智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

公平性机器学习研究综述

邓蔚，邢钰晗，李逸凡，李振华，王国胤

引用本文:
邓蔚, 邢钰晗, 李逸凡, 等. 公平性机器学习研究综述[J]. 智能系统学报, 2020, 15(3): 578–586.
DENG Wei, XING Yuhan, LI Yifan, et al. Survey on fair machine learning[J]. *CAAI Transactions on Intelligent Systems*, 2020, 15(3): 578–586.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202007004

DOI: 10.11992/tis.202007004

公平性机器学习研究综述

邓蔚[1,2], 邢钰晗[1], 李逸凡[1], 李振华[3], 王国胤[2]

（1. 西南财经大学 统计研究中心, 四川 成都 611130; 2. 重庆邮电大学 计算智能重庆市重点实验室, 重庆 400065; 3. 西南财经大学 金融学院, 四川 成都 611130）

摘　要: 随着机器学习在社会中的广泛使用, 带来的歧视问题引发广泛的社会争议, 这逐步引起了产业界和学术界对机器学习算法公平性问题的浓厚兴趣。目前对公平性度量和机器学习公平性机制的研究仍然处于初级阶段。本文对公平性机器学习的研究进行了调研, 首先从公平性的定义出发, 比较了衡量公平性指标的方法, 然后调研了公平性数据集, 对公平性问题的产生进行了分析, 接下来对现有的公平性机器学习算法进行归类和比较, 最后总结了当前公平性机器学习研究中存在的问题, 并对关键问题和重大挑战进行了讨论。

关键词: 算法伦理; 算法偏见; 公平性; 公平性机器学习; 公平性指标; 公平性设计; 公平性数据集; 动态性

中图分类号: TP181　　**文献标志码:** A　　**文章编号:** 1673−4785(2020)03−0578−09

中文引用格式: 邓蔚, 邢钰晗, 李逸凡, 等. 公平性机器学习研究综述 [J]. 智能系统学报, 2020, 15(3): 578–586.

英文引用格式: DENG Wei, XING Yuhan, LI Yifan, et al. Survey on fair machine learning[J]. CAAI transactions on intelligent systems, 2020, 15(3): 578–586.

Survey on fair machine learning

DENG Wei[1,2], XING Yuhan[1], LI Yifan[1], LI Zhenhua[3], WANG Guoyin[2]

(1. Center of Statistical Research, Southwestern University of Finance and Economics, Chengdu 611130, China; 2. Chongqing Key Laboratory of Computational Intelligence, Chongqing University of Posts and Telecommunications, Chongqing 400065, China; 3. School of Finance, Southwestern University of Finance and Economics, Chengdu 611130, China)

Abstract: With the widespread applications of machine learning in our society, the problems of discrimination have caused widespread social controversy. It gradually arouses strong interests in fair machine learning in the industry and academia. Nowdays the deep understanding of the basic issues related to fairness and mechanism of fair machine learning is still in their infancy. We makes a survey on fair machine learning. Starting from the definitions of fairness, it compares the different difinitions on fairness in different problems. Common datasets are also summarized. And the issues of fairness is analyzed. We classify and compare the existing methods of achieving fairness. Finally, we summarizes the problems in current fairness machine learning research and propose the key problems and important challenges in the future.

Keywords: algorithmic ethics; algorithmic discrimination; fairness; fair machine learning; fair indicator; fair design; fair dataset; dynamicity

随着机器学习算法被应用于金融、反恐、征税、司法、医疗、保险等公共领域, 算法不仅为经济社会带来了许多进步, 还决定着公民的福祉[1-3]。然而近些年来, 由于种种原因导致不公平甚至是歧视现象的出现, 如信用评分[4]、犯罪预测[5]和贷款评估[6]等。算法伦理特别是机器学习的公平性问题引起国家、社会和公众的广泛关注[7], 如IEEE(国际电气电子工程师学会) 于 2017 年底发布的《人工智能设计的伦理准则》(第 2 版)[8]、欧盟于 2018 年发布的《人工智能、机器人与自动系统宣言》[9]、第 40 届数据保护与隐私专员国际大会于 2018 年发布的《人工智能伦理与数据保护宣言》[10], 世界经济合作与发展组织 (OED) 于 2019 年发布的《人工智能原则》[11]、G20 组织于 2019 年发布的 *Human-centred AI Principles*[12]、国

收稿日期: 2020−07−02.
基金项目: 国家自然科学基金重点项目 (61936001).
通信作者: 王国胤. E-mail: wanggy@cqupt.edu.cn.

家新一代人工智能治理专业委员会于 2020 年发布的《新一代人工智能治理原则——发展负责任的人工智能》[13]。在以上宣言或原则中, 都不同程度强调了算法的公平性问题。所谓公平性机器学习是将公平性植入到模型设计中, 使得算法对敏感属性 (如种族、性别宗教等) 的分类预测结果对人来说是公平或非歧视的。虽然模型的结果必须考虑准确性指标, 但是也必须考虑其社会影响, 因此对机器学习的公平性的评估和分析显得尤为重要[14-15]。

1　公平性的定义

公平性机器学习算法需要考虑两个密切相关的方面: 首先是在特定社会场景中如何定义公平性, 其次是社会可接受程度。通常假定敏感属性为二元属性, 以下对本文中使用的符号进行定义: S 表示敏感属性或受保护属性; X 表示除了受保护属性外的其他可观测属性; Y 为要被预测的标签; $\hat{Y}$ 为对于 Y 的预测, 是取决于 S、X 的随机变量。

1.1　**无意识公平** (fairness through unawareness, FTU)

如果在算法中不使用受保护属性 S 进行训练及预测, 则这个算法的公平满足 FTU[16]。FTU 模型简单, 但是加入 X 包含类对 S 的歧视性信息, 导致不公平。

1.2　**个体公平** (individual fairness, IF)

IF 是由 Dwork 在 2012 年提出的[17]。如果一个算法对于相似个体的预测结果是相同的, 就称其实现了个体公平。给定一个度量, 如果个体 i 和 j 是相似的, 则对其预测也应该相似[18-20]。Kim 等[21] 改进了告知个人偏好的个体公平 (preference-informed individual fairness, PIIF), 即放松个体公平, 允许偏离 IF 但是要符合个人偏好, 可以为个人提供更为有利的解决方案。

1.3　**人口均等** (demographic parity, DP)/**不同影响** (disparate impact, DI)

如果预测值 $\hat{Y}$ 满足 $P(\hat{Y}|S=0)=P(\hat{Y}|S=1)$, 则算法实现了人口均等 (DP)[21]。

DI 定义为 $P(\hat{Y}|S=0)/P(\hat{Y}|S=1)$。在二分类中, 所有非特权类被分组成一个单独的值 S=0 (例如, “非白种人”), 与特权类 S=1 (例如, “白种人”) 进行比较。在多分类中, 对特权组进行成对 DI 计算, 并取这些计算的平均值作为最终值[22]。

Zafar 等[23] 在将不同影响解释为决策系统的不同结果, 对于具有敏感属性的某些人群, 所产生的不同有益或有害的影响。Beretta 等[24] 将 5 种民主思想与现存的公平概念结合, 提出在不同社会背景中某一特定的民主思想背景下, 如何选取公平性的评价标准, 并指出反事实公平、无意识公平及基于组条件的公平更适用于竞争民主, 个体公平更适用于自由民主, 基于偏好的公平更适用于平等民主。

1.4　**机会均等** (equality of opportunity, EO)

如果预测值满足 $P(\hat{y}=1|S=0,Y=1)=P(\hat{y}|S=1,Y=1)$, 则称算法实现了机会均等。比较在同类别 Y 中的基于不同敏感属性 S, 预测标签的概率是否相等, 称为基于组条件精度的公平。不同的 EO 的衍生定义如表 1 所示。

表 1　不同的机会均等定义

Table 1　Different definitions of EO

名称	数学定义
s的准确率	$P(\hat{Y}=y\mid S=s,Y=y)$
s的召回率	$P(\hat{Y}=1\mid S=s,Y=1)$
s-TNR	$P(\hat{Y}=0\mid S=s,Y=0)$
s-BCR	$\left[P\left(\hat{Y}=0\middle\vert S=s,\ Y=0\right)+P\left(\hat{Y}=1\middle\vert S=s,\ Y=1\right)\right]/2$

Chouldechova[25] 的目标是在敏感群体中实现 1−s-TPR 和 1−s-TNR 值相等, 即错误率平衡。在敏感群体中如果实现了相同的 s-TPR 和 1−s-TNR, 则称机会均衡。

1.5　**不同对待** (disparate treatment)

通常法律通过两个不同概念来评估决策过程的公平性: 不同对待和不同影响。如果决策过程部分或全部基于受试者的敏感属性信息, 则决策过程将导致不同对待。当决策系统为具有相同 (或类似) 非敏感属性 (或特征) 值, 但不同敏感属性值的人群提供不同的输出结果时, 称为不同对待。

Zafar 等[23] 参照上述两种概念定义, 通过从决策过程中删除敏感属性避免不同对待, 以及增加公平约束消除不同影响进行研究。观察到标准的公平约束是非凸的, 其引入协方差将非凸问题转化为凸型, 并通过参数来衡量输出结果及敏感属性, 研究了多分类的敏感属性及多个敏感属性分析问题。

1.6　**不同误判率** (disparate mistreatment, DM)

在决策实现公平性中, 如果历史数据中存在偏见, 那么不同影响及不同对待适合作为公平衡量标准, 但是当训练数据是基于历史事实而来, 使用不同误判率指标来衡量公平性会更准确。当分类器对于不同敏感属性输出错误率不同时, 称为不同误判率 DM, 公式为[23]

$$P(\hat{Y}=1|Y=0,\ S=0)+P(\hat{Y}=0|Y=1,S=0)=P(\hat{Y}=1|Y=0,S=1)+P(\hat{Y}=0|Y=1,S=1)$$

1.7　基于组条件的校准度量

在敏感值均衡中引入校准的概念，即如果 $p(Y=1|\hat{Y}=p)=p$，则输出事件概率 $\hat{Y}$ 的预测因子被称为经过良好校准。可以通过调节校准功能来确定公平性测量[25-26]。

1.8　反事实公平 (counterfactual fairness)

由 Kusner 等[15] 提出，建立在给定结构方程模型 (U,V,F) 的基础上，其中 U 表示潜变量，$V\equiv S\cup X$，F 为一系列方程。如果预测值 $\hat{Y}$ 满足式 (1)，则称为反事实公平：

$$P(\hat{Y}_{S\leftarrow s}(U)=y|X=x,S=s)=P(\hat{Y}_{S\leftarrow s'}(U)=y|X=x,S=s) \tag{1}$$

该定义针对个体层面，如果在现实世界和反事实世界中预测相同，那么对个人是公平的。

1.9　事前公平 (ex-ante fairness) 及事后公平 (ex-post fairness)

事前公平指一个算法 A 满足对于任意一对候选人 x_{ij} 和 $x_{i'j'}$，其累积分布函数 $F_j(x_{ij})>F'_j(x_{i'j'})$，则概率满足 $\mathrm{E}[A(X,x_{ij})]\geqslant\mathrm{E}[A(X,x_{i'j'})]$。

事后公平指一个算法 A 满足对于任意一对候选人 x_{ij} 和 $x_{i'j'}$，其累积分布函数 $F_j(x_{ij})>F'_j(x_{i'j'})$，个体 $x_{i'j'}$ 只当 x_{ij} 也被选中时才被选中[27]。

1.10　综合分析比较

许多不同公平性定义方法在本质上是相关的。Friedler 等[14] 通过分析许多算法的公平性度量，度量了不同定义的相关性，发现不同的公平性指标之间有着非常密切的相关性，分别在 Ricci 和 Adult 数据集上进行了实验，如图 1 所示。

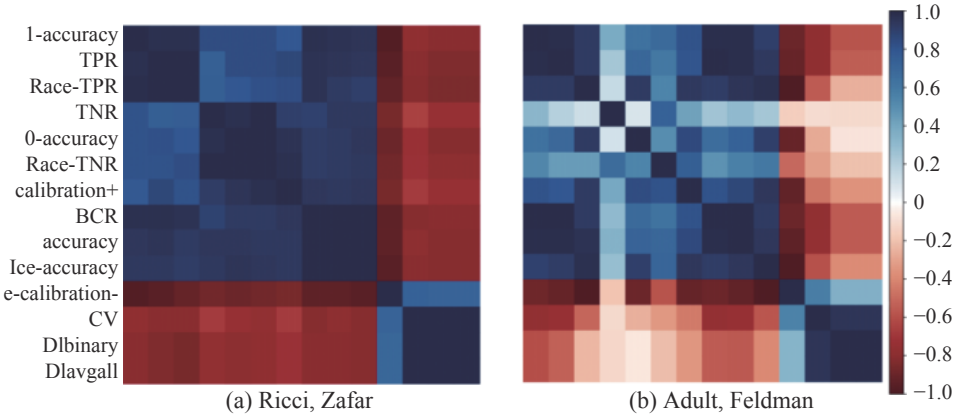

图 1　不同公平性标准之间的关系

Fig. 1　Examining the relationships between different measures of fairness

图 1 表示数据集和算法的相关性分析，展示了不同的公平性度量如何相互关联。各种群体条件下的公平指标之间的关系非常密切。值得注意的是，对负结果的群条件校准测量 (s-calibration) 与其他组条件测量相比，它与基准率测量的关系更为密切。

此外准确性指标与群体条件下的公平指标相关性，表明公平性−准确性的权衡与基准率公平度量更相关。

在某些情况下，期望在不同的公平性度量之间进行权衡。Chouldechova[25] 和 Kleinberg 等[28] 研究表明，假设不同人群比率不相等，则不可能同时实现校准和错误率平衡 (组间相同的假阳性率和相同的假阴性率)。Friedler 等[14] 通过一个实例检验了这种权衡。每种算法的 s-calibration 与 s-TPR 之间存在明显的权衡，且不同的算法位于不同的权衡线上，如图 2 所示。

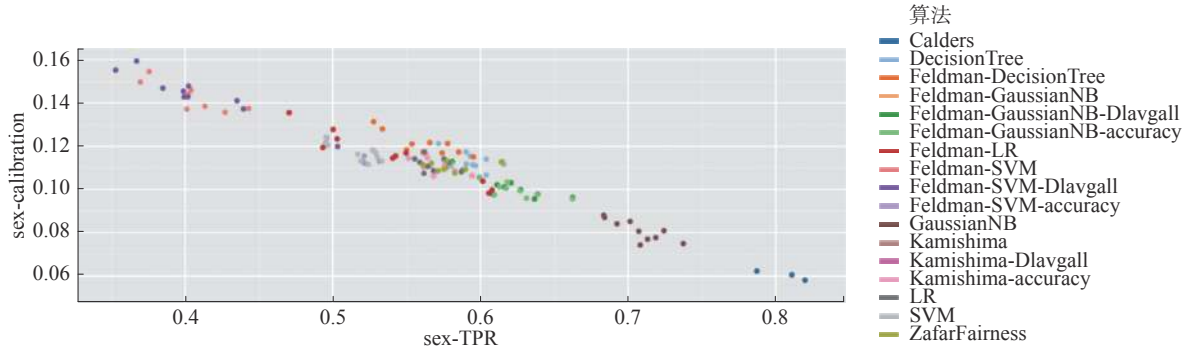

图 2　对于所有算法在 Adult 数据集中 sex-calibration 和 sex-TPR 指标之间的权衡

Fig. 2　Trade-off between s-calibration and s-TPR for all algorithms on the Adult dataset

2 公平性测试数据集

公平性测试数据集和普通数据集的差别在于具有敏感属性, 目前国际上对公平性机器学习算法的测试大多基于以下几个数据集。

2.1 消防员晋升 (ricci)

该数据集来源于美国最高法院诉讼案, 是关于消防员是否会获得晋升的测试数据集, 包括 118 条记录和 5 个属性, 其中有一个敏感属性 (种族), 目标是预测晋升, 同时实现对敏感属性的公平[29]。

2.2 成年人收入 (adult income)

该数据集包含 1994 年美国人口普查的个人信息, 有 45 222 条记录, 14 个属性 (包括年龄、种族、性别、受教育程度等), 还包含有一个二分类标签, 表示每个个体的收入是否超过 50 000 美元[30]。

2.3 德国人信用贷款 (german credit)

该数据集包含 1 000 个记录和 20 个属性, 标签描述了每个人的信用分类为好或坏。敏感属性包含性别、年龄, 性别不直接包含在数据中, 但可以从给定的信息中得到[31]。

2.4 预测再次被捕率

该数据是关于佛罗里达州布劳沃德县使用 COMPAS 风险评估工具评估的数据, 包括 6 167 人的少年重罪数量、当前逮捕的指控程度等信息, 以及敏感属性 (种族和性别), 预测在第一次被捕后的两年内被再次逮捕的概率[32]。

2.5 预测再犯暴力

该数据描述了与上述累犯数据相同的情况, 但预测结果是两年内再次暴力犯罪的结果。预处理后共有 4 010 人, 敏感属性为种族和性别[32]。

2.6 银行营销数据

该数据集有 45 211 条记录, 17 个属性, 还包含一个二分类标签, 表示每个客户是否有定期存款[33]。

2.7 纽约市阻止和搜身 (NYC stop and frisk)

该数据集是阻止、搜索和搜身数据集, 该网站提供了纽约警察局阻止的司机的人口学信息和其他信息, 总共有 45 787 条记录, 具有 73 个特征, 受保护属性为种族[34]。

2.8 美国司法部数据

该数据集包含 1990~2009 年在美国 75 个人口最多的县中 40 个保留了 151 461 名重罪被告的法院处理信息。

3 公平性机器学习设计

从算法的基本定义出发, 用输入、过程及输出不同阶段描述解决问题的策略机制。在输入阶段, 存在问题是输入数据集中存在偏差; 在过程阶段, 存在程序的黑箱问题; 而输出阶段存在算法决策产生的歧视性影响。

3.1 预处理公平性设计

训练数据中存在的不公平是算法能学习的, 如使训练算法不能学习该偏差就能得到预处理公平, 可归为两类: 1) 更改训练数据中单个项目的敏感属性或类标签的值; 2) 将训练数据映射到转换空间, 在该空间中敏感属性和类标签之间的依赖关系消失。

Feldman 等[35] 对每个属性进行修改, 使得基于给定敏感属性子集的边际分布都相等,并且这种改变不会影响其他变量, 转换后的数据保留了非敏感属性的大部分特征信号, 还提出交叉敏感属性, 并且两种敏感属性的影响不叠加。

其他方法包括具有二元敏感属性和二分类问题, 对预处理技术进行改进, 抑制敏感属性, 通过更改类标签来调整数据集, 重新加权或重新采样数据以消除歧视, 而无需重新标记实例[36]。Calmon 等[37] 提出了一种用于学习数据转换的凸优化, 其目标有控制歧视、限制单个数据样本中的失真度以及保持效用。

3.2 过程公平性设计

对特定机器学习算法的改进中, 最常见的是给算法附加约束条件。Kusner 等[15] 将因果模型引入到算法中, 并给出了 3 种实现不同等级算法公平性的方法。1) 运用与敏感属性无直接或间接关联的属性来构建模型; 2) 通过潜在变量来构建模型, 潜在变量是可观测变量的非确定性因素; 3) 通过具有潜变量的确定性模型 (如加性误差模型) 进行建模。Zafar 等[16] 对不同误判率下的逻辑回归及支持向量机算法进行改进, 基于历史信息不存在偏见, 在基于不同误判率的公平性和准确性之间提供了灵活的权衡。当敏感属性信息不可用时, 此方法效果较好。Zemel 等[20] 结合预处理和算法修改, 学习规范数据表示以实现在分类中实现高效, 同时实现不受敏感属性值的影响。Kearns 等[27] 结合事前公平及事后公平, 利用不同个体的累积分布函数, 给定一组个体的得分, 根据候选人的经验值来提供置信区间, 然后将使用的偏差界限为候选人分配分数, 运行 NoisyTop 算法提供近似的公平性。Kamishima 等[38] 引入以公平性为中心的正则化项, 并应用于逻辑回归分类算法中。Calders 等[39] 为敏感属性的每个值构建

了单独的模型，并根据相应输入的属性值来适当选择模型，在 CV 度量下评估迭代组合模型的公平性。Bose 等[40] 针对现有的图嵌入算法无法处理公平约束的问题，在确保学习表示与敏感属性不相关的条件下，通过引入对抗框架来对图嵌入进行公平性约束，使用复合框架去除掉更多的敏感信息。

3.3　后处理公平性设计

Hardt 等[41] 考虑在敏感属性情况下，对不公平类别的概率估计进行后处理，学习不同敏感属性下的不同决策阈值，并在决策时应用这些特定阈值。Kamiran 等[42] 通过在训练后修改决策树中的叶子标签，以满足公平约束。Woodworth 等[43] 采取了一阶矩的统计和计算理论学习非歧视预测，提出了统计上最佳的二阶矩程序，同时在二阶矩上对非歧视性较为松弛，使得算法易于学习。

3.4　公平性算法分类

部分算法比较如表 2 所示。Corbett-davies[44] 等将公平性算法定义分为 3 类，即反分类 (anti-classfication)、分类均等 (classification parity)、校准 (calibration)，并指出 3 种公平性算法的定义都受到统计上的限制，如表 3 所示。

表 2　算法之间的比较

Table 2　Comparison between algorithms

算法	*S*是否是多分类?	一次能处理多个*S*吗?	其他分类数量	分类器
Feldman[35]	√	√	数值型	任何
Kamiran, Calders[36]	×	×	数值型	基于打分的分类器
Calmon[37]	√	×	分类及数值	任何
Kamishima[38]	×	×	数值型	逻辑回归
Calders,Verwer[39]	×	×	分类及数值	朴素贝叶斯
Zafar[23]	√	√	数值型	基于凸边缘分布的分类器
Kusner[15]	×	×	分类及数值	构建结构方程
Kamiran[42]	×	×	分类及数值	决策树
Hardt[41]	√	×	分类及数值	任何
Woodworth[43]	×	×	分类及数值	任何

表 3　公平性算法的分类

Table 3　Algorithm classifications of fair machine learning

公平性算法分类	特点	缺陷
反分类	规定算法不考虑受保护的特征，如种族、性别或其代理人	1) 此策略可确保决策不明确依赖于组成员身份；2) 即使不使用受保护的特征，明显的歧视行为是可能的
分类均等	要求某些常见的预测性能指标在被保护属性限制的群体中是相等的	1) 当风险分布不同时，强制执行分类均等常常会降低所有群体的效用；2) 误解假正率是一个群体总体福祉的合理代表
校准	要求结果是独立于保护属性，控制估计的风险	不足以确保风险分数是准确的或决定是公平的，在评估离散风险评分的校准时必须小心谨慎

4　公平性机器学习应用研究

4.1　环境演化与动态性公平

大多数的公平性算法任务关注于一次性的分类任务，但是在现实的算法系统中包含着很多相互影响的因素。在某些场景中，学习算法本身会作用于复杂环境的动态演化，而公平性也会相应地在不同群体间变化。

在静态环境中，Kannan 等[45] 进行了初步尝试，展示了如何在一个简单、程式化的环境中，使用小额货币支付来激励个人公平的特定概念。Davies 等[46] 在分类中采用公平的福利最大化观点，并描述了附加统计公平约束的成本。

在动态环境中，Google 提出了 ML-fairness-gym 仿真框架[47-48]，可以帮助理解算法如何动态地影响信息系统的环境以及人类行为者的动机。以公平性为导向适用于马尔可夫决策过程框架，具有很高的泛化能力。

4.2　数据的复杂性

某些统计数据十分敏感, 受到法律条款保护。通过加入不同程度的噪音保护敏感数据, 往往会降低机器学习的准确性。

在某些情况下, 常见的隐私算法对个体公平的影响可以忽略不计。但当隐私更严格时或决策涉及人口较少时可能出现严重的不公平现象[49]。

在数据很少甚至是没有数据时, 难以做出准确预测, 其一是公平警告 (fair warnings), 当将系统应用到一个相似数据集时, 可以解释产生不公平性的边界条件; 其二是模型不可知的元学习 (model agnostic meta learning, MAML), 该算法在强化学习和图像识别的方面用很少的数据达到一个很好的效果[50]。

4.3　特定场景下的公平性问题

实现某种程度上的公平需要一定的公平代价 (price of fairness, PoF)[51], 需要在公平性与其他待优化因素之间进行权衡。机器学习的准确率降低, 在资源分配问题中则体现为对资源的不充分利用, 考虑公平性下的资源最大利用率, 小于不考虑公平性时的最大利用率。在需求不确定时分配资源, 存在某些自然分布族, 如指数分布和威尔分布, 能够使在考虑公平性条件下的资源最大可利用率达到无该约束条件下的最大可利用率; 幂律分布虽然不能消除最优解与公平解之间的距离 PoF, 但是能够被一个独立于分布参数的常数所限制[52]。

在计算机视觉领域, 由于 ImageNet 数据大多为人为标注, 并且类别绝大部分是欧美人[53], 因此比如在人脸识别系统中, 对某些种族的人错误率有明显偏高, 因此产生公平性问题。比如基于 ILSVRC 数据集为场景, 这是 ImageNet 的子集[54-56], 有 1 000 种类别, 其中人只有 3 种类别 (司机、棒球员、新娘)。使用 ImageNet 来研究人的类别标签, 发现导致图像识别不公平的根本的 3 个原因: 1)WordNet 中的停滞词汇; 2) 所有类别图像的详尽描述; 3) 图像类别的不平衡。对 ImageNet 的 2832 种类别进行过滤, 只有 158 种类别可以用于图像识别[45]。

在推荐系统中, 需要保证推荐是公平的而不依赖于用户的种族或性别, 系统能够在不暴露自身属性的前提下学习节点嵌入表示以保证隐私。Bose 等[40] 聚焦于对社会图加入不变性约束的可行性, 即生成对特定敏感信息不变的图嵌入。首先训练得到一组过滤器, 以防止对抗式的甄别者将敏感信息与过滤后的嵌入信息进行分类。然后将过滤器以不同的方式组合在一起, 灵活生成对任何敏感属性子集不变的嵌入。

语言建模是一项非常容易受到性别偏见影响的自然语言处理任务, 同时也非常具有实际应用价值, 例如屏幕键盘中的单词预测。Bordia 等[57] 评估了性别偏见对于在文本语料库中训练的单词级别的语言模型的性能影响。

在司法智能应用中, 需要保证算法的公平性。Green 等[58] 认为不应过于偏向研究算法的精确性, 而应该研究怎么设计一个系统使得人们利用和系统的交互, 综合做出公平性判断, 并提出了一个循环算法的框架。

5　关键问题和重大挑战

国际上对公平性机器学习的研究中, 如何设计更好的公平性指标, 如何针对实际问题对公平性机器学习算法进行设计、分析、测试和评估, 这是当前面临的关键问题。本文对公平性机器学习领域面临的重大挑战, 进行了进一步的总结和思考。

1) 数据的复杂性

在许多情况下, 获取代表性数据可能过于困难。同时公平性算法对于输入的变化特别敏感。数据挑战包括在数据标注时的错误、测量误差、偏差和样本不平衡。数据标签错误是实现公平机器学习最严重障碍之一。数据类别不平衡也会影响不公平性学习效果。只有通过尽量准确的标注数据, 以及针对训练数据的特点设计机器学习算法实现公平性。同时, 建设评估公平性算法的数据集也很关键。

2) 公平性表示学习

公平性表示学习包含数据和模型的公平性表示。数据的公平性表示, 是指一方面能够表达出包含数据点相关的大量有用信息, 同时这一表示是公平的, 即包含有关敏感属性信息, 目的是防止机器学习算法产生歧视性行为。模型的公平性表示, 是指把公平性嵌入到机器学习模型中, 成为不可分割的一部分。结合认知理论, 特别是多粒度认知计算进行公平性表示学习是值得探索的方向[59]。

3) 公平性算法的鲁棒性

已经有相当多的研究专注于实验条件下公平性算法的准确性, 然而算法在实际应用中会面临攻击行为, 如何设计鲁棒性公平性机器学习方法是一个公开问题。

4) 公平性算法的动态性

目前大部分研究主要是在静态分类的场景下进行机器学习中的公平性问题研究, 并没有研究这些决定会如何随着时间流逝而影响未来。通常认为给机器学习模型施加公平性的限制以后, 能够更好地改进不公平性, 然而实际情况并非如

此。如何设计动态公平性机器学习算法，实现有反馈和长期的公平性非常重要。

5) 公平性算法的可解释性

以往的研究过于重视机器学习算法的准确性，特别是深度学习的广泛使用，在图像、视频和文本领域实现高准确性的同时，模型越来越复杂。在实际应用中，算法的可解释性变得越来越重要，在设计可解释性模型的同时，务必保证公平性结果的可解释性。

6) 平衡公平性与其他指标

实现公平性会对算法的其他性能指标带来损耗，如何平衡公平性和效率，如何平衡公平性与准确性，如何平衡公平性与可解释性等[60]。如何达到平衡，综合考虑多个指标的影响是值得研究的问题。

7) 建设公平性软件工具箱

目前只有 Google 发布的工具箱 ML-fairness-gym，用来探索在社交网络环境中部署智能决策系统的影响。公平性机器学习软件工具的开发和应用是丰富生态系统的重要举措。

8) 社会环境适用性

公平性设计需要结合各国的法律，结合各国的规章制度，以及不同应用系统的要求，如为非歧视性雇佣、量刑指导风险评估和贷款分配，实现应用领域的公平性。算法的公平性需要与社会公正的平等概念建立更深层次的联系，才能避免狭隘的技术解决方案，建立更公平的机器学习模型。

6　结束语

随着机器学习算法的广泛应用，当前人工智能伦理和人工智能治理引起了社会的广泛关注，算法公平性成为研究热点，不过研究和应用仍处于起步阶段，还有很多问题值得深入研究。

本文针对公平性机器学习的研究现状进行了综述，对公平性算法的实现机制进行了调研，并总结了当前面临的重大问题和挑战，希望给相关研究人员提供一些参考。只有实现了机器学习算法在应用中的公平性，人工智能才能更好地服务于人类，才能更好地发挥其社会和经济价值。

参考文献：

[1] 高庆吉, 赵志华, 徐达, 等. 语音情感识别研究综述 [J]. 智能系统学报, 2020, 15(1): 1–13.
GAO Qingji, ZHAO Zhihua, XU Da, et al. Review on speech emotion recognition research[J]. CAAI transactions on intelligent systems, 2020, 15(1): 1–13.

[2] YOCHUM P, 常亮, 古天龙, 等. 基于位置和开放链接数据的旅游推荐系统综述 [J]. 智能系统学报, 2020, 15(1): 25–32.
YOCHUM P, CHANG Liang, GU Tianlong, et al. A review of linked open data in location-based recommendation system in the tourism domain[J]. CAAI transactions on intelligent systems, 2020, 15(1): 25–32.

[3] 常乐, 杨忠, 张秋雁, 等. 悬挂负载空中机器人的抗摆控制 [J]. 应用科技, 2020, 47(2): 17–22.
CHANG Le, YANG Zhong, ZHANG Qiuyan, et al. Anti-swing control research of aerial robot with suspended load[J]. Applied science and technology, 2020, 47(2): 17–22.

[4] KHANDANI A E, KIM A J, LO A W. Consumer credit-risk models via machine-learning algorithms[J]. Journal of banking and finance, 2010, 34(11): 2767–2787.

[5] BRENNAN T, DIETERICH W, EHRET B. Evaluating the predictive validity of the compas risk and needs assessment system[J]. Criminal justice and behavior, 2009, 36(1): 21–40.

[6] MAHONEY J F, MOHEN J M. Method and system for loan origination and underwriting[P]. US: 7287008.1, 2007-10-23.

[7] KEARNS M, ROTH A. The ethical algorithm: the science of socially aware algorithm design[M]. New York: Oxford University Press, 2019: 11.

[8] IEEE 新版“人工智能设计的伦理准则”白皮书全球重磅发布 [EB/OL]. (2017-12-15)[2020-07-26] https://www.sohu.com/a/210646713_468720.

[9] Publications Office of the EU[EB/OL]. (2018-03-09)[2020-07-26] https://op.europa.eu/en/publication-detail/-/publication/dfebe62e-4ce9-11e8-be1d-01aa75ed71a1/language-en/format-PDF/source-78120382.

[10] 吴沈括, 周洁, 杨滢滢. 人工智能伦理与数据保护宣言 [EB/OL]. (2018-10-30)[2020-07-26]. http://www.yidianzixun.com/m/article/0KOD5oLY.

[11] OECD Principles on AI[EB/OL]. [2020-07-26] https://www.oecd.org/going-digital/ai/principles/.

[12] G20 ministerial statement on trade and digital economy[EB/OL]. (2019-06-09)[2020-07-26] http://www.g20.utoronto.ca/2019/2019-g20-trade.html.

[13] 国家新一代人工智能治理专业委员会. 发展负责任的人工智能: 新一代人工智能治理原则发布 [EB/OL]. (2019-06-17)[2020-07-26] http://www.most.gov.cn/kjbgz/201906/t20190617_147107.htm.

[14] FRIEDLER S A, SCHEIDEGGER C, VENKATASUBRAMANIAN S, et al. A comparative study of fairness-enhancing interventions in machine learning[C]//Proceedings of the Conference on Fairness, Accountability, and Transparency. New York, USA, 2019: 329−338.

[15] KUSNER M, LOFTUS J, RUSSEL C, et al. Counterfactual fairness[C]//Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017). Long Beach, USA, 2017.

[16] GRGIĆ-HLAČA N, ZAFAR M B, GUMMADI K P, et al. The case for process fairness in learning: feature selec-

tion for fair decision making[C]//Symposium on Machine Learning and the Law at the 29th Conference on Neural Information Processing Systems. Barcelona, Spain, 2016: 1.

[17] DWORK C, HARDT M, PITASSI T, et al. Fairness through awareness[C]//Proceedings of the 3rd Innovations in Theoretical Computer Science Conference. New York, USA, 2012: 214−226.

[18] JOSEPH M, KEARNS M, MORGENSTERN J, et al. Rawlsian fairness for machine learning [DB/OL]. (2017-06-29)[2020-08-07] arXiv preprint arXiv:1610. 09559V2, arxiv.org/abs/1610.09559v2, 2016.

[19] LOUIZOS C, SWERSKY K, LI Yujia, et al. The variational fair autoencoder[C]//Proceedings of the 4th International Conference on Learning Representations. San Juan, Puerto Rico, 2016.

[20] ZEMEL R, WU Yu, SWERSKY K, et al. Learning fair representations[C]//Proceedings of the 30th International Conference on International Conference on Machine Learning. Atlanta, USA, 2013: 325−333.

[21] KIM M P, KOROLOVA A, ROTHBLUM G N, et al. Preference-informed fairness[C]//Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. New York, USA, 2020: 546.

[22] ZAFA M B, VALERA I, ROGRIGUEZ M G, et al. Fairness constraints: mechanisms for fair classification[C]//Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. Lille, France, 2017: 962−970.

[23] ZAFAR M B, VALERA I, RODRIGUEZ M G, et al. Fairness beyond disparate treatment & disparate impact: learning classification without disparate mistreatment[C]//Proceedings of the 26th International Conference on World Wide Web. Perth, Australia, 2017: 1171−1180.

[24] BERETTA E, SANTANGELO A, LEPRI B, et al. The invisible power of fairness. How machine learning shapes democracy [DB/OL]. (2019-03-22)[2020-07-26] arXiv preprint arXiv:1903.09493v1, https://arxiv.org/abs/1903.09493, 2019.

[25] CHOULDECHOVA A. Fair prediction with disparate impact: a study of bias in recidivism prediction instruments[J]. Big data, 2017, 5(2): 153–163.

[26] BAROCAS S, SELBST A D. Big data's disparate impact[J]. California law review, 2016, 104: 671–732.

[27] KEARNS M, ROTH A, WU Z S. Meritocratic fairness for cross-population selection[C]//Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia, 2017: 1828−1836.

[28] KLEINBERG J, MULLAINATHAN S, RAGHAVAN M. Inherent trade-offs in the fair determination of risk scores[C]//Proceedings of the 8th Innovations in Theoretical Computer Science Conference. Dagstuhl, Germany, 2017.

[29] Supreme Court of the United States. Ricci v. DeStefano [EB/OL]. (2009-06-29)[2020-08-07]. 557 U.S. 557,https://supreme.justia.com/cases/federal/us/557/557/, 2009.

[30] Adult data[EB/OL]. [2020-07-26]. http://tinyurl.com/UCI-Adult, 1996.

[31] LICHMAN M. UCI machine learning repository[EB/OL]. (2013)[2020-07-26]. http://archive.ics.uci.edu/ml, 2013.

[32] ANGWIN J, LARSON J, MATTU S, et al. Machine bias. risk assessments in criminal sentencing[EB/OL]. (2016-05-23)[2020-07-26] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, 2016.

[33] Bank Marketing Data Set [EB/OL]. (2012-02-14) [2020-07-26] https://archive.ics.uci.edu/ml/datasets/ Bank%2BMarketing, 2012.

[34] KHADEMI A, LEE S, FOLEY D, et al. Fairness in algorithmic decision making: an excursion through the lens of causality[C]//The World Wide Web Conference. San Francisco, USA, 2019: 2907−2914.

[35] 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. New York, USA, 2015: 259−268.

[36] KAMIRAN F, CALDERS T. Data preprocessing techniques for classification without discrimination[J]. Knowledge and information systems, 2012, 33(1): 1–33.

[37] CALMON F P, WEI D, VINZAMURI B, et al. Optimized pre-processing for discrimination prevention[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, USA, 2017: 3995−4004.

[38] KAMISHIMA T, AKAHO S, ASOH H, et al. Fairness-aware classifier with prejudice remover regularizer[M]// FLACH P A, DE BIE T, CRISTIANINI N. Machine Learning and Knowledge Discovery in Databases. Berlin: Springer, 2012: 35−50.

[39] CALDERS T, VERWER S. Three naive Bayes approaches for discrimination-free classification[J]. Data mining and knowledge discovery, 2010, 21(2): 277–292.

[40] BOSE A J, HAMILTON W. Compositional fairness constraints for graph embeddings [DB/OL]. (2019-07-16)[2020-07-07] https://arxiv.org/abs/1905.10674, 2019.

[41] HARDT M, PRICE E, SREBRO N. Equality of opportunity in supervised learning[C]//Proceedings of the 30th International Conference on Neural Information Processing Systems. Red Hook, USA, 2016: 3315−3323.

[42] KAMIRAN F, CALDERS T. Classifying without discriminating[C]//Proceedings of 2009 2nd International Conference on Computer, Control and Communication. Karachi, Pakistan, 2009.

[43] WOODWORTH B, GUNASEKAR S, OHANNESSIAN M I, et al. Learning non-discriminatory predic-

tors [EB/OL]. (2017-11-01)[2020-07-07] https://arxiv.org/abs/1702.06081, 2017.

[44] CORBETT-DAVIES S, GOEL S. The measure and mismeasure of fairness: a critical review of fair machine learning [DB/OL]. (2018-08-14)[2020-07-07] https://arxiv.org/abs/1808. 00023, 2018.

[45] KANNAN S, KEARNS M, MORGENSTERN J, et al. Fairness incentives for myopic agents[C]//Proceedings of the 2017 ACM Conference on Economics and Computation. New York, USA, 2017: 369−386.

[46] CORBETT-DAVIES S, PIERSON E, FELLER A, et al. Algorithmic decision making and the cost of fairness[C]//Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA, 2017: 797−806.

[47] D'AMOUR A, SRINIVASAN H, ATWOOD J, et al. Fairness is not static: deeper understanding of long term fairness via simulation studies[C]//Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. Barcelona, Spain, 2020: 525−534.

[48] Google/ml-fairness-gym[EB/OL]. [2020-07-26] https://github.com/google/ml-fairness-gym/.

[49] KUPPAM S, MCKENNA R, PUJOL D, et al. Fair decision making using privacy-protected data [DE/OL]. (2020-01-24)[2020-08-07] https://arxiv.org/abs/1905.12744, 2020.

[50] SLACK D, FRIEDLER S A, GIVENTAL E. Fairness warnings and fair-MAML: learning fairly with minimal data[C]//Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. Barcelona, Spain, 2019: 200−209.

[51] GANCHEV K, KEARNS M, NEVMYVAKA Y, et al. Censored exploration and the dark pool problem[C]//Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence. Arlington, USA, 2009: 185−194.

[52] DONAHUE K, KLEINBERG J. Fairness and utilization in allocating resources with uncertain demand[C]//Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. New York, USA, 2020: 658−668.

[53] DEVRIES T, MISRA I, WANG C, et al. 2019. Does object recognition work for everyone? [EB/OL]. (2019-06-18)[2020-07-07] https://arxiv.org/abs/1906.02659, 2019.

[54] STOCK P, CISSE M. ConvNets and ImageNet beyond accuracy: understanding mistakes and uncovering biases[C]//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany, 2018: 498−512.

[55] DULHANTY C, WONG A. Auditing imageNet: towards a model-driven framework for annotating demographic attributes of large-scale image datasets [EB/OL]. (2019-06-04)[2020-07-07] https://arxiv.org/abs/1905.01347, 2019.

[56] YANG Kaiyu, QINAMI K, FEI-FEI L, et al. Towards fairer datasets: filtering and balancing the distribution of the people subtree in the ImageNet hierarchy[C]//Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. New York, USA, 2020: 547−558.

[57] BORDIA S, BOWMAN S R. Identifying and reducing gender bias in word-level language models[C]//Proceedings of the 9th American Chapter of the Association for Computational Linguistics. Minneapolis, Minnesota, 2019: 7−15.

[58] GREEN B, CHEN Yiling. Disparate interactions: an algorithm-in-the-loop analysis of fairness in risk assessments[C]//Proceedings of the Conference on Fairness, Accountability, and Transparency. Atlanta, USA, 2019: 90−99.

[59] SONG Jiaming, KALLURI P, GROVER A, et al. Learning Controllable Fair Representations[C]//Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics. Naha, Japan, 2019: 2164−2173.

[60] LIU L T, DEAN S, ROLF E, et al. Delayed impact of fair machine learning[C]//Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden, 2018: 3150−3158.

作者简介：

邓蔚，讲师，博士后，主要研究方向为知识图谱、机器行为学、计算社会科学与算法伦理。近年来参与国家自然科学基金重点项目、国家重点研发计划等国家级项目 3 项。申请国家发明专利 10 余项，发表学术论文 30 余篇，出版学术著作 1 部。

邢钰晗，硕士研究生，主要研究方向为公平性机器学习和数据科学。

王国胤，教授，博士生导师，重庆邮电大学副校长，研究生院院长，人工智能学院院长，中国人工智能学会副理事长，主要研究方向为粗糙集、粒计算和认知计算。近年来承担多个国家重点研发计划、国家自然科学基金重点项目等。发表学术论文 300 余篇，出版专著 10 余部。