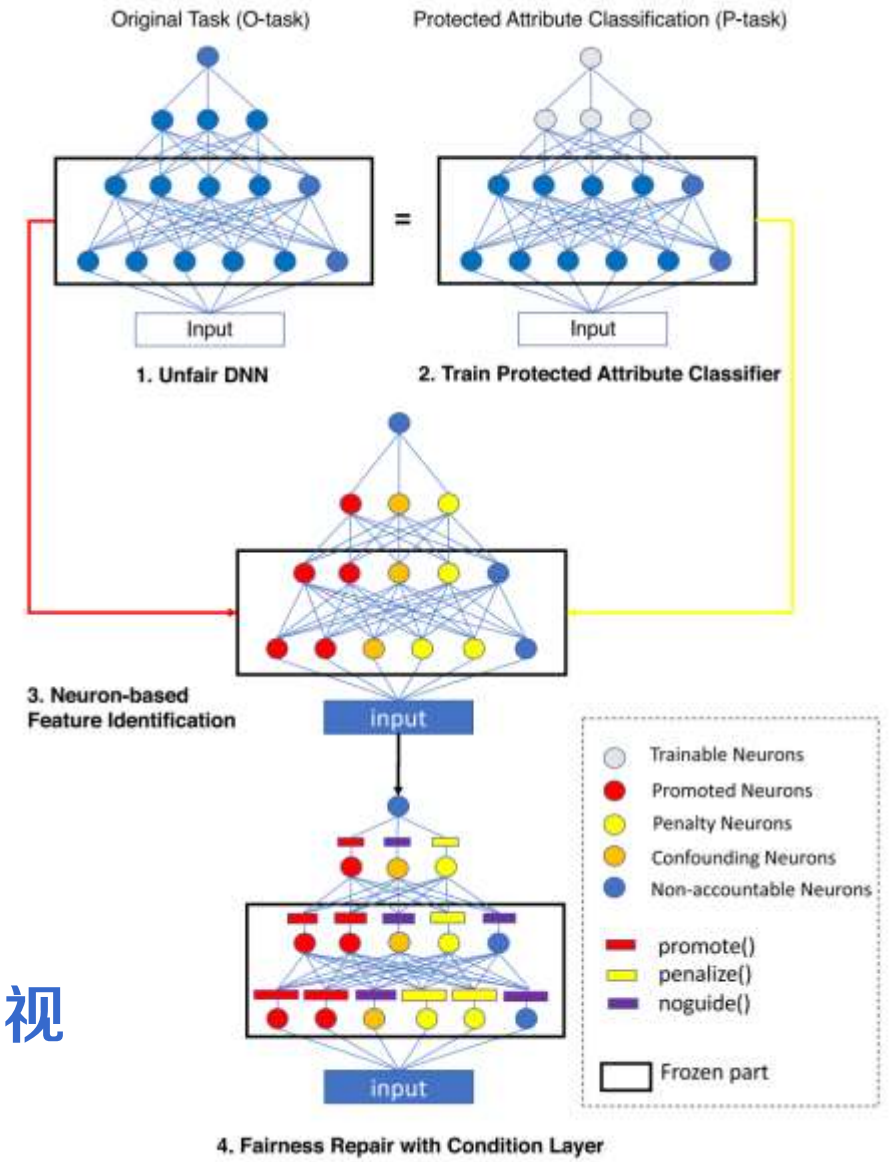
步骤1：基于原模型训练保护特征分类器

步骤2：对数据集的特征进行神经元层面的分析

步骤3：通过添加隐藏层继续训练以提高公平性

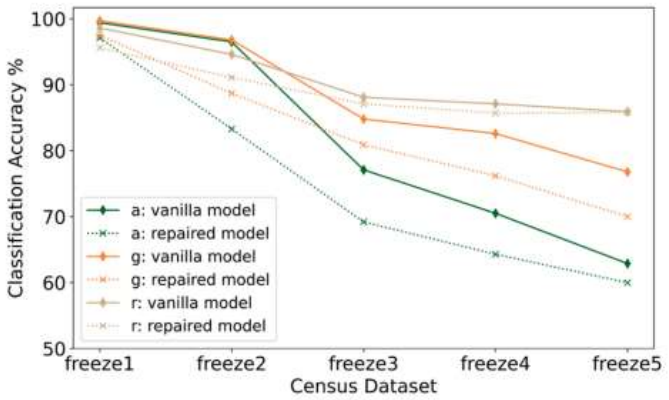
- 原始任务O的关键特征
- 分类任务P的关键特征
- 同时负责P任务和O任务（混淆神经元）
- 非P任务也非O任务的关键特征

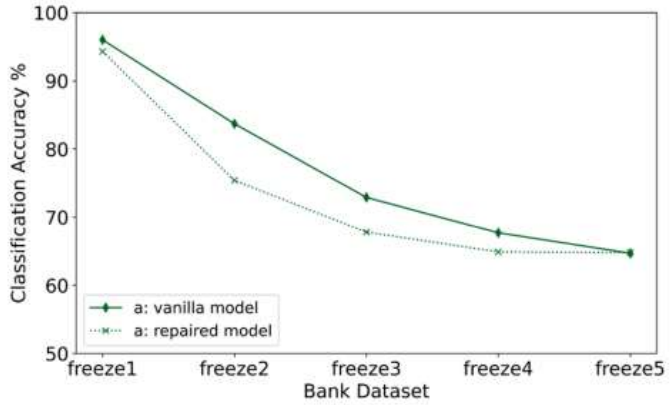
促进●以保证原模型的功能，抑制●以消除歧视

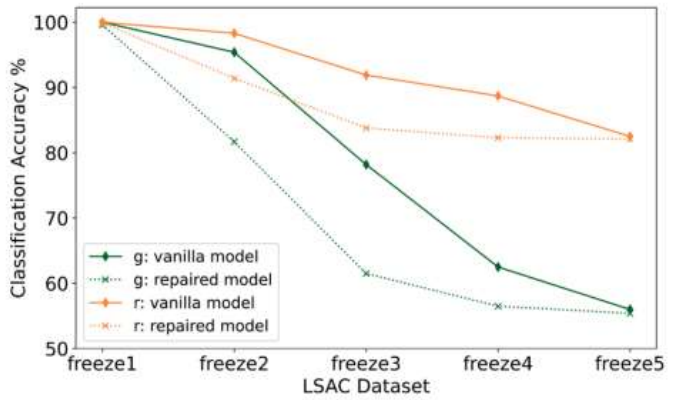


创新点验证　模型使用保护特征是不公平的根本原因

- 证明保护特征（PF）与最终预测公平性之间的联系
  - 直观经验：保护特征分类器的分类精度越高，模型中使用的PF越多
    - 用性别标签(即男性和女性)重新标记训练数据
    - 重用和冻结原始模型的前几层，并重新训练模型的其余部分以识别新标签
    - 再训练模型的精度越高，意味着相应原始模型的前几层往往提供更多的PF







当更多的层被冻结时，通常更难根据P-task训练高性能分类器，
这表明PF往往可能被逐层丢弃

创新分析　神经元状态分析

- 神经元的四种分类
  - 处罚神经元
    - 负责保护属性的分类，但不负责原始分类，对模型正确性影响较小
    - 惩罚这些神经元的输出，但不惩罚所有神经元
  - 促进神经元
    - 负责原始任务，但不负责识别受保护的属性
    - 预测应该更多地依赖于这些神经元，需要促进其输出
  - 混淆神经元
    - 这些神经元可能在这两项任务中都发挥着重要作用，不能简单地惩罚或促进
  - 非可问责神经元
    - 不清楚这些神经元是应该受到促进还是惩罚，但其往往不会对结果产生太大影响

实验设计　数据资源

- 数据与模型资源

数据集名称	Census Income	COMPAS	German Credit	LSAC	Bank Marketing
数据集用途	人口收入普查	再犯罪风险预测	评估个人信用	律师资格预测	银行营销
样本数量	48842	10000+	1000	/	45000
特征总数	14	9	20	/	16
模型准确率	84.7%	67.5%	78.2%	86.3%	89.2%

- 对比方法
  - EIDIG [M_{dis} / M_a] (2021)
  - 翻转数据增强 [M_{flip}]
  - 多任务学习 [M_{flip}] (2020)

- 评价指标
  - Acc：模型的准确率
  - RR：模型的修复率
  - GS：测试工具能够成功生成歧视性实例的数量
  - RG：输入实例的歧视实例比例

实验设计　对比实验

- 对比实验：评估生成的歧视性实例集的修复率
- 实验结论
  - Faire的修复率远高于基线
  - 单个属性的修复率高于两个属性组合的修复率
  - Faire修复后的准确率有所下降，但不太多（1%）
  - 对于图片数据集的修复效果不如表格数据

公平性和正确性的权衡

Dataset	Attr	M_{van}	M_{dis}		M_a		M_{mt}		M_{flip}		*Faire*			
		ACC	ACC	RR	ACC	RR	ACC	RR	ACC	RR	ACC	RR	*lb*	*ub*
Census	a	0.847	0.841	0.941	0.838	0.968	**0.846**	0.435	**0.846**	0.451	0.831	**0.999**	−0.30	0.15
	g	0.847	0.841	0.976	0.840	0.948	**0.849**	0.710	0.845	0.856	0.837	**0.995**	−0.60	0.10
	r	0.847	0.841	0.978	0.844	0.955	0.842	0.853	**0.846**	0.767	0.839	**0.994**	−0.95	0.80
	a&g	0.847	0.841	0.908	0.843	0.962	0.844	0.347	**0.845**	0.293	0.833	**0.990**	−0.30	0.25
	a&r	0.847	0.841	0.924	0.841	0.958	**0.847**	0.369	0.845	0.433	0.832	**0.987**	−0.35	0.25
	r&g	0.847	0.841	0.957	0.842	0.942	**0.846**	0.382	0.843	0.545	0.835	**0.972**	−0.90	0.80
Bank	a	0.892	0.890	0.916	0.890	0.977	**0.892**	0.838	0.889	0.478	0.890	**0.998**	−0.15	0.30
LSAC	g	0.863	0.859	0.964	**0.860**	0.947	**0.860**	0.914	0.857	0.679	0.832	**0.997**	−0.85	0.20
	r	0.863	0.859	0.918	0.848	0.963	0.858	0.884	**0.861**	0.806	0.833	**0.998**	−0.90	0.05
	g&r	0.863	0.859	0.878	0.844	0.951	**0.861**	0.841	0.850	0.448	0.830	**0.993**	−0.55	0.45
MNIST	box	0.995	–	–	–	–	**0.995**	0.333	**0.995**	0.500	0.993	**1.000**	−0.05	0.05
Credit	a	0.782	0.745	0.951	0.758	0.943	**0.762**	0.807	0.752	0.627	0.743	**0.958**	−0.50	0.60
	g	0.782	0.745	0.941	**0.772**	0.955	0.752	0.820	0.748	0.728	0.765	**0.997**	−0.10	0.60
	a&g	0.782	0.745	**0.991**	**0.762**	0.988	0.752	0.719	0.730	0.677	0.743	0.913	−1.00	0.25
COMPAS	g	0.675	0.668	0.733	0.658	0.703	**0.677**	0.653	0.657	0.270	0.637	**0.995**	−0.10	0.40
	r	0.675	0.668	0.883	**0.676**	0.859	0.673	0.787	0.661	0.871	0.659	**0.997**	−0.10	0.50
	g&r	0.675	0.668	0.672	0.671	0.642	0.672	0.481	**0.675**	0.128	0.645	**0.953**	−0.10	0.45

We run five times with different seeds and the average results are reported.

实验设计　对比实验

- 对比实验：公平性修复的有效性

Data	M_{van}			M_{dis}			M_a			M_{mt}			M_{flip}			*Faire*		
	GS_A	GS_E	RG	GS_A	GS_E	RG	GS_A	GS_E	RG	GS_A	GS_E	RG	GS_A	GS_E	RG	GS_A	GS_E	RG
C-a	0.464	0.654	0.111	0.305	0.218	0.023	0.220	0.132	0.016	0.540	0.749	0.183	0.736	0.743	0.176	**0.001**	**0.001**	**0.001**
C-g	0.187	0.282	0.039	0.105	0.070	0.016	0.353	0.255	0.064	0.252	0.398	0.054	0.262	0.301	0.034	**0.006**	**0.010**	**0.000**
C-r	0.203	0.324	0.105	0.168	0.122	0.014	0.281	0.194	0.030	0.200	0.323	0.077	0.378	0.460	0.098	**0.013**	**0.017**	**0.003**
C-a&g	0.518	0.717	0.151	0.360	0.279	0.031	0.161	0.074	0.013	0.594	0.788	0.285	0.828	0.822	0.298	**0.028**	**0.034**	**0.003**
C-a&r	0.608	0.740	0.211	0.413	0.339	0.036	0.121	0.087	0.009	0.616	0.802	0.272	0.758	0.773	0.331	**0.019**	**0.021**	**0.005**
C-r&g	0.355	0.486	0.127	0.257	0.182	0.031	0.297	0.218	0.038	0.273	0.527	0.130	0.711	0.707	0.214	**0.035**	**0.047**	**0.009**
avg	0.389	0.534	0.124	0.268	0.202	0.025	0.239	0.160	0.028	0.413	0.598	0.167	0.612	0.634	0.192	**0.017**	**0.022**	**0.004**
B-a	0.679	0.795	0.118	0.407	0.345	0.023	0.093	0.072	0.011	0.738	0.653	0.055	0.896	0.869	0.145	**0.003**	**0.003**	**0.002**
L-g	0.417	0.330	0.022	0.246	0.138	0.013	0.258	0.176	0.014	0.379	0.239	0.014	0.622	0.512	0.055	**0.005**	**0.005**	**0.003**
L-r	0.728	0.730	0.062	0.494	0.331	0.040	0.187	0.130	0.023	0.651	0.441	0.029	0.816	0.801	0.056	**0.020**	**0.020**	**0.005**
L-g&r	0.844	0.822	0.091	0.606	0.441	0.035	0.184	0.112	0.011	0.597	0.454	0.032	0.904	0.888	0.191	**0.028**	**0.028**	**0.009**
avg	0.663	0.627	0.058	0.449	0.303	0.029	0.210	0.139	0.016	0.542	0.378	0.025	0.781	0.734	0.101	**0.018**	**0.018**	**0.006**
Cre-a	0.363	0.434	0.266	0.120	0.080	0.032	0.082	0.042	0.017	0.186	0.267	0.070	0.225	0.349	0.220	**0.019**	**0.019**	**0.009**
Cre-g	0.156	0.192	0.115	0.075	0.040	0.041	0.157	0.089	0.050	0.199	0.254	0.118	0.125	0.216	0.075	**0.003**	**0.003**	**0.031**
Cre-a&g	0.408	0.469	0.407	0.175	0.131	0.074	0.112	0.051	0.033	0.273	0.389	0.181	0.475	0.453	0.387	**0.059**	**0.059**	**0.002**
avg	0.309	0.365	0.263	0.123	0.084	0.049	0.117	0.061	0.033	0.219	0.303	0.123	0.275	0.339	0.227	**0.027**	**0.027**	**0.014**
Com-g	0.692	0.655	0.003	0.498	0.379	0.014	0.506	0.431	0.017	0.638	0.641	0.127	0.735	0.658	0.707	**0.010**	**0.010**	**0.002**
Com-r	0.554	0.542	0.075	0.476	0.465	0.002	0.467	0.486	0.006	0.339	0.494	0.067	0.480	0.554	0.011	**0.007**	**0.007**	**0.0**
Com-g&r	0.698	0.682	0.348	0.486	0.385	0.014	0.537	0.526	0.025	0.723	0.724	0.145	0.749	0.742	0.611	**0.109**	**0.109**	**0.011**
avg	0.142	0.648	0.626	0.487	0.410	0.010	0.503	0.481	0.016	0.567	0.620	0.113	0.655	0.651	0.443	**0.042**	**0.042**	**0.004**

- 实验结论
  - Faire在提高模型公平性方面明显优于其他方法
  - 尽管其他方法可以修复许多原始的歧视性实例，但新模型的公平性可能不会真正增强

- **Faire的效率**

Dataset	Attr	M_{dis}			M_a			M_{flip}	M_{mt}	*Faire*			
		T_{data}	T_{train}	T_{total}	T_{data}	T_{train}	T_{total}	T_{train}	T_{train}	T_{train_p}	$T_{analysis}$	T_{repair}	T_{total}
Census	a	301,341.1	254.3	301,595.4	49,538.0	137.9	49,675.9	58.0	52.4	87.4	52.3	36.4	176.1
	g	301,341.1	254.3	301,595.4	10,724.1	132.2	10,856.3	51.2	52.8	94.3	53.7	32.5	180.5
	r	301,341.1	254.3	301,595.4	18,555.7	148.7	18,704.4	49.3	52.8	88.6	59.2	29.8	177.6
	a&g	301,341.1	254.3	301,595.4	44,060.5	168.5	44,229.0	59.8	54.2	181.7	114.6	42.3	338.6
	a&r	301,341.1	254.3	301,595.4	110,551.9	233.2	110,785.1	79.5	66.3	176	117.4	45.3	338.7
	r&g	301,341.1	254.3	301,595.4	67,910.9	172.5	68,083.4	62.2	56.5	182.9	109.8	41.2	333.9
Bank	a	94,645.6	126.3	94,771.9	94,645.6	126.2	94,771.8	55.9	51.2	80.7	51.0	36.9	168.6
LSAC	g	158,461.2	88.5	158,549.7	17,755.8	111.8	17,867.6	28.5	35.8	54.4	46.8	22.1	123.3
	r	158,461.2	88.5	158,549.7	56,930.7	125.3	57,056.0	30.7	33.9	50.0	47.2	24.2	121.4
	g&r	158,461.2	88.5	158,549.7	83,774.7	176.1	83,950.8	42.5	38.9	104.4	92.9	26.5	223.8

- **实验结论**
  - 尽管Faire需要搜索最优超参数，但可以很容易地并行完成，从而减少引入的时间开销
  - 虽然基于翻转的再训练和多任务学习方法非常快速，但其有效性不足

综合考虑Faire更加高效

- 修复模型的深度分析

Data	layer2			layer3			layer4			layer5			layer6		
	M_{van}	M_{dis}	*Faire*	M_{van}	M_{dis}	*Faire*	M_{van}	M_{dis}	*Faire*	M_{van}	M_{dis}	*Faire*	M_{van}	M_{dis}	*Faire*
C-a	0.296	0.157	**0.070**	0.138	0.069	**0.031**	0.273	0.055	**0.049**	0.170	0.017	**0.012**	0.138	0.021	**0.000**
C-r	0.237	**0.063**	0.236	0.152	**0.039**	0.146	0.233	**0.032**	0.169	0.150	**0.016**	0.119	0.106	0.011	**0.002**
C-g	0.135	0.055	**0.028**	0.091	**0.033**	0.044	0.169	**0.025**	0.038	0.112	**0.011**	0.025	0.082	0.008	**0.001**
C-a&r	0.393	**0.186**	0.213	0.169	**0.097**	0.210	0.145	**0.069**	0.153	0.093	**0.030**	0.035	0.141	0.029	**0.004**
C-a&g	0.368	0.169	**0.102**	0.154	0.076	**0.040**	0.240	0.061	**0.021**	0.149	0.028	**0.011**	0.138	0.025	**0.004**
C-r&g	0.289	**0.085**	0.239	0.175	**0.055**	0.188	0.162	**0.036**	0.226	0.099	**0.015**	0.143	0.112	0.016	**0.009**
B-a	0.433	**0.191**	0.226	0.260	**0.092**	0.206	0.246	**0.041**	0.082	0.240	**0.018**	0.023	0.120	0.029	**0.000**
L-g	**0.047**	0.111	0.068	0.077	0.037	**0.023**	0.121	0.027	**0.026**	0.084	0.026	**0.010**	0.089	0.017	**0.001**
L-r	**0.141**	0.177	0.145	0.220	**0.066**	0.067	0.317	0.050	**0.047**	0.198	0.050	**0.042**	0.178	0.036	**0.001**
L-g&r	**0.169**	0.220	0.249	0.231	**0.101**	0.166	0.356	**0.067**	0.093	0.180	0.064	**0.055**	0.165	0.047	**0.002**

- 实验结论
  - 歧视性实例再训练方法一般可以减少每层的平均距离
  - 在不使用歧视性实例的情况下，Faire并没有一致地减少中间层的特征距离，而是大大减少了最后一层的距离
  - 最后一层的结果会直接影响公平性，这也解释了为什么Faire比其他方法更有效

算法总结 BET

- 算法流程
  - 基于原模型训练保护特征分类器
  - 对数据集的特征进行神经元层面的分析
  - 通过添加隐藏层继续训练以提高公平性
- 算法优势
  - 适用范围不局限于表格数据集，对于图像数据集的CNN模型都适用
  - 无需引入额外的歧视样本进行训练
- 算法不足
  - 相比其他方法，正确率下降的更多
  - 并没有很好地减小中间层的特征距离，仍然存在细微的个体歧视

特点总结与未来展望

特点总结与未来展望

- **FairNeuron**
  - 仅使用训练样本，完全不生成歧视样本进行模型公平性修复
  - 侧重于群体公平性，采用群体公平性指标进行评估
  - 侧重于数据集公平，使用Dropout层对有偏数据进行训练
- **Faire**
  - 方法不涉及歧视样本生成
  - 侧重于神经元层面，引入隐藏层对神经元促进或者抑制
  - 适用于图像分类CNN模型，但效果欠佳
- 未来发展
  - 将相关方法拓展至图像分类CNN模型中
  - 将仅微调模型适当与歧视样本生成结合，改善微调数据

预期收获　回顾分析

- 预期收获
  - 掌握人工智能模型的公平性测试相关知识
    - 个体公平与群体公平的定义
    - 歧视样本的生成方法
  - 了解两种人工智能模型的公平性修复方法
    - 几种公平性评价指标和模型修复指标
    - 歧视路径和可问责神经元的选择
    - 模型修复的几种方式

参考文献

[1] Gao X, Zhai J, Ma S, et al. FairNeuron: improving deep neural network fairness with adversary games on selective neurons[C].Proceedings of the 44th International Conference on Software Engineering. 2022: 921-933.

[2] Li T, Xie X, Wang J, et al. Faire: Repairing fairness of neural networks via neuron condition synthesis[J]. ACM Transactions on Software Engineering and Methodology, 2023, 33(1): 1-24.

道德经

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！