计算机科学
COMPUTER SCIENCE

机器学习公平性指标:现状、挑战和展望

张文琼, 李云

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230500224

机器学习公平性指标:现状、挑战和展望

张文琼　李　云
南京邮电大学计算机学院、软件学院、网络空间安全学院　南京 210023
江苏省大数据安全与智能处理重点实验室　南京 210023
(18763370336@163.com)

摘　要　随着机器学习应用的日益普及,机器学习公平性问题引起了学术界和工业界的广泛关注,成为了可信人工智能的重要组成部分。为了评估和改善机器学习应用的公平性,研究人员提出了一系列公平性指标,这些指标有助于保障机器学习模型在不同个体、群体间的公平决策,并为改善和优化模型提供指导。但各界对于指标之间的区别与联系仍没有形成共识,对不同场景、不同任务的公平性定义没有明确的划分,公平性指标缺乏完善的分类体系。文中对公平性指标进行了全面的整理和归类,从指标的数学定义出发,根据是否基于概率统计将公平性指标分为两类,然后分别对这两类指标进行进一步的细粒度划分和阐述。为了便于读者理解和运用,结合一个实际案例,从适用场景和实现条件等方面指出各类指标的优势和面临的挑战,还结合数学定义讨论了指标之间的关系,并对未来趋势进行了展望。

关键词:机器学习;机器学习公平性;可信人工智能;公平性指标;公平决策
中图分类号　TP391

Fairness Metrics of Machine Learning:Review of Status,Challenges and Future Directions

ZHANG Wenqiong and LI Yun
School of Computer Science,Nanjing University of Posts and Telecommunications,Nanjing 210023,China
Jiangsu Key Laboratory of Big Data Security and Intelligent Processing,Nanjing 210003,China

Abstract　With the increasing popularity of machine learning applications,fairness of machine learning has attracted widespread attention from academia and industry,and has become an important component of trust-worthy artificial intelligence. To evaluate and improve the fairness of machine learning applications,a series of fairness metrics have been proposed by researchers. These metrics help to ensure fair decision-making of machine learning models among different individuals and groups,and provide guidance for improving and optimizing the model. However,there is still no consensus on the difference and correlation between these metrics,which are not clearly divided in different scenarios and tasks. This means that these fairness metrics lack a comprehensive classification system. In this paper,the fairness metrics are comprehensively organized and classified. Starting from the mathematical definition of these metrics,they are divided into two categories according to whether they are based on probability statistics. The two types of metrics are then further divided and elaborated separately. In order to facilitate readers' understanding and application,combined with a practical case,the advantages and challenges of various metrics are pointed out in terms of application scenarios and implementation conditions,and the relationship between metrics is also discussed in conjunction with mathematical concepts,and possible future research directions are prospected.

Keywords　Machine learning,Fairness of machine learning,Trust-worthy artificial intelligence,Fairness metrics,Fair decision

1　引言

近年来,机器学习技术被广泛应用于医疗[1]、广告[2]、金融[3]等具有重要影响力的领域。从社会影响来看,这些应用不可避免地存在对种族、性别等敏感属性的偏见或者歧视[4]。歧视指基于某一类别的成员身份而非个人价值产生的对个人或群体的不公平或不平等待遇[5]。歧视性行为影响了弱势群体(又称受保护群体)在就业、教育等社会活动中的合法权益。

因此,研究如何保障机器学习应用的公平性成为了社会各界的共同关注点。2022年3月中共中央办公厅、国务院办公厅印发了《关于加强科技伦理治理的意见》,将人工智能作为科技伦理治理的重点领域之一。在实际应用中,机器学习公平性指标作为一个重要工具,可以帮助我们评估和监测机器学习应用的公平性,并为改善和优化模型提供方向。学术

到稿日期:2023-05-30　返修日期:2023-09-27
基金项目:江苏省高校自然科学基金重大项目(21KJA520003)
This work was supported by the Natural Science Fund for Colleges and Universities in Jiangsu Province(21KJA520003).
通信作者:李云(liyun@njupt.edu.cn)

界对机器学习公平性指标的研究逐步深入并覆盖到不同层面。群体公平性指标旨在保证不同群体在模型中受到公平待遇，个体公平性指标则关注个体在模型中是否受到公平待遇。然而，机器学习公平性指标也面临着一些挑战，因此对公平性指标进行归类分析，有利于辅助研究人员根据任务目标选择适合的指标。

本文旨在梳理当前机器学习公平性指标研究的现状，并为后续研究提供可借鉴的思路。Verma 等[6]在同一个数据集上进行实验，对部分公平性指标进行了对比介绍。Mehrabi 等[7]也列举了一些传统的公平性指标，起到了概述梳理作用，但对各指标之间的联系和区别分析较浅，没有明确当前公平性指标面临的挑战且缺乏对未来的展望。不同于以往依据群体、个体分类，本文从指标的数学定义出发，根据是否基于概率统计将指标分为两类。为了便于读者理解和运用，我们结合汽车保险定价案例对其进行说明，分析其优势和面临的挑战，以独特的视角介绍了各类指标的特点及它们之间的联系，并对未来进行了展望。

2 背景介绍

机器学习公平性领域主要研究在预定义敏感属性前提下的二分类有监督学习任务，即样本标签分正类和负类两类。通常假定敏感属性为二元属性，根据敏感属性（如性别）可将数据集划分成受保护群体（女性）和非受保护群体（男性）。

保险定价系统用于评估车主保险费用，它对车主的驾龄、理赔记录、购车年份等信息进行量化，将车主预测为驾驶高风险（正类）或驾驶低风险（负类），进而影响保费定价。在本例中，我们以性别为敏感属性，研究当系统满足相应公平性指标时，男性车主和女性车主的分类结果需要满足的关系。

本文中使用的符号的定义如下：训练样例 D 表示为（$\boldsymbol{X}$，Y，Z），其中 $\boldsymbol{X}$ 表示除敏感属性外的其他可观测属性，$Z\in\{0,1\}$表示敏感属性，$Y\in\{0,1\}$表示真实标签，$\hat{Y}\in\{0,1\}$表示预测标签。

3 基于概率统计的公平性指标

实现基于概率统计的公平性需要结合经典的统计方法，通常要求敏感属性取值不同的群体之间具有相似的准确率[8]、误判率[9]、真阳率[10]等，对数据标签的完整性要求较高。根据标签与敏感属性的关系，可以将其分为预测公平性、测试公平性、精度公平性和误判公平性 4 类，其中精度公平性和误判公平性是基于混淆矩阵衡量的。在机器学习领域，混淆矩阵（Confusion Matrix）是一种评价模型精度的形象化展示工具，矩阵的每一列表示模型的预测值，每一行表示样本的真实值。图 1 给出了混淆矩阵和与之相关的公平性指标。对于二分类任务，真实标签和预测标签均有两个取值，下面对图 1 中的定义进行说明：1）True Positive（TP）表示样本的真实标签和预测标签均为正类；2）True Negative（TN）表示样本的真实标签和预测标签均为负类；3）False Negative（FN）表示样本的真实标签为正类，预测标签为负类；4）False Positive（FP）表示样本的真实标签为负类，预测标签为正类。

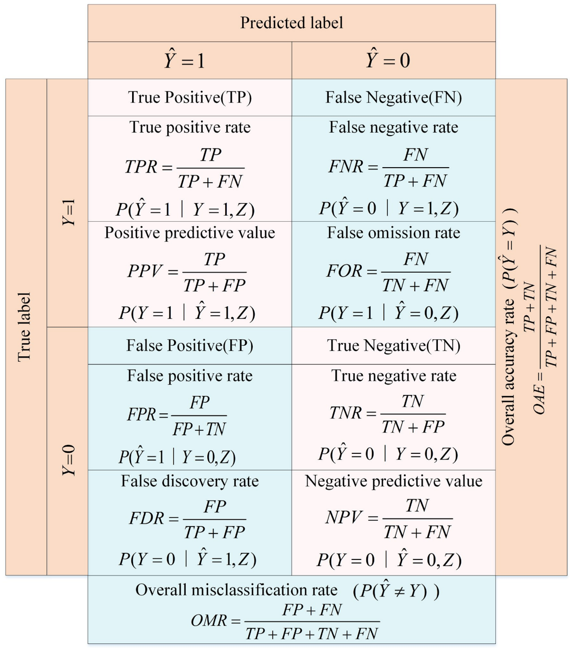

图 1 混淆矩阵中的公平性定义

Fig. 1 Definitions of fairness in confusion matrix

3.1 预测公平性

预测公平性主要基于数据的预测标签和敏感属性来衡量，要求不同群体被预测为正类的概率近似。

1）统计均等[11]（Demographic Parity）

$$P(\hat{Y}=1|Z=1)=P(\hat{Y}=1|Z=0) \tag{1}$$

该指标也被称为 Statistical Parity，如果预测标签$\hat{Y}$独立于敏感属性 Z，即敏感属性取值不同的两个群体被预测为正类的概率相同，则满足此指标。在保险定价案例中，要求男性和女性群体被预测为高风险的概率相同。该指标为群体公平提供了一些保障，但不能反映预测准确率，不同群体实际为正类的比例可能不同，强制满足该指标可能会偏袒某些负类样本，导致新的歧视。Jiang 等[12]把该指标推广到敏感属性为连续值的场景，拓宽了该指标的使用范围。

2）条件统计均等[13]（Conditional Statistical Parity）

$$P(\hat{Y}=1|L=l,Z=1)=P(\hat{Y}=1|L=l,Z=0) \tag{2}$$

该指标是对统计均等指标的拓展，其中 L 为合法属性，其与敏感属性 Z 的相关性较小，与预测标签$\hat{Y}$的相关性较大。如果不同群体在合法属性 L 取值相同时被预测为正类的概率相同，则满足此指标。在保险定价案例中，可以选择驾龄为合法属性，要求在平均驾龄相同的条件下，男性和女性群体被预测为高风险的概率相等。从另一个角度分析，如果训练集中男性群体的平均驾龄比女性群体长，则女性群体有更大概率被预测为驾驶高风险或许是合理的，这不是对性别的直接歧视。相比统计均等，该指标在分析数据时考虑到除敏感属性以外的其他条件，提高了公平性指标的可解释性，然而这要基于选择不受争议的合法属性和敏感属性组合。与之类似，文献[14]提出了条件公平性（Conditional Fairness，CF）指标，将原始特征映射到表示空间，以寻找合法属性 L。

3)差异性影响[15](Disparate Impact)

$$\min\left(\frac{P(\hat{Y}=1|Z=0)}{P(\hat{Y}=1|Z=1)},\frac{P(\hat{Y}=1|Z=1)}{P(\hat{Y}=1|Z=0)}\right)\geqslant\frac{p}{100} \tag{3}$$

该指标一般用美国平等就业机会委员会提出的 $P\%$ 规则度量[5]。$P\%$ 规则规定，不同群体被预测为正类的概率比值应不小于 P∶100，通常 P 值取 80。在保险定价案例中，假设女性群体有 30% 的人被预测为高风险，则男性群体至少有 24% 的人被预测为高风险。该指标可以根据任务目标灵活调整 P 值，更具应用价值，但仍可能出现为追求实质平等而对特定群体给予不必要优待的现象，这被称为反向歧视(Reverse Discrimination)。

3.2 测试公平性

测试公平性是基于数据的真实标签和标签概率预测分值 S 衡量的，标签概率预测分值 S 表示样本被预测为正类的概率，取值范围为[0,1]。

1)标定均等[16](Calibration)

$$P(Y=1|S=s,Z=0)=P(Y=1|S=s,Z=1) \tag{4}$$

在预测分值 S 相同的情况下，不同群体真实标签为正类的概率相等则满足此指标。在保险定价案例中，对于某一确定的预测分值 S，要求男性和女性群体中实际为高风险的概率相同。该指标可以灵活选取预测分值，适用于决策阈值不确定的情况，但是易受数据标签偏差的影响。

2)预测均等[16](Predictive Parity)

$$P(Y=1|S>s',Z=0)=P(Y=1|S>s',Z=1) \tag{5}$$

该指标中的 s' 表示阈值，如果预测分值 S 大于阈值 s'，则预测标签 $\hat{Y}=1$。以预测分值 S 大于阈值 s' 为条件，该指标要求不同群体中的真实标签为正类的概率相等。在保险定价案例中，当预测分值 S 大于阈值 s' 时，男性和女性群体中实际为高风险的概率应相同。当预测分值 S 是连续值时，该指标对阈值 s' 的选取依赖性较强，需针对特定问题进行调整；当预测分值 S 是离散二元属性时，满足该指标等同于要求不同群体阳性预测值(Positive Predictive Value, PPV)相等。

3)正类平衡[17](Balance for Positive Class)

$$E(S|Y=1,Z=1)=E(S|Y=1,Z=0) \tag{6}$$

首先分别计算不同群体中正类样本的预测分值 S，然后对其求期望，如果相等则满足该指标。在保险定价案例中，要求男性和女性群体中实际为高风险的样本预测分值 S 的期望相同。与之类似，负类平衡[17](Balance for Negative Class)指标针对不同群体的实际标签为负类的样本。

3.3 精度公平性

精度公平性基于数据的真实标签和预测标签以及敏感属性来衡量，这一类公平性指标要求不同群体预测准确率近似。

1)机会均等[10](Equal Opportunity/True Positive Rate Parity)

$$P(\hat{Y}=1|Y=1,Z=0)=P(\hat{Y}=1|Y=1,Z=1) \tag{7}$$

在真实标签为正类 $Y=1$ 的情况下，不同群体被预测为正类的概率相等，即满足真阳率(True Positive Rate, TPR)相等，则满足该指标。在保险定价案例中，该指标要求实际为高风险的男性和女性群体被预测为高风险的概率相同。如果只关注不同群体的正类样本是否公平，则可使用该指标。预测平等[18]指标与该指标相对应，要求不同群体满足假阳率(False Positive Rate, FPR)相等。

在实际应用中，可以根据任务需求选择其一，如果被预测为负类对用户有利，如解雇、风险评估等，系统则对假阳率更敏感，则适合使用预测平等指标；反之，在招聘、贷款等场景下，通常使用机会均等指标。如果要求真阳率和假阳率均相等，则选用概率均等[10](Equalized Odds)，该指标为两者的合并形式。

2)整体精度均等[8](Overall Accuracy Equality)

$$P(\hat{Y}=Y|Z=0)=P(\hat{Y}=Y|Z=1) \tag{8}$$

该指标要求不同群体的预测精度相等，在保险定价案例中，要求男性和女性群体被正确预测的概率相同。该指标考虑到了准确性，但是没有将真阳率和真阴率区分开，如果男性群体中真阳率远高于真阴率，反之女性群体中真阴率远高于真阳率，仍存在歧视现象。

3)条件使用精度均等[8](Conditional Use Accuracy Equality)

$$\begin{aligned}&P(Y=1|\hat{Y}=1,Z=0)=P(Y=1|\hat{Y}=1,Z=1)\\&P(Y=0|\hat{Y}=0,Z=0)=P(Y=0|\hat{Y}=0,Z=1)\end{aligned} \tag{9}$$

在已知标签预测值的情况下，如果不同群体真实标签值和预测标签值一致，则满足此指标，即要求模型同时满足不同群体之间的阳性预测值(PPV)和阴性预测值(NPV)相等。在保险定价案例中，对于男性和女性群体，其被预测为高风险的人中实际为驾驶高风险的概率应该相同，被预测为低风险的群体同理。概率均等指标是以真实标签值为条件，而该指标以预测标签为条件。

3.4 误判公平性

Zafar 等[9]提出不同的误判(Disparate Mistreatment)指标，要求不同群体的误判率相同。下面介绍与误判率相关的指标。

1)测试均等[19](Test Fairness/Predictive Rate Parity)

$$\begin{aligned}&P(Y=1|\hat{Y}=0,Z=0)=(Y=1|\hat{Y}=0,Z=1)\\&P(Y=0|\hat{Y}=1,Z=0)=P(Y=0|\hat{Y}=1,Z=1)\end{aligned} \tag{10}$$

该指标要求不同群体同时满足错误遗漏率(False Omission Rate, FOR)和错误发现率(False Discovery Rate, FDR)相等，在保险定价案例中，要求被预测为低风险的男性和女性群体中实际为高风险的概率相同，在被预测为高风险的群体中依然成立。

2)整体误判率均等[9](Overall Misclassification Rate Equality)

$$P(\hat{Y}\neq Y\ |Z=0)=P(\hat{Y}\neq Y\ |Z=1) \tag{11}$$

该指标要求不同群体预测错误的概率相等，在保险定价案例中，要求男性和女性群体被预测错误的概率相同。与 3.3 节中的整体精度均等指标类似，该指标没有将假阳率和假阴率区分开，因此仍可能存在歧视现象。

3)待遇均等[8](Treatment Equality)

$$\frac{P(\hat{Y}=0 \mid Z=0,Y=1)}{P(\hat{Y}=1 \mid Z=0,Y=0)}=\frac{P(\hat{Y}=0 \mid Z=1,Y=1)}{P(\hat{Y}=1 \mid Z=1,Y=0)} \tag{12}$$

如果不同群体中假阴率(False Negative Rate,FNR)和假阳率(FPR)的比值相同,则满足该指标。在保险定价案例中,分别计算男性和女性群体中高风险车主和低风险车主被预测错误的比值,如果比值相同则满足该指标。

3.5 指标之间的关系

如图 2 所示,混淆矩阵中的评价指标之间存在一些联系,例如真阳率和假阴率之间存在互推关系,$TPR=\frac{TP}{TP+FN}=1-\frac{FN}{TP+FN}=1-FNR$。因此如果各群体之间满足真阳率相等,那么假阴率也相等,与之相关的公平性指标之间也会存在关联。例如,如果模型满足机会均等指标,则说明不同群体的真阳率相等,则可推出他们的假阴率相等。同样地,如果模型满足概率均等,那么一定满足机会均等和预测平等[18],进一步可推出满足不同群体的假阴率和真阴率也相等;如果满足整体精度均等,则可推出满足整体误判率均等;如果满足条件使用精度均等,则可推出满足测试均等。

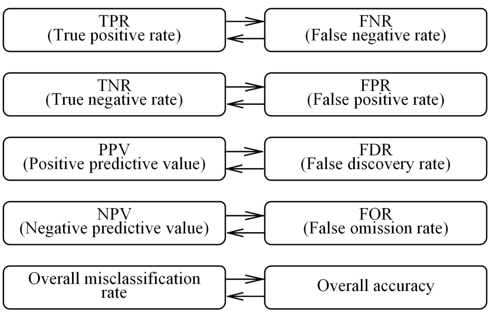

图 2 混淆矩阵评价指标之间的互推关系

Fig. 2 Interpolations between confusion matrix evaluation indicators

4 非概率统计的公平性指标

非概率统计的公平性关注指标是否符合直觉,常见的方法是在数据预处理阶段直接去除敏感属性,或根据距离度量方法判定样本的相似度,使相似的样本满足预测标签相近,以及结合广义熵指数或 K 近邻等方法衡量个体公平性[20],还可以基于一些因果假设减少敏感属性对预测结果的影响。

4.1 无意识公平性[21]

$$P(\hat{Y} \mid X)=P(\hat{Y} \mid X,Z) \tag{13}$$

无意识公平性(Fairness Through Unawareness)要求模型在预测样本标签的过程中不依赖敏感属性。不同的对待(Disparate Treatment)指标属于这一类,仅要求在实验中将训练集中的敏感属性删除即可,因此该指标具有高度简单性、易于使用的优点。在保险定价案例中,将车主信息数据中的性别属性剔除即可满足该指标。然而,数据的属性通常是相关的,例如年龄和储蓄相关、种族和姓名相关等,删除敏感属性并不能保证与之相关的所有信息都被删除,模型仍能将敏感属性预测出来,模型的输出结果仍可能存在偏差。可以定义一个相关性阈值来排除部分与敏感属性相关的特征,但是这可能会在一定程度上损失准确率[22]。

4.2 相似度公平性[23]

$$D(M(x_i),M(x_j))\leqslant d(x_i,x_j) \tag{14}$$

相似度公平性属于有意识公平性(Fairness Through Awareness)的范畴,设 x_i 和 x_j 表示训练集中的两个样本,$M(\cdot)$ 表示从样本输入到输出的映射,$D(\cdot)$ 和 $d(\cdot)$ 分别表示在输出空间和输入空间自定义的距离度量方法,距离越小表示两个个体越相似。在保险定价案例中,要求相似的个体在输出空间中的距离相近。该指标为个体公平(Individual Fairness)提供了一些保障,但是如何定义个体之间的相似度是一项比较复杂的任务。

K 近邻是衡量相似度公平性的一种常见策略,如式(15)所示,对于每个样本$(x_i,\hat{y}_i)$,该指标希望目标样本与其 K 个最近邻样本的预测标签相似,衡量了目标样本预测值 $\hat{y}_i$ 与其 K 近邻样本预测结果的接近程度[22]。

$$consistency=1-\frac{1}{n}\left(\sum_{i=1}^{n}\left|\hat{y}_i-\frac{1}{k}\sum_{x_j\in kNN(x_i)}\hat{y}_j\right|\right) \tag{15}$$

4.3 因果公平性

$$P(\hat{Y}_{Z\leftarrow z}(U)=y \mid X=x,Z=z)=P(\hat{Y}_{Z\leftarrow z'}(U)=y \mid X=x,Z=z) \tag{16}$$

因果公平性的定义基于因果图,因果图[24-25]是一个有向无环图,节点表示属性,有向箭头表示因果关系。在因果图中,解析变量(Resolving Attribute)表示不会受敏感属性影响的变量,代理变量(Proxy Attribute)则与敏感属性高度相关,并且是敏感属性的后代[26]。为了便于说明,我们依据保险定价任务构建了一个因果图,如图 3 所示。图 3 中,性别为敏感属性,驾龄为解析变量;汽车颜色为代理变量,其与性别高度相关。依据历史数据,选购红色汽车的人性格相对奔放,更易有激烈驾驶行为(急刹车、急加速等),传统风险评估系统更容易将红色汽车车主预测为高风险。而选择红色汽车的人中女性居多,这可能会引起对女性群体的间接歧视。

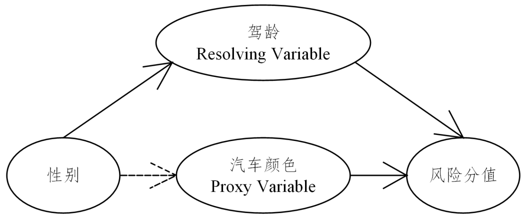

图 3 保险案例因果图示例

Fig. 3 Example of a causality diagram for an insurance case

在公平性领域,只修改敏感属性而不改变其他属性的现象被称为反事实[27],反事实公平性[28](Counterfactual Fairness)是应用最广泛的因果公平性指标。如果反事实样本与原样本预测结果一致,即预测标签独立于敏感属性,则满足该指标。该指标可以与条件统计均等联系起来,两者都要求在其他属性相同的情况下,敏感属性值的改变不会影响模型的预测结果,然而反事实公平性针对个体层面,后者针对群体层面。如图 3 所示,风险分值依赖于驾龄和汽车颜色,而汽车颜色是敏感属性性别的后代,因此不满足该指标。因果公平性

具有良好的可解释性，但高度依赖因果图的结构，因此需要根据专家知识厘清属性之间的因果关系，构建一个高可信度的因果图。

5 分析与讨论

机器学习公平性的基本任务是将一般的机器学习应用扩展为最大化公平性的应用，度量公平性的前提是根据具体问题和应用场景选择适合的指标。而在实际应用中，不同机器学习任务的要求和适用性不同，各类指标也有不同的使用优势，因此很难确定一种通用的公平性指标。表 1 列出了从指标的实现条件、优势和处理对象 3 方面对公平性指标进行对比归类的结果。实现条件表明了敏感属性与标签和其他属性之间的独立性约束，处理对象区分了指标的使用对象为个体公平或群体公平。

表 1 常见的公平性指标对比

Table 1 Comparison of common fairness metrics

公平性指标分类		公平性指标名称	参考文献	实现条件	优势			处理对象
					形式简单	可解释性强	考虑背景变量	
基于概率统计	预测公平性	Demographic parity	[11]	$\hat{Y}\perp Z$	√	×	×	群体
		Generalized demographic parity	[12]					
		Conditional statistical parity/ Conditional fairness	[13-14]	$\hat{Y}\perp Z\|L$				
		Disparate impact	[15]	$\hat{Y}\perp Z$				
	测试公平性	Calibration	[16]	$Y\perp Z\|S$	×	×	√	群体
		Predictive parity	[16]					
		Balance for positive class	[17]	$S\perp Z\|Y$				
		Balance for negative class	[17]					
	精度公平性	Equal opportunity	[10]	$\hat{Y}\perp Z\|Y$	×	×	√	群体
		Predictive equality	[18]					
		Equalized odds	[10]					
		Overall accuracy equality	[8]	$\hat{Y}\perp Z$				
		Conditional use accuracy equality	[8]	$Y\perp Z\|\hat{Y}$				
	误判公平性	Test fairness	[19]	$Y\perp Z\|\hat{Y}$	×	×	√	群体
		Overall misclassification rate equality	[9]	$\hat{Y}\perp Z$				
		Treatment equality	[8]	$\hat{Y}\perp Z\|Y$				
非概率统计	无意识公平性	Disparate treatment	[21]	$\hat{Y}\perp Z$	√	√	×	个体
	相似度公平性	Fairness through awareness	[23]	相似性度量	×	√	×	个体
	因果公平性	Counterfactual fairness	[28]	因果图	×	√	√	个体/群体

其中预测公平性指标和无意识公平性指标对公平性的度量只涉及预测标签与敏感属性，无须考虑其他属性的影响，因此具有简单易于实现的优点；非概率统计的公平性指标不采用概率统计的方式，而是从用户主观上容易理解的角度分析属性与预测标签的关系，具有可解释性强的优点；测试公平性、精度公平性、误判公平性等充分考虑到所有属性与真实标签、预测标签等之间的联系，而非简单分析敏感属性与预测标签的关系，具有惩罚惰性学习方法的优点。我们结合具体案例对目前常见的公平性指标进行了归纳和对比，对公平性指标的理解和选择有一定启示作用。除此之外，结合不同应用场景，一些新的公平性指标也逐渐被提出[29-30]，例如在推荐系统领域，Gao 等[31]将公平性考量纳入传统的信息检索指标，并据此对相应信息进行排名。还有许多用于公平性评估的平台和工具，正被逐步完善。

如 3.5 节所述，各类公平性指标之间存在一些关联，Friedler 等[32]也证实了这一点。也有研究发现有些公平性指标之间互不相容，Kleinberg 等[33]指出只有在严格约束的特殊情况下才能近似同时满足校准（Calibration）和平衡正负类（Balance the Positive and Negative Classes）指标，因此探究多个公平性指标之间的关联与权衡也是未来一项有意义的工作。另外，Liu 等[34]提出的几个常见的公平性指标只解决了静态歧视，长久来看可能会对弱势群体产生消极影响。例如，在银行信贷案例中[35]，强制使用统计均等指标，使得受保护群体和其他群体被授予贷款的比例相同，造成对受保护群体的过度偏袒。因此，在进行公平性评估时，不仅要考虑个人属性，也要考虑个人与他人、社会的联系，要关注领域中的关系结构[36]。同时还应关注机器学习公平性研究的社会影响[37]，将相应的理论落实到教育[38]、医疗[39]、金融[40]等领域。例如，Baker 等[41]总结了教育偏见的原因，并分析了哪些人口亚群体更容易受到不公平对待。Reddy 等[42]提出了一种面向健康护理领域的 AI 治理框架，主要强调 AI 治理中的公平性、透明度、可信度和问责能力。Delobelle 等[43]将自然语言处理领域的公平性指标进行归类，并分析了不同指标之间的区别与联系。另外，不能忽略弱势群体的长期利益，寻找新的动态公平性指标来度量长期不公平问题也是未来一项有意义的工作。

6 问题与展望

机器学习公平性指标有助于评估和改进机器学习应用的公平性，但仍存在一些问题，这些问题也为未来指明了方向。

1）指标之间存在依赖与矛盾

目前公平性指标还未形成共识，对不同领域、不同任务的公平性定义没有明确的划分，在某些情况下，采用不同的公平性指标会得到不同的反馈，有些公平性指标之间甚至存在

冲突,未来应建立完善的公平性指标评估体系。

2)缺乏可解释性

多数公平性指标仅提供数学表达式,缺乏对该评估结果含义和影响因素的说明,使得这些指标难以被实际应用和解释,这可能会产生误导从而引入新的偏见。未来需要深入研究公平性指标的可解释性,提高评估结果的可信度。

3)敏感属性数量和维度受限

通常假设敏感属性为二元属性,当前的公平性指标没有涉及多个群体之间的比较,并且通常限制在一个维度,未来需要针对多维敏感属性集合,设计更加具备应用价值的指标。

4)缺乏隐私保护机制

在使用公平性指标时,需要访问敏感属性,这可能会产生用户隐私泄漏的问题。未来可以结合差分隐私、加密等技术手段将隐私保护纳入公平性指标的设计中。

5)忽略长期影响

目前公平性指标都是基于已有数据集进行评估,忽略了受保护群体的长期利益。未来不仅要考虑个人属性,也要考虑个人与他人、社会的联系,寻找新的动态公平性指标。

结束语 越来越多的研究人员对机器学习公平性指标进行了深入研究,但总体上这方面的研究和应用还处于初级阶段。本文对公平性指标的定义、挑战、未来展望等相关问题进行了科学的分析,旨在为相关研究人员提供一些参考,推动机器学习公平性研究的进一步发展。机器学习公平性指标并不是忽略个体本身特性而强制使不同群体或个人实现待遇均等,而是督促研究人员更细致地了解不同群体、个体之间的差异,从不同视角反向审视研究相关的数据和模型,关注模型的社会意义与价值,从而开发出更具广泛实用性的指标。

参 考 文 献

[1] MAHAJAN S,CARABALLO C,LU Y,et al. Trends in differences in health status and health care access and affordability by race and ethnicity in the United States,1999-2018[J]. Jama,2021,326(7):637-648.

[2] CAUFFMAN C. Discrimination in online advertising[J]. Maastricht Journal of European and Comparative Law,2021,28(3):283-286.

[3] LEE M S,FLORIDI L. Algorithmic fairness in mortgage lending:from absolute conditions to relational trade-offs[J]. Minds and Machines,2021,31(1):165-191.

[4] XU G N,CHEN Y P,CHEN Y M,et al. Data Debiasing Method Based on Constrained Optimized Generative Adversarial Networks[J]. Computer Science,2022,49(6A):184-190.

[5] PEDRESHI D,RUGGIERI S,TURINI F. Discrimination-aware data mining[C]//International Conference on Knowledge Discovery and Data Mining. ACM,2008:560-568.

[6] VERMA S,RUBIN J. Fairness definitions explained[C]//International Workshop on Software Fairness(Fairware). ACM,2018:1-7.

[7] 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.

[8] BERK R,HEIDARI H,JABBARI S,et al. Fairness in criminal justice risk assessments:The state of the art[J]. Sociological Methods Research,2021,50(1):3-44.

[9] ZAFAR M B,VALERA I,GOMEZ R M,et al. Fairness beyond disparate treatment disparate impact:learning classification without disparate mistreatment[C]//International Conference on World Wide Web. ACM,2017:1171-1180.

[10] HARDT M,PRICE E,SREBRO N. Equality of opportunity in supervised learning[J]. Advances in Neural Information Processing Systems,2016,29(1):3323-3331.

[11] CALDERS T,VERWER S. Three naive Bayes approaches for discrimination-free classification[J]. Data Mining and Knowledge Discovery,2010,21(2):277-292.

[12] JIANG Z,HAN X,FAN C,et al. Generalized demographic parity for group fairness[C]//International Conference on Learning Representations. OpenReview. net,2022.

[13] KAMIRAN F,ŽLIOBAITÈ I,CALDERS T. Quantifying explainable discrimination and removing illegal discrimination in automated decision making[J]. Knowledge and Information Systems,2013,35(3):613-644.

[14] XU R,CUI P,KUANG K,et al. Algorithmic decision making with conditional fairness[C]//International Conference on Knowledge Discovery and Data Mining. ACM,2020:2125-2135.

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

[16] STEWART R T. Identity and the limits of fair assessment[J]. Journal of Theoretical Politics,2022,34(3):415-442.

[17] HEDDEN B. On statistical criteria of algorithmic fairness[J]. Philosophy and Public Affairs,2021,49(2):209-231.

[18] CORBETT D S,PIERSON E,FELLER A,et al. Algorithmic decision making and the cost of fairness[C]//International Conference on Knowledge Discovery and Data Mining. ACM,2017:797-806.

[19] ZHAO H,GORDON G. Inherent tradeoffs in learning fair representations[J]. Machine Learning Research, 2022,23(57):1-26.

[20] DEHO O B,ZHAN C,LI J,et al. How do the existing fairness metrics and unfairness mitigation algorithms contribute to ethical learning analytics? [J]. British Journal of Educational Technology,2022,53(4):822-843.

[21] BAROCAS S,SELBST A D. Big data's disparate impact[J]. Calif. L. Rev. ,2016,104:671-732.

[22] CASTELNOVO A,CRUPI R,GRECO G,et al. A clarification of the nuances in the fairness metrics landscape[J]. Scientific Reports,2022,12(1):4209.

[23] DWORK C,HARDT M,PITASSI T,et al. Fairness through awareness[C]//International Conference on Innovations in Theoretical Computer Science. ACM,2012:214-226.

[24] LI A,PEARL J. Unit selection with causal diagram[C]//International Conference on Artificial Intelligence. AAAI, 2022,36(5):5765-5772.

[25] GEBHART V,SMERZI A. Extending the fair sampling assumption using causal diagrams[J]. Quantum,2023,7:897-906.

[26] KILBERTUS N, ROJAS C M, PARASCANDOLO G, et al. Avoiding discrimination through causal reasoning[J]. Advances in Neural Information Processing Systems, 2017, 30(1): 656-666.

[27] KIM H, SHIN S, JANG J H, et al. Counterfactual fairness with disentangled causal effect variational autoencoder[C]// International Conference on Artificial Intelligence. AAAI, 2021, 35(9): 8128-8136.

[28] KUSNER M J, LOFTUS J, RUSSELL C, et al. Counterfactual fairness[J]. Advances in Neural Information Processing Systems, 2017, 30(1): 4066-4076.

[29] PUJOL D, MCKENNA R, KUPPAM S, et al. Fair decision making using privacy-protected data[C]// International Conference on Fairness, Accountability, and Transparency. ACM, 2020: 189-199.

[30] CZARNOWSKA P, VYAS Y, SHAH K. Quantifying social biases in nlp: a generalization and empirical comparison of extrinsic fairness metrics[J]. Transactions of the Association for Computational Linguistics, 2021, 9(1): 1249-1267.

[31] GAO R, GE Y, SHAH C. FAIR: Fairness-aware information retrieval evaluation[J]. Association for Information Science and Technology, 2022, 73(10): 1461-1473.

[32] FRIEDLER S A, SCHEIDEGGER C, VENKATASUBRAMANIAN S, et al. A comparative study of fairness-enhancing interventions in machine learning[C]// International Conference on Fairness, Accountability, and Transparency. ACM, 2019: 329-338.

[33] KLEINBERG J, MULLAINATHAN S, RAGHAVAN M. Inherent trade-offs in the fair determination of risk scores[C]// International Conference on Innovations in Theoretical Computer Science. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017: 43: 1-43: 23.

[34] LIU L T, DEAN S, ROLF E, et al. Delayed impact of fair machine learning[C]// International Conference on Machine Learning. PMLR, 2018: 3150-3158.

[35] BOUVATIER V, EL OUARDI S. Credit gaps as banking crisis predictors: a different tune for middle-and low-income countries [J]. Emerging Markets Review, 2023, 54(C): 101001.

[36] FARNADI G, BABAKI B, GETOOR L. Fairness in relational domains[C]// International Conference on AI, Ethics, and Society. ACM, 2018: 108-114.

[37] CHAO L M, YIN X L. AI Governance and System: Current Situation and Trend[J]. Computer Science, 2021, 48(9): 1-8.

[38] HUGGINS-MANLEY A C, BOOTH B M, D'ELLO S K. Toward Argument-Based Fairness with an Application to AI-Enhanced Educational Assessments[J]. Journal of Educational Measurement, 2022, 59(3): 362-388.

[39] GICHOYA J W, MCCOY L G, CELI L A, et al. Equity in essence: a call for operationalising fairness in machine learning for healthcare[J]. BMJ Health & Care Informatics, 2021, 28(1): e100289.

[40] KALLUS N, ZHOU A. Fairness, welfare, and equity in personalized pricing[C]// Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. 2021: 296-314.

[41] BAKER R S, HAWN A. Algorithmic bias in education[J]. International Journal of Artificial Intelligence in Education, 2022, 32(1): 1052-1092.

[42] REDDY S, ALLAN S, COGHLAN S, et al. A governance model for the application of AI in health care[J]. Journal of the American Medical Informatics Association, 2020, 27(3): 491-497.

[43] DELOBELLE P, TOKPO E K, CALDERS T, et al. Measuring fairness with biased rulers: A comparative study on bias metrics for pre-trained language models[C]// The 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2022). 2022: 1693-1706.

ZHANG Wenqiong, born in 1998, Ph.D candidate. Her main research interest is fairness of machine learning.

LI Yun, born in 1974, Ph.D, professor, Ph.D supervisor. His main research interests include machine learning and pattern recognition.

（责任编辑：喻藜）