软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2023,34(9):4037−4055 [doi: 10.13328/j.cnki.jos.006872]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

深度学习模型中的公平性研究*

王昱颖[1,2], 张 敏[1,2], 杨晶然[1,2], 徐晟恺[1,2], 陈仪香[1,2]

[1](华东师范大学 软件工程学院, 上海 200062)

[2](上海市高可信重点实验室 (华东师范大学), 上海 200062)

通信作者: 张敏, E-mail: mzhang@sei.ecnu.edu.cn

摘 要: 近几年深度神经网络正被广泛应用于现实决策系统, 决策系统中的不公平现象会加剧社会不平等, 造成社会危害. 因此研究者们开始对深度学习系统的公平性展开大量研究, 但大部分研究都从群体公平的角度切入, 且这些缓解群体偏见的方法无法保证群体内部的公平. 针对以上问题, 定义两种个体公平率计算方法, 分别为基于输出标签的个体公平率 (IFR_b), 即相似样本对在模型预测中标签相同的概率和基于输出分布的个体公平率 (IFR_p), 即相似样本对的预测分布差异在阈值范围内的概率, 后者是更严格的个体公平. 更进一步, 提出一种提高模型个体公平性的算法 IIFR, 该算法通过余弦相似度计算样本之间的差异程度, 利用相似临界值筛选出满足条件的相似训练样本对, 最后在训练过程中将相似训练样本对的输出差异作为个体公平损失项添加到目标函数中, 惩罚模型输出差异过大的相似训练样本对, 以达到提高模型个体公平性的目的. 实验结果表明, IIFR 算法在个体公平的提升上优于最先进的个体公平提升方法. 此外 IIFR 算法能够在提高模型个体公平性的同时, 较好地维持模型的群体公平性.

关键词: 深度学习; 模型偏见; 个体公平; 群体公平

中图法分类号: TP18

中文引用格式: 王昱颖, 张敏, 杨晶然, 徐晟恺, 陈仪香. 深度学习模型中的公平性研究. 软件学报, 2023, 34(9): 4037–4055. http://www.jos.org.cn/1000-9825/6872.htm

英文引用格式: Wang YY, Zhang M, Yang JR, Xu SK, Chen YX. Research on Fairness in Deep Learning Models. Ruan Jian Xue Bao/Journal of Software, 2023, 34(9): 4037–4055 (in Chinese). http://www.jos.org.cn/1000-9825/6872.htm

Research on Fairness in Deep Learning Models

WANG Yu-Ying[1,2], ZHANG Min[1,2], YANG Jing-Ran[1,2], XU Sheng-Kai[1,2], CHEN Yi-Xiang[1,2]

[1](Software Engineering Institute, East China Normal University, Shanghai 200062, China)

[2](Shanghai Key Laboratory of Trustworthy Computing (East China Normal University), Shanghai 200062, China)

Abstract: In recent years, deep neural networks have been widely employed in real decision-making systems. Unfairness in decision-making systems will exacerbate social inequality and harm society. Therefore, researchers begin to carry out a lot of studies on the fairness of deep learning systems, where as most studies focus on group fairness and cannot guarantee fairness within the group. To this end, this study defines two individual fairness calculation methods. The first one is individual fairness rate IFR_b based on labels of output, which is the probability of having the same predicted label for two similar samples. The second is individual fairness rate IFR_p based on distributions of output, which is the probability of having similar predicted output distribution for two similar samples respectively, and the latter has stricter individual fairness. In addition, this study proposes an algorithm IIFR to improve the individual fairness of these models. The algorithm employs cosine similarity to measure the similarity between samples and then selects similar sample pairs via the similarity threshold decided by different applications. Finally, the output difference of the similar sample pairs is added to the objective function as

* 基金项目: 国家自然科学基金 (61672012); 科技部重点研发项目 (2020AAA0107800); 国家自然科学基金中以国际合作项目 (6216 1146001)
本文由“AI 软件系统工程化技术与规范”专题特约编辑张贺教授、夏鑫博士、蒋振鸣副教授、祝立明教授和李宣东教授推荐.
收稿时间: 2022-08-23; 修改时间: 2022-10-13, 2022-12-14; 采用时间: 2022-12-28; jos 在线出版时间: 2023-01-13
CNKI 网络首发时间: 2023-07-05

an individual fairness loss item during the training, which penalizes the similar training samples with large differences in model output to improve the individual fairness of the model. The experimental results show that the proposed IIFR algorithm outperforms the state-of-the-art methods on individual fairness improvement, and can maintain group fairness of models while improving individual fairness.

Key words: deep learning; model bias; individual fairness; group fairness

近几年深度神经网络发展迅速, 强大的特征提取能力以及高维度数据处理能力让深度神经网络在许多应用中表现出优越的性能. 但随着深度学习应用逐渐渗透到生活中, 比如自动驾驶[1,2]、医疗诊断[3]、欺诈检测[4]等, 研究者们开始追求可信的深度学习模型, 而不是将精度作为模型的唯一度量标准. 公平性作为可信模型中的重要指标, 即确保没有任何群体或者个人因其固有或后天获得的特征而受到偏见或青睐[5]. 由于深度学习模型依赖于数据, 在学习过程中模型可能会有意或无意的偏向某个群体或个人, 从而导致模型中存在偏见. 例如: 使用统计数据预测亚马逊招聘应用[6]中, 出现了 AI 招聘系统对男性求职者更青睐的现象, 这是对性别的歧视; 使用个人行为数据预测累犯概率[7]时, 会显示出错误的、有种族歧视的预测结果, 这是对种族的歧视. 这些带有偏见的现实应用会加剧社会不平等, 造成更严重的社会危害.

为了评估模型公平性, 深度学习领域首先对其定义展开了一系列的研究[8]. Grgić-Hlača 等人[9]首先提出了忽略受保护属性公平, 即在决策过程中不显式地使用受保护属性. 然而这并不是避免歧视的充分条件, 因为数据样本中仍可能存在与受保护属性有着强相关性的其他属性, 如个人住址通常与种族相关. 目前公平性定义大致分为两类, 第 1 类为个体公平[10], 即模型对相似的个体应给出相似的预测, 基于这一类公平性定义的应用, 其挑战在于如何计算个体之间及预测之间的相似度. Zemel 等人[11]和 Lahoti 等人[12]提出了利用聚类的技术寻找相邻样本, 这一方法需事先确定原型样本集合, 原型样本集合的选择会直接影响聚类的结果. Zhang 等人[13]从预处理的角度出发, 通过扰动找到潜在的歧视样本来增强数据集, 再使用较为公平的数据重训练模型. 另一类公平为群体公平[14−17], 该类公平首先需根据特定的受保护属性对样本进行分组, 其次计算不同受保护组在模型预测中的统计数据, 并比较组间差异. 例如, Zafar 等人[15]提出了统计均等, 它要求深度神经网络的预测应独立于受保护属性, 也就是统计均等要求不同群体应具有相似的输出结果. 基于上述公平性定义, 深度学习研究人员分别从 3 个角度来缓解模型中存在的偏见和不公平, 包括预处理机制、处理中机制和后处理机制. 预处理机制是通过修改原始样本来消除数据中有关于受保护属性的信息, 比如因果公平[18−20]. 这种预处理机制需要大量的背景信息, 而这些信息并不总可以访问. 处理中机制仅修改机器学习算法, 如在模型中增加额外的公平性约束来得到公平的样本表示, 以消除算法中存在的偏见, 对抗学习[21−23]是常见的处理中机制. 后处理机制认为歧视性决策通常在决策边界附近, 因此该方法认为可以直接修改模型的输出来提高公平性, 如阈值化[24], 但这类方法较难权衡准确性和公平性.

本文发现大多数公平性研究都基于群体公平展开, 且 Speicher 等人[25]注意到这些缓解群体偏见的方法只处理群体之间的问题, 并没有考虑到群体内部的变化. 例如他们证明了仅使用最小化群体差距的缓解方法会增加群体内部的不公平概率, 从而导致整体不公平性的增加. 因此本文基于“相似的个体应该有相似的结果”这一直觉, 针对公平任务中的分类模型, 通过计算非歧视样本占测试样本的比例得到个体公平率 (individual fairness rate, *IFR*) 包括标签级别的个体公平率 IFR_b 和概率分布级别的个体公平率 IFR_p. 具体而言, 本文针对分类模型定义了相似测试样本对及相似输出结果的判定方法, 其中相似测试样本对指只有受保护属性不同的两个样本. 当相似测试样本对在模型中有不同的预测结果时, 将被判定为歧视样本. 本文提出的个体公平率 IFR_b 和 IFR_p 是不同的相似输出结果的判定方法, 前者使用分类任务中相似测试样本对的输出标签做比较, 后者在前者的基础上基于余弦相似度对相似测试样本对的输出分布作结果判定.

在实验过程中, 本文使用上述指标衡量了不同分类模型的个体公平性, 实验结果与 Grgić-Hlača 等人[9]的研究结论相同, 缓解群体偏见的方法会导致 IFR_p 值的显著下降, 即会损害模型的个体公平性. 因此本文提出了一种提高模型个体公平性的算法 IIFR, 其主要思想是使用正则化技术惩罚存在过大模型预测差异的相似训练样本对, 来达到相似样本具有相似预测结果的目的. 具体做法如下: 首先通过余弦相似度筛选出每批数据中与每个样本最相似的另一个样本, 其次根据数据特征设置的相似临界值 ϵ 过滤不满足相似条件的样本对, 得到相似训练样本对. 再次通过当前模型得到相似训练样本对的输出分布, 利用 JS 散度计算两个输出分布的差异程度, 该差异程度就作为个体公平损失

项添加到目标函数中, 最后通过梯度下降更新模型参数以得到满足个体公平的分类模型. 本文将 IIFR 算法在 3 个真实数据集上进行实验, Adult 数据集中的实验结果表明, IIFR 算法在损失一定正确性的情况下, 可以将个体公平率 IFR_p 从 61.4% 提高到 90.6%, 并将统计均等 (该群体公平指标越小越好) 从 0.192 降至 0.059. 另外, 本文比较了 IIFR 算法与最新的个体公平性提升方法 EIDIG 在基准模型上的性能表现, 还使用 IIFR 算法优化了最新的群体公平性提升方法 CFAIR 的个体公平性性能. 实验结果表明, 在相同的基准模型上, 相较于 EIDIG 方法, IIFR 算法能够更好地提升模型的个体公平性 (尤其是 IFR_p) 和群体公平性. 相较于对抗模型 CFAIR, IIFR 算法优化后的对抗模型不仅能够维持其原有的群体公平性能, 还能够有效地提高模型整体的个体公平性和群体内的个体公平性.

综上, 本文的主要工作包括以下内容.

(1) 定义了两种个体公平率, 分别为要求相似测试样本对输出标签一致的个体公平率 IFR_b, 和要求相似测试样本对输出分布接近的个体公平率 IFR_p.

(2) 提出了一个提高模型个体公平性的算法 IIFR, 该算法基于正则化技术惩罚模型输出差异过大的相似训练样本对, 以达到提高模型个体公平性的目的.

(3) 3 个真实数据集上的实验结果表明 IIFR 算法能够在维持较好的群体公平性下, 有效地提高模型的个体公平性, 从而在个体公平和群体公平间达到较好的权衡.

本文第 1 节介绍缓解个体偏见的相关方法, 及实验中所测试的缓解群体偏见的方法. 第 2 节介绍本文所需的基础知识, 包括深度神经网络和相似度计算方法. 第 3 节介绍本文提出的个体公平率 *IFR* 并给出示例. 第 4 节介绍本文提出的优化模型个体公平性的算法 IIFR. 第 5 节通过实验说明 IIFR 算法的有效性. 最后总结全文.

1 相关工作

● 缓解个体偏见措施. 关于个体公平性的研究较少, 最初 Dwork 等人[10]认为实现群体间的简单统计均等可能会在个体层面上产生直觉上的不公平. 他们提出如果强行保证两个群体在分类模型中预测为积极的概率相等, 那么可能会导致某些原本标签为负类的个体预测为积极的结果, 这是不公平的. 因此一些研究[10,12,26]提出了个体层面上的公平性标准, 这些标准都基于“相似的个体应该得到相似的预测或决定”这一直觉, 但个体公平至今没有统一的标准, 并且 Kearns 等人[27]认为距离函数应该根据具体任务由专家判断得出. 因此文本针对深度学习中的分类模型, 对模型中的输入距离和输出距离定义了具体判定方法. 另一些研究探索了优化模型个体公平性性能的方法, 如 2016 年 Joseph 等人[26]使用强化学习 bandit 方法对公平的选择给予奖励, 对不公平选择给予惩罚, 例如在招聘中, 录用不太合格的申请人而不录用合格的申请人是不公平的. 2019 年 Lahoti 等人[12]使用公平表示对模型去偏, 他们将输入的样本映射到另一表示空间, 目的是去除原始样本中的受保护属性信息. 2021 年 Zhang 等人[13]提出了基于梯度搜索的高效白盒公平性测试方法, 该方法属于缓解模型个体偏见的预处理机制, 通过梯度搜索生成个体差异的测试样例, 并利用生成的单个歧视性实例进行数据增强, 最后重训练原始模型以达到缓解偏见的目的.

● 缓解群体偏见措施. 对抗学习作为处理中机制的重要方法, 是由生成对抗网络框架[28]引起的热潮. 2016 年 Edwards 等人[14]提出了 ALFR 模型来缓解深度学习中的偏见, 该模型基于一个对手网络判断训练过程是否公平, 若不公平则使用对手的反馈改进模型. ALFR 模型对对抗网络的输入较为敏感, 只有平衡的输入才能显著提升模型的公平性[21]. 2018 年 Madras 等人[22]提出了 LAFTR 模型, 该模型将不同的群体公平性度量融合到对抗损失目标函数中, 从而让模型更有针对性. 2020 年 Zhao 等人[29]提出了一种公平表示算法 CFAIR, 该算法扩展了对抗网络模型并使用 BER 计算目标损失, 以同时实现近似的准确率均等和几率均等. 上述对抗模型通过设置分类网络预测标签, 并阻止对手网络预测受保护属性, 这在实践中较难优化[30].

2 基础知识

2.1 深度学习系统

深度学习 (deep learning, DL) 系统一般定义为包括至少一个深度神经网络 (deep neural network, DNN) 的任何

软件系统. DL 系统与传统软件系统在开发过程中的区别如图 1 所示, 两者之间的主要不同点在于开发人员是否直接指定系统逻辑, 在传统软件系统开发过程中, 系统的决策依赖开发人员编写的逻辑, 而 DL 系统的开发人员只需编写数据的处理过程, 确定 DNN 的结构及不断优化 DNN 参数, 通过大批数据训练得到一个具体的 DNN 模型.

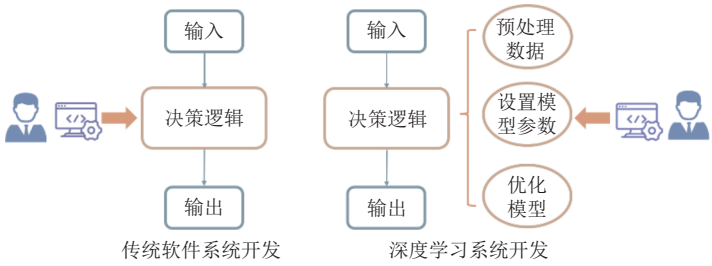

图 1 传统软件系统与 DL 系统对比

DNN 通常包含一个输入层, 多个隐藏层和一个输出层. 如图 2 所示, 从网络层面来看, 每一层的神经元都会与下一层的神经元连接, 其中每个神经元又是一个单独的计算单元, 它们通过不同的权值和激活函数将结果传递给与其连接的下一层的神经元. 从数据层面来看, 输入层接收到数据后, 通过隐藏层提取重要特征, 最后在输出层预测各个类别的概率. 这里对图 2 中的全连接二分类深度神经网络作出形式化定义, 本文将输入空间定义为 $X \subseteq R^n$, 隐藏层对应的表示空间为 $Z \subseteq R^m$, 输出空间为 $Y = \{0,1\}$. 输入 $x \in X$ 表示实例的特征向量, 表示 $z \in Z$ 是实例 x 特征提取后的中间表示, 输出 $y \in Y$ 表示实例 x 对应的类别. 需要注意的是, 对于含有多个隐藏层的 DNN, 其具有多个表示空间, 一般情况下各个表示空间的大小也不相同.

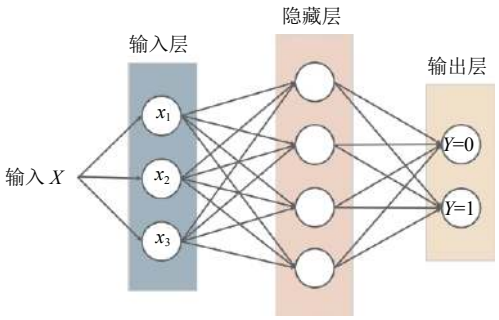

图 2 一个二分类深度神经网络

在经典的全连接 DNN 中, 每一个神经元都与下一层的所有神经元相连接, 每一条边都有一个权值和一个偏置项, 表明神经元之间的连接强度, 即每个特征的重要程度. 每一层的前向传播可表示为:

$$z_i^{(l)} = \varphi\left(\sum_{j=1}^{n} W_{ij}^{(l-1)} z_j^{(l-1)} + b_i^{(l-1)}\right) \tag{1}$$

其中, $\varphi(\cdot)$ 表示激活函数, 常用的激活函数有 ReLU, Sigmoid 和 tanh. $z_i^{(l)}$ 表示第 l 层的第 i 个神经元, $W_{ij}^{(l)}$ 表示第 l 层的第 j 个神经元连接第 $l+1$ 层第 i 个神经元的权值, $b_i^{(l)}$ 表示第 $l+1$ 层第 i 个神经元的偏置项. 这里将 W 和 b 统称为神经网络参数 θ. 综上所述, 一个深度神经网络是建立从输入空间 X 到输出空间 Y 的映射, 可表示为 $f_\theta : X \to Y$. 并且 DNN 的输出是由每个神经元和 θ 中的每个权值共同决定的.

2.2 相似度

余弦相似度也称余弦距离, 它通过计算向量空间中两个向量的余弦值来衡量它们之间的差异程度. 余弦相似度经常用于计算两段文本和两个个体用户的相似度, 计算公式如下:

$$\cos X\cdot Y=\frac{X\cdot Y}{\|X\|\cdot\|Y\|}=\frac{\sum_{i=1}^{n}(x_i\times y_i)}{\sqrt{\sum_{i=1}^{n}(x_i)^2}\times\sqrt{\sum_{i=1}^{n}(y_i)^2}} \tag{2}$$

若余弦值接近 1, 则两个向量的夹角接近 0 度, 表明两个向量越相似; 若余弦值接近 0, 则两个向量的夹角接近 90°, 表明两个向量不相似.

Kullback-Leibler (KL) 散度又称相对熵, 它表示同一随机变量的两个概率分布 P 和 Q 之间的差异. 在大多数机器学习任务中, P 往往表示样本的真实分布, Q 表示模型预测的分布. KL 散度的原理是基于 Q 的编码来编码 P 样本平均所需的比特个数, 计算公式如下:

$$D_{\mathrm{KL}}(P\|Q)=\sum_{i=1}^{n}P(x_i)\log\left(\frac{P(x_i)}{Q(x_i)}\right) \tag{3}$$

当 Q 的分布接近 P 的分布时, 那么 KL 散度指标值小, 即模型的预测较准确. 但考虑到 KL 散度选取不同的编码基准会导致不同的结果, 如 $D_{\mathrm{KL}}(P\|Q)\neq D_{\mathrm{KL}}(Q\|P)$, 而距离度量一般需具有对称性. 因此本文衡量相似样本对输出分布的差异时选用 KL 散度的变体 JS 散度, 其计算公式如下:

$$D_{\mathrm{JS}}(P\|Q)=\frac{1}{2}D_{\mathrm{KL}}\left(P\|\frac{P+Q}{2}\right)+\frac{1}{2}D_{\mathrm{KL}}\left(Q\|\frac{P+Q}{2}\right) \tag{4}$$

JS 散度的取值范围为 [0, 1], 当相似训练样本对 x 和 x' 的输出分布距离 $D_{\mathrm{JS}}(f(x)\|f(x'))$ 为 0 时, 则表明两个输出结果完全相等, 反之则表明两个输出结果差距较大.

3 个体公平率 IFR

本节将介绍本文提出的个体公平率 (*IFR*) 的两种计算方法, 并举例说明该指标的有效性. *IFR* 是针对深度学习中的分类任务设定的, 本文根据 Dwork 等人[10]提出的“相似的个体应该有相似的结果”这一个体公平直觉进行具体化说明和计算方法设计.

3.1 公平性模型的概念表示

在深度学习的公平性相关任务中, 输入空间 X 包含非敏感属性集合 U 及受保护属性集合 A 即 $X=\{U,A\}$. 受保护属性集合定义为 $A\in\{0,\ldots,s-1\}$, 其中 s 表示受保护属性的取值个数, 如性别属性经常作为公平性任务中的受保护属性, 当样本 x 中的属性 $a=0$ 时, 表示该样本为男性数据. 在第 2.1 节中已经介绍深度神经网络是将输入空间 X 映射到输出空间 Y 的函数, 而本文第 3.2 节在计算个体公平率时, 运用 DNN 中输出层的输出结果进行计算, 即使用每个类别的分类概率. 因此本文将二分类模型定义为 $f(x,\theta):X\to P$, 其中 θ 是该模型的权重和偏置项参数, P 表示模型预测的概率空间. 也就是本文的分类模型 f 将一个输入样本 x 映射为一个二维的向量 $p=[p_0,p_1]$, 其中 p_0 和 p_1 分别表示样本 x 属于负类和正类的概率. 最终将向量 p 中最大值的索引作为样本 x 预测结果. 分类模型的预测 $\hat{y}$ 可表示为:

$$\hat{y}=\arg\max p=\arg\max f(x,\theta) \tag{5}$$

3.2 个体公平率计算方法

3.2.1 相似测试个体

个体公平率 *IFR* 作为一项测试指标, 通常使用测试集 $Test=\{(t_i,y_i)\}_{i=1}^{M}$ 对给定模型进行评估, 这里为了区分训练数据和测试数据, 分别使用 x_i 和 t_i 表示. 首先定义测试样本中的“相似个体”. 在测试环节中本文只考虑受保护属性的扰动对模型预测结果的影响, 不希望模型因为受保护属性的改变而做出不一样的决策. 因此本文对测试集中每一个测试样本的受保护属性在其取值范围内进行扰动, 得到一个相似的测试集 $Test'$, 其中每一个样本都与原始测试集中的样本一一对应, 本文将对应的每组样本称为相似测试样本对.

相似测试样本对: 对于测试集 $Test$ 中任意一个样本 $t=\{u,a\}\in X$, 其相似测试样本表示为 $t'=\{u,a'|a'\in A\backslash a\}$. 本文将这样的 t 和 t' 称为相似测试样本对. 测试集 $Test$ 中所有样本的相似样本组成相似测试集 $Test'$.

这里, 给出一对相似测试样本对的具体示例. 对于一个拥有 10 个特征 (包括 9 个非敏感属性和 1 个受保护属性) 和 2 个类别标签的样本数据集 D, 一对相似的测试样本对示例如下:

$$\begin{cases} t:[56,1,9,15,6,2,2,0,5,3] \\ t':[56,1,9,15,6,2,2,1,5,3] \end{cases},$$

其中, 第 8 个属性为受保护属性, 相似测试样本对 (t,t') 只有受保护属性不同.

3.2.2 相似结果的判定方法

本节将定义相似测试样本对的“相似结果”. 本文针对分类任务的预测输出定义了两种相似结果的判定方法. 第 1 种是根据预测的标签进行判断, 当一对相似测试样本对 (t,t') 的预测类别 $(\hat{y},\hat{y}')$ 一致时, 则称 t 和 t' 是一对满足个体公平性的测试样本, 反之则为一对歧视样本. 最后遍历测试集 $Test$ 和 $Test'$, 满足个体公平性的测试样本对数量占测试样本总数的比例即基于输出标签的个体公平率, 记为 IFR_b.

第 2 种相似结果的判定方法进一步严格约束了 IFR_b. 本文考虑到以下情况: 虽然两个相似测试样本对的输出标签一致, 但它们的预测概率相差过大, 例如在一个二分类模型中, 测试样本 t 的预测概率分布为 [0.98, 0.02], 对应的相似测试样本 t' 的预测概率分布为 [0.55, 0.45], 虽然这两个相似测试样本最终都会分为第 1 类, 但它们的预测概率却具有较大差距. 因此第 2 种判断方法将这种预测标签相同但预测概率差距过大的相似测试样本对也定义为歧视样本对.

为了进一步衡量输出概率间的距离, 本文基于 JS 散度计算两个预测概率分布之间的差异. JS 散度能够考虑到所有类别预测结果之间的差距, 这在多分类任务中效果显著, 因为对于两个相似的测试样本, 其相似的预测结果不但追求样本对在预测概率最大的类别中相似, 且在其他类别预测中也追求相似的预测概率. 因此 JS 散度适合于这类判断. 此外, 在使用 JS 散度得到相似测试样本对预测概率之间的相似程度后, 还需针对不同的分类任务设置不同的阈值 τ 来进一步判定是否为相似结果. 第 2.2 节中已经介绍 JS 值接近 0 表明预测结果相似, 因此本文定义当 JS 值大于阈值 τ 时, 则表示两个预测概率分布不相似, 即对应的相似测试样本对是歧视样本对. 同样地, 非歧视样本对数量占测试样本总数的比例就是第 2 种更严格的基于输出分布的个体公平率, 记为 IFR_p. 个体公平率 IFR_p 的具体计算过程如算法 1 所示.

算法 1. 个体公平率计算 IFR_p.

输入: 测试集 $Test=(t_i,y_i)_{i=1}^{M}$, 深度学习模型 f, 模型参数 θ, 阈值 τ;
输出: 个体公平率 IFR_p.

1. $ind\_num=0$ // 初始化满足个体公平性的测试样本对数量
2. $Test'=(t_i',y_i)_{i=1}^{M}$ // 通过扰动样本中的受保护属性得到每个测试样本的相似测试样本
3. **for** $t,t'\in Test,Test'$ **do**
4. $p,p'=f(t,\theta),f(t',\theta)$ // 计算相似样本各自的输出分布
5. $\hat{y},\hat{y}'=\arg\max p,\arg\max p'$
6. **if** $\hat{y}==\hat{y}'$ **then**
7. $js\_dis=D_{\mathrm{JS}}(p\|p')$ // 通过 JS 散度计算相似样本输出分布的距离
8. **if** $js\_dis\leqslant\tau$ **then**
9. // 同时满足预测标签一致, 并且所有类别预测概率相近
$ind\_num=ind\_num+1$
10. **end if**
11. **end if**
12. **end for**
13. **return** ind_num/M

4 提高模型个体公平性算法 IIFR

为了缓解模型中存在的个体偏见, 本文提出了一种提高深度学习模型个体公平性的算法 (improved individual fairness rate, IIFR). IIFR 是基于真实的训练样本数据, 在训练过程中使用余弦相似度找到与每个训练样本最相似的另一个训练样本. 再根据不同数据集拥有的特征确定一个相似界限 ϵ, 使用 ϵ 去除一些不满足相似要求的训练样本对. 最终使用 JS 散度计算所有相似训练样本对的输出差异和作为模型的个体公平损失添加到目标函数中, 使用梯度下降进行参数更新, 以提高模型的个体公平性. IIFR 算法的流程如图 3 所示.

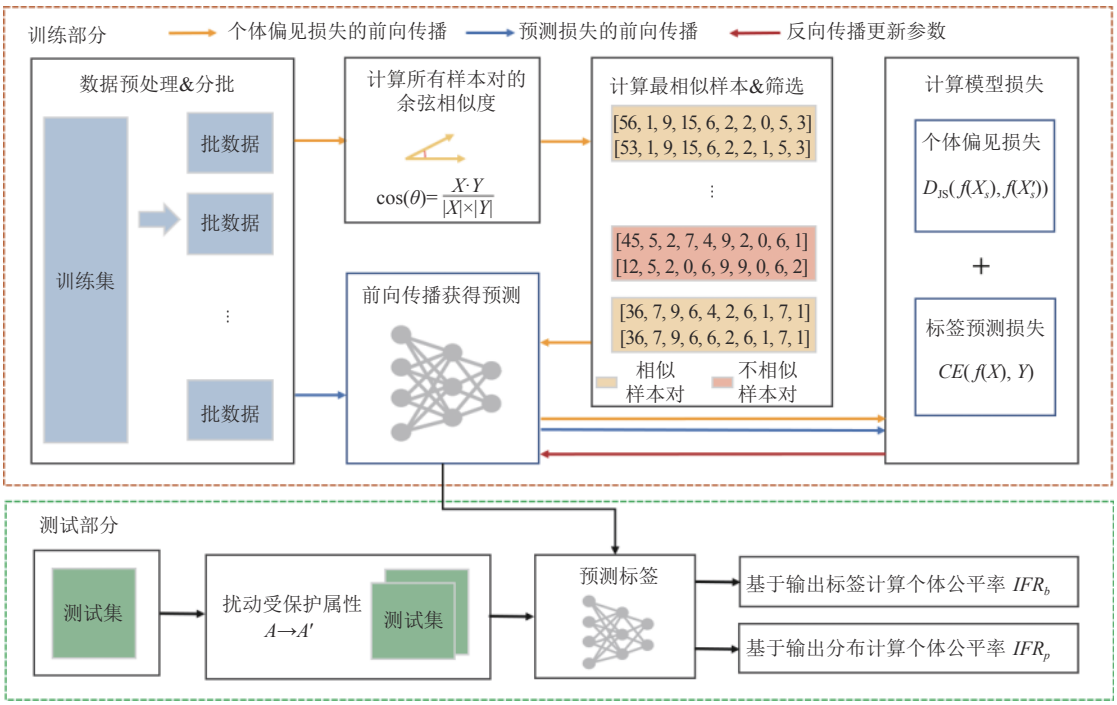

图 3 提高个体公平率算法流程

4.1 相似训练个体

不同于测试部分的相似个体, 本文在训练过程中允许个体样本在所有特征上进行扰动, 但为了防止扰动造成无效样本或不合理样本的情况, 本文将训练过程中的相似样本对定义为满足一定条件的最相似训练样本对, 即在训练数据中找到训练样本 x 最相似的另一个训练样本 x', 并根据它们的相似度进行筛选. 具体方法如图 3 所示, 第 1 步对原始的结构化数据 (特征均为连续变量类型或者类别数据类型) 进行预处理操作, 通过读热码技术和归一化将其转换为向量表示, 合理的向量表示有利于确定样本特征之间的距离. 其次需对训练集进行分批操作以防大量数据在训练过程中造成大量内存的消耗. 第 2 步是利用余弦相似度计算每批数据中所有样本之间的距离来得到最相似训练样本对. 转化为向量表示的两个样本可以直接使用 cos 相似度得到表示个体相似程度的余弦值. 第 2.2 节中已介绍余弦值越大表示样本越相似. 具体而言, 在一批训练数据中, 对于一个样本 x, 其最相似的训练样本 x' 是该批数据中与 x 的余弦值最大的另一个样本, 本文将这样的 (x,x') 称为最相似训练样本对, 需要注意的是每一个训练样本都有其对应的最相似训练样本, 即一批数据中有 k 个样本就存在 k 个最相似训练样本对. 利用余弦值判定两个样本的相似程度是受自然语言处理领域中文本相似度计算[31,32]启发, 文本相似度的计算是将给定的两个文本根据合并的词列表转化为两个词频向量, 对于两个相似的词频向量, 余弦值会趋近于 1, 即两个文本具有较高的相似度, 当余弦值趋于 0 则表明两个文本不具有相似性. 第 3 步是对最相似训练样本对进行进一步筛选, 因为一批数

据是有限且随机的, 即使是样本 x 最相似训练样本 x', 它们也可能不满足相似要求, 因此本文使用相似界限 ϵ 来对相似训练样本作进一步限制, 即只有余弦值大于 ϵ 的最相似样本对才能用来指导模型提高个体公平性.

相似训练样本对: 对于一批训练数据中任意一个样本 $x=\{u,a\}\in X$, 其最相似训练样本是同批数据中与 x 余弦值最大的另一个训练样本, 最相似训练样本对记为 (x,x'), 其中 $x'=X^{(i)}[\arg\max\{\cos x\cdot x'|x'\in X^{(i)}\backslash x\}]$, $X^{(i)}$ 表示数据集中的第 i 批数据. 当最相似训练样本对的余弦值大于相似界限 ϵ 时, 即 $\cos x\cdot x'>\epsilon$, 这样的 (x,x') 为相似训练样本对, 记为 (x_s,x_s').

这里使用 3 个数据样本对的余弦相似度进行举例说明 (为了简化样本表示, 这里的样本没有转化为读热码形式和进行归一化操作), 且假设相似界限 ϵ 为 0.9.

$$\begin{cases}x_1:[56,1,9,15,6,2,2,0,5,3]\\ x_2:[56,1,9,15,6,2,2,1,5,3]\\ x_3:[26,0,5,40,9,2,2,1,5,3]\end{cases},$$

在上述 3 个数据中, x_2 是 x_1 的最相似训练样本, 其余弦相似度为 0.999, 大于相似界限 ϵ, 本文将这样的两个训练样本称为相似训练样本对; x_2 是 x_3 的最相似训练样本, 但其余弦相似度为 0.752, 小于相似界限 ϵ, 本文将这样的两个训练样本称为不相似训练样本对.

4.2 模型训练

IIFR 算法分别从两个方面训练模型, 第 1 个方面是从正确性的角度出发, 通过不断缩小预测标签和真实标签之间的差距来完成分类任务. 具体流程为图 3 中蓝色线部分, 将预处理和分批后的数据向量输入到深度神经网络中, 进行前向传播, 在网络的输出层得到样本预测, 最后通过交叉熵 CE[33]计算样本预测与真实标签之间的差距, 得到标签预测损失 $Loss_{\text{pred}}$.

$$Loss_{\text{pred}}=CE(f(X^{(i)},\theta),Y^{(i)}) \tag{6}$$

另一方面是从个体公平性的角度出发, 通过不断缩小相似训练样本对的预测结果来达到“相似的个体应该有相似的结果”这一目的, 即达到个体公平的要求. 具体流程为图 3 中橙色线部分, 通过第 4.1 节的相似训练个体计算方法得到相似训练样本对, 将相似训练样本对输入到深度神经网络中, 分别得到训练样本 x 的预测概率 p 和相似训练样本 x' 的预测概率 p', 其次使用 JS 散度计算预测概率 p 和 p' 之间的差距, 得到该对相似训练样本的偏见值, 最后所有相似训练样本对的预测差距之和即个体偏见损失 $Loss_{\text{indi}}$.

$$Loss_{\text{indi}}=D_{\text{JS}}(f(X_S^{(i)},\theta)\|f(X_S^{(i')},\theta)) \tag{7}$$

本文基于正则化技术利用个体损失权重参数 λ 权衡上述两种损失, 即训练模型时的目标函数为 $Loss_{\text{pred}}+\lambda\cdot Loss_{\text{indi}}$, 通过 AdaDelta 梯度下降法[34], 分别降低标签预测损失和个体偏见损失, 并不断地进行参数优化直至收敛, 最终得到具有较好正确性和个体公平性的深度学习模型. 其中参数优化过程如下.

$$\theta=\theta-lr\cdot\frac{\partial Loss_{\text{pred}}+\lambda\cdot Loss_{\text{indi}}}{\partial\theta} \tag{8}$$

IIFR 算法如算法 2 所示.

算法 2. 提高个体公平性算法 IIFR.

输入: 训练集 $Train=\{(x_i,y_i)\}_{i=1}^{N}$, 最大训练次数 max_epoch, 批大小 $batch\_size$, 正则化参数 λ, 阈值 ϵ;

1. **for** *epoch* from 0 to *max_epoch* **do**
2. $L=N/batch\_size$
3. 将训练集分为 L 批数据, 分别为 $X^{(0)},\ldots,X^{(L)}$
4. **for** i from 0 to L **do**
5. $Loss_{\text{indi}}=0$
6. **for** x in $X^{(i)}$ **do**

$similar\_x = X^{(i)}[\arg\max\{\cos x \cdot x' | x' \in X^{(i)} \backslash x\}]$

if $\cos x \cdot similar\_x > \epsilon$ **then**

$Loss_{\text{indi}} = Loss_{\text{indi}} + D_{\text{JS}}(f(x,\theta)||f(similar\_x,\theta))$

end for

$Loss_{\text{pred}} = CE\left(f\left(X^{(i)}\right), Y^{(i)}\right)$

//根据以下目标函数更新模型参数

$L = Loss_{\text{pred}} + \lambda \cdot Loss_{\text{indi}}$

end for

end for

5 实验分析

在本节中, 本文将在 3 个流行数据集 Adult、COMPAS 和 German 上评估本研究提出的 2 种个体公平衡量指标 IFR_b 和 IFR_p 以及提高模型个体公平性的算法 IIFR 的有效性. 本部分的实验研究了以下 4 个问题.

- RQ1: 不同基准模型及公平性提升方法在个体公平衡量指标 IFR_b 和 IFR_p 上表现如何?
- RQ2: IIFR 算法是否能提高模型的个体公平性? 与 EIDIG 方法相比 IIFR 算法优化的模型性能是否更好?
- RQ3: IIFR 算法能否缓解对抗模型造成的群体内部的不公平?
- RQ4: 针对个体公平的 EIDIG 方法和 IIFR 算法在群体公平指标上表现如何? 应用 IIFR 算法缓解对抗模型群体内部的不公平后, 新模型的群体公平性有何变化?

5.1 实验设置

5.1.1 实验数据集

本文在公开的公平性数据集上进行实验, 包括以性别作为敏感属性的 Adult 数据集、以种族作为敏感属性的 COMPAS 数据集和以年龄作为敏感属性的 German 数据集. 它们是公平性测试研究中最常用的结构化数据集. 表 1 和表 2 分别给出了 3 个数据集的详细信息.

表 1 二类的敏感属性 Adult 和 COMPAS 数据集的统计信息

数据集	样本数	$P(A=0)$	$P(Y=1)$	$P(Y=1\|A=0)$	$P(Y=1\|A=1)$
Adult	45 222	0.675	0.248	0.312	0.114
COMPAS	6 172	0.486	0.455	0.383	0.523

表 2 多类的敏感属性 German 数据集的统计信息

样本数	$P(A=0)$	$P(A=1)$	$P(A=2)$	$P(A=3)$	$P(A=4)$
1 000	0.190	0.398	0.226	0.115	0.071
$P(Y=1)$	$P(Y=1\|A=0)$	$P(Y=1\|A=1)$	$P(Y=1\|A=2)$	$P(Y=1\|A=3)$	$P(Y=1\|A=4)$
0.3	0.421	0.296	0.243	0.243	0.268

Adult 数据集是 1994 年美国人口普查数据库中的人口统计数据 (https://archive.ics.uci.edu/dataset/2/adult), 每条数据包括职业、性别、受教育程度等 14 个属性, 其中受保护属性是性别, 包括男性 (A=0), 女性 (A=1). 其标签表示每个人每年的收入是否超过 50k. 该数据集存在严重的数据偏斜, 比如: 67.5% 的数据为男性, 且男性数据中 31.2% 的男性每年收入超过 50k, 而女性数据中收入超过 50k 的只有 11.4%. 此外 Adult 数据集的标签分布也不均衡, 仅 24.8% 的人具有较高的工资.

COMPAS 数据集是美国佛罗里达州布劳沃德县的被告记录 (https://github.com/propublica/compas-analysis), 记

录着每个被告的先前犯罪次数、种族、年龄等 12 个属性, 其中受保护属性是种族, 包括白种人 (A=0), 黑种人 (A=1). 其任务是预测被告在两年内是否会再次犯罪. 从表 1 可知, COMPAS 数据集在受保护属性和标签的分布上都较为均衡.

German 数据集是由德国 Hofmann 博士收集制作的德国信用卡数据 (https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data), 每条数据包括拥有者的年龄、当前就业状态、历史信用记录、房产状态等 20 个属性, 其中受保护属性是年龄, 包括 19–75 岁, 本文根据年龄的分布情况, 将其分类为 19–25 岁 (A=0), 25–35 岁 (A=1), 35–45 岁 (A=2), 45–55 岁 (A=3), 55–75 岁 (A=4). 该数据集的任务是预测信用卡的拥有者是否会称为潜在的坏用户即违约用户. 从表 2 可知 German 数据集在标签分布上较不均衡, 但在受保护属性上分布较为均衡.

5.1.2 评估指标

本文主要从正确性和公平性两个方面评估模型的性能. 与之前的工作相同[12,22], 本文采用 ACC 来衡量模型的正确性, 说明模型的预测表现, 正确性计算公式如下:

$$ACC = \frac{TP+TN}{TP+FP+FN+TN} \tag{9}$$

其中, TP 指模型预测结果中, 被正确识别的正实例数量, TN 表示模型正确识别负实例的数量, FP 和 FN 则分别表示模型将负实例和正实例识别错误的数量.

在公平性方面, 本文使用个体公平性与群体公平性两类公平性指标来考察模型性能. 个体公平性采用的是本文提出的基于输出标签的个体公平率 IFR_b 和基于输出分布的个体公平率 IFR_p, 这两个度量指标在第 3 节中已详细介绍.

群体公平性指标同样与之前的工作[22,29]相同, 采用统计均等 DP 和几率均等 EO 来衡量模型的群体公平性. 统计均等 (statistical parity/demographic parity, DP) [15]要求输出的预测 $\hat{Y}$ 和受保护属性 A 相独立, 即统计均等保证不同群体间输出为正类的概率相等:

$$P(\hat{Y}=1|A=0)=P(\hat{Y}=1|A=1)=\ldots=P(\hat{Y}=1|A=s) \tag{10}$$

但该指标在偏斜的真实数据上不可能达到绝对相等, 这是由历史数据本身的分布不均衡造成的. 因此在之前的工作中通常使用这两个概率的差值作为评估模型公平性的标准, 记作 ΔDP . ΔDP 的值越小表明模型越公平. 本文针对多类的敏感属性数据集, 将 ΔDP 定义为所有群体间的 $P\left(\hat{Y}=1|A=a\right)$ 的标准差.

几率均等 (equalized odds/positive rate parity, EOdd)[16]要求输出的预测 $\hat{Y}$ 和受保护属性 A 在所有类别 Y 下条件独立, 换言之, 不同群体间的输出需具有相同的假阳性和真阳性:

$$P(\hat{Y}=1|A=0,Y=y)=P(\hat{Y}=1|A=1,Y=y)=\ldots=P(\hat{Y}=1|A=s,Y=y),\forall y\in\{0,1\} \tag{11}$$

本文使用正类样本和负类样本在不同群体间的预测差值作为群体公平性评估指标, 分别记作 $\Delta EO_{y=0}$ 和 $\Delta EO_{y=1}$. 同样 $\Delta EO_{y=0}$ 和 $\Delta EO_{y=1}$ 的值越小表明模型越公平. 针对多类的敏感属性数据集, $\Delta EO_{y=0}$ 和 $\Delta EO_{y=1}$ 分别为所有群体间的 $P\left(\hat{Y}=1|A=a,Y=0\right)$ 的标准差和 $P\left(\hat{Y}=1|A=a,Y=1\right)$ 的标准差.

5.1.3 实验模型

本节将介绍本文使用的 4 个实验模型, 包括 1 个未使用任何公平性算法的基准模型 MLP 和 3 个缓解群体偏见的对抗模型 ALFR, LAFTR, CFAIR. 其中 MLP 模型作为基准模型, 给出无偏见措施的性能基准, 用于衡量 IIRF 算法的有效性, 并与领域最新的个体公平性提升方法 EIDIG 进行比较, 即在同一个 MLP 基准模型上分别使用 IIFR 算法以及 EIDIG 算法进行重训练后, 两个最终的优化模型的性能比较. 其他模型作为群体公平模型, 用于研究 IIFR 算法能否缓解群体及群体内部的偏见.

4 个模型的网络信息如下所示.

- MLP: 多层的感知机模型, 即带有 ReLU 激活函数的多层全连接网络. 该模型使用交叉熵损失进行训练.
- ALFR[14]: 多任务对抗模型, 即带有一个共享隐藏层和两个分类层的网络, 其中两个分类层分别为最小化分类任务预测损失的全连接层和最小化受保护属性预测损失的全连接层. 该模型的两个任务均使用交叉熵损失进行训练.
- LAFTR[22]: 多任务对抗模型, 与 ALFR 结构相似, 区别在于该模型将不同的群体公平度量融合到对抗损失目

标函数中 (即融合到预测受保护属性的损失中), 且在预测类别的任务中使用交叉熵损失, 在预测受保护属性的任务中使用 L1 损失.

● CFAIR[29]: 多任务对抗模型, 带有一个共享隐藏层和 $|A|+1$ 个分类层的网络, 分类层包含一个最小化分类任务预测损失的全连接层和 $|A|$ 个最小化某个群体内的受保护属性预测损失的全连接层. 该模型的任务均使用交叉熵损失进行训练.

3 个数据集 Adult、COMPAS 和 German 的其他参数设置如表 3 所示.

表 3 超参数设置

参数	Adult	COMPAS	German
基准模型结构	[114, 30, 20, 15, 15, 10, 2]	[11, 30, 10, 5, 2]	[30, 50, 30, 15, 10, 5, 2]
优化算法	AdaDelta	AdaDelta	AdaDelta
学习率	0.1	1	1
批大小	256	128	30
训练轮数	10	20	5
相似测试结果阈值 τ	0.001	0.001	0.001
相似训练样本界限	0.8	0.8	0.8

5.2 实验结果与分析

5.2.1 IFR 指标和 IIFR 算法的有效性评估

RQ1: 不同基准模型及公平性提升方法在个体公平衡量指标 IFR_b 和 IFR_p 上表现如何?

为了回答这个问题, 本文比较了基准感知机模型 MLP、针对个体公平的 EIDIG 方法和针对群体公平的 3 个对抗模型的个体公平性性能. 一个模型的个体公平性能由个体公平率指标体现, 即 IFR_b 和 IFR_p 的值越高, 则表明模型的个体公平性越好, 相反则表明该模型存在严重的个体歧视. 表 4 给出了 4 个基准模型及 EIDIG 的预测表现, 对于未使用 IIFR 算法的模型, 训练后得到的预测模型中存在大量预测标签相同但预测概率相差较大的相似测试样本对. 例如基于 Adult 数据集训练的 MLP 模型中, 标签相同的相似测试样本对占 97.3%, 然而这些测试样本中只有 61.4% 的样本满足预测概率相近, 这说明 MLP 模型对相似样本仍然存在歧视的情况. 甚至这种歧视在对抗模型中更为严重, 表明缓解群体偏见的对抗网络会进一步放大模型的个体偏见. 例如 ALFR 对抗模型的个体公平率 IFR_b 可以达到 87.1%, 但在更严格的 IFR_p 指标下, 其个体公平率下降至 30.5%. 其他对抗模型也存在不同程度的个体歧视. 因此相似训练样本对在模型中的预测概率相近是提高模型个体公平性的必要条件. 使用 EIDIG 方法优化的 MLP 模型的个体公平率 IFR_b 相较于原模型均有提升, 但该方法在 IFR_p 指标上表现并不稳定, 说明 EIDIG 优化方法训练的模型也可能导致潜在的歧视.

RQ2: IIFR 算法是否能提高模型的个体公平性? 与 EIDIG 方法相比 IIFR 算法优化的模型性能是否更好?

为了回答这个问题, 本文对比了 4 个基准模型采用 IIFR 算法前后的性能, 并比较了在同一基准模型 MLP 上使用 IIFR 算法与 EIDIG 方法得到的优化模型的性能. 从基准模型的角度看, 表 4 展现出模型使用 IIFR 算法后, 它们的个体公平性均有明显提升. 例如在 COMPAS 数据集中, MLP 模型的个体公平率 IFR_b 指标从 93.7% 提升至 95.1%, 且更严格的 IFR_p 指标从 73.5% 提升至 93.4%. 对于 ALFR 等基准对抗模型, 本文的 IIFR 算法同样能帮助其提升个体公平性性能, 且在 IFR_p 指标上表现显著. 例如 LAFTR 模型在无个体公平算法时, IFR_p 值为 59.4%, 这说明只有 59.4% 的相似测试样本对满足相近的预测概率分布. 但使用本文提出的 IIFR 算法后, 满足 IFR_p 个体公平性的测试样本数量达到了 82.6%. 在 German 数据集中, 基准模型 MLP 已具有较好的个体公平性, 而本文的 IIFR 算法可以在较小的准确率损失下, 进一步提高模型的个体公平性性能. 表 4 中 EIDIG 方法与 IIFR 算法的对比展现了 EIDIG 方法能够在正确性指标上优于本文的 IIFR 算法, 但在个体公平性指标上本文的 IIFR 算法对模型的提升更优. 另外 EIDIG 方法在 IFR_p 指标上表现并不稳定, 这是由于修正个体歧视实例标签仍是从标签的层面进行优化, 该方法并未考虑歧视样本对的输出分布的差异. 而本文的 IIFR 方法则是从输出分布的角度出发, 有效地近似相似训练样本对的输出分布, 达到提高个体公平性的目的.

表 4 不同模型在使用 IIFR 算法前后的正确性和个体公平性对比 (%)

数据集	模型	参数	*ACC*	个体公平	
				IFR_b	IFR_p
Adult (性别)	基准模型	MLP	85.5	97.3	61.4
	基准对抗模型	ALFR	83.6	87.1	30.5
		LAFTR	83.1	88.3	25.3
		CFAIR	84.4	92.0	41.5
	对比模型	MLP+EIDIG	**84.1**	98.5	82.0
	基准模型+IIFR算法	MLP+IIFR	80.1	**98.6**	**90.6**
		ALFR+IIFR	80.3	97.6	61.5
		LAFTR+IIFR	79.9	96.6	43.9
		CFAIR+IIFR	82.0	98.3	83.7
COMPAS (种族)	基准模型	MLP	68.4	93.7	73.5
	基准对抗模型	ALFR	67.0	79.9	9.5
		LAFTR	68.9	79.9	59.4
		CFAIR	68.3	84.7	29.9
	对比模型	MLP+EIDIG	**68.2**	94.3	68.2
	基准模型+IIFR算法	MLP+IIFR	64.4	**95.1**	**93.4**
		ALFR+IIFR	63.9	78.6	45.7
		LAFTR+IIFR	64.4	82.6	82.6
		CFAIR+IIFR	67.8	85.4	79.4
German (年龄)	基准模型	MLP	81.5	98.0	93.0
	基准对抗模型	ALFR	81.2	96.0	73.5
		LAFTR	80.5	97.3	90.7
		CFAIR	81.2	97.7	82.3
	对比模型	MLP+EIDIG	77.0	94.8	55.0
	基准模型+IIFR算法	MLP+IIFR	78.8	98.8	**98.8**
		ALFR+IIFR	**79.5**	98.8	76.0
		LAFTR+IIFR	78.8	99.0	98.0
		CFAIR+IIFR	78.5	**100.0**	89.8

由于 EIDIG 通过修正个体歧视实例的标签, 让模型在训练过程中学习更加公平的样本, 从而不需要更改预测结果来缓解模型偏见. 因此该方法能够在有限的精度损失下提升其个体公平性. 而文本的方法并未对数据集进行修正, IIFR 算法通过近似原数据集中的相似样本的输出分布来获得公平的输出分类, 从而达到缓解模型偏见的目的, 但该公平操作可能会导致预测与样本标签不一致. 本文从数据集的角度进行分析, 通过 cos 相似度找到 Adult 数据集中相似训练样本, 表 5 展示了一对相似但标签不同的训练样本.

表 5 Adult 数据集中带有歧视的一对相似训练样本

年龄	工作性质	受教育程度	受教育时间	婚姻状态	工作	亲属关系	种族	性别	资本盈利	资本损失	工作时长 (h)	国家	收入
21	私人	大学未毕业	10	未婚	销售	未婚	白种人	**女性**	0	0	**45**	美国	≤**50K**
22	私人	大学未毕业	10	未婚	销售	未婚	白种人	**男性**	0	0	**25**	美国	>**50K**

一位 21 岁的一周工作时长为 45 小时的女大学生, 其收入低于工作时长更短的 22 岁男大学生. 由于 MLP 模型是以实际收入和预测收入的差异作为损失函数, 因此 MLP 模型将该 21 岁女性分类为≤50K, 将 22 岁男性分类为>50K. 而在本文提出的 IIFR 算法下, 这两个样本的余弦相似度为 0.808, 在训练过程中会将他们识别为相似训练样本, IIFR 为了消除这两个样本的偏见, 通过近似他们的输出分布以提高模型个体公平性, 最终使得两者均分为≤50K 这一类别. 因此这导致了一个错误的样本分类, 造成模型正确性的下降.

RQ3: IIFR 算法能否缓解对抗模型造成的群体内部的不公平?

表 6 给出了不同模型基于 Adult 数据集的群体间的个体公平性表现, 其中 CFAIR 等对抗模型的个体公平率低于无公平措施的 MLP 模型, 这说明对抗模型会增加群体内部的歧视, 尤其增加了女性群体内部的歧视. EIDIG 方法和 IIFR 算法均可以很大程度上缓解了群体内部的不公平. 结合表 4 的整体性能展示, 这两种个体公平方法均能有效提升模型的整体的个体公平性和群体内的个体公平性. 并且本文的 IIFR 算法在 IFR_p 指标上表现更优. 在 CFAIR 模型中, 相较 IIFR 算法对男性群体内部的 IFR_p 从 30.5% 提升至 73.4%, 女性群体内的 IFR_p 只提升了 19%, 这说明女性群体内部仍存在不公平, 本研究认为这是由数据偏斜造成的, 应对数据集进行重采样[35]、类别均衡采样[36]等操作来得到高质量的训练数据.

表 6 不同模型在使用 IIFR 算法前后的群体内部的个体公平性对比 (%)

数据集	参数	男性		女性	
		IFR_b	IFR_p	IFR_b	IFR_p
Adult (无IIFR)	MLP	97.4	71.2	97.3	41.5
	ALFR	83.8	19.6	86.8	5.8
	LAFTR	88.3	28.3	88.3	19.3
	CFAIR	85.0	30.5	91.4	18.0
	MLP+EIDIG	97.9	89.2	99.6	78.4
Adult (有IIFR)	MLP+IIFR	98.4	94.4	98.8	82.8
	ALFR+IIFR	97.7	73.4	97.1	37.0
	LAFTR+IIFR	96.9	54.5	95.9	22.1
	CFAIR+IIFR	97.7	73.4	97.1	37.0

5.2.2 IIFR 算法对模型群体公平性的影响

RQ4: 针对个体公平的 EIDIG 方法和 IIFR 算法在群体公平指标上表现如何? 应用 IIFR 算法缓解对抗模型群体内部的不公平后, 新模型的群体公平性有何变化?

为了回答这个问题, 本文比较了在同一基准模型 MLP 上分别采用 EIDIG 方法与 IIFR 算法得到的优化模型的群体公平性表现, 并且比较了原对抗模型和使用 IIFR 算法优化个体公平性后的对抗模型的群体公平性表现. 一个模型的群体公平性能主要表现在不同群体间预测结果的相似度以及同一类别下不同群体间预测结果的相似度, 即 ΔDP 和 ΔEO 越接近 0, 则表明模型的群体公平性越好, 反之则表明该模型存在群体偏见, 如性别偏见.

最近的几项[17,37]研究表明, 不同的公平性概念间存在不相容性. 例如, 当组间的基本比例不相等时, 则各群体间不可能同时满足相同的假阳性率和相同的假阴性率. 由于本研究群体间的样本基数不同, 群体公平指标 DP 和 EO 同样具有不可能调和性[17], 但在一个极端的情况下, 群体公平可以达到完全均等, 如模型将所有的样本均分为正类, 则 ΔDP 与 ΔEO 均为 0, 而这种情况破坏了模型的分类性能, 具体表现为模型的正确率 ACC 等于模型标签为正类的比例. 因此模型在追求公平性的同时, 需保证一定的正确性及防止各群体的假阳性指标全为 0 或 1 的情况. 结合表 1、表 2 的样本分布和表 4 中模型的正确性数据, 说明了本文的 IIFR 算法仍具有分类能力.

相较于 EIDIG, IIFR 算法较好的平衡不同群体公平指标间的性能. 例如在 COMPAS 数据集中, EIDIG 方法仅显著提升了 $\Delta EO_{y=0}$. 而 IIFR 算法可将 MLP 基准模型的统计均等 ΔDP 从 0.325 降至 0.155, 将几率均等 $\Delta EO_{y=0}$ 从 0.303 降至 0.087, $\Delta EO_{y=1}$ 从 0.258 降至 0.177, 该性能提升表明 IIFR 算法能够显著降低白人群体和黑人群体被预测为累犯之间的差异使模型能够更公平的预测白人和黑人的行为. 在 Adult 和 German 数据集中, IIFR 算法相比于 EIDIG, 能够更好地权衡不同群体公平指标. 后文表 7 中 IIFR 算法在对抗模型中的表现说明了该算法在提升对抗模型个体公平性的同时, 能够保持对抗模型对原模型群体公平的提升.

5.2.3 个体损失权重参数 λ 对模型的影响

为了充分讨论 IIFR 算法受模型参数的影响, 本文使用不同的个体损失权重参数 λ (算法 2 第 12 行中的参数)进行实验, 观察模型性能变化趋势. 图 4–图 6 分别展示了 IIFR 算法在 Adult、COMPAS、German 数据集上, 变化其个体损失权重 λ 对模型的正确性、个体公平性以及群体公平性的影响.

表 7 不同模型在使用 IIFR 算法前后的群体公平性对比

数据集	模型	参数	群体公平		
			ΔDP	$\Delta EO_{y=0}$	$\Delta EO_{y=1}$
Adult (性别)	基准模型	MLP	0.192	0.083	0.038
	基准对抗模型	ALFR	0.040	0.038	0.272
		LAFTR	0.023	0.046	0.296
		CFAIR	0.111	0.014	0.078
	对比模型	MLP+EIDIG	0.027	0.088	0.159
	基准模型+IIFR算法	MLP+IIFR	0.059	0.011	0.063
		ALFR+IIFR	0.015	**0.005**	0.192
		LAFTR+IIFR	**0.004**	0.014	0.259
		CFAIR+IIFR	0.082	0.016	**0.032**
COMPAS (种族)	基准模型	MLP	0.325	0.303	0.258
	基准对抗模型	ALFR	0.041	0.010	0.007
		LAFTR	0.009	0.019	0.028
		CFAIR	0.084	0.014	0.064
	对比模型	MLP+EIDIG	0.266	0.187	0.265
	基准模型+IIFR算法	MLP+IIFR	0.155	0.087	0.177
		ALFR+IIFR	0.028	0.040	0.087
		LAFTR+IIFR	**0.009**	**0.019**	**0.027**
		CFAIR+IIFR	0.076	0.034	0.052
German (年龄)	基准模型	MLP	0.038	0.051	0.074
	基准对抗模型	ALFR	0.068	0.077	0.056
		LAFTR	0.037	0.052	0.064
		CFAIR	0.035	0.030	0.064
	对比模型	MLP+EIDIG	0.127	0.073	0.229
	基准模型+IIFR算法	MLP+IIFR	0.028	0.062	0.084
		ALFR+IIFR	0.043	0.046	0.065
		LAFTR+IIFR	0.019	0.031	0.068
		CFAIR+IIFR	**0.017**	**0.004**	**0.045**

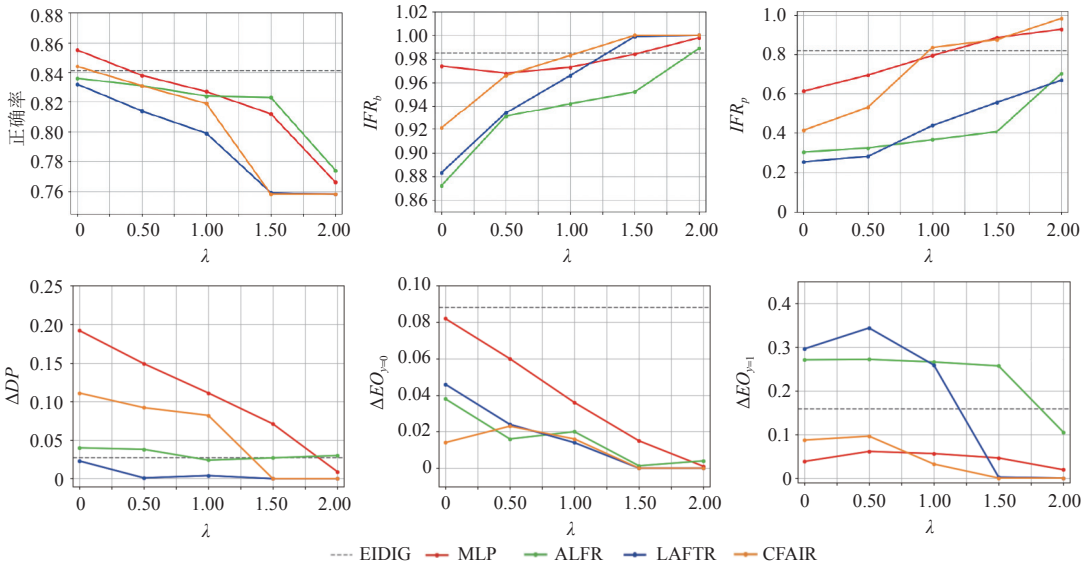

图 4 Adult 数据集中个体损失权重对不同模型性能的影响

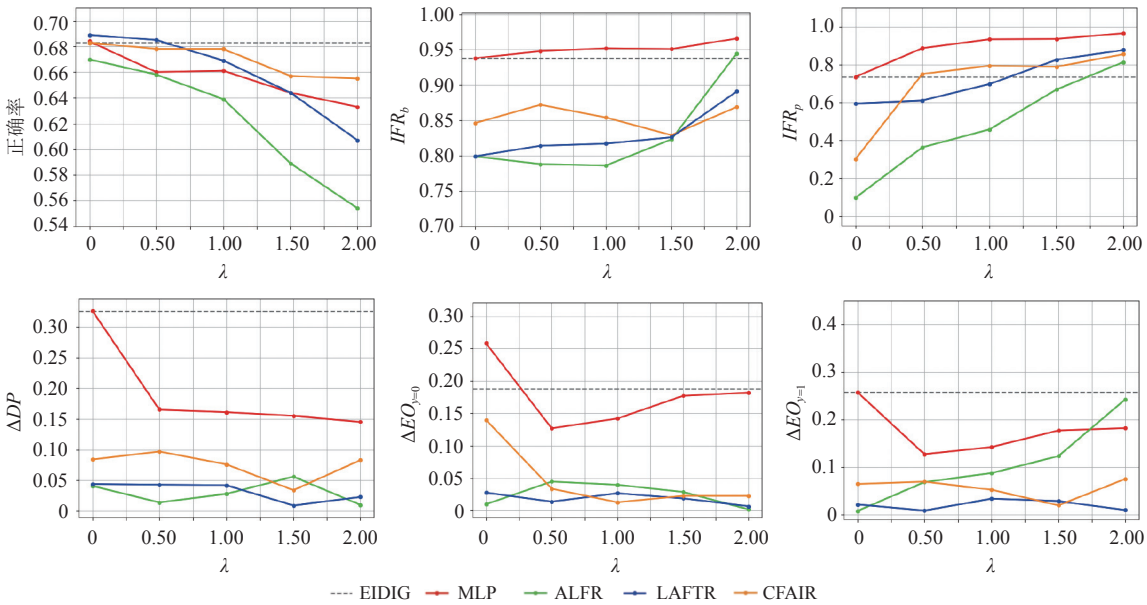

图 5 COMPAS 数据集中个体损失权重对不同模型性能的影响

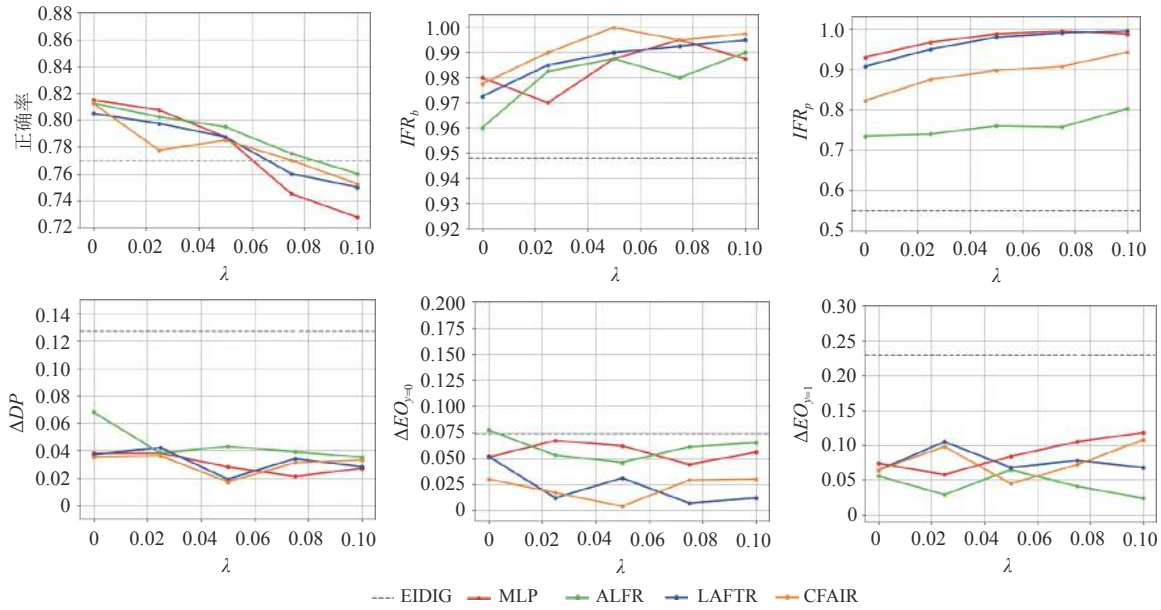

图 6 German 数据集中个体损失权重对不同模型性能的影响

图 4 展示了 Adult 数据集的实验结果, IIFR 算法能够在一定的正确性损失下有效提升模型的个体公平性和群体公平性. 当$\lambda = 1$时, 与同基准模型 MLP 下 EIDIG 方法 (黑色虚线) 相比, IIFR 算法 (红色线) 在个体公平性能上大致相似, 而在几率均等上ΔEO 上 IIFR 明显优于 EIDIG. 图 5、图 6 展示了不同模型与方法在相对均衡的 COMPAS 和 German 数据集上的实验表现, IIFR 算法对模型个体公平性的优化优于 EIDIG, 并且 IIFR 算法在群体公平性上也表现出更好的性能. 当λ增大时, IIFR 算法可以持续提升模型的个体公平性能, 这说明本文的算法能够有效地缓解模型中存在的个体间歧视, 让模型对不同群体但相似的人作出相同的预测结果. 图 4–图 6 的第 2 行展

示了不同程度的 IIFR 算法对模型群体公平性的影响, 结果表明 IIFR 能够较好地保持甚至优化不同基准的群体公平性, 例如在 German 数据集中, IIFR 算法能够将 ΔDP 稳定在 0.001–0.07 之间, 将 $\Delta EO_{y=0}$ 稳定在 0–0.076 之间, 将 $\Delta EO_{y=1}$ 稳定在 0.02–0.14 之间, 且该性能表现均优于 EIDIG.

在基准对抗模型 ALFR (绿线)、LAFTR (蓝线)、CFAIR (黄线) 上使用 IIFR 的实验结果表明, IIFR 算法能够有效地提升模型的个体公平性. 并且当 λ 较小时, CFAIR 模型能够在较小的正确性损失下显著的提升模型的 IFR_p. 例如 Adult 数据集中 CFAIR 对抗模型在 IIFR 算法下能够将 1% 的准确性损失转化为 4% 的个体公平性 IFR_b 及 12% 的 IFR_p 提升, 且其群体公平性能够维持在 0.05 内波动.

不同数据集中的实验结果表明, 当数据集不均衡且数据本身存在歧视时, 正确性对 λ 的变化较为敏感, 且完全的公平性可能会导致模型将所有数据分为一类的现象, 因此选择公平性提升算法时需要权衡不同任务追求的性能. 对于追求正确性较高的任务, 应选择以正确性为损失函数的模型, 并从数据集的角度出发, 不断增强数据和消除数据中潜在的错误样本. 对于追求公平性较高的任务, 由于不同的公平性标准之间存在冲突[17], 因此需确认模型期望的公平性类别. 同时追求个体公平性和高正确性时, EIDIG 是较好的选择, 而对于同时追求较高的个体公平和群体公平性的任务, IIFR 算法是目前最好的选择.

5.2.4 其他超参数的选择及其对模型的影响

对于相似测试结果阈值 τ 的取值, 该值限定了个体公平性的强度, 若 τ 的取值为 0, 表示相似测试样本 (只有受保护属性不同的两个样本) 的训练结果需要具有完全相同的输出分布才满足个体公平性 IFR_p. 当 τ 逐渐增大, 表示允许输出分布具有 τ 范围内的差异. τ 越大, 模型能接受的相似测试样本的输出分布之间的差异就越大. 当 τ 的取值为 1 时, 此时基于输出分布的 IFR_p 指标将转换为基于标签的 IFR_b 指标. 文本实验中的 τ 值设定是由决策边界处相似测试样本的输出分布决定, 因为决策边界处的相似样本易预测为不同的输出结果, 从而造成歧视, 因此本文计算了决策边界处二类预测分布差异为 5% 的 JS 散度, 例如样本 t 的输出分布为 [0.53, 0.47], 而其相似测试样本 t' 的输出分布为 [0.48, 0.52], 该对样本具有不同的预测结果, 他们之间的 JS 散度为 0.001 3. 本文通过比较大量决策边界处二分类输出分布间的 JS 散度, 最终将 τ 值设为 0.001, 表示相似测试样本对的输出分布的距离大于 0.001 且预测标签相同时, IFR_p 指标会将其判定为歧视样本对. 由于阈值仅为 IFR_p 的判断依据, 指示了模型的个体公平性的强度, 因此该参数的调整不会影响模型的其他性能.

对于相似训练样本界限 ϵ 的取值, 该值与样本间的 cos 相似度有关, 本文使用该参数限制指导模型个体公平性的训练样本, 即满足 $\cos x \cdot x' > \epsilon$ 的样本 x 和 x' 需要在训练过程中指导模型优化, 使两者的预测结果尽可能相近. 若 ϵ 的取值为 1, 表示只有完全相同的两个样本被定义为相似训练样本对, 此时的模型优化转化为基准模型的优化训练. 当 ϵ 逐渐减小, 表示具有一定差异的两个训练样本会被判定为相似训练样本对, 且 ϵ 越小, 差异较大但在 ϵ 范围内的两个样本会参与指导模型的梯度方向, IIFR 算法会在训练过程中不断缩小这两个样本预测结果之间的距离.

表 8 为 Adult 数据集中不同相似训练样本界限 ϵ 的取值对模型性能的影响.

表 8 相似训练样本界限 ϵ 的取值对模型的影响

ϵ	ACC (%)	IFR_b (%)	IFR_p (%)	ΔDP	$\Delta EO_{y=0}$	$\Delta EO_{y=1}$
0.75	75.8	100.0	94.4	0.000	0.000	0.000
0.80	**80.1**	**98.6**	**90.6**	**0.059**	**0.011**	**0.063**
0.85	83.5	97.5	83.1	0.136	0.047	0.095
0.90	84.2	97.5	68.4	0.152	0.059	0.057
0.95	85.1	97.1	57.6	0.194	0.086	0.074
1.00	85.5	97.4	61.4	0.192	0.082	0.038

较大的 ϵ 对相似训练样本的要求较严格, 识别出的相似训练样本对的数量较少, 此时的模型会专注少量的相似训练样本对的分类结果, 这可能会影响其他样本的分类结果. 逐渐增大 ϵ 后, 通过近似更多的相似训练样本对的输出分布, 使得带有歧视标签的样本预测为更公平结果, RQ2 中说明了公平的结果会造成某些样本错误的预测, 因此正确率逐渐减小, 但相较于原模型, 通过 IIFR 算法优化后其个体公平性和群体公平性均有显著提升. 当 ϵ 较

大时, 将会存在不相似的训练样本指导模型的学习, 导致模型学习了错误的公平性, 造成模型正确性的大幅降低. 因此使用 IIFR 算法时, 需要根据不同任务的需求选择合适的 ϵ 来平衡正确性和公平性之间的关系, 例如在表 8 中, 当 ϵ 为 0.8 时, IIFR 算法能够在一定正确性损失的情况下取得较高个体公平性和较好的群体公平性.

6 总结与展望

本文提出了两种个体公平率指标, 分别是基于输出标签计算的 IFR_b 和基于输出分布计算的 IFR_p, 这两种指标均可以有效衡量深度学习模型满足个体公平性的程度, 由于个体公平率 IFR_p 要求相似样本对不同类别的预测概率相近, 因此该指标还能够识别出模型潜在的偏见. 本文的实验通过不同模型的个体公平性比较, 证实了缓解群体偏见的对抗网络会导致模型个体公平性下降. 因此本文提出了一个可以缓解模型整体和群体内部偏见的个体公平性算法 IIFR, 该算法在较小的精度损失下, 维持较好的群体公平性的同时, 有效地提高了模型的个体公平性, 从而在个体公平和群体公平间达到较好的平衡.

未来的工作将继续探究 IIFR 算法在深度学习模型上的优化. 目前的 IIFR 算法已较好地权衡了个体公平和群体公平, 但模型的正确性仍有一定损失, 后期我们将从数据集预处理机制 (如: 因果推理、重新标记、扰动、重新加权等) 和后处理机制 (如: 后验正则化、广义期望最大算法等) 两个方面进行优化, 让基于 IIFR 算法的模型保持现有或更好的公平性下进一步提升模型的效率. 其次, 目前的 IIFR 算法仅通过实验证实了在结构化数据上的有效性, 下一阶段我们将迁移 IIFR 算法应用于自然语言处理领域的文本分类和计算机视觉领域的目标检测中.

References:

[1] Bojarski M, Del Testa D, Dworakowski D, Firner B, Flepp B, Goyal P, Jackel LD, Monfort M, Muller U, Zhang JK, Zhang X, Zhao J, Zieba K. End to end learning for self-driving cars. arXiv:1604.07316, 2016.

[2] Goodall NJ. Can you program ethics into a self-driving car? IEEE Spectrum, 2016, 53(6): 28–58. [doi: 10.1109/MSPEC.2016.7473149]

[3] Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, van der Laak JAWM, van Ginneken B, Sánchez CI. A survey on deep learning in medical image analysis. Medical Image Analysis, 2017, 42: 60–88. [doi: 10.1016/j.media.2017.07.005]

[4] Fu K, Cheng DW, Tu Y, Zhang LQ. Credit card fraud detection using convolutional neural networks. In: Proc. of the 23rd Int'l Conf. on Neural Information Processing. Kyoto: Springer, 2016. 483–490. [doi: 10.1007/978-3-319-46675-0_53]

[5] Saxena NA, Huang KR, DeFilippis E, Radanovic G, Parkes DC, Liu Y. How do fairness definitions fare? Examining public attitudes towards algorithmic definitions of fairness. In: Proc. of the 2019 AAAI/ACM Conf. on AI, Ethics, and Society. Honolulu: ACM, 2019. 99–106. [doi: 10.1145/3306618.3314248]

[6] Dastin J. Amazon scraps secret AI recruiting tool that showed bias against women. Ethics of Data and Analytics. New York: Auerbach Publications, 2022. 296–299.

[7] Angwin J, Larson J, Mattu S, Kirchner L. Machine bias. Ethics of Data and Analytics. New York: Auerbach Publications. 2016. 254–264. [doi: 10.1201/9781003278290]

[8] Liu WY, Shen CY, Wang XF, Jin B, Lu XJ, Wang XL, Zha HY, He JF. Survey on fairness in trustworthy machine learning. Ruan Jian Xue Bao/Journal of Software, 2021, 32(5): 1404−1426 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6214.htm [doi: 10.13328/j.cnki.jos.006214]

[9] Grgić-Hlača N, Zafar MB, Gummadi KP, Weller A. The case for process fairness in learning: Feature selection for fair decision making. In: Proc. of the 2016 Symp. on Machine Learning and the Law at the 29th Conf. on Neural Information Processing Systems. Barcelona, 2016. 1–11.

[10] Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R. Fairness through awareness. In: Proc. of the 3rd Innovations in Theoretical Computer Science Conf. Cambridge: ACM, 2012. 214–226. [doi: 10.1145/2090236.2090255]

[11] Zemel R, Wu Y, Swersky K, Pitassi T, Dwork C. Learning fair representations. In: Proc. of the 30th Int'l Conf. on Machine Learning. Atlanta: JMLR.org, 2013. III-325–III-333.

[12] Lahoti P, Gummadi KP, Weikum G. iFair: Learning individually fair data representations for algorithmic decision making. In: Proc. of the 35th IEEE Int'l Conf. on Data Engineering. Macao: IEEE, 2019. 1334–1345. [doi: 10.1109/ICDE.2019.00121]

[13] Zhang LF, Zhang YL, Zhang M. Efficient white-box fairness testing through gradient search. In: Proc. of the 30th ACM SIGSOFT Int'l Symp. on Software Testing and Analysis. Denmark: ACM, 2021. 103–114. [doi: 10.1145/3460319.3464820]

[14] Edwards H, Storkey A. Censoring representations with an adversary. arXiv:1511.05897, 2016.
[15] Zafar MB, Valera I, Rogriguez MG, Gummadi KP. Fairness constraints: Mechanisms for fair classification. In: Proc. of the 20th Int'l Conf. on Artificial Intelligence and Statistics. Florida: JMLR, 2017. 962–970.
[16] Hardt M, Price E, Srebro N. Equality of opportunity in supervised learning. In: Proc. of the 30th Int'l Conf. on Neural Information Processing Systems. Barcelona: Curran Associates Inc., 2016. 3323–3331.
[17] Berk R, Heidari H, Jabbari S, Kearns M, Roth A. Fairness in criminal justice risk assessments: The state of the art. Sociological Methods & Research, 2021, 50(1): 3–44. [doi: 10.1177/0049124118782533]
[18] Galhotra S, Brun Y, Meliou A. Fairness testing: Testing software for discrimination. In: Proc. of the 11th Joint Meeting on Foundations of Software Engineering. Paderborn: ACM, 2017. 498–510. [doi: 10.1145/3106237.3106277]
[19] Kilbertus N, Rodriguez MG, Schölkopf B, Muandet K, Valera I. Fair decisions despite imperfect predictions. In: Proc. of the 23rd Int'l Conf. on Artificial Intelligence and Statistics. Palermo: PMLR, 2020. 277–287.
[20] Kusner MJ, Loftus J, Russell C, Silva R. Counterfactual fairness. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems, Long Beach: Curran Associates Inc., 2017. 4069–4079.
[21] Beutel A, Chen JL, Zhao Z, Chi EH. Data decisions and theoretical implications when adversarially learning fair representations. arXiv:1707.00075, 2017.
[22] Madras D, Creager E, Pitassi T, Zemel R. Learning adversarially fair and transferable representations. In: Proc. of the 35th Int'l Conf. on Machine Learning. Stockholm: PMLR, 2018. 3384–3393.
[23] Feng R, Yang Y, Lyu Y, Tan CH, Sun YZ, Wang CP. Learning fair representations via an adversarial framework. arXiv:1904.13341, 2019.
[24] Menon AK, Williamson RC. The cost of fairness in classification. arXiv:1705.09055, 2017.
[25] Speicher T, Heidari H, Grgic-Hlaca N, Gummadi KP, Singla A, Weller A, Zafar MB. A unified approach to quantifying algorithmic unfairness: Measuring individual & group unfairness via inequality indices. In: Proc. of the 24th ACM SIGKDD Int'l Conf. on Knowledge Discovery & Data Mining. London: ACM, 2018. 2239–2248. [doi: 10.1145/3219819.3220046]
[26] Joseph M, Kearns M, Morgenstern J, Roth A. Fairness in learning: Classic and contextual bandits. In: Proc. of the 30th Int'l Conf. on neural Information Processing Systems. Barcelona: Curran Associates Inc., 2016. 325–333.
[27] Kearns M, Neel S, Roth A, Wu ZS. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In: Proc. of the 35th Int'l Conf. on Machine Learning. Stockholm: PMLR, 2018. 2564–2572.
[28] Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. In: Proc. of the 27th Int'l Conf. on Neural Information Processing Systems. Montreal: MIT Press, 2014. 2672–2680.
[29] Zhao H, Coston A, Adel T, Gordon GJ. Conditional learning of fair representations. arXiv:1910.07162, 2020.
[30] Blei DM, Kucukelbir A, McAuliffe JD. Variational inference: A review for statisticians. Journal of the American statistical Association, 2017, 112(518): 859–877. [doi: 10.1080/01621459.2017.1285773]
[31] Gabrilovich E, Markovitch S. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In: Proc. of the 20th Int'l Joint Conf. on Artifical Intelligence. Hyderabad: Morgan Kaufmann Publishers Inc., 2007. 1606–1611.
[32] Potthast M, Stein B, Anderka M. A Wikipedia-based multilingual retrieval model. In: Proc. of the 30th European Conf. on IR Research. Glasgow: Springer, 2008. 522–530. [doi: 10.1007/978-3-540-78646-7_51]
[33] De Boer PT, Kroese DP, Mannor S, Rubinstein RY. A tutorial on the cross-entropy method. Annals of Operations Research, 2005, 134(1): 19–67. [doi: 10.1007/s10479-005-5724-z]
[34] Zeiler MD. ADADELTA: An adaptive learning rate method. arXiv:1212.5701, 2012.
[35] Burnaev E, Erofeev P, Papanov A. Influence of resampling on accuracy of imbalanced classification. In: Proc. of the 8th Int'l Conf. on Machine Vision. Barcelona: SPIE, 2015. 423–427. [doi: 10.1117/12.2228523]
[36] Cui Y, Jia ML, Lin TY, Song Y, Belongie S. Class-balanced loss based on effective number of samples. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 9260–9269. [doi: 10.1109/CVPR.2019.00949]
[37] Chouldechova A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, 2017, 5(2): 153–163. [doi: 10.1089/big.2016.0047]

附中文参考文献:

[8] 刘文炎, 沈楚云, 王祥丰, 金博, 卢兴见, 王晓玲, 查宏远, 何积丰. 可信机器学习的公平性综述. 软件学报, 2021, 32(5): 1404−1426. http://www.jos.org.cn/1000-9825/6214.htm [doi: 10.13328/j.cnki.jos.006214]

王昱颖(1998－), 女, 硕士生, CCF 学生会员, 主要研究领域为深度学习公平性.

张敏(1977－) 女, 博士, 教授, CCF 专业会员, 主要研究领域为复杂系统的量化分析与验证, AI 系统的测试与分析验证.

杨晶然(1999－), 女, 硕士生, 主要研究领域为机器学习测试.

徐晟恺(1998－), 男, 硕士生, 主要研究领域为机器学习可解释性.

陈仪香(1961－), 男, 博士, 教授, CCF 杰出会员, 主要研究领域为物联网与信息物理融合系统, 实时软件系统, 软件形式化方法与可信评估, 软硬件协同设计与优化技术.