计算机科学
COMPUTER SCIENCE

# 对抗生成式的多敏感属性数据去偏方法

王文鹏，葛洪伟，李婷

引用本文

王文鹏，葛洪伟，李婷. 对抗生成式的多敏感属性数据去偏方法[J]. 计算机科学，2025，52(11)：90-97.

WANG Wenpeng, GE Hongwei, LI Ting. Adversarial Generative Multi-sensitive Attribute Data Biasing Method [J]. Computer Science, 2025, 52(11): 90-97.

---

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240900061

# 对抗生成式的多敏感属性数据去偏方法

王文鹏　葛洪伟　李　婷
康养智能化技术教育部工程研究中心(江南大学)　江苏 无锡 214122
江南大学人工智能与计算机学院　江苏 无锡 214122
(2734847275@qq.com)

**摘　要**　针对消除数据中敏感属性与非敏感属性之间的相关性、减轻实现公平性对模型准确性的损失以及多敏感属性去偏的问题,提出一种对抗生成式的多敏感属性数据去偏方法。在多敏感属性去偏问题上,该方法通过多个敏感属性的组合值来划分群组,并通过消除各群组与多敏感属性组合的相关性来提升各群组预测结果的公平性。在消除数据中敏感属性与非敏感属性之间的相关性问题上,采用自编码器与预测敏感属性的网络进行对抗式训练,这种训练机制能够深入挖掘并消除群组中潜藏的与敏感属性相关的信息,从而在保留数据有用性的同时,显著降低偏见。在减轻实现公平性对模型准确性损失,最大化准确性与公平性之间平衡的问题上,通过引入预测网络,并利用其损失函数作为约束,优化编码器的信息提取能力,确保在数据编码过程中能够更精准地捕捉关键信息,避免数据在去偏过程中过度牺牲模型的预测性能。在3个真实数据集上进行数据去偏实验,将经编码器编码的数据应用于逻辑回归模型,公平性提升50.5%～84%,验证了该数据去偏方法的有效性。综合考虑公平性、准确性以及公平性与准确性的平衡,该去偏方法优于其他去偏算法。

**关键词**:数据去偏;机器学习;对抗学习;自编码器
**中图分类号**　TP391

## Adversarial Generative Multi-sensitive Attribute Data Biasing Method

WANG Wenpeng,GE Hongwei and LI Ting
Engineering Research Center of Intelligent Technology for Healthcare,Ministry of Education,Jiangnan University,Wuxi,Jiangsu 214122,China
School of Artificial Intelligence and Computer Science,Jiangnan University,Wuxi,Jiangsu 214122,China

**Abstract**　This paper proposes a method for multi-sensitive attribute data debiasing,leveraging adversarial learning and autoencoder to eliminate correlations between sensitive and non-sensitive attributes,minimize the impact on model accuracy when striving for fairness,and address the issue of multi-sensitive attribute debiasing. In addressing multi-sensitive attribute debiasing,this method groups based on the combined values of multiple sensitive attributes,enhancing the fairness of each group's predictions by eliminating group correlations with these sensitive attribute combinations. To eliminate correlations between sensitive and non-sensitive attributes,an adversarial training approach is employed,utilizing auto-encoders alongside networks predicting sensitive attributes. This training effectively uncovers and eliminates latent sensitive attribute-related information within the groups,significantly reducing bias while retaining data utility. To mitigate the impact on model accuracy from striving for fairness and optimize the balance between accuracy and fairness,a prediction network is introduced. Its loss function is used as a constraint to enhance the encoder's ability to extract information,ensuring more precise capture of key information during data encoding and preventing excessive sacrifice of predictive performance during the debiasing process. Data debiasing experiments on three real datasets are conducted,applying the encoded data to logistic regression models. The fairness improvements range from 50.5% to 84%,validating the effectiveness of the debiasing method. Considering fairness,accuracy,and their balance,this debiasing method outperforms other debiasing algorithms.

**Keywords**　Data depolarization,Machine learning,Adversarial learning,Auto-encoder

机器学习,凭借其卓越的特征提取与高效的决策判断能力,已广泛渗透至社会关键决策领域[1],随之而来的,是其公平性问题日益凸显。传统机器学习算法往往侧重于优化预测性能指标,如准确性、F1分数及精度等,未充分考量其决策过程中的公平性及其社会影响。这种做法可能不经意间延续或加剧了训练数据中固有的偏见,进而对弱势群体造成不利影

到稿日期:2024-09-10　返修日期:2024-12-16
基金项目:国家自然科学基金(61806006)
This work was supported by the National Natural Science Foundation of China(61806006).
通信作者:葛洪伟(ghw8601@163.com)

响。鉴于非敏感属性与敏感属性间的高度相关性，单纯从训练集中剔除敏感属性，并不足以确保模型的公平性，此现象即所谓的“红线效应”，值得深入探究并提出应对之策[2]。

公平机器学习的一个内在挑战是公平性和准确性之间的平衡，然而减轻不公平的行为往往会损害模型的预测性能[1,3]，对弱势群体偏见较小的预测模型，其预测结果可能会偏离真正的类别。在现实实践中，有时需要同时考虑多个敏感属性[4]，这些属性的组合为各种可能的子组创建了不同级别的特权或劣势，例如黑人女性可能容易受到性别歧视和种族主义的影响[5]。

为解决数据偏见问题，提出了一种结合对抗学习与自编码器的创新方法，专门对数据集进行多敏感属性数据去偏。该方法利用编码器与预测敏感属性网络的对抗学习机制，有效削弱非敏感属性组与敏感属性组间的关联，从而消除“红线效应”，确保基于非敏感属性训练的预测模型具备公平性。值得注意的是，敏感属性的数量依据具体场景任务灵活设定，本文方法全面考虑了多敏感属性场景下的公平性需求。

为平衡公平性与模型预测性能，引入了预测网络作为辅助。在预测网络和解码器的协同作用下，编码器朝着保留数据原始信息和有利于模型预测的方向编码。进一步，编码器（生成器）与预测敏感属性网络（对抗器）进行对抗式去偏，旨在去除编码后的数据与多敏感属性的相关性。通过上述方法，能够减轻追求公平性对预测准确性的潜在负面影响。

本文方法的核心贡献概述如下：

1）创新去偏策略：提出了一种融合对抗学习与自编码器的技术，针对多敏感属性进行去偏处理。该方法通过多敏感属性的组合值划分群组，利用对抗学习机制有效剥离群组与多敏感属性的相关性，实现数据集的多敏感属性去偏。

2）优化编码器设计：采用自编码器与预测网络的协同工作，结合预测敏感属性的对抗式约束，优化编码器的性能。此设计旨在平衡模型的准确性与公平性，同时减少在追求公平过程中可能导致的预测性能损失。

3）整合预处理与深度学习：将数据预处理技术与深度学习算法相结合，通过编码器输出处理后的去偏数据集。这一策略确保了在后续模型训练过程中能够避免学习并放大原始数据集中的偏见，从而提升整体模型的公正性和可靠性。

## 1 相关工作

机器学习的公平性问题受到学术界的广泛关注，现有的机器学习去偏算法可以分为数据预处理去偏方法（Pre-processing Algorithms）、模型去偏方法（In-processing Algorithms）以及后验去偏方法（Post-processing Algorithms）3种。

预处理去偏方法是对原始数据进行处理，消除潜在的歧视或偏见，例如修改训练样本标签[6-7]、修改训练样本权重[8-9]和合成数据去偏[10]。Kamiran等[5]提出了Re-Weighing方法，通过修改数据集中样本的权重，再用于模型训练，以提高后续分类的公平性。

模型去偏方法是对模型添加公平性约束或正则化项等来优化模型。例如：Petrovic′等[11]提出蒙特卡罗策略梯度法，直接优化组公平性和预测性能指标；Zhang等[12]提出对抗去偏方法，其优化目标是最大化分类器的准确性和降低判别器从预测结果中确定敏感属性的能力，进而实现预测结果的公平性。上述例子并没有考虑公平性与准确性平衡的问题，因此学者提出使用多目标优化的思想来寻找公平性与准确性的最优解，如公平性与准确性的帕累托优化[13]和多目标优化及混合自适应优先级重加权方法[14]。

后验去偏方法是对分类后的结果再次进行处理，例如添加约束或规则重新生成结果。Kamiran等[15]提出了Reject Option Classification方法，该方法给予无特权群体有利的结果，并对具有最高不确定性的决策边界周围置信区内的特权群体产生不利结果。

上述方法主要用于解决单个敏感属性的公平性问题，但考虑到现实实践问题，多敏感属性公平性问题亟需解决。Chakraborty等[16]提出了结合预处理方法和模型处理方法的Fairway框架，通过多个敏感属性来划分群组，根据情景测试思想删除有偏数据，来实现多敏感属性的公平性。d’Aloisio等[17]提出Debiaser for Multiple Variables，考虑了敏感变量的值和标签值的所有可能组合，以定义所谓的敏感组，然后将各个敏感组的样本数与其期望的个数（公平情况下）的比值作为平衡的准则，增删数据集以达到数据集去偏。Canalli等[18]提出的Fairness Trainsition Loss利用某些标签噪声的特征，在群组中重新分配噪声的概率，重新加权预测概率以减少不同群组中有利与不利结果的差异，并采用多目标优化的方法，来选择最佳的准确率与公平性。

其他关于机器学习公平性的研究方向还有通过平衡各个群组的损失而不是平衡预测结果来实现公平性预测[19-20]，通过对抗学习的方式实现预测公平性[21-22]。此外，联邦学习隐私场景下公平性问题[23-24]也备受关注。

当前去偏技术多聚焦于单一敏感属性的公平性保障，面对多元敏感属性的数据集时，仅能确保模型预测针对某一敏感属性的公正性，在多元敏感属性的背景下，去偏效果较差。更为关键的是，这些方法往往忽视了非敏感属性中潜藏的敏感属性信息，使得模型在训练过程中可能会从非敏感特征中捕捉到敏感信息，从而引发偏见问题。此外，多目标优化策略的去偏方法虽具潜力，却也伴随着计算复杂度的激增与参数设置的繁琐挑战。鉴于此，本文独辟蹊径，聚焦于直接剥离非敏感属性组与敏感属性组之间的关联性，充分利用对抗学习的强大力量，旨在攻克多敏感属性场景下的公平性难题，并力求在提升公平性的同时，最小化对模型准确性的潜在影响。

## 2 对抗生成式多敏感属性数据去偏方法

### 2.1 模型网络

模型网络共分为3个部分：自编码器（生成器）、预测分类网络和预测敏感属性网络（判别器）。自编码器分为两个部分：编码器和解码器。考虑到数据降维和编码器具有很好的信息提取能力，则结合解码器损失、预测分类网络损失和预测敏感属性网络损失对编码器进行约束优化，为其提供梯度优化方向。本文方法所用网络均为全连接神经网路，具体方法框架如图1所示。

原数据输入编码器，编码得到中间隐变量，中间隐变量分

别输入解码器、预测分类网络和预测敏感属性分类网络。将预测分类网络损失、预测敏感属性分类网络损失和自编码器损失作为编码器的约束优化。预测敏感属性网络的标签是由敏感属性的个数决定的，例如，对于性别属性＝{男人，女人}和种族属性＝{白人，非白人}，有以下 4 种取值：{(男人，白人)，(男人，非白人)，(女人，白人)，(女人，非白人)}。对其编码可得{3，1，2，0}，因此预测敏感属性网络是个四分类网络。

预测敏感属性网络的输入为编码器的输出，通过对抗学习的方式使得编码器的输出与敏感属性组独立；解码器的输入为编码器的输出，使得经编码器降维编码的数据能够学习到原始数据的主要特征；预测分类网络的输入为编码器的输出，将准确性指标加入编码器的约束优化过程，以减轻追求公平性对预测准确性的潜在负面影响。

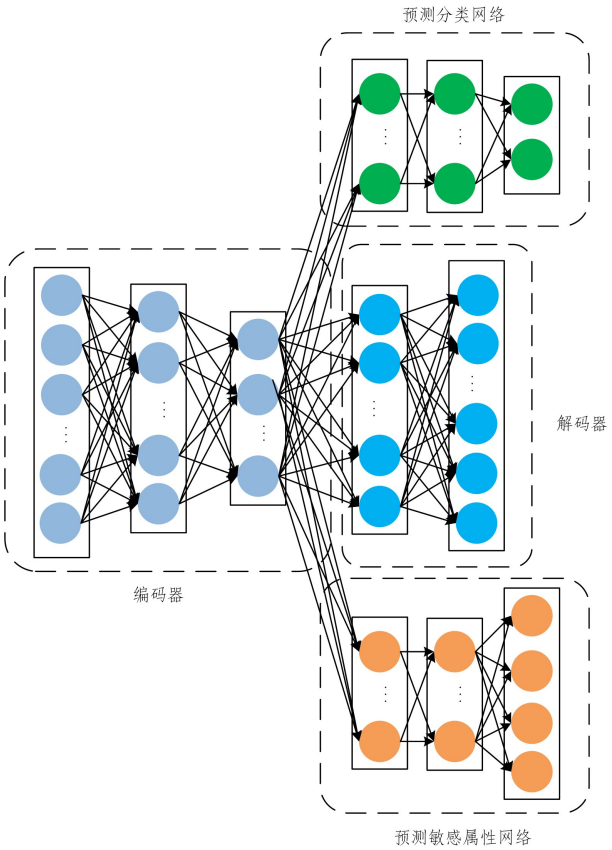


图 1 对抗生成式多敏感属性去偏方法网络

Fig. 1 Adversarial generative multi sensitive attribute depolarization method network

### 2.2 编码数据的噪声问题

自编码器和对抗生成数据都存在数据质量的问题。本文方法仅对原始数据的非敏感属性进行编码，并没有对标签进行编码，因此仅期望编码器能够学习到某种映射，使得使用编码器编码的数据训练的预测模型能在较高准确性的前提下实现预测的公平性。因此本文方法不需要考虑编码后的数据的噪声问题。

### 2.3 损失函数设计

在本方法模型中，编码器、解码器、预测分类网络和预测敏感属性网络都需要一个损失函数来提供梯度更新优化方向，包括作用于预测分类网络的损失函数 $L_C$、作用于预测敏感属性网络的损失函数 $L_S$、作用于编码器的损失函数 $L_E$、作用于解码器的损失函数 $L_D$。设数据集 $D=\{X,S,Y\}$，$X$ 是非敏感属性组，$\hat{X}$ 是解码器输出，$S$ 是敏感属性标签值，$\hat{S}$ 是敏感属性预测值，$Y$ 为真实分类标签值，$\hat{Y}$ 表示分类预测标签。损失函数 $L_C$ 用于优化预测分类网络的准确率，具体计算式如下：

$$L_C=\sum_i^n L(\hat{Y}_i,Y_i)$$

损失函数 $L_S$ 用于优化预测敏感属性网络的准确率，具体计算式如下：

$$L_S=\sum_i^n L(\hat{S}_i,S_i)$$

损失函数 $L_D$ 使用均方误差(MSE)来优化解码器，具体计算式如下：

$$L_D=\sum_i^n MSE(\hat{X}_i,X_i)$$

损失函数 $L_E$ 结合预测分类网络、预测敏感属性网络和解码器的损失来约束优化编码器，则编码器的优化目标函数为：

$$\min L_E=\min_{L_D+L_C}\max_{L_S}[(L_D+L_C)-\lambda L_S]$$

从上式可以看出，编码器朝着最小化预测分类网络损失和解码器损失，以及最大化预测敏感属性网络损失的方向优化，预测敏感属性网络则最小化预测敏感属性的损失，显然编码器与预测敏感属性网络之间存在对抗的零和博弈关系。

根据文献[12]，为了防止编码器在优化方向意外地协助对抗网络减少其损失，需要在编码器的优化目标中明确添加一个映射项。因此，最终的优化目标调整为包含该映射项的形式：

$$\nabla w(L_D+L_C)-proj_{\nabla_W L_S}\nabla(L_D+L_C)-\lambda\nabla_W L_S$$

$L_S$ 项添加 $\lambda$ 参数，用于调整去偏的权重，当模型预测的准确性与公平性同等重要时，取 $\lambda=1$。

### 2.4 对抗学习去相关性理论保证

命题：将训练数据集 $D$ 划分为 $\{X,S,Y\}$，$X$ 为非敏感属性组，$S$ 为敏感属性，$Y$ 为实际标签，并有以下假设：

(1)敏感属性 $S$ 是离散的。

(2)对抗器的输入只有编码器(生成器)的输出 $\hat{X}$。

(3)对抗器能够收敛，通过学习一个随机函数 $A$ 来最小化交叉熵损失 $E_{(x,y,s)\sim D}[-\log p(A(\hat{x})=s)]$，即对抗器的优化目标是能从编码器(生成器)的输出 $\hat{X}$ 预测敏感属性 $S$。

(4)编码器(生成器)能完全欺骗对抗器，则对抗器的交叉熵损失是敏感属性 $S$ 的信息熵 $H(S)$。

结论：$\hat{X}\perp S$。

证明：对抗器根据分布 $S|\hat{X}=\hat{x}$ 生成 $A(\hat{x})$，则其损失正是条件熵：

$$\begin{aligned}H(S|\hat{X})&=E_{(x,y,s)\sim D}[-\log P(S=s|\hat{X}=\hat{x})]\\&=E_{(x,y,s)\sim D}[-\log P(A(\hat{x})=s|\hat{X}=\hat{x})]\end{aligned}$$

使用反证法来证明，假设 $\hat{X}$ 与 $S$ 不独立，则 $H(S|\hat{X})<H(S)$，那么对抗器能够实现比 $H(S)$ 更小的损失值，这与命

题假设(4)矛盾。所以 $\hat{X}$ 与 $S$ 独立。

上述命题指出，当编码器(生成器)生成的数据难以被对抗器准确预测时，这标志着经过编码器处理的数据与多个敏感属性之间的相关性显著降低。更进一步，当生成器能够完全迷惑对抗器，使其无法从数据中识别出敏感属性时，编码器的输出便实现了与敏感属性的完全独立。

## 3 实验

### 3.1 实验数据集

实验采用公平性文献中的 3 个著名二分类数据集 Adult[25]，Compas[26] 和 German[26]。对于每个数据集，我们考虑由两个敏感变量组成的敏感属性组。Adult 数据集的预测目标是预测个体的年收入是否超过 5 万美元，敏感属性为种族和性别。Compas 数据集的预测目标是预测个体在未来两年内是否会再次犯罪，敏感属性是种族和性别。German 数据集的预测目标是评估个体信用风险的好坏，敏感属性为性别和年龄，其中年龄大于 25 岁是有利属性值。

表 1 数据预处理后数据集信息

Table 1 Dataset information after data preprocessing

| 数据集 | 样本数 | 特征属性个数 | 敏感属性组合 | 标签 |
|---|---|---|---|---|
| Adult | 48842 | 102 | (Race,Sex) | 收入是否大于 5 万元 |
| Compas | 7214 | 9 | (Race,Sex) | 是否再次犯罪 |
| German | 1000 | 58 | (Sex,Age) | 信用状况是否良好 |

在数据处理过程中，对多个敏感属性进行了组合，并对这些组合后的敏感属性进行了独热编码。随后，将独热编码后的敏感属性组视为一个新的敏感属性，用于替代原先的敏感属性组，且将此新属性的值设定为预测敏感属性网络的标签值。表 1 列出了 Adult，Compas 和 German 这 3 个数据集经过预处理后的结果，对数据集中所有数值型特征(如收入、年龄、学历等)进行标准化处理，以确保它们遵循均值为 0、标准差为 1 的标准正态分布；对所有的类别型特征(如性别、种族、职业等)进行了独热编码转换，以避免在模型训练中引入不必要的顺序关系。

### 3.2 评价指标

在评估模型的分类性能时，采用准确率作为核心指标来衡量其分类能力。而对于模型的公平性考量，则需综合考虑多个敏感属性及其交叉影响。为此，引入了文献[27-28]中定义的交叉公平性指标平均机会差异(Average Odds Difference，AOD)和机会均等差异(Equal Opportunity Difference，EOD)，来量化评估任意两个子组之间的最大差异。这两个指标的取值范围均为[0，1]，且数值越小，表示模型的公平性表现越优秀。定义 $A$ 表示敏感属性，$S$ 为敏感属性的所有可能组合的集合，$s$ 是 $S$ 的一个子组，$Y$ 表示实际的标签值，$\hat{Y}$ 表示预测的标签值，其中 1 为有利标签，0 为非有利标签，则 EOD 与 AOD 的公式表示如下：

$$EOD=\max_{s\in S}P(\hat{Y}=1|A=s,Y=1)-\min_{s\in S}P(\hat{Y}=1|A=s,Y=1)$$

$$AOD=\frac{1}{2}[\max_{s\in S}(P(\hat{Y}=1|A=s,Y=0)+P(\hat{Y}=1|A=s,Y=1))-\min_{s\in S}(P(\hat{Y}=1|A=s,Y=0)+P(\hat{Y}=1|A=s,Y=1))]$$

为了更准确地量化模型的准确性与公平性之间的权衡，采用文献[19]中定义的 FAT(Fairness-Accuracy Trade-off)指标进行量化分析。FAT 值越大，意味着模型在提升准确性时利用数据中固有社会偏见的程度越低，体现出模型在公平性与准确性之间达到了更佳的权衡状态。

$$FAT_{f_p}=\frac{1}{\alpha\sum_{p\in P}\left(\frac{1}{1-f_p}\right)+(1-\alpha)\frac{1}{ACC}}$$

其中，$f_p$ 是关于敏感属性 $p$ 的公平性指标；$ACC$ 为准确率；$\alpha$ 是平衡公平性和准确性的权重参数，在公平性和准确性同等重要的情景下，本文所有实验中将其设置为 0.5。

### 3.3 对比算法

实验选取逻辑回归模型作为基础分类模型，使用去偏算法改进其预测的公平性。去偏算法选取 3 个经典去偏算法 EG，DIR，RW 和 3 个最新算法 DEMV，Fairway，FTL，具体如下：

(1)逻辑回归模型 LR(Logist Regression)。选用逻辑回归模型作为基础分类器的原因是它非常高效，同时也是一种白盒方法。参数使用 scikit-learn 默认参数设置，$max\_iter=1000$。

(2)DEMV(Debiaser for Multiple Variables)[17]。该方法的主要目标是有效地提高分类器在预处理过程中的公平性，因此需要考虑敏感变量的值和标签值的所有可能组合，以定义所谓的敏感组，然后将各个敏感组的个数与其期望个数的比值作为平衡的准则，来平衡数据集。使用文献[17]提供的方法，参数采用默认设置。

(3)EG(Exponentiated Gradient)[29]。它是一种添加线性约束的规约算法。其与在线文档 Fairlearn[30] 一样，使用绝对统计奇偶性(Absolute Statistical Disparity)或平等机会(Equal Opportunity)作为约束条件，选取最优结果。

(4)Fairway[16]。将预处理去偏方法与模型去偏方法相结合的一种公平性学习方法。预处理去偏的主要思想是使用特权组数据训练的模型与使用非特权组训练的模型对同一数据进行预测，根据它们预测结果是否匹配，来判断该数据是否有偏。模型去偏方法采用一种优化器方法，根据评价指标来选择最佳的模型训练参数。此方法的目标是使模型尽可能地公平，同时也不降低其他性能指标，如准确率等。

(5)RW(Re-weighting)[5]。通过重新调整训练数据中不同样本的权重，RW 算法试图减少模型对敏感属性的依赖，从而提高预测的公平性。

(6)DIR(Disparate Impact Remover)[31]。DIR 关注于减少模型预测结果中不同子群体间的差异影响，确保模型决策不会过度偏向某一特定群体。将此方法应用于多敏感属性去偏时，采用文献[31]中的做法，在考虑性别和种族两个敏感属性时，将白人男性作为特权组，非白人男性、白人女性和非白人女性作为非特权组，然后计算特权组与非特权组的 DI 值，

再取其平均值作为准则，来进行数据去偏。

(7)FTL(Fair Transition Loss)[18]。FTL 通过利用某些标签噪声的特征，在不同类别中重新分配不平衡噪声的概率，重新加权预测概率以减少不同社会群体中有利与不利结果的差异，并采用多目标优化的方法，来选择最佳的准确率与公平性。

### 3.4 实验环境

为确保对比的公正性，所有参与评估的算法均基于表 1 中预处理完毕的数据集进行训练集与测试集的分割。本文统一采用十折交叉验证法作为评估标准，且在数据分割过程中，设定了一个固定的随机数种子，以确保所有算法在训练与测试过程中，数据集的交叉验证划分保持一致，从而保障实验结果的可靠性与可比性。

对于 DEMV、RW、DIR 和本文方法，将去偏后的数据用于逻辑回归模型的训练预测。为了确保实验的一致性和公平性，所有逻辑回归模型在参数设置上均保持一致。此外，逻辑回归模型的训练集和测试集与去偏前的数据划分保持一致，从而确保实验条件的一致性。对于 FTL，EG 和 Fairway 这 3 种方法，它们各自将逻辑回归模型作为优化的基石。我们严格遵循了相关文献中的实验设置与指导，对每种方法进行了细致的调参与优化，以训练出各自框架下的最佳逻辑回归模型。

为了确保公正性，本文为所有预测分类模型采用了统一的评价指标。同时，严格设定优化算法的公平性优化目标，以确保与模型评估时所采用的公平性指标保持高度一致，从而全面而准确地衡量模型的性能与公平性表现。

### 3.5 详细网络结构

鉴于实验数据集之间的差异，经过特征提取后所得的特征数量可能会有所不同，导致自编码器的网络结构需相应调整。具体而言，在 Adult，Compas 和 German 数据集中，自编码器的网络结构分别设计为[101,64,12,64,101]，[8,16,7,16,8]和[57,64,32,18,32,64,57]。对于预测网络而言，其输入层节点数被设定为自编码器中间隐藏层的节点数，隐藏层则统一采用[64,32]的结构。由于本文聚焦于二分类任务，因此预测网络的输出层节点数被设定为 2。同样地，预测敏感属性网络的结构与预测网络相似，其输入层节点数也基于自编码器的中间隐藏层节点数确定，隐藏层结构同样为[64,32]。考虑到本文涉及两个敏感属性，且这两个属性的组合值共有 4 种，属于四分类问题，因此设定预测敏感属性网络的输出层节点数为 4。

### 3.6 模型训练

模型训练使用学习率为 0.001 的 Adam 优化器，采用指数衰减对学习率进行调整，gamma 值为 0.9；采用 L2 正则化防止过拟合，超参数取值为 0.001；公平性与准确性平衡超参数 $\lambda$ 取值为[0,10]，步长为 0.5，通过训练选取合适的 $\lambda$。模型训练过程中，编码器与预测敏感属性网络进行双向博弈的对抗式训练，更新两者优化目标。原数据集 $X$ 先输入自编码器，并进行梯度更新，经过一次梯度更新后的编码器再次对原数据集 $X$ 进行编码，编码后的数据作为预测分类网络和预测敏感数据网络的输入，以各自的损失函数进行梯度更新，最后编码器以损失函数 $L_E$ 进行梯度更新，以上述步骤作为一次迭代，选择合适的训练批次进行训练。

选择 Adam 优化器是因为它结合了动量和自适应学习率调整的优点，在处理稀疏梯度和高噪声的数据时表现出色。Adam 优化器在深度学习领域得到了广泛应用和验证，具有稳定性和快速收敛的优势。gamma 值设置为 0.9，是因为在指数衰减学习率策略中，该值能够在保证模型训练初期快速学习的同时，在后期逐步减小学习率，避免模型陷入局部最优。通过多次实验验证，发现 gamma 值为 0.9 时，模型的收敛效果最佳。在超参数的具体调优过程中，对学习率、正则化参数和去偏权重参数进行了初步设置，分别为 0.001，0.001 和 0.5。然后通过网格搜索法和交叉验证技术，对这些参数进行细化调整。在每个参数组合下分别训练模型并评估其在验证集上的表现，最终选择了在公平性和准确性之间取得最佳平衡的参数设置。

### 3.7 实验训练和测试流程

本文实验采用 10 折交叉验证法应用于所有算法，旨在防止模型因复杂度过高而出现过拟合现象。通过计算 10 次模型评价结果的均值来全面评估模型的性能。在实验流程中，首先，将训练集数据输入本文方法模型中，经过训练得到相应的编码器。随后，分别将训练集和测试集输入该编码器，以获取去偏处理后的训练集和测试集。这些去偏后的数据集随后被用于逻辑回归模型的训练和评估。最终，统计了逻辑回归模型在 10 次评估中的均值结果，以此作为本文方法的最终性能评价指标。

对于 DEMV，RW 和 DIR 去偏算法，处理后的数据集被进一步用于逻辑回归模型的训练和测试阶段；而对于 FTL，EG 和 Fairway 方法，则直接采用逻辑回归模型作为基础分类模型，并在此基础上进行参数调优以提升性能。在所有涉及逻辑回归模型的场景中，均采用了 scikit-learn 库的默认参数设置，并特别指定了迭代次数 $max\_iter=1000$。在 FTL 算法中，遵循了文献中推荐的默认参数配置，同时，将准确率和交叉公平性(AOD 和 EOD)作为多目标优化的关键指标，以全面评估和优化模型的性能。

### 3.8 实验结果与分析

本节详细展示了 7 种不同方法在 3 个数据集上的实验结果。Zafar 等[32]指出，公平性的提升往往伴随着准确率的牺牲。因此，在本文中，努力将去偏算法的准确率控制在与其他算法相近的水平，以便更公正地评估本文算法在去除偏见方面的效果。实验结果如表 2 所列，其中粗体表示最优结果，而下划线表示次优结果。可以看出，本文方法显著优于其他方法，具体体现在以下几个方面：

1)显著的公平性提升：本文提出的去偏算法在 3 个不同的数据集(Adult，Compas，German)上均展现出在公平性指标 EOD 和 AOD 上的显著提升，平均提升幅度超过 50%，最高达到 84.0%。这表明该算法在去除数据偏见方面取得了显著成效。

2)保持竞争力的准确性：尽管在追求公平性的同时，算法在某些情况下需要对准确性做出一定妥协，但本文方法通过精细的设计，成功地将准确率控制在与其他算法相近甚至更

优的水平。在 Adult 和 German 数据集中，本文方法实现了次优的准确率，并相较于基线模型有显著提升。

3）公平性与准确性的平衡：本文方法在多个评估指标上实现了公平性与准确性的良好平衡。在 6 个评估指标中，有 5 个达到了最优，1 个为次优，充分证明了本文算法在平衡两者之间的权衡时具有显著优势。

表 2 实验结果

Table 2 Experimental results

| 数据集 | 算法 | Accuracy | EOD | AOD | $FAT_{EOD}$ | $FAT_{AOD}$ |
|---|---|---|---|---|---|---|
| Adult | LR | 0.852 | 0.1306 | 0.1765 | 0.8606 | 0.8375 |
| | DIR | 0.850 | 0.1095 | 0.1663 | 0.8698 | 0.8418 |
| | DEMV | 0.847 | 0.0837 | 0.1763 | 0.8803 | 0.8352 |
| | Fairway | **0.882** | 0.1267 | 0.2176 | 0.8776 | 0.8292 |
| | FTL | 0.841 | 0.0627 | 0.0878 | 0.8865 | 0.8752 |
| | EG | 0.834 | 0.1965 | 0.3400 | 0.8638 | 0.7735 |
| | RW | 0.852 | 0.0826 | 0.1280 | 0.8835 | 0.8619 |
| | Ours | 0.853 | **0.0415** | **0.0874** | **0.9027** | **0.8818** |
| Compas | LR | 0.681 | 0.3278 | 0.3981 | 0.6766 | 0.6390 |
| | DIR | 0.661 | 0.1264 | 0.1480 | 0.7526 | 0.7444 |
| | DEMV | 0.671 | 0.1162 | 0.1585 | **0.7628** | 0.7466 |
| | Fairway | **0.715** | 0.2606 | 0.3658 | 0.7270 | 0.6722 |
| | FTL | 0.649 | 0.2409 | 0.2565 | 0.6997 | 0.6930 |
| | EG | 0.651 | 0.1166 | 0.1444 | 0.7496 | 0.7394 |
| | RW | 0.678 | 0.2091 | 0.2727 | 0.7301 | 0.7018 |
| | Ours | 0.661 | **0.1105** | **0.1179** | 0.7584 | **0.7557** |
| German | LR | 0.743 | 0.4001 | 0.4719 | 0.6638 | 0.6174 |
| | DIR | 0.719 | 0.3058 | 0.2171 | 0.7064 | 0.7496 |
| | DEMV | 0.753 | 0.3328 | 0.2879 | 0.7075 | 0.7320 |
| | Fairway | **0.837** | 0.3847 | 0.1707 | 0.7092 | 0.8331 |
| | FTL | 0.709 | 0.2600 | 0.2403 | 0.7242 | 0.7335 |
| | EG | 0.752 | 0.3811 | 0.2868 | 0.6790 | 0.7321 |
| | RW | 0.753 | 0.3777 | 0.4175 | 0.6814 | 0.6569 |
| | Ours | 0.764 | **0.1689** | **0.0753** | **0.7961** | **0.8367** |

4）与准确性最高的 Fairway 方法的对比：虽然 Fairway 方法在所有数据集中取得了最高的准确率，但其通过剔除预测分类错误的样本数据来提升准确性的方式，牺牲了数据的多样性和完整性，并且 Fairway 方法在 $FAT_{EOD}$ 和 $FAT_{AOD}$ 等公平性指标上的表现与本文方法相比明显逊色，这暗示了 Fairway 方法可能通过利用数据中固有的歧视偏见来换取准确性的提升。相比之下，本文方法在保证合理准确性的同时，显著提升了公平性，更加符合实际应用的需求。

5）与其他去偏算法的对比：通过实验结果可以看出，本文提出的去偏算法在公平性和准确性的表现上，均显著优于其他去偏算法。这进一步验证了本文算法在去除数据偏见、提升模型公平性方面的有效性和优越性。

综上所述，本文提出的去偏算法在保持合理准确性的同时，成功实现了对多敏感数据的有效去偏，显著提升了后续模型预测的公平性。

### 3.9 对抗生成式去偏方法的局限性分析

与机器学习算法一样，对抗生成式方法对数据的质量和多样性持有严格要求，这一特性在某种程度上确实限定了其应用的广泛性。然而，在机器学习的实际应用场景中，所采集到的用于模型训练的数据，往往经过了数据预处理与验证，从而在一定程度上确保了其质量和多样性。本文方法在此基础上，对数据进行深度的去相关性优化处理。因此，本文方法的应用范畴，自然而然地与那些适宜于机器学习训练的数据集保持了高度的一致性。

值得注意的是，本文提出的多敏感数据去偏方法，其焦点在于剥离非敏感属性与敏感属性之间的关联性，然而，这一过程中并未将数据的标签纳入考量范畴。具体而言，存在这样一种情况：除敏感属性不同而其他均相同的数据存在决策结果的差异，这显然揭示了偏见问题。对于这种偏见问题，今后需要进行进一步研究并加以解决。

### 3.10 消融实验

为了研究对抗生成式的多敏感数据去偏方法各个模块的作用，以基础 LR 分类模型为基准，$\lambda$ 取值为 1，本文方法实验采用 10 折交叉验证。因为对比实验的训练数据和测试数据划分为 9:1，所以消融实验的训练数据和测试数据也划分为 9:1。为防止过拟合，训练过程中测试集也用作验证集。综上设计了如下实验：

w/o Predict：包含自编码器模块和预测敏感属性模块（去偏模块），编码器受最小化自编码器损失和最大化预测敏感属性网络损失的联合约束优化，以编码器输出作为 LR 输入。

w/o Debias：包含自编码器模块和预测分类网络模块（提高准确性模块），编码器受最小化自编码器损失和最小化预测分类网络损失的联合约束优化，以编码器输出作为 LR 输入。

w/ All：提出的多敏感属性去偏方法，研究自编码器在预测分类网络模块和预测敏感属性模块共同作用的表现。

实验结果如表 3 所列。w/o Predict 与 LR 在 3 个数据集上进行对比，w/o Predict 可以有效提升模型预测的公平性，说明自编码器与预测敏感属性网络的对抗式训练可以有效实现多敏感属性的去偏。w/o Debias 与 LR 对比，w/o Debias 可以有效提升模型预测的准确率，说明预测网络可以有效约束优化编码器，提高使用去偏数据模型的预测性。w/o Predict 与 w/o Debias 相比，w/o Debias 比 w/o Predict 拥有更高的准确率，w/o Predict 比 w/o Debias 拥有更好的公平性，这说明这两部分的侧重点不同，w/o Predict 侧重于提高公平性而不考虑准确性，w/o Debias 侧重于提高准确性而不考虑公平性。w/o Predict 与 w/o Debias 方法在性能上展现了截然不同的特点，这些差异正是它们设计理念的直接体现。具体而言，w/o Predict 方法在无预测分类网络的约束下，专注于剥离敏感属性信息以最大化公平性，但提高公平性的过程中往往牺牲了模型的准确性[32]。其结果是，该方法在公平性方面表现优异，而在准确性方面则相对不足。相反，w/o Debias 方法则完全聚焦于提高模型的准确性，其策略在于通过预测网络和解码器的双重优化编码机制，强化数据中那些对模型预测具有正面贡献的特征，从而有效提升模型的预测性能。然而，这种方法忽视了公平性的考量，导致使用其编码的数据进行模型预测时，公平性结果呈现出不确定性。而 w/ All 方法则巧妙地融合了 w/o Debias 和 w/o Predict 的优势，将准确性和公平性作为两个并行目标进行优化。通过构建对抗性自编码机制，该方法不仅追求模型预测的准确性，同时也兼顾了公平性，实现了两者之间的平衡与协同提升。

从消融实验结果上看，w/ All 与 w/o Debias 相比，去掉去偏模块后，模型的公平性变差，在 3 个数据集上 EOD 指标分别减少了 9.9 个百分点、3.51 个百分点和 1.23 个百分点，

AOD 指标减少 13.15 个百分点、11.54 个百分点和 5.62 个百分点，说明了去偏模块对 w/ All 公平性的重要性。w/ All 与 w/o Predict 相比，去掉提高准确性模块后，模型的准确性在 3 个数据集上分别下降了 3.2 个百分点、1.2 个百分点和 11 个百分点，说明了提高准确性模块对 w/ All 准确性的重要性。结合两者的 w/ All 并没有各取所长，总体上是两者优点的折中，这是因为在追求模型预测准确性和公平性的问题上，提升其中一个指标经常会损害另一个指标，较高的准确性是利用了数据中固有的社会偏见来获得的。但是 w/ All 与 LR，w/o Debias，w/o Predict 在 3 个数据集上相比，总体上有更好的公平性与准确性的平衡，这说明在提高准确性模块和去偏模块的作用下，w/ All 朝着最大化准确性和公平性的方向优化，减轻了实现公平性对准确性的影响。

在去偏问题上，追求在尽可能高准确性的前提下实现公平性。w/o Predict 有较高的公平性，但准确性最差。w/o Debias 的准确性最优，但公平性较低或更差。例如在 Adult 和 Compas 数据集上，AOD 指标差于 LR。w/ All 与 LR 基线相比拥有同一水平甚至更高的准确性，而公平性也得到大幅度的提升，并且拥有更好的准确性与公平性的平衡。因此综合考虑，w/ All 最优。

表 3 消融实验结果

Table 3 Ablation experiment results

| 数据集 | 方法 | 准确率 | EOD | AOD | $FAT_{EOD}$ | $FAT_{AOD}$ |
|---|---|---|---|---|---|---|
| Adult | LR | 0.855 | 0.1546 | 0.2200 | 0.8502 | 0.8158 |
| | w/o Predict | 0.822 | **0.0430** | **0.0720** | 0.8844 | 0.8718 |
| | w/o Debias | **0.862** | 0.1547 | 0.2247 | 0.8536 | 0.8164 |
| | w/All | 0.854 | 0.0557 | 0.0932 | **0.8969** | **0.8796** |
| Compas | LR | 0.682 | 0.2546 | 0.2559 | 0.7123 | 0.7117 |
| | w/o Predict | 0.670 | 0.1216 | **0.0919** | 0.7602 | **0.7711** |
| | w/o Debias | **0.693** | 0.1472 | 0.2641 | 0.7646 | 0.7138 |
| | w/All | 0.682 | **0.1121** | 0.1487 | 0.7714 | 0.7573 |
| German | LR | 0.700 | 0.3566 | 0.2632 | 0.6711 | 0.7179 |
| | w/o Predict | 0.650 | 0.0930 | **0.0526** | 0.7573 | 0.7710 |
| | w/o Debias | **0.770** | 0.0956 | 0.1579 | **0.8318** | 0.8044 |
| | w/All | 0.760 | **0.0833** | 0.1053 | 0.8310 | **0.8219** |

### 3.11 超参数研究

本节对编码器损失函数 $L_E$ 中的 $\lambda$ 超参数进行研究。$\lambda$ 用于控制公平性的程度，本实验通过改变 $\lambda$ 参数的取值，来研究 $\lambda$ 对模型性能的影响。实验设置 $\lambda$ 参数的取值范围为[0，10]，步长为 0.5，使用 Compas 数据集进行实验，得到如图 2 所示的折线图，左侧纵坐标表示准确率，右侧纵坐标表示公平性，横坐标表示 $\lambda$ 的取值。

从图中可以看出，随着 $\lambda$ 的增大，整体上模型的准确性在降低，公平性在提高，这是因为 $\lambda$ 取值越大，公平性对编码器的训练影响也越强，而模型预测准确性对编码器相应的影响就越弱。当公平性降低到一定程度时，模型的准确性和公平性会有所起伏，例如当 $\lambda$ 在[7，10]时，准确性有波动上升的趋势，公平性有波动下降的趋势，这是因为当非敏感属性组与敏感属性组的相关性去除达到极限，此时增大 $\lambda$ 值不会再提升模型的公平性，而会让公平性和准确性在一个范围内波动。本文去偏方法需要调试出最佳 $\lambda$ 取值，以适应不同场景下对公平性和准确性的要求。

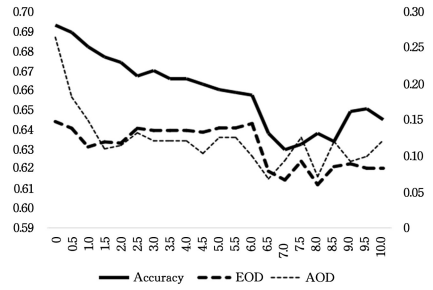


图 2 超参数 $\lambda$ 实验结果

Fig. 2 Experiment results of Hyper-parameter

**结束语** 本文深入探讨了对抗生成式的多敏感属性数据去偏方法，成功构建了一个能够有效去除数据中潜在偏见并同时兼顾公平性与准确性的模型框架。在探索公平性与准确性这一复杂权衡关系的过程中，我们充分认识到两者之间的固有张力，并通过自编码器(生成器)结合预测网络与预测多敏感属性网络(对抗器)的对抗训练策略，有效减轻了增强模型公平性对预测性能的负面影响。实验结果表明，本文方法在多敏感数据二分类任务中表现出色，不仅取得了较高的准确率和公平性，还在两者的平衡上展现了显著优势，明显优于其他去偏方法。未来，将继续深入研究敏感属性、非敏感属性和数据标签之间的关系，解决历史决策导致的标签偏见难题，完善本文方法，同时研究本文方法在不同领域的应用潜力，特别是在医疗、金融等包含复杂敏感属性的数据领域，以期进一步推动数据公平性和可靠性的提升。

## 参 考 文 献

[1] MEHRABI N, MORSTATTER F, SAXENA N, et al. A survey on bias and fairness in machine learning[J]. ACM computing surveys, 2021, 54(6): 1-35.

[2] PEDRESHI D, RUGGIERI S, TURINI F. Discrimination-aware data mining[C]// Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2008: 560-568.

[3] CATON S, HAAS C. Fairness in machine learning: A survey [J]. ACM Computing Surveys, 2024, 56(7): 1-38.

[4] CHEN Z, ZHANG J M, SARRO F, et al. MAAT: a novel ensemble approach to addressing fairness and performance bugs for machine learning software[C] // Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering. 2022: 1122-1134.

[5] KAMIRAN F, CALDERS T. Data preprocessing techniques for classification without discrimination[J]. Knowledge and Information Systems, 2012, 33(1): 1-33.

[6] FELDMAN M, FRIEDLER S A, MOELLER J, et al. Certifying and removing disparate impact[C] // Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2015: 259-268.

[7] ZHANG L, WU Y, WU X. Achieving non-discrimination in data release[C] // Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2017: 1335-1344.

[8] CHAI J,WANG X. Fairness with adaptive weights[C]// International Conference on Machine Learning. PMLR,2022:2853-2866.

[9] LI P,LIU H. Achieving fairness at no utility cost via data reweighing with influence[C]// International Conference on Machine Learning. PMLR,2022:12917-12930.

[10] XU D,YUAN S,ZHANG L,et al. Fairgan:Fairness-aware generative adversarial networks[C]// 2018 IEEE International Conference on Big Data(Big Data). IEEE,2018:570-575.

[11] PETROVIĆA,NIKOLIĆ M,JOVANOVIĆ M,et al. Fair classification via Monte Carlo policy gradient method[J]. Engineering Applications of Artificial Intelligence,2021,104:104398.

[12] ZHANG B H,LEMOINE B,MITCHELL M. Mitigating unwanted biases with adversarial learning[C]// Proceedings of the 2018 AAAI/ACM Conference on AI,Ethics,and Society. 2018:335-340.

[13] WEI S,NIETHAMMER M. The fairness-accuracy Pareto front[J]. Statistical Analysis and Data Mining:The ASA Data Science Journal,2022,15(3):287-302.

[14] HU Z,XU Y,TIAN X. Adaptive priority reweighing for generalizing fairness improvement[C]// International Joint Conference on Neural Networks(IJCNN 2023). IEEE,2023:1-8.

[15] KAMIRAN F,MANSHA S,KARIM A,et al. Exploiting reject option in classification for social discrimination control[J]. Information Sciences,2018,425:18-33.

[16] CHAKRABORTY J,MAJUMDER S,YU Z,et al. Fairway:a way to build fair ML software[C]// Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering. 2020:654-665.

[17] D'ALOISIO G,D'ANGELO A,DI MARCO A,et al. Debiaser for Multiple Variables to enhance fairness in classification tasks[J]. Information Processing & Management,2023,60(2):103226

[18] CANALLI Y,BRAIDA F,ALVIM L,et al. Fair Transition Loss:From label noise robustness to bias mitigation[J]. Knowledge-Based Systems,2024,294:111711.

[19] KIM D,PARK S,HWANG S,et al. Fair classification by loss balancing via fairness-aware batch sampling[J]. Neurocomputing,2023,518:231-241.

[20] KHALILIM M,ZHANG X,ABROSHAN M. Loss balancing for fair supervised learning[C]// International Conference on Machine Learning. PMLR,2023:16271-16290.

[21] LIANG Y,CHEN C,TIAN T,et al. Fair classification via domain adaptation:A dual adversarial learning approach[J]. Frontiers in Big Data,2023,5:129.

[22] GRARI V,LAMPRIER S,DETYNIECKI M. Adversarial learning for counterfactual fairness[J]. Machine Learning,2023,112(3):741-763.

[23] CHEN H,ZHU T,ZHANG T,et al. Privacy and fairness in Federated learning:on the perspective of Tradeoff[J]. ACM Computing Surveys,2023,56(2):1-37.

[24] VUCINICH S,ZHU Q. The Current State and Challenges of Fairness in Federated Learning[J]. IEEE Access,2023,11:80903-80914.

[25] ANGWIN J,LARSON J,MATTU S,et al. Machine bias[M]// Ethics of Data and Analytics. Auerbach Publications,2022:254-264.

[26] ASUNCION A,NEWMAN D. UCI machine learning repository[DB/OL]. https://archive.ics.uci.edu/ml.

[27] FOULDS J R,ISLAM R,KEYA K N,et al. An intersectional definition of fairness[C]// 2020 IEEE 36th International Conference on Data Engineering(ICDE). IEEE,2020:1918-1921.

[28] GHOSH A,GENUIT L,REAGAN M. Characterizing intersectional group fairness with worst-case comparisons[C]// Artificial Intelligence Diversity,Belonging,Equity,and Inclusion. PMLR,2021:22-34.

[29] AGARWAL A,BEYGELZIMER A,DUDÍK M,et al. A reductions approach to fair classification[C]// International Conference on Machine Learning. PMLR,2018:60-69.

[30] BIRD S,DUDÍK M,EDGAR R,et al. Fairlearn:A toolkit for assessing and improving fairness in AI:MSR-TR-2020-32 [R]. Microsoft,2020.

[31] FELDMAN M,FRIEDLER S A,MOELLER J,et al. Certifying and removing disparate impact[C]// Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2015:259-268.

[32] ZAFAR M B,VALERA I,ROGRIGUEZ M G,et al. Fairness-constraints:Mechanisms for fair classification[C]// Artificial Intelligence and Statistics. PMLR,2017:962-970.

**WANG Wenpeng**, born in 1998, postgraduate. His main research interests include recommendation system and machine learning.

**GE Hongwei**, born in 1967, Ph. D, professor, Ph. D supervisor. His main research interests include artificial intelligence, pattern recognition, machine learning, image processing and analysis.

(责任编辑:喻藜)