软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2021,32(5):1404−1426 [doi: 10.13328/j.cnki.jos.006214]

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

可信机器学习的公平性综述*

刘文炎[1], 沈楚云[2], 王祥丰[2,3], 金 博[2,3], 卢兴见[2], 王晓玲[2,3], 查宏远[2,3], 何积丰[1,3]

[1](华东师范大学 软件工程学院,上海 200062)

[2](华东师范大学 计算机科学与技术学院,上海 200062)

[3](上海自主智能无人系统科学中心 可信人工智能研究所,上海 200092)

通讯作者: 王祥丰, E-mail: xfwang@cs.ecnu.edu.cn; 王晓玲, E-mail: xlwang@cs.ecnu.edu.cn

摘 要: 人工智能在与人类生活息息相关的场景中自主决策时,正逐渐面临法律或伦理的问题或风险.可信机器学习是建立安全人工智能系统的核心技术,是人工智能领域的热门研究方向,而公平性是可信机器学习的重要考量.公平性旨在研究机器学习算法决策对个人或群体不存在因其固有或后天属性所引起的偏见或偏爱.从公平表征、公平建模和公平决策这 3 个角度出发,以典型案例中不公平问题及其危害为驱动,分析数据和算法中造成不公平的潜在原因,建立机器学习中的公平性抽象定义及其分类体系,进一步研究用于消除不公平的机制.可信机器学习中的公平性研究在人工智能多个领域中处于起步阶段,如计算机视觉、自然语言处理、推荐系统、多智能体系统和联邦学习等.建立具备公平决策能力的人工智能算法,是加速推广人工智能落地的必要条件,且极具理论意义和应用价值.

关键词: 可信人工智能;可信机器学习;公平性;统计公平;因果公平;公平表征;公平建模;公平决策

中图法分类号: TP18

中文引用格式: 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

英文引用格式: 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). http://www.jos.org.cn/1000-9825/6214.htm

Survey on Fairness in Trustworthy Machine Learning

LIU Wen-Yan[1], SHEN Chu-Yun[2], WANG Xiang-Feng[2,3], JIN Bo[2,3], LU Xing-Jian[2], WANG Xiao-Ling[2,3], ZHA Hong-Yuan[2,3], HE Ji-Feng[1,3]

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

[2](School of Computer Science and Technology, East China Normal University, Shanghai 200062, China)

[3](Institute of Trusted Artificial Intelligence, Shanghai Research Institute for Intelligent Autonomous Systems, Shanghai 200092, China)

Abstract: Artificial intelligence raises legal and ethical issues or risks when used to automated decision-making in areas closely related to daily life. Trustworthy machine learning is the core technology in artificial intelligence safety. It is a trending research direction, of which fairness is an essential aspect. Fairness is the absence of any prejudice or favoritism towards an individual or a group based on their inherent or acquired characteristics which are irrelevant in the particular context of decision-making. A comprehensive and structured

* 基金项目: 上海市科委创新行动计划人工智能科技支持专项(20DZ1100300, 20511101100); 国家自然科学基金(61972155, 61672231, 12071145); 上海市自然科学基金(19ZR1414200); 国家重点研发计划(2020AAA0107400, 2018YFB2101300)

Foundation item: AI Project of Innovation Action Plan of Science and Technology Commission of Shanghai Municipality (STCSM) (20DZ1100300, 20511101100); National Natural Science Foundation of China (61972155, 61672231, 12071145); Natural Science Foundation of Shanghai Municipality (19ZR1414200); National Key Research and Development Program of China (2020AAA0107400, 2018YFB2101300)

收稿时间: 2020-07-19; 修改时间: 2020-10-02, 2020-11-07; 采用时间: 2020-12-01; jos 在线出版时间: 2021-01-15

overview of three research contents is provided, namely, fair representation, fair modeling, and fair decision-making algorithm. The potential causes and harmful consequences of unfairness are first identified in data and algorithm processing. Then, the abstract definition and primary mechanisms for eliminating unfairness are summarized. The research on fairness is at its early stage in fields such as computer vision, natural language processing, recommender systems, multi-agent systems, and federated learning. Fairness is a prerequisite for the application of machine learning, and constructing fair algorithms has theoretical significance and practical values.

Key words: trustworthy artificial intelligence; trustworthy machine learning; fairness; statistical fairness; causal fairness; fair representation; fair modeling; fair decision-making

机器学习通过计算的手段,利用以数据形式存在的经验来改善系统的能力与性能[1].机器学习是智能计算的核心技术,受到了学术界和产业界的广泛关注,在计算机视觉、自然语言处理、语音识别、数据挖掘和信息检索等应用领域取得了巨大突破.随着人类社会被机器学习逐渐渗透,机器学习技术影响着人们生活,如果利用不当,甚至会损害人类的利益.人类和机器学习的关系也引发了新的法律、伦理以及技术问题.例如:优步无人驾驶系统设计中没有考虑到不守规则、横闯马路的行人,导致致命的交通事故;脸书用户资料遭剑桥分析公司窃取,该公司基于窃取的用户资料有针对性地推送政治广告,涉嫌操纵舆论宣传.在这样的背景下,机器学习的可信属性孕育而出[2,3],即公平性、隐私性[4,5]、透明性、鲁棒性和可解释性[6]等,并受到国际各界的重视.欧洲联盟委员会于 2019 年 4 月发布《可信人工智能的伦理指南》(Ethics guidelines for trustworthy AI);美国国家科学技术委员会于 2019 年 6 月更新《国家人工智能研究与发展战略规划》(The national artificial intelligence research and development strategic plan),重点关注机器学习算法的合法性、道德性和鲁棒性;中国科技部于同月发布《新一代人工智能治理原则——发展负责任的人工智能》,提出人工智能治理的框架和行动指南.

公平指处理事情合情合理,不偏袒任何一方.公平机器学习算法指在决策过程中,对个人或群体不存在因其固有或后天的属性所引起的偏见或偏爱[7].机器学习算法因数据驱动,可能在无意中编码人类偏见.一个典型案例是 ProPublica 组织发现:美国法院使用的替代性制裁犯罪矫正管理剖析软件(correctional offender management profiling for alternative sanctions,简称 COMPAS)将非裔美国被告人与高风险累犯评分联系在一起,从而给予更严厉的监禁判决.除此之外,雇佣、保险和广告等领域也发现了类似问题.

算法公平性是机器学习向善的重要主题之一,建立合理的模型保证算法的决策客观,是加速推广机器学习落地的必要条件,具有理论意义和应用价值.美国计算机学会 ACM 于 2018 年开始专门设立 FAccT 会议(ACM Conf. on Fairness, Accountability, and Transparency),研讨包括计算机科学、统计学、法律、社会科学和人文科学等交叉领域的公平性、问责制和透明度问题.此外,包括 ICML、NeurIPS 和 AAAI 在内的多个人工智能重要国际会议专门设置研究专题讨论公平机器学习.

在政府机构指导性原则引导下,学术界和产业界正着力推动公平机器学习理论、技术及应用发展.本文的主旨是梳理目前机器学习公平性研究的现状,并为后续研究提供可借鉴的思路.机器学习算法通常包括以下关键环节:数据所有者采集数据,模型提供者设计算法,算法使用者运行并做出决策.公平机器学习研究的关键问题是如何建立以法律、伦理、社会学为引导的公平性定义,以及如何设计公平性定义驱动的公平机器学习算法.图 1 总体介绍了公平机器学习算法设计流程框架.

- 首先,明确公平目标,即确定符合应用需求的公平性目标,常用的公平性目标包括感知公平性、统计公平性和因果公平性.
- 其次,明确公平任务,即确定面向公平目标的算法公平性提升任务,包括公平表征任务、公平建模任务和公平决策任务:公平表征任务旨在建立公平数据集或提取公平数据特征;公平建模任务旨在建立公平机器学习模型;而公平决策任务将机器学习模型视为黑盒模型,旨在利用其输出结果进行公平决策.处理机制的选择通常对应机器学习算法的关键环节,具体包括预处理机制、处理中机制和后处理机制.
- 最后,从理论分析和实验评估两个角度分别验证公平机器学习的表现.

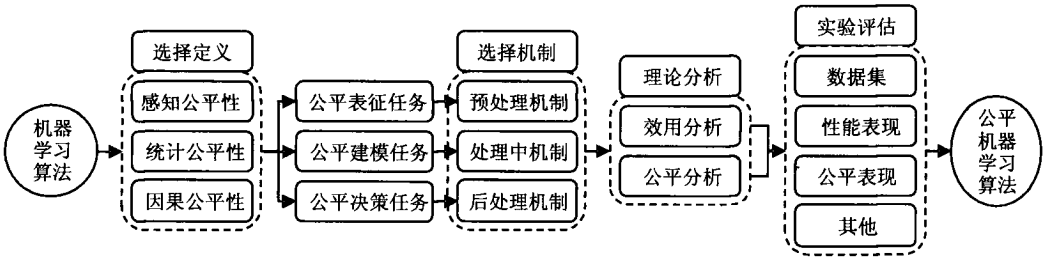

Fig.1 From machine learning algorithm to fair machine learning algorithm

图 1 从机器学习算法到公平机器学习算法

本文的主要贡献如下:明确公平机器学习算法的设计流程框架,形式化公平性定义,给出公平性定义的分类体系,总结并综述 3 类公平性任务,系统性梳理未来的研究方向,有助于指导后续研究者针对公平性理论的研究和探索.

本文第 1 节列举算法不公平产生的危害,探讨造成该现象的潜在原因,提供消除算法偏差的机制.第 2 节提取机器学习中公平性定义的抽象模型,比较现有机器学习的公平性定义.第 3 节详述解决公平表征任务、公平建模任务和公平决策任务的具体方法.第 4 节举例说明公平机器学习的应用,并列举供研究的数据集和检测工具.第 5 节指出公平机器学习的研究问题及其面临的挑战.第 6 节对全文进行总结.

1 公平机器学习的问题定义

本节从机器学习算法辅助决策的多个典型案例入手,分析可能受到不公平对待的对象以及对他们造成的潜在危害,进一步讨论出现这种现象的潜在原因,并梳理列举在算法中消除偏差的 3 种机制,以指导后续的文献分析.表 1 列举本文使用的符号及其意义.

Table 1 Notations

表 1 符号表

符号	描述	符号	描述	符号	描述	符号	描述
$\boldsymbol{S}$	输入空间	$\boldsymbol{X}$	非受保护属性集合	$\mathcal{A}$	受保护属性集合	$\boldsymbol{Z}$	数据特征表示
$\boldsymbol{s}$	$\in\boldsymbol{S}$,输入元素	$\boldsymbol{x}$	$\in\boldsymbol{X}$,非受保护属性	$\boldsymbol{a}$	$\in\mathcal{A}$,受保护属性	π	因果路径
$\boldsymbol{O}$	观察变量集合	Π	因果路径集合	y	$\in\boldsymbol{Y}$,输出元素	$\hat{y}$	$\in\hat{\boldsymbol{Y}}$,预测标记
$\boldsymbol{Y}$	输出空间	$\hat{\boldsymbol{Y}}$	预测标记集合	$\tilde{\boldsymbol{Y}}$	符合公平性定义的输出结果	$\tilde{\boldsymbol{Y}}_a$	干预 a 后的输出结果
f	机器学习模型	f'	符合公平性定义的机器学习模型	g	数据特征变换	h	输出结果变换
δ	公平参数	$\varepsilon_1,\varepsilon_2$	近似控制参数	τ	阈值变量	n	样本数量
$\boldsymbol{A}\perp\boldsymbol{B}\|\boldsymbol{C}$	$\boldsymbol{A}$ 和 $\boldsymbol{B}$ 在给定 $\boldsymbol{C}$ 发生时条件独立	$D(\cdot),d(\cdot)$	距离度量	do(·)	do 算子,干预观察变量	$\mathbb{E}(\cdot)$	数学期望

机器学习的目标是,从训练集中学得数据的潜在规律.以监督学习的预测任务[1]为例,以包含 n 个样本的训练集 $\{(\boldsymbol{s}_i,\boldsymbol{y}_i)\}_{i=1}^n$ 为基础,建立从输入空间 $\boldsymbol{S}$ 到输出空间 $\boldsymbol{Y}$ 的映射:

$$f:\boldsymbol{S}\mapsto\boldsymbol{Y}.$$

通过学习到的模型 f 对 $\boldsymbol{s}$ 进行预测,得到其预测标记 $\hat{\boldsymbol{y}}=f(\boldsymbol{s})$.在样本的属性 $\boldsymbol{s}$ 中,部分属性 $\boldsymbol{a}\in\mathcal{A}$ 记为受保护属性,且 $\mathcal{A}\subseteq\mathbb{R}^{|\boldsymbol{a}|}$ 为受保护属性集合,其中,$|\boldsymbol{a}|$ 表示受保护属性的维度,$\mathbb{R}^{|\boldsymbol{a}|}$ 表示 $|\boldsymbol{a}|$ 维实数空间.$\boldsymbol{x}\in\boldsymbol{X}$ 指不含受保护属性的部分,即 $\boldsymbol{X}\cup\mathcal{A}=S$ 且 $\boldsymbol{X}\cap\mathcal{A}=\varnothing$.此时,映射可重定义为

$$f:(\boldsymbol{X},\mathcal{A})\mapsto\boldsymbol{Y}.$$

不失一般性,不妨假设受保护属性 $\boldsymbol{a}=0$ 和 $\boldsymbol{a}=1$ 的群体分别是弱势群体和非弱势群体.为了叙述简洁,本文使用二分类任务来进行说明,即令 $\boldsymbol{Y}=\{0,1\}$,所得结论很容易推广到其他机器学习任务上.

1.1 不公平算法的危害及典型案例

机器学习算法正在参与到社会生活的各个层面,影响着人们的重要决策.为了直观地说明机器学习算法可能产生的不公平现象,如图 2 所示,本节以真实世界中累犯预判、电商杀熟和保险定价为例,观察理应受保护的对象在不考虑公平性的机器学习算法中受到的伤害.

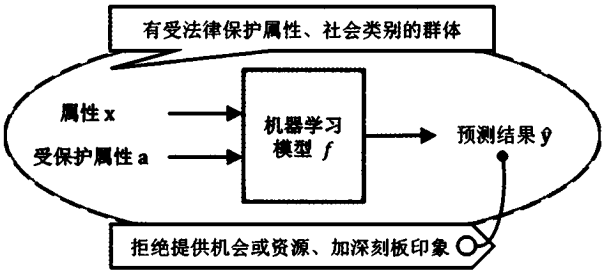

Fig.2 Conceptual graph of unfair machine learning algorithms

图 2 不公平机器学习算法概念图

1.1.1 累犯预判

美国的 COMPAS 风险评估工具 f 根据被告的犯罪记录、犯罪类型、和社区的联系记录以及未能出庭的历史记录等信息 $\boldsymbol{x}$ 来预判其构成累犯的风险 $\hat{y}$,以协助法官做出保释的决定,如图 2 所示.ProPublica 团队发现:该系统将非裔(a=0)被告人错误地标记为高风险人群的概率,几乎是欧裔(a=1)被告人的 2 倍.

在这个案例中,受到区别对待的对象处于弱势地位,具有受法律保护的属性 A,如种族、性别、年龄等.在不考虑公平性的机器学习算法中,依据有偏差的 Y 做出决策,剥夺了他们平等的机会或资源.

1.1.2 电商杀熟

某在线旅游经营者对同一商品或服务在相同条件 $\boldsymbol{x}$ 下设置了差异化的价格,老客户看到的价格 $\hat{y}$ 比新客户看到的高,如图 2 所示.这些电子商务平台故意构建消费习惯评估模型 f,或根据客户的搜索记录预测其购买需求,或根据客户过往的消费记录绘制用户画像,或根据客户对优惠券的操作习惯判断其对价格的敏感程度.从而 f 辅助电子商务平台实现对潜在消费目标涨价.在考虑公平性的机器学习算法中,搜索记录、消费记录以及操作习惯均可被设定为受保护属性 $\boldsymbol{a}$,变换甚至不使用 $\boldsymbol{a}$.在手机应用商店中也有类似的情况发生,即同一个应用的开发者对不同设备类型 $\boldsymbol{a}$ 的用户的消费能力有不同的评估,从而导致定价 $\hat{y}$ 不同.

在这些案例中,有不同消费特征的旅游者或手机用户受到了价格歧视,他们的知情权和公平交易权被侵犯.在不考虑公平性的机器学习算法中,这些群体具有不同的受保护社会属性 A,如语言、文化、位置、收入等,依据有偏差的 Y 做出决策,他们得到的服务质量不同.

1.1.3 保险定价

某汽车保险公司建立事故理赔评估模型 f,通过预测车主的事故发生率来进行保险定价 $\hat{y}$[8],如图 2 所示.假设发生攻击性驾驶行为的车主有较大几率有红色汽车,且某性别的车主偏爱红色汽车(此时,受保护属性 $\boldsymbol{a}$ 是汽车颜色和性别).实际上,在有相同颜色汽车的车主中,该性别的车主没有比其他车主引发更多的交通事故,但是却被收取了更多的保险费用 $\hat{y}$.

这个案例中,保险公司在使用机器学习算法的过程中,并非故意加深对某性别车主的刻板印象,而是错误地认识汽车颜色、性别 A 和事故发生率 Y 间的因果关系,从而导致了诋毁特定车主的结果发生.

以上应用案例敦促我们反思机器学习算法造成有害结果的潜在原因,以帮助人们做出更公平的决策.

1.2 算法不公平的潜在原因

在机器学习算法中,造成不公平的原因是多方面的.例如,不正确地解读并使用算法的结果可能导致不公平的发生.文献[9,10]详细阐述了机器学习算法的偏差来源及其分类描述.本节重点关注机器学习算法的不公平,

即机器学习的偏差.通过梳理归纳机器学习中普遍存在的偏差,我们可以将其归纳为数据的偏差和模型的偏差.

1.2.1 数据的偏差

随着计算设备的普及,与日常生活相关的海量应用落地,人们产生并存储数据信息愈加方便.因为人们的认知水平不同,所以收集到的数据质量也不相同.这些数据可能包含现实世界中人们的认知偏差.根据数据中偏差的状态,可分为静态的历史偏差和动态的交互偏差.

(1) 历史偏差[9]:历史偏差是现实世界中早已长期存在的,体现在数据的属性和标记中,可能导致下游学习任务有偏或不准确的预测.在累犯预判的案例中,审前释放、量刑和假释等决策都是在人类直觉和个人偏见下产生的.如果机器学习算法不加甄别地学习这些潜在规律,那么它将编码对数据主体的偏差,其预测结果将反应不公平.

(2) 交互偏差[11,12]:交互偏差通常来自有偏差策略的使用、用户有偏差的行为以及有偏差的反馈.这些有偏差的交互产生的数据集是倾斜的,这种倾斜可能随着时间而加剧.在电商杀熟的案例中,电子商务平台的记录来自那些已完成的交易,平台倾向于对价格敏感程度低的客户投放更多高价的商品广告,导致该客户群体更可能产生高额的支付记录.未来观察(客户产生高额的消费)证实预测结果(客户对价格的敏感程度低)的可能性增高.因此,使用这些训练数据的机器学习算法倾向于错误地评估客户真实的消费意图,减少与预测结果不同的观察机会.

1.2.2 模型的偏差

在机器学习建模过程中,有多个步骤依赖人们参与并做出决定,而人们的决定对结果的公平与否有着重要的影响:描述样本的特征需要由人类专家设计,这可能引入属性偏差;在模型运行过程中,可能引入探索偏差;观察并解释实验现象,可能引入因果偏差;而在实验评估中,可能引入归纳偏差.

(1) 属性偏差[12,13]:属性偏差通常发生在选择和利用属性的过程中.面向不同任务,相同的属性变量应采取不同的处理方式以适应任务.在保险定价场景中,包含性别属性的机器学习算法可能引起歧视;而在医疗场景中,排除性别属性的机器学习算法却可能削弱辅助诊疗的效果.因此,对属性的排除、包含和加权等操作均可能引起机器学习算法的偏差.

(2) 探索偏差[14]:探索偏差指的是决策者有时会采用次优的行动以获取更多的数据,而这些行动可能导致部分受众承担不成比例的探索代价.

(3) 因果偏差[10,15]:因果偏差通常是由于因果关系不合理构建引起的.保险定价的案例中存在因果偏差,为了刻画车主发生攻击性驾驶行为的概率,保险公司希望找到能够支持这一结论的数据.汽车颜色是易怒心理的外在形式,公司选取的性别属性只是部分地影响汽车颜色.保险公司没有认识到性别、汽车颜色和攻击性驾驶行为间的因果关系,在构建机器学习模型时引起因果偏差.

(4) 归纳偏差[16,17]:归纳偏差发生在机器学习算法的测试评估阶段.机器学习算法的目标函数通常设定为整体最小化均方误差,那么如果从样本数量的角度理解,拟合多数群体比拟合少数群体更重要(对极小化误差更有利),极端情况下,与多数群体的数据分布显著不同的少数群体甚至可能被视为离群数据样本.

总体而言,上述偏差并不是孤立存在的.例如,归纳偏差和交互偏差是相关的.在归纳偏差中,少数群体的误差偏高是因为代表性的样本不足.以有偏的方式采样交互过程中产生的数据,即使增大数据体量,也无助于提高模型的准确率.因此,需要在机器学习算法的全生命周期中重视偏差问题以及不同偏差间的相互影响,并尝试提供针对性的解决方案.

1.3 消除偏差的机制

根据机器学习算法的阶段不同,分别可以使用预处理、处理中和后处理机制,介入算法以实现公平机器学习.表 2 比较了不同的消除偏差的机制.当能够参与数据生成或修改采集到的数据时,采用预处理机制清洗数据;当对算法有完全控制时,采用处理中机制以符合公平性定义的方式调整算法;如果对数据和算法都没有能力改变,采用后处理机制修改算法的输出结果.

Table 2 Comparison of the mechanisms for eliminating unfairness

表 2 消除偏差的机制对比

机制	描述	优势	挑战
预处理	消除原始数据中与受保护属性相关的偏差信息	灵活适应下游任务; 测试时无需访问受保护属性	需要保证结果准确度
处理中	在机器学习模型中增加约束或正则项	算法准确度和算法公平性间灵活权衡;测试时无需访问受保护属性	依赖机器学习算法
后处理	修改机器学习算法的输出结果	灵活适应机器学习算法	测试时需要访问受保护属性; 需要权衡准确度和公平性

1.3.1 预处理机制

预处理机制(pre-processing mechanism)旨在建立一种消除原始数据 x 中与受保护属性 a 相关的偏差信息的数据预处理方法.机器学习算法作用于消除偏差后的数据,以获得公平结果,如图 3 所示.基于预处理机制,我们可以发布合成数据集或原始数据的去偏特征,并不需要修改机器学习算法,而且在测试时不需要访问受保护属性.但是,预处理机制是一种通用机制,提取出的特征广泛适用于下游任务,以损失机器学习算法结果的准确度为代价,换取较高的灵活度.

Fig.3 Conceptual graph of pre-processing mechanism

图 3 预处理机制概念图

1.3.2 处理中机制

处理中机制(in-process mechanism)通过在机器学习模型中增加约束或正则项,以促进偏差消除,如图 4 所示.该机制可以实现算法准确度和算法公平性之间的平衡,并且在测试时不需要访问受保护属性.但是,处理中机制依赖机器学习算法且需要修改算法.

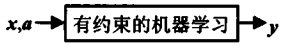

Fig.4 Conceptual graph of in-processing mechanism

图 4 处理中机制概念图

1.3.3 后处理机制

后处理机制(post-processing mechanism)直接修改机器学习算法的输出结果以满足公平性,如图 5 所示.后处理机制不需要修改机器学习算法,且将其视为黑盒模型,因此,该机制能够消除任意算法输出的偏差.但是,后处理机制在测试时需要访问受保护属性,并且较难权衡算法准确度和公平性.

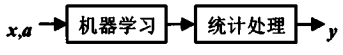

Fig.5 Conceptual graph of post-processing mechanism

图 5 后处理机制概念图

根据机器学习算法做出的决策,可能对有不同受保护属性的群体产生不公平影响.本节讨论了机器学习算法的不公平来源,即数据偏差和模型偏差,认为偏差不是孤立存在的,需要在机器学习的全生命周期中重视偏差问题及其间的相互影响.此外,本章归纳了消除偏差的 3 种机制.

2 机器学习算法的公平性定义

公平机器学习算法的基本任务是,将一般的机器学习算法(见第 1 节)扩展到保证公平性的算法.机器学习模

型f称为δ-公平的,如果满足:

$$\left|\ln\frac{P(\boldsymbol{Y}\mid \boldsymbol{a}=0,\boldsymbol{O},\boldsymbol{\Pi})}{P(\boldsymbol{Y}\mid \boldsymbol{a}=1,\boldsymbol{O},\boldsymbol{\Pi})}\right|\leqslant\delta,$$

其中,$\boldsymbol{O}$表示观察变量,$\boldsymbol{\Pi}$表示因果路径,δ是常数.$\delta\geqslant 0$可以看作在观察到$\boldsymbol{O}$和$\boldsymbol{\Pi}$的条件下,弱势群体和非弱势群体输出结果概率分布间距离的度量;δ越大,不同受保护属性$\boldsymbol{A}$对应 $\boldsymbol{Y}$ 的概率分布差距越悬殊,对不同群体实行差别待遇行为的程度越严重,公平性越低;δ为 0 是一种理想情况,即 $P(\boldsymbol{Y}|\boldsymbol{a}=0,\boldsymbol{O},\boldsymbol{\Pi})=P(\boldsymbol{Y}|\boldsymbol{a}=1,\boldsymbol{O},\boldsymbol{\Pi})$,此时,受保护属性$\boldsymbol{A}$不对输出结果产生影响.换言之,公平机器学习算法满足受保护属性与输出结果之间的独立性假设:对给定条件,受保护属性$\boldsymbol{A}$和输出结果 $\boldsymbol{Y}$ 独立.即

$$(\boldsymbol{Y}\perp\boldsymbol{A})|(\boldsymbol{O},\boldsymbol{\Pi}),$$

其中,$\perp$表示 $\boldsymbol{Y}$ 中的元素和$\boldsymbol{A}$中的元素在 $\boldsymbol{O}$ 和$\boldsymbol{\Pi}$中元素的条件下是相互不影响的,即

$$\frac{P(\boldsymbol{Y},\boldsymbol{A},\boldsymbol{O},\boldsymbol{\Pi})}{P(\boldsymbol{O},\boldsymbol{\Pi})}=\frac{P(\boldsymbol{Y},\boldsymbol{O},\boldsymbol{\Pi})}{P(\boldsymbol{O},\boldsymbol{\Pi})}\cdot\frac{P(\boldsymbol{A},\boldsymbol{O},\boldsymbol{\Pi})}{P(\boldsymbol{O},\boldsymbol{\Pi})}.$$

在现实世界中,不同机器学习任务关注的焦点不同,很难确定一种通用的公平性定义.本节总结现有文献中提出的公平性定义,根据受保护属性在实现公平机器学习算法过程中的作用,现有公平机器学习算法中公平性定义可划分为 3 类:感知公平性、统计公平性和因果公平性,如图 6 所示.当不使用受保护属性时,得到忽略受保护属性公平;当利用受保护属性度量样本间距离时,得到感知受保护属性公平;当使用受保护属性作为条件概率的依据时,得到频率统计的统计公平和贝叶斯统计的因果公平.

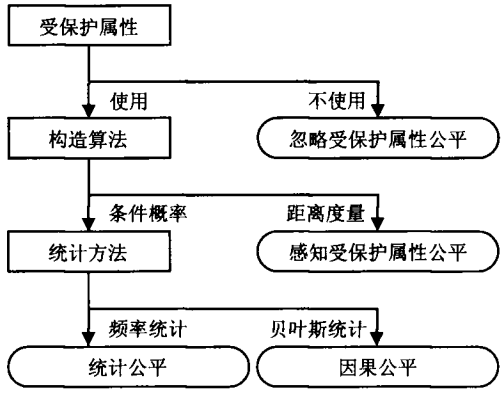

Fig.6 Fairness definitions in machine learning

图 6 机器学习的公平性定义

2.1 感知公平性

感知公平性关注如何直接处理受保护属性以获得公平.这类公平符合直觉,其或在决策过程中直接剔除受保护属性,或将受保护属性视为度量样本间距离的依据.

2.1.1 忽略受保护属性公平

如果不在决策过程中显式地使用受保护属性,那么该算法是忽略受保护属性公平(fairness through unawareness)[8]的,即 $\hat{y}=f(\boldsymbol{x})$.该定义简单而直观,但是数据集中可能存在其他属性与受保护属性高度相关(如街道与种族相关),如果利用这些属性,算法的输出结果中仍然存在偏差.

2.1.2 感知受保护属性公平

如果包含受保护属性在内的属性相似个体得到相似的对待,那么该算法是感知受保护属性公平(fairness through awareness/individual fairness)[18]的,即

$$D(f(\boldsymbol{x},\boldsymbol{a}),f(\boldsymbol{x}',\boldsymbol{a}'))\leqslant d((\boldsymbol{x},\boldsymbol{a}),(\boldsymbol{x}',\boldsymbol{a}')),$$

其中,$D(\cdot)$和$d(\cdot)$分别是在输出空间和输入空间中定义的距离度量.该定义相当于对输出结果直接施加约束,即要

求在确定的距离度量下,相似个体(即 $d((\boldsymbol{x},\boldsymbol{a}),(\boldsymbol{x}',\boldsymbol{a}'))$较小)间输出结果的距离(即 $D(f(\boldsymbol{x},\boldsymbol{a}),f(\boldsymbol{x}',\boldsymbol{a}'))$)小,反之亦然.但是,针对特定任务的个体间距离度量的选择相对比较困难.

2.2 统计公平性

统计公平性要求受保护群体的待遇与非弱势群体或整个群体相似.在累犯预判案例中,ProPublica 对比了非裔美国被告人群体和欧裔美国被告人群体的风险评估结果假阳率和假阴率,即实际不再犯罪却被标记为高风险的概率和再犯却被标记为低风险的概率.对比结果发现,两个群体在上述的统计量上差距大,从而判断该算法对非裔美国被告人群体存在偏见.统计公平无需对数据做出额外假设且容易验证,但该定义无法在个体层面提供公平性保证.根据使用度量的不同,现有统计公平性可分为基本率统计公平、精度统计公平和校准统计公平[19],图 7 表示统计公平定义间的关系.

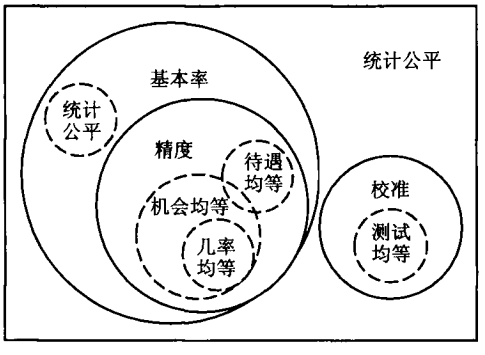

Fig.7 Relationship between statistical fairness definitions

图 7 统计公平定义间的关系

2.2.1 基本率统计公平

基本率(base rate)统计公平不以观察变量为条件,而是直接度量不同群体输出结果的分布差异,并要求差异满足:

$$\left|\ln\frac{P(\boldsymbol{Y}=\hat{\boldsymbol{y}}\mid\boldsymbol{a}=0,\boldsymbol{O}=\varnothing,\boldsymbol{\Pi}=\varnothing)}{P(\boldsymbol{Y}=\hat{\boldsymbol{y}}\mid\boldsymbol{a}=1,\boldsymbol{O}=\varnothing,\boldsymbol{\Pi}=\varnothing)}\right|\leqslant\delta.$$

基本率统计公平的代表性定义有统计均等,其定义如下:

如果输出结果 $\hat{\boldsymbol{y}}$ 在任意情况下均独立于受保护属性 $\boldsymbol{a}$,那么 $\hat{\boldsymbol{y}}$ 满足统计均等(statistical parity/demographic parity)[20−25],即

$$P(\hat{\boldsymbol{y}}\mid\boldsymbol{a}=0)=P(\hat{\boldsymbol{y}}\mid\boldsymbol{a}=1) \tag{1}$$

该定义要求不同群体以相同概率获得相同输出预测结果.进一步,通常采用如下两种近似形式[26]来近似地表达公式(1):

$$\begin{cases}\dfrac{P(\hat{\boldsymbol{y}}\mid\boldsymbol{a}=0)}{P(\hat{\boldsymbol{y}}\mid\boldsymbol{a}=1)}\geqslant 1-\varepsilon_1\\ |P(\hat{\boldsymbol{y}}\mid\boldsymbol{a}=0)-P(\hat{\boldsymbol{y}}\mid\boldsymbol{a}=1)|\leqslant\varepsilon_2\end{cases} \tag{2}$$

其中,ε_1 和 ε_2 均代表近似控制参数.公式(2)的近似形式有法律中"五分之四原则"[25,27]支撑.但是,该定义在 $\hat{\boldsymbol{y}}$ 和 $\boldsymbol{a}$ 相关的情况下会削弱算法效用.例如,在一个群体中选择有资质的成员,在另一个群体中以相同概率随机选择成员,虽然这种惰性做法符合统计均等定义,但是没有现实意义.因此,下面介绍精度和校准统计公平,它们在基本率统计公平的基础上额外考虑数据标记.

2.2.2 精度统计公平

精度统计公平度量每个群体输出结果的错误率和正确率的差异,并要求差异满足:

$$\left|\ln\frac{P(\boldsymbol{Y}=\hat{\boldsymbol{y}}\mid \boldsymbol{a}=0,\boldsymbol{O}=\boldsymbol{y},\boldsymbol{\Pi}=\varnothing)}{P(\boldsymbol{Y}=\hat{\boldsymbol{y}}\mid \boldsymbol{a}=1,\boldsymbol{O}=\boldsymbol{y},\boldsymbol{\Pi}=\varnothing)}\right|\leqslant\delta.$$

精度统计公平的代表性定义有几率均等、机会均等和待遇均等,其定义如下:

如果输出结果 $\hat{y}$ 和受保护属性 a 在给定标记 y 时条件独立,那么 $\hat{y}$ 满足关于 a 和 y 的几率均等(equalized odds/positive rate parity)[15],即

$$P(\hat{y}=1\mid a=0,y)=P(\hat{y}=1\mid a=1,y).$$

如果我们有:

$$P\left(\hat{y}=1\middle|\begin{matrix}a=0\\y=1\end{matrix}\right)=P\left(\hat{y}=1\middle|\begin{matrix}a=1\\y=1\end{matrix}\right),$$

那么输出结果 $\hat{y}$ 满足关于受保护属性 a 和标记 y 的机会均等(equal opportunity/true positive rate parity)[15].

如果我们有:

$$P\left(\hat{y}=0\middle|\begin{matrix}a=0\\y=1\end{matrix}\right)=P\left(\hat{y}=0\middle|\begin{matrix}a=1\\y=1\end{matrix}\right),$$

且同时满足

$$P\left(\hat{y}=1\middle|\begin{matrix}a=0\\y=0\end{matrix}\right)=P\left(\hat{y}=1\middle|\begin{matrix}a=1\\y=0\end{matrix}\right),$$

那么输出结果 $\hat{y}$ 满足关于受保护属性 a 和标记 y 的待遇均等(treatment equality)[28].

几率均等要求受保护和不受保护群体输出结果的真阳率和假阳率是相等的.机会均等更关心真阳率,也就是每个群体中正向类别的成员有正向结果的概率.机会均等同几率均等一样,在输出结果中引入标记的信息.但是标记本身可能是有偏的,该定义无法缩小不同群体间的差异.进一步,待遇均等以相同思路分析均等问题,其同时考虑不同群体输出结果的假阳率和假阴率.

2.2.3 校准统计公平

校准统计公平[29]度量每个群体输出结果的置信度的差异,并要求差异满足:

$$\left|\ln\frac{P(\boldsymbol{Y}=\boldsymbol{y}\mid \boldsymbol{a}=0,\boldsymbol{O}=\hat{\boldsymbol{y}},\boldsymbol{\Pi}=\varnothing)}{P(\boldsymbol{Y}=\boldsymbol{y}\mid \boldsymbol{a}=1,\boldsymbol{O}=\hat{\boldsymbol{y}},\boldsymbol{\Pi}=\varnothing)}\right|\leqslant\delta.$$

校准统计公平的代表性定义有测试均等,其定义如下:

如果标记 y 和受保护属性 a 在给定输出结果 $\hat{y}$ 时条件独立,即

$$P(y=1\mid a=0,\hat{y})=P(y=1\mid a=1,\hat{y}),$$

那么 y 满足关于 a 和 $\hat{y}$ 的测试均等(test fairness/predictive rate parity)[30].

该定义反映群体属于正向类别(y=1)的概率相等.

2.3 因果公平性

因果公平性是基于因果模型[31]的公平性定义,通过干预因果模型,研究受保护属性对输出结果的影响,并消除该影响.因果图模型是表示属性间因果关系的有向无环图,因果图中,节点表示属性,箭头表示因果关系,代表因的属性节点指向代表果的属性节点.因果图中常常使用 do 操作来得到干预后的因果图,例如,干预受保护属性A意味着:删除原因果图中所有指向A的箭头,并且对A赋值,从而得到干预后的因果图.通常使用 do(a=0)表示对A进行干预,且对属性A赋值为 0.

因果公平与统计公平不同,它并不完全由观察到的数据驱动,还需要引入额外的因果关系假设.因果图模型存在局限性,其结构本身来自于领域知识,可能出现假设不一致的情况;另外,基于观测数据,因果图模型通常存在模型不唯一的情况.在某些特定的情况下,受保护属性导致的偏差不一定是不公平的[32].因果公平可以有针对性地消除系统中不公平的影响,同时保留公平的部分.因果公平会使用到这些符号:$\hat{Y}_{a=0}$ 表示干预为 do(a=0)的标签预测值,do(a=1)的标签预测值,π表示指定的因果路径.图 8 表示因果公平定义间的关系.

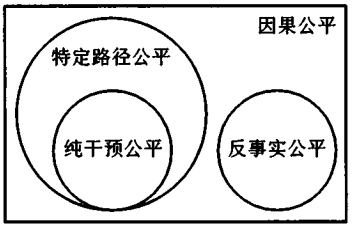

Fig.8 Relationship between causal fairness definitions

图 8 因果公平定义间的关系

2.3.1 纯干预公平

纯干预公平(purely interventional fairness)[33]需要输出结果 $\hat{y}$ 满足:

$$P(\hat{y}_{a=0}) = P(\hat{y}_{a=1}).$$

纯干预定义度量同一群体干预前和干预后的输出结果分布的差异,并要求差异满足:

$$\left|\ln\frac{P(\boldsymbol{Y}=\hat{y}_{a=0}|\boldsymbol{A}=\varnothing,\boldsymbol{O}=\varnothing,\boldsymbol{\Pi}=\varnothing)}{P(\boldsymbol{Y}=\hat{y}_{a=1}|\boldsymbol{A}=\varnothing,\boldsymbol{O}=\varnothing,\boldsymbol{\Pi}=\varnothing)}\right|\leqslant\delta.$$

由于敏感属性通常是因果图的根节点,满足 $P(\hat{y}_{a=0}|a=0)=P(\hat{y}|a=0)$ 和 $P(\hat{y}_{a=1}|a=1)=P(\hat{y}|a=1)$,所以纯干预公平通常比较简单、直观且容易实现.

2.3.2 反事实公平

反事实公平(counterfactual fairness)[8]需要输出结果 $\hat{y}$ 满足:

$$P(\hat{y}_{a=0}|\boldsymbol{x},a) = P(\hat{y}_{a=1}|\boldsymbol{x},a).$$

该定义度量同一个体或者群体的输出结果在现实世界中和反事实世界中分布的差异,并要求差异满足:

$$\left|\ln\frac{P(\boldsymbol{Y}=\hat{y}_{a=0}|\boldsymbol{A}=\boldsymbol{a},\boldsymbol{O}=\boldsymbol{x},\boldsymbol{\Pi}=\varnothing)}{P(\boldsymbol{Y}=\hat{y}_{a=1}|\boldsymbol{A}=\boldsymbol{a},\boldsymbol{O}=\boldsymbol{x},\boldsymbol{\Pi}=\varnothing)}\right|\leqslant\delta.$$

通过选择 $\boldsymbol{x}$ 的子集作为观察变量 $\boldsymbol{O}$,反事实公平保护的对象可从个体扩展到群体.但是,因果图的不唯一性导致反事实公平有不可测试性.

2.3.3 特定路径公平

特定路径公平(path-specific fairness)[33]需要输出结果 $\hat{y}$ 满足:

$$P(\hat{y}_{a=0}|\pi) = P(\hat{y}_{a=1}|\pi).$$

该定义度量同一群体在采用不同的特定路径干预后的预测结果的分布的差异,并要求差异满足:

$$\left|\ln\frac{P(\boldsymbol{Y}=\hat{y}_{a=0}|\boldsymbol{A}=\varnothing,\boldsymbol{O}=\varnothing,\boldsymbol{\Pi}=\pi)}{P(\boldsymbol{Y}=\hat{y}_{a=1}|\boldsymbol{A}=\varnothing,\boldsymbol{O}=\varnothing,\boldsymbol{\Pi}=\pi)}\right|\leqslant\delta.$$

特定路径公平可以保留受保护属性带来的合理差异.

表 3 总结比较了各种公平性定义的不同.从受到公平保护对象的角度,可分为个体公平和群体公平.

- 个体公平要求相似个体有相似输出结果,对每个个体做有意义的保证,但是需要对个体间距离度量做出与任务相关的假设;
- 而群体公平要求少数受保护群体与任务相关的统计分布和所有群体的分布近似相等.

公平机器学习算法的基本任务是,将一般的机器学习算法扩展到保证公平性的算法.本节给出了通用公平性定义的方式,总结并梳理了现有文献中的 3 类公平性,即感知公平性、统计公平性和因果公平性.

Table 3 Comparison of the fairness definitions in machine learning algorithms

表 3 机器学习算法中典型的公平性定义对比

定义		个体/群体	公式	描述	优势	挑战
感知公平性	忽略受保护属性公平	群体	$\hat{y}=f(x)$	不显式地使用受保护属性	普适性	难以剔除属性中受保护信息
	感知受保护属性公平	个体	$D(f(x,a),f(x',a'))\leqslant d((x,a),(x',a'))$	包括受保护属性在内的属性相似个体得到相似的对待	度量意义下公平	度量距离不易确定
统计公平性	统计均等	群体	$P(\hat{y}\mid a=0)=P(\hat{y}\mid a=1)$	每个群体的输出结果相似	有法律原则支撑	易造成算法效用降低;导致惰性做法
	几率均等	群体	$P(\hat{y}=1\mid a=0,y)=P(\hat{y}=1\mid a=1,y)$	每个群体输出结果的错误率和正确率相等	惩罚惰性做法	易受带偏差的标记影响
	测试均等	群体	$P(y=1\mid a=0,\hat{y})=P(y=1\mid a=1,\hat{y})$	每个群体输出结果的置信度相等		
因果公平性	纯干预公平	群体	$P(\hat{y}_{a=0})=P(\hat{y}_{a=1})$	干预前后输出结果的分布相同	可解释	依赖假设的因果图
	反事实公平	个体/群体	$P(\hat{y}_{a=0}\mid x,a)=P(\hat{y}_{a=1}\mid x,a)$	现实世界和反事实世界的输出结果在给定可观察属性时相同	可解释;适用不同粒度的保护对象	
	特定路径公平	群体	$P(\hat{y}_{a=0}\mid \pi)=P(\hat{y}_{a=1}\mid \pi)$	干预沿指定路径传播,干预前后输出结果的分布相同	可解释;保留受保护属性本身的差异	

3 公平机器学习

公平机器学习算法旨在发布近似准确的算法模型,且输出结果符合某种公平性定义.对机器学习算法模型进行公平性提升时,公平性定制(fairness-tailored)的正则项或者强约束条件常被用于改进机器学习算法,而上述介绍的公平性定义为公平性提升方法的设计提供理论支撑.在实际的机器学习任务中,3 种实现满足给定公平性定义的机制已经在第 1.3 节介绍.根据使用机制以及对机器学习过程的理解,本文主要从公平表征任务、公平建模任务和公平决策任务这 3 个角度展开对现有工作的梳理:公平表征任务建立公平数据集或提取公平数据特征;公平建模任务建立公平机器学习模型;公平决策任务使用机器学习模型的输出结果进行公平决策.

综上所述,解决 3 项任务可以实现机器学习模型到公平机器学习模型的衍化.

3.1 公平表征任务

数据所有者从公平表征任务切入,努力寻找一种作用到 X 和 A 的特征变换方法,实现保留与 Y 有关的信息且近似地与 A 无关的特征 Z,即 $(X,A)\xrightarrow{g}Z\xrightarrow{f'}Y$,其中,$f'$ 表示与 f 有相同学习目标的模型.与 f 不同的是,f' 的输入是 X 和 A 由特征变换 g 得到的 Z.直观地理解:如果不同群体中成员的表征形式是相似的,那么任何基于这些表征建立的算法所做决策应独立于受保护属性 A[23].得益于深度神经网络强大的表达能力,现有工作通常将公平表征任务建模为双人博弈模型[34],即数据所有者和攻击者做对抗学习.攻击者的核心目标是推断出群体成员的受保护属性;而数据所有者的目标是去除与群体受保护属性有关的信息,且同时保留与效用有关的信息来进行准确预测.表 4 总结了公平表征任务的具体信息.公平表征任务重点集中在数据预处理阶段,下面分别介绍基于信息论、生成对抗网络和解耦学习的公平表征任务.

Table 4 Fair representation task

表 4 公平表征任务

目标	数据表征学习
输入	与任务相关的训练样本
输出	公平的合成数据集,公平的特征表示
公平性	统计公平,因果公平
偏差消除机制	预处理机制

3.1.1 基于信息论的公平表征任务

本节从信息论的角度量化分析公平表征任务,从而刻画公平性和效用间的权衡.文献[35]量化分析了统计均等和不同群体效用之间的平衡,其建立准确率均等(即群体间正确率相等)和几率均等间的关系,证明了一系列由群体间基本率和均衡错误率(balanced error rate,简称 BER)引导的群体输出结果误差界.当群体间基本率不同时,证明了任何满足统计均等的分类器至少在一个群体上产生较大的误差.当群体表征间的总变分距离(total variation distance)[36]接近时,这种表征满足准确率均等.

CFair 算法[37]利用文献[35]的理论,实现了同时满足近似的准确率均等和几率均等的目的.其目标损失和对抗损失使用均衡错误率,即,对不同群体的分类错误给予同等重视(相同权重),从而提取同时满足近似的准确率均等和几率均等的公平表征.LAFTR[24]关联多种统计公平定义与对抗性学习,学习适用于下游分类任务的表征.CFair[37]在实现群体间几率均等方面比 LAFTR[24]更有效,这是因为 LAFTR 使用平均绝对误差定义目标函数,而 CFair 使用渐进有效和最优的交叉熵损失定义目标函数;另外,LAFTR 使用一组对抗惩罚表征中包含的受保护属性信息,而 CFair 使用两组对抗分别惩罚表征中包含弱势群体的受保护属性信息和非弱势群体的受保护属性信息,更细致地控制表征的公平程度.

3.1.2 基于生成对抗网络的公平表征任务

本节主要介绍利用深度生成模型并结合因果公平,提高表征可解释性.FairGAN[38]设计了两个对抗网络,使得生成网络生成的表征与原数据集相似的同时,不同群体的属性分布也相似.FairGAN$^+$[39]在 FairGAN 的基础上,将预测任务添加到学习框架中,再增加一个对抗网络来保证生成数据的分类结果与受保护属性独立.LAFTR、FairGAN、FairGAN$^+$这 3 种算法都可以在预处理阶段实现统计公平,其中,LAFTR 算法是从真实数据学习公平的特征表示,FairGAN 和 FairGAN$^+$算法学习的是随机噪声到公平数据集的映射,可以生成不同规模的合成数据集.

CFGAN[40]基于 Causal GAN[41]学习属性之间的因果关系,使用生成对抗网络促使预测值与反事实世界预测值相同,从而实现纯干预公平、反事实公平和特定路径公平.CFGAN 学习了属性之间的因果关系,比基于统计公平的表征学习算法 LAFTR、FairGAN、FairGAN$^+$有更好的可解释性.

3.1.3 基于解耦学习的公平表征任务

本节针对下游任务中受保护属性不确定性,应用解耦思想学习公平表征任务.FFVAE 算法[42]突破在提取表征前需要完全确定受保护属性的限制,同时考虑多个受保护属性组合的公平表征任务.其学习目标的奖励项包括建模非受保护的观测变量、对齐隐变量和受保护属性、将各受保护属性隔离到独立的子空间以及匹配隐变量和先验分布.从而在测试时,使用单一表征适应各种不同标记和定义的分类任务,实现提取满足统计公平的多个受保护属性及其组合的公平表征.

除前文调研的公平表征任务外,现有工作[43]还基于最优传输理论,利用 Wasserstein 质心技术,以随机方式改变原始观察到的变量分布.

3.2 公平建模任务

公平建模任务既关心预测准确率,又关注如何减少预测结果的歧视.公平建模任务以符合某种公平性为目标,调整原有算法获得 f',使得 $(\boldsymbol{X},\mathcal{A})\xrightarrow{f'}\boldsymbol{Y}$.公平建模任务适用于模型提供者对算法模型有完全控制的情况,公平建模任务重点介入处理中阶段.表 5 总结了公平建模任务的主要信息,且根据机器学习任务的不同,我们将现有公平建模任务从解决任务的角度进行分类,即分类任务、回归任务、组合优化任务、集成学习任务、图计算任务和聚类任务等.

Table 5 Fair modeling task

表 5 公平建模任务

目标	输入	输出	公平性	偏差消除机制
算法模型改进	与任务相关的训练样本	公平的机器学习算法模型	统计公平,因果公平	处理中机制

3.2.1 分类任务的公平建模任务

AVD/SD Penalizers 算法[44]为计算群体间假阳率和假阴率近似的最优分类器,其分别惩罚不同群体到决策边界的距离之差的绝对值或方差,从而得到满足准确率均等的分类结果.

Counterfactual Fairness[8]使用因果关系来解决公平问题,提出了反事实公平的概念,通过估计隐变量的后验分布来减少受保护属性对预测的影响,从而在处理中实现反事实公平的预测.PSCF 算法[32]为应对敏感属性可能沿公平和不公平路径影响决策的场景,通过修正敏感属性的后代属性在不公平因果路径上的观测值,从而实现特定路径反事实公平的预测,比 Counterfactual Fairness 算法适用范围更广.

针对基于因果模型的公平算法存在难以度量和因果模型不唯一的情况,Muti-World Fairness 算法[45]结合多个可能的因果模型做出近似公平的预测,从而使预测结果满足反事实公平,同时可缓解确定因果模型困难的问题.PC-Fairness 算法[46]认为特定路径反事实公平涵盖了基于因果关系的公平概念,其采用响应变量函数,以衡量不可识别情况下的特定路径的反事实影响,进一步给出特定路径反事实影响紧确界.实验证明,PC-Fairness 算法在反事实影响的度量上比 CF[47]算法更准确.

3.2.2 回归任务的公平建模任务

Fair Regression 算法[48]为应对公平回归任务中无穷个约束的挑战,首先离散化实值预测空间,并进一步简化为受约束的代价敏感的分类问题,将约束每个群体的损失最小化问题转换为加权损失最小化的问题,从而使回归结果满足有界的群体损失(bounded group loss,简称 BGL).

3.2.3 组合优化任务的公平建模任务

GF1A/B 算法[49]均考虑将一批不可分割的商品公平地分配给一组有不同偏好参与者的问题.如果没有一组参与者宁愿重新分配并接受任意其他组根据组规模调整资源后的商品来替代原先分配给本组的商品,那么该状态是满足帕累托最优的.GF1A[49]在重新分配商品后,移除弱势组中每位参与者的一件商品;GF1B[49]在重新分配商品前,移除更受优待的组中每位参与者的一件商品,从而使用本地纳什最优分配解决该不可分割商品的公平分配问题.进一步,为了解释现实世界中决策的因果路径以及系统中他人的决策如何影响某个个体,Counterfactual Privilege 算法[50]的目标是最大化总体效益,同时防止某个个体因受保护属性而获得超过阈值的有益影响,从而提供满足反事实公平的分配方案.

3.2.4 集成学习任务的公平建模任务

TREE01 算法[51]考虑为每个组生成最准确的模型,而不损害任何组分类器的准确度,即使用组内的数据训练满足偏好的解耦分类器,使每个组的大多数个人更喜欢为其分配的分类器,而不是忽略组成员的集中模型(即合理性)或者分配给任何其他组的模型(即无嫉妒).TREE01 自适应地选择解耦的组属性,递归地在训练数据上生成满足合理性和无嫉妒性的候选分类器树;剪枝在泛化误差方面违反公平性的候选分类器树,从而实现满足合理性和无嫉妒公平的分类器.

3.2.5 图计算任务的公平建模任务

Maximin Fairness 算法[52]和 Diversity Constraints 算法[52]回答社交网络中不同的社群是否公平地分配了干预措施的效益,解决影响力最大化的公平问题.Maximin Fairness 的策略是最大化受影响比例最小的组;Diversity Constraints 使每个社群受到的影响至少等于为其按比例分配干预措施受到的影响,从而选择满足公平影响的目标集.

3.2.6 聚类任务的公平建模任务

Proportionally Fair Clustering 定义[53]:如果不存在联合簇,那么簇标记是均衡的.Greedy Capture 算法[53]使簇半径围绕簇标记增长,贪心地包含满足公平聚类定义的元素.但是,Greedy Capture 始终产生近似的公平聚类结果.Local Capture Heuristic[53]为了搜索更精确的解,迭代地将违反公平聚类定义的簇标记替换为当前包含样本最少的簇标记,从而生成满足公平聚类定义的聚类结果.

3.3 公平决策任务

算法使用者需要承担公平决策任务,以确保机器学习算法输出结果对每个群体是公平的,即

$$(\boldsymbol{X},\mathcal{A},\hat{\boldsymbol{Y}})\xrightarrow{h}\tilde{\boldsymbol{Y}},$$

其中,h 表示输出结果变换,$\tilde{\boldsymbol{Y}}$ 表示符合公平定义的输出结果.表 6 总结了公平决策任务的主要性质.

Table 6 Fair decision-making task

表 6 公平决策任务

目标	决策结果调整
输入	训练样本和决策结果
输出	公平的决策结果
公平性	统计公平,因果公平
偏差消除机制	后处理机制

公平决策任务重点集中在后处理阶段.同样,根据机器学习任务的不同,将现有公平决策任务从解决任务的角度进行分类,即分类任务和隐私保护任务.

3.3.1 分类任务的公平决策任务

为了突破现有算法只在选定的阈值上保证公平的限制,Wass-1 Penalty/Post-Process 算法[54]通过最小化 Wasserstein-1 的距离来提升分类结果的输出和受保护属性间的独立性,从而实现在后处理阶段使分类结果满足强统计均等(strong demographic parity,简称 SDP).即,对于任意阈值 τ:

$$\mathbb{E}_{\tau\sim\boldsymbol{Y}}(|P(\hat{\boldsymbol{y}}>\tau\,|\,\boldsymbol{a}=0)-P(\hat{\boldsymbol{y}}>\tau\,|\,\boldsymbol{a}=1)|)=0,$$

其中,τ 是输出空间 $\boldsymbol{Y}$ 中的随机变量,$\mathbb{E}(\cdot)$ 是其数学期望.

CF 算法[47]基于因果模型,采用 C 成分分解,确定造成反事实影响度量中不可量化的项,度量了反事实影响的上下界.在后处理阶段,给定数据集和任意分类器的预测结果 $\hat{\boldsymbol{Y}}$,在反事实影响在一定范围内的约束下,最小化最终预测与真实标签的误差,从而使预测结果满足反事实公平.

3.3.2 隐私保护的公平决策任务

现实应用可能限制收集和使用敏感属性,需要在精准度、公平性和隐私性间做权衡.在技术上,如果没有隐私性的需求,那么增加算法的迭代轮次能够将误差减少到任何期望的程度;当存在隐私性的需求时,增加算法的迭代轮次需要放大梯度扰动的范围,从而增大误差.

DP-postprocessing 算法[55]在不使用敏感属性的情况下训练分类算法,使用拉普拉斯机制扰动其概率分布,在测试时,显式地使用受保护属性,从而在后处理阶段生成满足几率均等的输出结果.

但是,即使不考虑为保护隐私引入的噪音,DP-postprocessing 这种后处理算法通常无法达到最优的分类准确度,且在现实世界的测试中,可能无法收集并使用敏感属性.进一步,DP-oracle-learner[55]通过学习器和审计器间的零和博弈寻找最优的公平分类器.学习器通过代价敏感的分类预言机,使用指数机制计算无约束的学习问题;审计器使用拉普拉斯机制扰动的梯度下降惩罚违反公平性的程度,从而实现在处理中阶段训练满足几率均等的分类器,且该分类器在测试时,不需要访问受保护群体的成员属性.

本节介绍了公平表征任务、公平建模任务和公平决策任务的相关工作.公平表征任务的关键是如何在降低特征和受保护属性之间关联性的同时,保留下游机器学习任务所需要的信息.我们进一步需要探索合成更多数据类型的公平数据集以满足不同应用的需求,并尝试对同一数据集提取不同公平表征以适应不同任务.公平建模任务的关键问题是如何对目标函数做调整,以实现公平机器学习的目标.现有工作主要关注分类任务,我们需要进一步探索更多复杂任务以适应不同应用的需求.公平决策任务无需调整机器学习算法,现有工作较少;对受保护属性和敏感属性有交集的应用,我们进一步需要权衡精准度、公平性和隐私性.

4 公平机器学习的应用

在学术界与产业界,机器学习公平性的研究处于起步阶段.本节重点介绍目前已出现的公平机器学习典型应用案例以及常用数据集和检测工具.

4.1 公平机器学习的典型应用案例

- 计算机视觉的公平机器学习

在视觉识别任务中,模型可能放大标记和受保护属性间的联系.adv@image/conv4/conv5 算法[56]尝试利用对抗学习消除这种联系,获得消除受保护属性相关信息的掩码,保留与目标标记相关的信息.文献[57]认为,学习生成合成数据集比学习隐式嵌入有更好的可解释性.为解决缺少目标数据的问题,最大化残差(公平表征和原数据的差异)和受保护属性相关性,同时最小化表征与受保护属性相关性,从而得到保留语义信息的公平表征.文献[58]观察目标识别系统对家庭用品的识别准确率,发现对于部分目标识别系统,在低收入国家中,日常用品的识别错误率比富裕国家中的高 10%.其研究结果表明,需要进一步探索使目标识别系统适用于不同国家和不同收入水平的群体.文献[59]针对视觉识别分类器存在的标记和受保护属性有伪关系的问题,设计了数据偏见的衡量基准.

- 自然语言处理的公平机器学习

自然语言的文本中存在对性别的刻板印象,而用于刻画词间语义和句法关系的词嵌入易受到语料库的影响,其偏差通常表现为某些属性(如职业)与某个群体的关联较其他群体更紧密.文献[60]观察到新闻和百科全书等力求客观的文本中存在主观偏见,其提出了 MODULAR 和 CONCURRENT 算法,使用降噪自动编码器生成有类似含义的中立文本.文献[61]认为词嵌入在其几何结构中包含刻板印象,其在嵌入中使用性别词学习性别子空间,通过平衡等方法消除中性词中的偏见.文献[62]发现多标签分类语料库和语义角色标注的数据和模型中存在偏见.RBA 算法[62]基于拉格朗日松弛的近似推理算法,在语料库级别从结构化预测模型中校准不公平.文献[63]发现共指解析任务中存在性别偏见,其提出了 WinoBias 基准,并提出了数据扩充技术,结合 word2vec,使用基于规则的方法生成辅助数据集,交换男女实体,消除现有共指消解技术中的偏差.文献[64]提供大规模翻译中性别偏见的多语言定量证据,表明在 8 种目标语言中,所有 4 种被试的商业系统和两种学术翻译系统倾向于根据刻板印象而不是语境进行翻译.现有基于词嵌入的偏见评分依赖少量人工标记的词语.为了量化评估大量词语中性别偏见程度,Bias-in-wat 算法[65]基于提示和响应的记录构建词语联想图,通过随机游走观察性别偏见在词语间的传播路径,从而实现由词语联想图的内在结构获取词语的偏见评分.文献[66]发现词嵌入偏见传播到机器学习模型中的现象,分别从嵌入级别和算法级别缓解性别歧视、仇外心理等偏见.在预测阶段的不公平性研究中,MT-NLP 工具[67]识别任意语句中受保护的属性组,对自然语言处理模型执行蜕变测试,检测违反公平的输入.为了比较各公平性指标的相关性,WEFE 框架[68]封装现有公平性指标,输入由描述特征的属性词和描述群体的目标词组成的查询列表,构建公平性度量的抽象视图.

- 信息检索和推荐系统的公平机器学习

在推荐任务中,隐式反馈数据无需用户花费额外的时间提供评分、标注标签或评论,具有容易获取的优点.文献[69−72]主要从商品的角度考虑基于隐式反馈数据的推荐系统的选择性偏差,即用户只会对部分候选商品提供反馈,且他们的选择往往是非随机的.有序的推荐列表中最常见的选择性偏差是位置偏差,即排序在前的商品比在后的商品更可能收到用户的反馈.对于两个相关度相同的商品,排序在前的会收到更多的反馈,因此在学习到的算法中继续排在前面,这种现象是不公平的.FULTR 算法[69]将用户的关注(即曝光)视为资源向被排序的商品分配,其排序策略通过加权无偏效用估计量并增加基于价值的公平性约束,从而实现商品位置与曝光、相关度成正比的排序算法.文献[70]进一步考虑存在某类型商品曝光不足的情况,提出了 Pairwise Fairness 算法,如果不同类型的商品间或同类商品中,产生交互的商品排名在另一个相关但未产生交互的商品前的可能性不同,模型受到惩罚,从而满足逐对排序公平的推荐.Fair-PG-Rank 算法[71]将公平排序学习建模为受公平约束的策略学习问题,通过策略梯度法搜索满足商品曝光限制的公平排名.FairRec 算法[72]将公平推荐问题映射到有约束的公平分配不可分割的商品的问题,保证商品的曝光份额和用户推荐结果的无嫉妒.文献[73]关注底层平台算法频繁更新引起商品曝光度突变的不公平问题,提出了基于整数线性规划的在线增量更新机制,保证商品的平均曝光度平稳过渡.文献[74,75]从用户的角度考虑推荐系统中可能存在的不公平现象,即不活跃用户可能由于其缺乏足够的交互历史而更容易收到不令人满意的推荐结果,他们的推荐结果可能会因协同过滤的性质偏向活

跃用户而产生偏差.为了在推荐 Top K 时公平地汇总所有用户的偏好,文献[74]从社会选择理论得到启发,将 Top K 非个性化的众包推荐重构为定期重复的多重获胜者选举的结果,从而将 Top K 推荐中观察到的偏差归因于选举制度中沉默的大多数.其提出了基于可转移单票制的投票机制,自动推断用户的偏好排名,保证成比例表征(大多数用户的候选商品集至少包含他们最喜欢的商品之一)和反多元(拒绝大多数用户不喜欢的高度有偏的商品)公平.文献[75]针对推荐效用和解释多样性设计全体和个体公平标准,将其形式化为整数规划问题,提出了启发式重排序的公平约束方法.文献[76]为所有被排名用户创造平等机会,其量化和缓解个人排名机制中的算法偏差,根据一个或多个受保护属性的期望分布,对用户公平地重新排名.

在搜索任务中,FairCo[77]算法构建由位置偏差、内容的平均相关性和全局期望相关性决定的无偏统计量,控制曝光公平性,即每个内容单位偏好下的期望曝光尽可能相似.由于该统计量需要收集多轮迭代的搜索反馈,所以 FairCo 设计了控制器,防止系统启动时特定组的内容被曝光过多.

- 多智能体系统的公平机器学习

智能体通常优先最大化自身的效用,而多智能体系统中,资源分配与调度不仅需要考虑效用,还需要使资源分配得当、负载均衡.考虑公平性的多智能体系统能够提升系统的效用和稳定性,但这是个复杂的多目标联合策略优化问题.为解决社会困境,文献[78]假设智能体有不公平厌恶,在提高自身奖励的同时,保证自己的奖励和其他智能体间的偏差尽可能小.其通过回报函数引入不公平厌恶理论,促进多智能体间的协同.文献[79]为解决多智能体序列决策的公平问题,使用线性规划和博弈的方法计算策略,在保证整体性能的情况下,最大限度地提高智能体的最差性能.FEN[80]采用一个控制器和多个子策略的分层强化学习架构:控制器通过选择多种与环境交互行为的子策略来最大化奖励,指定其中一个子策略最大化环境奖励,其他子策略探索不同的公平行为来获取奖励;平均共识协议协调分布式的多智能体训练.

- 联邦学习的公平机器学习

联邦学习[81]是一种在各数据提供方与平台方互不信任的场景下,以保护任何一个参与方原始数据隐私的方式,用全局机器学习模型协同建模,并拟合异构网络中多个参与的远程设备或组织机构中本地数据的计算框架.现有方法朴素地最小化聚合损失,可能导致全局模型不成比例地利好部分设备的性能.q-FFL[82]设计了保证各个设备间联邦学习模型性能公平分布的优化方法,通过引入施加公平性的权重参数,实现对不同设备损失的加权计算,使损失较大的设备有较高的权重,从而减小性能分布的方差,从而实现在各设备上得到性能相近的联邦学习模型.AFL[83]针对由客户端的分布混合组成的任意目标分布,不以其他模型为代价过拟合特定的模型,最小化任意受保护类别的最大风险,产生良好意图的公平(good-intent fairness)概念.

4.2 常用数据集和仿真环境

数据集是公平机器学习的关键,本节介绍多个常用数据集和一个仿真环境.

- Adult 数据集:Adult 数据集[84]是从 1994 年美国的人口普查数据库中提取的数据,包含 48 842 条记录,每条记录有年龄、职业、受教育程度、种族、性别、婚姻状况、出生地、每周工作时长等属性,这些属性有连续的和离散的,其任务是预测个体每年的收入是否超过 50k.该数据集可以研究性别或种族对收入的影响.
- German Credit 数据集:German Credit 数据集包含 1 000 条记录,每条记录有个人身份、性别、信用评分、信用金额、住房状况等属性.该数据集可以研究性别对信用的影响.
- COMPAS 数据集:COMPAS 数据集包含来自美国佛罗里达州布劳沃德县的被告记录,显示了 2013 年至 2014 年间,监狱、服刑时间、人口统计数据、犯罪历史和 COMPAS 风险评分.该数据集可以研究种族和性别对累犯预判的影响.
- Communities&Crime 数据集:Communities and Crime 数据集来自美国社区和犯罪记录数据集,这些数据结合了 1990 年美国人口普查的社会经济数据、1990 年美国 LEMAS 调查的执法数据和 1995 年联邦调查局 UCR 的犯罪记录数据.
- LSAC National Bar Passage 数据集:Law School Admissions Council(LSAC)[85]跟踪调查了 1991 年到

1997 年约 2.7 万名法学学生,包括法学院学生、毕业学生和参加律师考试的学生,其记录了大量有抱负的律师的人口统计学、经验和成果.虽然这些数据有局限性,但它对研究与法律教育相关问题是有价值的.

- The UCI Bank Marketing 数据集:The UCI Bank Marketing 数据集包含 41 188 个个体的 20 个属性.这些数据与一家葡萄牙银行机构的直接营销活动(电话)有关,其目标是预测客户是否会认购定期存款.
- Tetrad 生成数据集:吴勇楷等人[46,47]建立的因果模型,给定模型中所有外生变量和函数,利用模型生成相应的人工数据库,可以用来验证公平算法的可靠性.
- Diversity in Faces (DiF)数据集:DiF 数据集[86]是 IBM 研究院为推进人脸识别技术的公平性和准确性研究而发布的具有多样性的平衡数据集,其包含 100 万张有标注的人脸图像,标注的编码方案有颅面特征(如头部长度、鼻子长度、前额高度等)、面部比例(对称性)、视觉属性(年龄、性别)、姿势以及分辨率等.DiF 数据集旨在加深对公平的人脸表征的理解,并改进人脸识别技术.
- iNaturalist 数据集:iNaturalist 数据集[87]包含 80 万张层次标注的动植物图像,其目标是物种分类,可能存在不同类别物种的分类正确率相差很大[88]的不公平现象.
- ML-fairness-gym 仿真环境:ML-fairness-gym[89]是用于构建模拟的一组组件,其探索机器学习决策的潜在长期影响.文献[90]发现:在简化的动态模拟环境中,长期影响实际上可能抵消预期目标.在 ML-fairness-gym 的模拟场景中,代理与环境互动,有助于帮助研究突破静态环境公平的局限,实现长期交互环境下的动态公平.

4.3 公平检测工具

本节重点介绍常用的公平检测或消除不公平的工具,见表 7.

Table 7 Fairness detection tools

表 7 公平检测工具

工具名称	差异检测	差异消除
InterpretML	✓	×
AI Fairness 360 Open Source Toolkit	✓	✓
IBM Watson OpenScale	✓	×
What if tool	✓	×
Microsoft Research Fairlearn	✓	✓

- InterpretML:InterpretML[91]是由微软开发、用于训练可解释机器学习模型和黑盒模型的开源工具包,其基于 Explainable Boosting Machine 算法使预测结果更精准,且有可解释性.
- AI Fairness 360 Open Source Toolkit:AI Fairness 360 Open Source Toolkit[92]是由 IBM 研究院开发的可扩展开源工具包,帮助开发者在机器学习应用程序生命周期中,检查、报告和减轻机器学习模型的歧视和偏见.其包含多种偏差检测机制和偏见缓解算法,旨在实现将算法研究从实验室推广到金融、人力资源管理、医疗保健和教育等领域实践中.
- IBM Watson OpenScale:IBM Watson OpenScale 跟踪和度量机器学习模型的输出结果,允许企业不考虑模型的构建和运行方式,确保机器学习模型公平、可解释且合规.另外,Watson OpenScale 能够向企业展示机器学习是如何构建、使用和执行的,企业可以自由地选择运行环境,将机器学习算法嵌入到新的或现有的商业应用和功能中.
- What if tool:What if tool 是由谷歌 AI 开发的,其提供易于使用的接口,提升对黑盒分类或回归机器学习模型的理解.使用该接口可对示例执行推理,并可视化结果.此外,可以手动或以编程方式编辑示例,并重新运行模型,以查看更改后的结果.它有研究模型性能和数据集子集公平性的功能,提供简单直观的方法,通过可视化界面在数据上使用机器学习模型,而不需要编程操作.
- Microsoft Research Fairlearn:Fairlearn 工具包允许人工智能系统开发人员评估其系统的公平性,并缓解观察到的不公平问题.Fairlearn 可以比较多个模型,例如由不同学习算法和不同缓解方法生成的模型.

现有的公平机器学习算法可以消除应用场景中存在的部分不公平现象,但尚缺乏公平性和效率、研发成本间的权衡,以使其能被更广泛地使用.现有的公平基准数据集覆盖个人收入、犯罪统计、入学录取等重要领域,可以帮助验证机器学习算法的有效性.产业界推出公平检测工具在一定程度上可以帮助促进人们对算法公平性的认识,还能降低厂商引入公平的研发成本,具有实际意义.

5 未来研究方向

现有的机器学习算法的公平试图解决下面 3 个问题:(1) 为现有的公平性定义提供更强的条件;(2) 试图涵盖现有问题中没有考虑到的歧视;(3) 合理松弛以提高预测模型准确度.尽管公平机器学习研究已经取得了瞩目的研究成果,但目前该研究还处于初级阶段,依然存在尚待解决的问题和挑战,包括:

(1) 提供更强的公平性定义

造成机器学习算法不公平的来源种类多且复杂,不同偏差对不公平造成的影响程度不同,且为了得到公平的机器学习算法,不同偏差的应对方式也不同,开发人员的先验知识有差异,所以基于人工的公平性定义成本高、效率低且不能主动发现新的、隐含的不公平风险.

公平机器学习的第 1 个挑战是如何提供综合性的公平定义:在对比不同公平机器学习算法的能力时,使用不同公平性度量往往会导致不同的评判结果.这意味着算法公平与否是相对的,这不仅取决于模型和数据,还取决于任务需求.现阶段缺乏完善、多维度的算法公平性评价指标和评估体系,无法对机器学习算法面临的公平性风险进行有效的量化评估,导致无法保证部署到生产环境中的机器学习模型的公平性.因此,我们进一步需要探索综合性公平定义,突破基于零先验知识的算法偏差自动挖掘和分析技术,研制算法公平性测试标准与工具,检测待测机器学习算法及系统的不公平性的存在及类型,着手建立完善的机器学习算法的多维度评估体系.此外,公平性定义的制定需要结合各国的法律法规和社会公平的概念,避免产生狭隘的技术解决方案.

公平机器学习的第 2 个挑战是如何适应公平性模型的动态性:现有工作主要在静态的、没有反馈的、短期影响的场景中研究机器学习中的公平性问题,而没有研究这些决定如何随着时间流逝影响未来应用中的公平性.现阶段机器学习的公平性研究呈现出动态演化的趋势,要求公平性定义和算法考虑决策系统动态的、有反馈的、长期的后果.机器学习算法的公平性可以看作一种博弈,当前,公平性研究正处于劣势,其具体表现为:现有的机器学习公平性研究提出的改进算法多针对被动的静态公平,无法有效地适应演化周期.因此,未来公平性研究应着手建立有反馈的长期公平性,结合对抗环境下公平性提升和破坏的动态博弈理论,研究辅助动态公平性的检测机制.

(2) 涵盖更真实的应用场景

机器学习模型的应用场景多元,实际应用中可能存在数据收集困难等难点,对公平机器学习带来挑战.

公平机器学习的第 3 个挑战是如何解决受保护属性与访问测试的矛盾:机器学习算法的意义不仅在于拟合训练集的分布,而在于拟合上线后真实世界的分布.然而在真实应用中,敏感属性往往不可访问,存在难以测试的问题.因此,我们进一步需要探索非侵入式测试或密文测试.

公平机器学习的第 4 个挑战是如何构建推理攻击的防御机制:公平性约束要求预测变量在某种程度上与群体成员的属性无关.隐私保护的成员攻击提出了相同的问题:是否可能保证即使是最强大的对手也无法通过推理攻击窃取个人的身份.因此,利用差分隐私技术可能会开发出更有效的公平学习算法.将现有的公平性机制与差分隐私结合,是未来研究中比较有前景的方向.

此外,随着公平性研究的理论和技术的深入,促进其进一步发展的目的在于促进建立更完善的生态系统.真实场景中的应用间存在公平性的相互影响:如在银行贷款中,对不同性别的群体在额度属性上存在不公平,而这种不公平可能是由于不同性别的群体在职场的薪资水平导致的;类似地,这些不同的性别群体在职场上遭受的不公平对待可能与其在入学时受到的歧视有关.所以,独立地解决银行贷款或职场薪酬中的偏差,无法从根本上缓解不公平现象.因此,我们需要进一步探索能够引导社会决策向公平发展的辅助算法,可以考虑引入联邦学习计算框架,实现跨领域、跨机构的协同公平解决算法.

(3) 提供更高的准确性

构建公平且可靠的算法是可信机器学习算法的基础.

公平机器学习的第 5 个挑战是如何权衡算法性能与公平:当受保护属性与预测结果相关时,如累犯预测,很难建立不包含与种族相关的分数,如果排除贫穷、失业和社会边缘化,准确率会下降.因此,我们需要进一步探索权衡准确度和公平性的方式.

6 结 论

公平性是一种具有相对性的社会概念,绝对意义上的公平是不存在的.公平机器学习算法通过探索消除不公平的机制,逐步完善机器学习算法的公平性.公平表征、公平建模和公平决策是可信机器学习公平性的 3 个关键环节,有效定位并解决这 3 个环节的不公平问题,对未来公平机器学习算法的研究和发展具有重要意义.公平具有在法律、社会层次的意义,不完全是一个技术问题,可信机器学习中的公平性研究可以认为是一个社会学与计算机科学的交叉研究领域.在未来工作中,需要探究技术、应用和伦理等多方面的公平问题,部署先进的公平机器学习算法于各应用领域,并形成统一且完整的公平性度量.

References:

[1] Zhou ZH. Machine Learning. Beijing: Tsinghua University Press, 2016 (in Chinese).

[2] He JF. Safe and trustworthy artificial intelligence. Information Security and Communications Privacy, 2019(10):5–8 (in Chinese).

[3] Meng XF, Wang LX, Liu JX. Data privacy, monopoly and fairness for AI. Big Data Research, 2020,6(1):35–46 (in Chinese with English abstract).

[4] Liu RX, Chen H, Guo RY, Zhao D, Liang WJ, Li CP. Survey on privacy attacks and defenses in machine learning. Ruan Jian Xue Bao/Journal of Software, 2020,31(3):866–892 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5904.htm [doi: 10.13328/j.cnki.jos.005904]

[5] Tan ZW, Zhang LF. Survey on privacy preserving techniques for machine learning. Ruan Jian Xue Bao/Journal of Software, 2020, 31(7):2127–2156 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6052.htm [doi: 10.13328/j.cnki.jos.006052]

[6] Cheng KY, Wang N, Shi WX, Zhan YZ. Research advances in the interpretability of deep learning. Journal of Computer Research and Development, 2020,57(6):1208–1217 (in Chinese with English abstract).

[7] Saxena NA, Huang K, 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 AIES. 2019. 99–106.

[8] Kusner MJ, Loftus JR, Russell C, Silva R. Counterfactual fairness. In: Proc. of the NIPS. 2017. 4066–4076.

[9] Suresh H, Guttag JV. A framework for understanding unintended consequences of machine learning. arXiv preprint arXiv:1901.10002v3, 2019.

[10] Mehrabi N, Morstatter F, Saxena N, Lerman K, Aram G. A survey on bias and fairness in machine learning. arXiv preprint arXiv:1908.09635v2, 2019.

[11] Baeza-Yates R. Bias on the Web. Communications of the ACM, 2018,61(6):54–61.

[12] Silva S, Kenney M. Algorithms, platforms, and ethnic bias. Communications of the ACM, 2019,62(11):37–39.

[13] Alessandro BD, Neil CO, LaGatta T. Conscientious classification: A data scientist's guide to discrimination-aware classification. arXiv preprint arXiv:1907.09013v1, 2019.

[14] Chouldechova A, Roth A. The frontiers of fairness in machine learning. arXiv preprint arXiv:1810.08810v1, 2018.

[15] Hardt M, Price E, Srebro N. Equality of opportunity in supervised learning. In: Proc. of the NIPS. 2016. 3315–3323.

[16] Chen IY, Johansson FD, Sontag DA. Why is my classifier discriminatory? In: Proc. of the NeurIPS. 2018. 3543–3554.

[17] Buolamwini J, Gebru T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Proc. of the FAT. 2018. 77–91.

[18] Dwork C, Hardt M, Pitassi T, Reingold O, Zemel RS. Fairness through awareness. In: Proc. of the ITCS. 2012. 214–226.

[19] Friedler SA, Scheidegger C, Venkatasubramanian S, Choudhary S, Hamilton EP, Roth D. A comparative study of fairness-enhancing interventions in machine learning. In: Proc. of the FAT. 2019. 329–338.

[20] Calders T, Verwer S. Three naive Bayes approaches for discrimination-free classification. Data Mining and Knowledge Discovery, 2010,21(2):277−292.

[21] Edwards H, Storkey AJ. Censoring representations with an adversary. In: Proc. of the ICLR (Poster). 2016.

[22] Louizos C, Swersky K, Li YJ, Welling M, Zemel RS. The variational fair autoencoder. In: Proc. of the ICLR. 2016.

[23] Zemel RS, Wu Y, Swersky K, Pitassi T, Dwork C. Learning fair representations. In: Proc. of the ICML. 2013. 325−333.

[24] Madras D, Creager E, Pitassi T, Zemel RS. Learning adversarially fair and transferable representations. In: Proc. of the ICML. 2018. 3381−3390.

[25] Adel T, Valera I, Ghahramani Z, Weller A. One-network adversarial fairness. In: Proc. of the AAAI. 2019. 2412−2420.

[26] Feldman M, Friedler SA, Moeller J, Scheidegger C, Venkatasubramanian S. Certifying and removing disparate impact. In: Proc. of the KDD. 2015. 259−268.

[27] Zafar MB, Valera I, Gomez-Rodriguez M, Gummadi KP. Fairness constraints: Mechanisms for fair classification. In: Proc. of the AISTATS. 2017. 962−970.

[28] Berk R, Heidari H, Jabbari S, Michael K, Roth A. Fairness in criminal justice risk assessments: The state of the art. In: Proc. of the Sociological Methods and Research. 2018. 3−44.

[29] Kleinberg JM, Mullainathan S, Raghavan M. Inherent trade-offs in the fair determination of risk scores. In: Proc. of the ITCS. 2017. 1−23.

[30] Chouldechova A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, 2017,5(2): 153−163.

[31] Pearl J. Causality. Cambridge University Press, 2009.

[32] Chiappa S. Path-specific counterfactual fairness. In: Proc. of the AAAI. 2019. 7801−7808.

[33] Loftus JR, Russell C, Kusner MJ, Silva R. Causal reasoning for algorithmic fairness. arXiv preprint arXiv:1805.05859v1, 2018.

[34] Beutel A, Chen JL, Zhao Z, Chi EH. Data decisions and theoretical implications when adversarially learning fair representations. arXiv preprint arXiv:1707.00075v2, 2017.

[35] Zhao H, Gordon GJ. Inherent tradeoffs in learning fair representations. In: Proc. of the NeurIPS. 2019. 15649−15659.

[36] Khosravifard M, Fooladivanda D, Gulliver TA. Confliction of the convexity and metric properties in f-divergences. IEICE Trans. on Fundamentals of Electronics, Communications and Computer Sciences, 2017,90(9):1848−1853.

[37] Zhao H, Coston A, Adel T, Gordon GJ. Conditional learning of fair representations. In: Proc. of the ICLR. 2020.

[38] Xu DP, Yuan SH, Zhang L, Wu XT. FairGAN: Fairness-aware generative adversarial networks. In: Proc. of the BigData. 2018. 570−575.

[39] Xu DP, Yuan SH, Zhang L, Wu XT. FairGAN$^+$: Achieving fair data generation and classification through generative adversarial nets. In: Proc. of the BigData. 2019. 1401−1406.

[40] Xu DP, Wu YK, Yuan SH, Zhang L, Wu XT. Achieving causal fairness through generative adversarial networks. In: Proc. of the IJCAI. 2019. 1452−1458.

[41] Kocaoglu M, Snyder C, Dimakis AG, Vishwanath S. CausalGAN: Learning causal implicit generative models with adversarial training. In: Proc. of the ICLR (Poster). 2018.

[42] Creager E, Madras D, Jacobsen JH, Weis MA, Swersky K, Pitassi T, Zemel RS. Flexibly fair representation learning by disentanglement. In: Proc. of the ICML. 2019. 1436−1445.

[43] Gordaliza P, Barrio E, Gamboa F, Loubes JM. Obtaining fairness using optimal transport theory. In: Proc. of the ICML. 2019. 2357−2365.

[44] Bechavod Y, Ligett K. Learning fair classifiers: A regularization-inspired approach. arXiv preprint arXiv:1707.00044v3, 2017.

[45] Chris R, Kusner MJ, Loftus JR, Silva R. When worlds collide: integrating different counterfactual assumptions in fairness. In: Proc. of the NIPS. 2017. 6414−6423.

[46] Wu YK, Zhang L, Wu XT, Tong HH. PC-fairness: A unified framework for measuring causality-based fairness. In: Proc. of the NeurIPS. 2019. 3399−3409.

[47] Wu YK, Zhang L, Wu XT. Counterfactual fairness: Unidentification, bound and algorithm. In: Proc. of the IJCAI. 2019. 1438−1444.

[48] Agarwal A, Dudík M, Wu ZWS. Fair regression: Quantitative definitions and reduction-based algorithms. In: Proc. of the ICML. 2019. 120–129.

[49] Conitzer V, Freeman R, Shah N, Vaughan JW. Group fairness for the allocation of indivisible goods. In: Proc. of the AAAI. 2019. 1853–1860.

[50] Kusner MJ, Russell C, Loftus JR, Silva R. Making decisions that reduce discriminatory impacts. In: Proc. of the ICML. 2019. 3591–3600.

[51] Ustun B, Liu Y, Parkes DC. Fairness without harm: Decoupled classifiers with preference guarantees. In: Proc. of the ICML. 2019. 6373–6382.

[52] Tsang A, Wilder B, Rice E, Tambe M, Zick Y. Group-fairness in influence maximization. In: Proc. of the IJCAI. 2019. 5997–6005.

[53] Chen XY, Fain B, Lyu L, Munagala K. Proportionally fair clustering. In: Proc. of the ICML. 2019. 1032–1041.

[54] Jiang R, Pacchiano A, Stepleton T, Jiang H, Chiappa S. Wasserstein fair classification. In: Proc. of the UAI. 2019. 862–872.

[55] Jagielski M, Kearns MJ, Mao JM, Oprea A, Roth A, Sharifi-Malvajerdi S, Ullman J. Differentially private fair learning. In: Proc. of the ICML. 2019. 3000–3008.

[56] Wang TL, Zhao JY, Yatskar M, Chang KW, Ordonez V. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In: Proc. of the ICCV. 2019. 5309–5318.

[57] Quadrianto N, Sharmanska V, Thomas O. Discovering fair representations in the data domain. In: Proc. of the CVPR. 2019. 8227–8236.

[58] DeVries T, Misra I, Wang CH, Maaten LVD. Does object recognition work for everyone? In: Proc. of the CVPR Workshops. 2019. 52–59.

[59] Wang ZY, Qinami K, Karakozis IC, Genova K, Nair P, Hata K, Russakovsky O. Towards fairness in visual recognition: Effective strategies for bias mitigation. In: Proc. of the CVPR. 2020. 8916–8925.

[60] Reid P, Martinez RD, Dass N, Kurohashi S, Jurafsky D, Yang DY. Automatically neutralizing subjective bias in text. In: Proc. of the AAAI. 2020. 480–489.

[61] Bolukbasi T, Chang KW, Zou JY, Saligrama V, Kalai AT. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In: Proc. of the NIPS. 2016. 4349–4357.

[62] Zhao JY, Wang TL, Yatskar M, Ordonez V, Chang KW. Men also like shopping: reducing gender bias amplification using corpus-level constraints. In: Proc. of the EMNLP. 2017. 2979–2989.

[63] Zhao JY, Wang TL, Yatskar M, Ordonez V, Chang KW. Gender bias in coreference resolution: Evaluation and debiasing methods. In: Proc. of the NAACL-HLT, Vol.2. 2018. 15–20.

[64] Stanovsky G, Smith NA, Zettlemoyer L. Evaluating gender bias in machine translation. In: Proc. of the ACL. 2019. 1679–1684.

[65] Du YP, Wu YB, Lan M. Exploring human gender stereotypes with word association test. In: Proc. of the EMNLP/IJCNLP. 2019. 6132–6142.

[66] Papakyriakopoulos O, Hegelich S, Serrano JCM, Marco F. Bias in word embeddings. In: Proc. of the FAT*. 2020. 446–457.

[67] Ma PC, Wang S, Liu J. Metamorphic testing and certified mitigation of fairness violations in NLP models. In: Proc. of the IJCAI. 2020. 458–465.

[68] Badilla P, Marquez FB, Pérez J. WEFE: The word embeddings fairness evaluation framework. In: Proc. of the IJCAI. 2020. 430–436.

[69] Yadav H, Du ZX, Joachims T. Fair learning-to-rank from implicit feedback. arXiv preprint arXiv:1911.08054v1, 2019.

[70] Beutel A, Chen JL, Doshi T, Qian H, Wei L, Wu Y, Heldt L, Zhao Z, Hong LC, Chi EH, Goodrow C. Fairness in recommendation ranking through pairwise comparisons. In: Proc. of the KDD. 2019. 2212–2220.

[71] Singh A, Joachims T. Policy learning for fairness in ranking. In: Proc. of the NeurIPS. 2019. 5427–5437.

[72] Patro GK, Biswas A, Ganguly N, Gummadi KP, Chakraborty A. FairRec: Two-sided fairness for personalized recommendations in two-sided platforms. In: Proc. of the WWW. 2020. 1194–1204.

[73] Patro GK, Chakraborty A, Ganguly N, Gummadi KP. Fair updates in two-sided market platforms: On incrementally updating recommendations. In: Proc. of the AAAI. 2020. 181–188.

[74] Chakraborty A, Patro GK, Ganguly N, Gummadi KP, Loiseau P. Equality of voice: Towards fair representation in crowdsourced top-*K* recommendations. In: Proc. of the FAT. 2019. 129–138.

[75] Fu ZH, Xian YK, Gao RY, Zhao JY, Huang QY, Ge YQ, Xu SY, Geng SJ, Shah C, Zhang YF, Melo GD. Fairness-aware explainable recommendation over knowledge graphs. In: Proc. of the SIGIR. 2020. 69–78.

[76] Geyik SC, Ambler S, Kenthapadi K. Fairness-aware ranking in search & recommendation systems with application to LinkedIn talent search. In: Proc. of the KDD. 2019. 2221–2231.

[77] Morik M, Singh A, Hong J, Joachims T. Controlling fairness and bias in dynamic learning-to-rank. In: Proc. of the SIGIR. 2020. 429–438.

[78] Hughes E, Leibo JZ, Phillips M, Tuyls K, Duéñez-Guzmán EA, Castañeda AG, Dunning I, Zhu T, McKee KR, Koster R, Roff H, Graepel T. Inequity aversion improves cooperation in intertemporal social dilemmas. In: Proc. of the NeurIPS. 2018. 3330–3340.

[79] Zhang CJ, Shah JA. Fairness in multi-agent sequential decision-making. In: Proc. of the NIPS. 2014. 2636–2644.

[80] Jiang JC, Lu ZQ. Learning fairness in multi-agent systems. In: Proc. of the NeurIPS. 2019. 13854–13865.

[81] Zhou J, Wang L, Wang L, Zheng XL. Shared learning: Ant financial's solution. Communications of the CCF, 2020,15(6):51–57 (in Chinese).

[82] Li T, Sanjabi M, Beirami A, Smith V. Fair resource allocation in federated learning. In: Proc. of the ICLR. 2020.

[83] Mohri M, Sivek G, Suresh AT. Agnostic federated learning. In: Proc. of the ICML. 2019. 4615–4625.

[84] Kohavi R. Scaling up the accuracy of Naive-Bayes classifiers: A decision-tree hybrid. In: Proc. of the KDD. 1996. 202–207.

[85] Wightman LF. LSAC national longitudinal bar passage study. Research Report, Law School Admission Council, 1998.

[86] Merler M, Ratha NK, Feris RS, Smith JR. Diversity in faces. arXiv preprint arXiv:1901.10436v6, 2019.

[87] Horn GV, Aodha OM, Song Y, Cui Y, Sun C, Shepard A, Adam H, Perona P, Belongie SJ. The INaturalist species classification and detection dataset. In: Proc. of the CVPR. 2018. 8769–8778.

[88] Bagdasaryan E, Poursaeed O, Shmatikov V. Differential privacy has disparate impact on model accuracy. In: Proc. of the NeurIPS. 2019. 15453–15462.

[89] D'Amour A, Srinivasan H, Atwood J, Baljekar P, Sculley D, Halpern Y. Fairness is not static: Deeper understanding of long term fairness via simulation studies. In: Proc. of the FAT*. 2020. 525–534.

[90] Liu LT, Dean S, Rolf E, Simchowitz M, Hardt M. Delayed impact of fair machine learning. In: Proc. of the ICML. 2018. 3156–3164.

[91] Nori H, Jenkins S, Koch P, Caruana R. InterpretML: A unified framework for machine learning interpretability. arXiv preprint arXiv:1909.09223v1, 2019.

[92] Bellamy RKE, Dey K, Hind M, Hoffman SC, Houde S, Kannan K, Lohia P, Martino J, Mehta S, Mojsilovic A, Nagar S, Ramamurthy KN, Richards JT, Saha D, Sattigeri P, Singh M, Varshney KR, Zhang TF. AI fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM Journal of Research and Development, 2019,63(4):1–15.

附中文参考文献:

[1] 周志华.机器学习.北京:清华大学出版社,2016.

[2] 何积丰.安全可信人工智能.信息安全与通讯保密,2019(10):5–8.

[3] 孟小峰,王雷霞,刘俊旭.人工智能时代的数据隐私、垄断与公平.大数据,2020,6(1):35–46.

[4] 刘睿瑄,陈红,郭若杨,赵丹,梁文娟,李翠平.机器学习中的隐私攻击与防御.软件学报,2020,31(3):866–892. http://www.jos.org.cn/1000-9825/5904.htm [doi: 10.13328/j.cnki.jos.005904]

[5] 谭作文,张连福.机器学习隐私保护研究综述.软件学报,2020,31(7):2127–2156. http://www.jos.org.cn/1000-9825/6052.htm [doi: 10.13328/j.cnki.jos.006052]

[6] 成科扬,王宁,师文喜,詹永照.深度学习可解释性研究进展.计算机研究与发展,2020,57(6):1208–1217.

[81] 周俊,王力,王磊,郑小林.蚂蚁金服共享智能实践.中国计算机学会通讯,2020,15(6):51–57.

刘文炎(1995－),女,博士生,主要研究领域为可信机器学习.

沈楚云(1997－),男,博士生,主要研究领域为可信机器学习,多智体强化学习,机器学习中的公平.

王祥丰(1987－),男,博士,副教授,CCF 专业会员,主要研究领域为分布式优化,多智能体强化学习,可信机器学习.

金博(1982－),男,博士,讲师,CCF 专业会员,主要研究领域为可信机器学习,多智体强化学习,计算机视觉技术及应用.

卢兴见(1986－),男,博士,副教授,CCF 专业会员,主要研究领域为可信机器学习,云计算.

王晓玲(1975－),女,博士,教授,博士生导师,CCF 专业会员,主要研究领域为可信机器学习,知识图谱,个性化推荐技术,机器学习及隐私保护技术.

查宏远(1963－),男,博士,教授,博士生导师,主要研究领域为机器学习.

何积丰(1943－),男,教授,博士生导师,CCF 会士,主要研究领域为可信人工智能,可信软件.