公平机器学习:概念、分析与设计

古天龙[1),2)] 李 龙[1),2)] 常 亮[2)] 罗义琴[1)]

[1)](暨南大学信息科学技术学院 广州 510632)

[2)](桂林电子科技大学广西可信软件重点实验室 广西 桂林 541004)

摘 要 随着人工智能的发展,机器学习技术越来越多地应用于社会各个领域,用以辅助或代替人们进行决策,特别是在一些具有重要影响的领域,例如,信用程度评级、学生质量评估、福利资源分配、疾病临床诊断、自然语言处理、个性信息推荐、刑事犯罪判决、无人驾驶等.如何在这些应用中确保决策公平或者无偏见?如何在这些应用中保护弱势群体的利益?这些问题直接影响到社会和公众对机器学习的信任,影响到人工智能技术的应用与系统的部署.通过系统梳理和全面剖析近年来的工作,对机器学习公平性或公平机器学习的定义及度量进行了解释及对比;从机器学习的全生命周期出发,对不同环节中出现的各类偏见及其发现技术进行了归类及阐释;从预处理、中间处理和后处理三个阶段,对公平机器学习的设计技术进行了介绍和分析;从可信赖人工智能全局出发,对公平性与隐私保护、可解释性之间的关系、影响及协同解决方案进行了阐述;最后对公平机器学习领域中亟待解决的主要问题、挑战及进一步研究热点进行了讨论.

关键词 机器学习;公平性;隐私保护;可解释;人工智能伦理

中图法分类号 TP18 **DOI号** 10.11897/SP.J.1016.2022.01018

Fair Machine Learning: Concepts, Analysis, and Design

GU Tian-Long[1),2)] LI Long[1),2)] CHANG Liang[2)] LUO Yi-Qin[1)]

[1)](*College of Information Science and Technology, Jinan University, Guangzhou* 510632)

[2)](*Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology, Guilin, Guangxi* 541004)

Abstract With the development of artificial intelligence, machine learning techniques is increasingly used in many social domains to assist or replace humankinds in decision-making, especially in some critical areas, such as, credit rating, students' qualification evaluation, welfare resource allocation, clinical diagnosis, natural language processing, personalized information recommendation, criminal judgment, autonomous vehicles and so on. Due to the intrinsic and technical characteristics of machine learning itself, its prediction and decision-making will inevitably produce a certain degree of bias or unfairness, which has gradually attracted the attention of scientific research, industry practitioners and the public. How to ensure fair or unbiased decisions in machine learning? How to protect the interests of disadvantaged groups in these applications? These issues have important impacts on the society and the public's confidence in machine learning and affect the application of artificial intelligence technology and the deployment of artificial intelligence systems. Fairness has been one of the basic supporting capabilities of trustworthy artificial intelligence, and machine learning with fairness is referred to as fair machine learning. In this paper, the concepts of fairness, the methods of discovering unfair or biased discrimination and the design techniques of fair machine

收稿日期:2021-08-11;在线发布日期:2022-03-09.本课题得到国家自然科学基金(U1711263,U1811264,61966009)资助.**古天龙**,博士,教授,中国计算机学会(CCF)高级会员,主要研究领域为形式化方法、可信人工智能、人工智能伦理、数据治理等.E-mail:gutianlong@jnu.edu.cn.**李 龙**(通信作者),博士,讲师,中国计算机学会(CCF)会员,主要研究方向为人工智能安全、公平机器学习、逻辑程序设计等.E-mail:lilong@guet.edu.cn.**常 亮**,博士,教授,主要研究领域为知识图谱、知识表示、形式化方法等.**罗义琴**,博士研究生,主要研究方向为公平表示学习、可信机器学习.

learning are reviewed and discussed. The detailed contents include the followings. Firstly, discrimination and bias are terminologies related to unfairness, and unfair behavior is known as biased behavior or discriminatory behavior. Since the taxonomy of discrimination and biases is helpful to understand and evaluate the fairness, direct discrimination, indirect discrimination, interpretable discrimination, uninterpretable discrimination, statistical discrimination and systematic discrimination are explained. In the framework of statistics, similarity and causal inference, the definitions and quantification of fairness in machine learning are categorized and explained. Secondly, the bias or prejudice is the main source of discrimination and unfairness. The training data and algorithms involved in machine learning can have biases that lead to unfair model predictions. From the perspectives of data, algorithm and human-computer interaction, the biases in the life cycle of machine learning are classified and discussed. The techniques to discover biases in machine learning, such as association rule mining, k-nearest neighbor classification, probabilistic causal network, and privacy attack and deep learning methods, are illustrated. Meanwhile, the design methodologies of fair machine learning have been undertaken roughly in three directions. On the view of specific applicable tasks, fair natural language processing, fair face recognition, fair recommendation system, fair classification, fair regression and fair clustering are elaborated. In light of particular machine learning algorithms, fair representation and fair adversarial learning are discoursed. From the life cycle of machine learning, preprocessing methods, intermediate processing methods and post-processing methods are expounded. Then, for the trustworthy artificial intelligence, the recent studies regarding anonymous protection, secure multi-party computing and security attack and defense for fair machine learning are promising works, which are briefly introduced. The explainability can help to discover algorithmic bias in machine learning models, on which some preliminary attempts are conducted, also being described. Finally, the main problems, challenges and hot topics in the research of fair machine learning, such as evaluation and testing of fair machine learning, novel modes of fair machine learning and ethically aligned machine learning, are presented.

Keywords　machine learning; fairness; privacy protection; interpretability; artificial intelligence ethics

《计算机学报》

1　引　言

机器学习(Machine Learning,ML)是人工智能(Artificial Intelligence,AI)的一个重要分支,是对通过数据或以往经验自动改进计算机系统或算法的性能的研究[1-3].随着数据的丰富与算力的提升,机器学习技术得到了长足发展,已经在与大众生活密切相关的诸多方面得到了广泛应用.受机器学习自身本质和技术特征的影响,其预测和决策会产生一定程度的偏见或不公平,这一问题逐渐引起科学研究、产业界从业人员和社会公众的关注[4-5].在预测和决策过程中,公平是指不存在基于个人或群体的内在或后天特征的任何偏见、偏好、歧视或不公正[6].因此,一个不公平的算法是指其决策对某一个体或特定群体存在偏见,由此引发对该个体或群体的不公正待遇,并使其利益受到损害.

人工智能应用中的偏见歧视已经出现在不少场景.例如,机票预订系统 SABRE 和 Apollo 存在的不公平和偏见[7-8],导致了航空公司之间的不公平竞争;许多推荐系统会放大数据中的偏见、引发不公平推荐[9-11],几乎所有的排名算法都采取了“短视”效用优化策略,导致了不公平[12];基于深度学习的人脸识别算法极大地提高了识别准确率,但大多数算法在男性面孔上的表现优于女性面孔,即人脸识别算法存在性别偏见[13-14];简历自动筛选系统通常会因应聘者无法控制的特质(如性别、种族、性取向等)而给出带有偏见的评测[9,15],这样的不公平不仅会对求职者产生歧视或偏见,也可能因错失优秀雇员而给雇主带来损失;对电子病历或医疗记录进行分析可预测(慢性)疾病,对于某些族群的错误率明显

高于其他族群,存在族群偏见或歧视[16-19];教师评价系统 IMPACT[9,20]通过教师的年龄、教育水平、经验、课堂观察、问卷调查等特征、学生考试成绩、学生问卷调查和学校的问卷调查、教师的问卷调查等来学习并分析教师的工作表现及应得工作报酬,对贫困社区教师可能产生系统性的较低评分①;GRADE、Kira Talent 等大学入学评估系统通过学习考生的就读学校、SAT 成绩、课外活动、GPA、面试成绩等,给出接收/拒绝考生的结果或者考生在相关领域的潜在表现评分[21-22],存在对特定种族群体的偏见和歧视[9,23];刑事风险评估系统 COMPAS 等[9,24-25],依据被捕次数、犯罪类型、家庭地址、就业状况、婚姻状况、收入、年龄、住房等,给出被告是否会再次犯罪的风险评分. ProPublica 曝光了这类系统评估中的不公平②和歧视[26];贷款发放评估系统 FICO、Equifax、TransUnion 等[9],给出的针对贷款人的贷款还款计划和贷款年利率的建议方案,会针对女性或者某些族群给出过高定价,造成系统性偏见[27];自然语言处理有放大社会对性别的已有成见的风险,导致对不同性别群体的不公平[28]. 共指消解系统 Stanford Deterministic Coreference System 等表现出了系统性的性别偏见. 类似的不公平现象也存在于在线新闻、信息检索、广告投送等领域[28-30].

机器学习中的不公平和偏见问题直接影响着社会和公众对其信任程度,影响着人工智能系统的应用部署,是机器学习技术研究与应用开发所面临的新挑战. 如何对公平性进行合理定义及度量?如何发现机器学习应用中的不公平?如何设计公平机器学习或者具有公平属性的机器学习?如何实现具有隐私保护或可解释性等能力的公平机器学习,并最终实现符合伦理的机器学习?为明确以上挑战的内涵并进行有效应对,本文对相关研究工作进行了系统性调研与剖析,并对公平机器学习的未来研究及值得关注的问题进行了讨论和展望. 图 1 为本文的组织架构图.

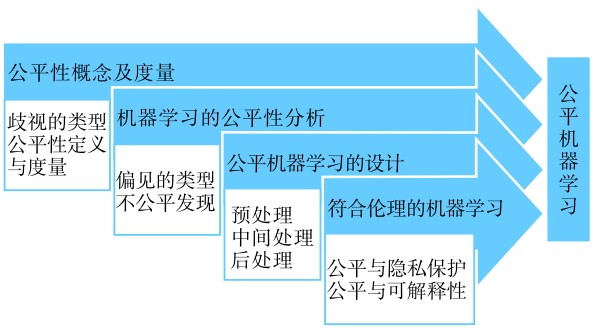


图 1 本文组织架构图

2 公平性概念及度量

公平性问题一直是哲学、政治、道德、法律等人文社科领域感兴趣的话题,公平性概念的提出和探讨始于 20 世纪 60 年代[31-32]. 能够确保每个人都有平等的机会获得一些利益的行为,称为公平的行为,或者称这样的行为具有公平性. 不能够确保每个人平等地获得一些利益,使得弱势群体的利益受到损害的行为,称为不公平的行为,或者称这样的行为具有不公平性. 歧视和偏见是与不公平相关联的概念,不公平的行为又称为具有偏见的行为或者歧视的行为. 如果机器学习的预测或决策结果能够确保每个人都有平等的机会获得一些利益,就称该机器学习具有公平性,并称之为公平机器学习. 公平性研究已经有 50 余年的历史,无论概念定义、还是度量评测都得到了极大的发展,不同文化具有不同偏好和观点视角,导致了人们对公平存在多种不同的理解方式. 目前还没有公平性的普适定义,为了满足各种应用需求,产生了各种各样的公平性定义、概念及度量. 对于歧视类型的了解,有助于各种公平性概念定义的理解[6]. 下面讨论歧视的类型、公平性定义及度量等.

2.1 歧视的类型

歧视可以由三个层次的从属概念来刻画[33]:(1)什么行为?(2)什么情况下?(3)对谁造成了歧视?行为是歧视的表现形式,情况是歧视的作用领域或场景,而歧视的理由描述了受到歧视的对象的特征. 从造成歧视的理由是否有明确表述的角度,歧视呈现直接性歧视和间接性歧视两种主要形式. 从歧视的行为是否能够被解释角度,歧视分为可解释性歧视和不可解释性歧视. 此外,系统性歧视刻画了文化和习俗等方面的负面影响所带来的歧视,统计性歧视刻画了社会成见的不良后果所导致的歧视.

(1)直接性歧视(Direct Discrimination). 由于受保护属性的明显原因导致了某个人或某群体的利益受到损失,由此所产生的歧视称为直接性歧视[34],也称为不平等对待(Disparate Treatment). 通常,法律规定基于某些特征的歧视行为是违法的,这些特征通常被认为是“受保护的”或“敏感的”属

① The unfair effects of impact on teachers with the toughest jobs. https://tcf.org/content/commentary/the-unfair-effects-of-impact-on-teachers-with-the-toughest-jobs, 2015,10,16

② Machine bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, 2016,5,23

《计算机学报》

性[35]. 例如，因为某人的宗教信仰特征，房屋户主不将房屋出租给该租房客的行为是一种直接性歧视.

(2) 间接性歧视(Indirect Discrimination). 基于看似中立和不受保护属性来对待某个人或某群体，然而，由于其受保护属性的隐性效应的间接影响，该个人或群体仍然会受到不公正的对待，由此所产生的歧视称为间接性歧视[34]，也称为不平等影响(Disparate Impact). 例如，某个俱乐部规定进入人员必须出示驾驶证来证明身份，这样就存在对视力受损人员的歧视，因为盲人不可能持有驾照，这是一种间接性歧视.

(3) 可解释性歧视(Explainable Discrimination). 在某些情况下，歧视可以通过某些案例中的某些属性得到合理解释，称之为可解释性歧视或者可解释的[36]. 例如，在 UCI 数据集中，男性平均年收入高于女性平均年收入，这是因为女性每周的平均工作时间少于男性. 如果在决策时不考虑每周工作时间，以至于男性和女性的平均工资收入相同，就会导致男性员工的工资低于女性员工的工资的反向歧视，该歧视是可解释性歧视.

(4) 不可解释性歧视(Unexplainable Discrimination). 与可解释性歧视相反，不可解释性歧视是指缺乏合理或正当理由来解释的歧视. 不可解释性歧视是不合理、甚至非法的. 一个不可解释性歧视，可能是一个直接性歧视，也可能是一个间接性歧视. 例如，某帆船爱好者俱乐部，接收会员申请，电子自动筛选系统根据某申请者的家庭住址的邮政编码，拒绝了其会员申请，这是不可解释性歧视，也是间接性歧视.

(5) 系统性歧视(Systematic Discrimination). 系统性歧视是受到根植于文化或政治制度中的某些政策和习俗的影响，所导致的对某些群体的长期或永久性的歧视[37]. 例如，一家餐厅为了满足顾客的偏好，定位了经营的文化特色，餐厅经理青睐于选择特征与自己餐厅文化相似的员工，对不具有这些特征而有能力从事餐厅服务工作的就业申请者就产生了歧视，导致了对不符合餐厅文化要求的应聘者群体的歧视，这是一个系统性歧视.

(6) 统计性歧视(Statistical Discrimination). 使用群体统计数据的平均特征，对属于该群体的个人进行评测并给出决策，由此所产生的歧视，称为统计性歧视[38]. 例如，某些用人单位依据应聘者所就读的大学选择新员工，因为，依据统计数据，985、211、双一流大学的毕业生总体上工作能力较高，但是，不太有名的大学的毕业生中也有许多较强实际能力的学生. 由此，基于就读学校选择应聘学生导致了统计性歧视.

2.2　公平性定义与度量

Cleary 首先给出了根据学生测试分数预测教育效果的不公平定义[39]. Hutchinson 和 Mitchell 从多学科角度对公平性概念定义进行了述评[31]，讨论了不同历史时期公平性定义产生的文化和社会背景. Verma 和 Rubin 将分类问题中 20 个主要的公平性定义划分为统计度量、相似度量、因果推理等三大类，并结合数据集 German Credit Dataset 中的信用评分进行了综述和讨论[40]. Makhlouf 等人基于机器学习决策的特征，从公平性定义与机器学习决策相适配的视角，系统地剖析了主要公平性概念及定义[9]. 下面从统计、相似和因果推理三个方面介绍机器学习的公平性定义与度量.

为了表述方便，引入如下符号：X 表示个体的所有属性；G 表示受保护或敏感属性；Y 表示真实(分类)结果(c 为 Y 中的一个元素)；S 表示某一分类 c 的预测概率 $P(Y=c|G,X)$；d 表示预测结果，通常由 S 导出，例如，当 S 超过某一阈值时，$d=1$. 同时，引入混淆矩阵中相关术语[3,40](参见表 1).

表 1　混淆矩阵相关术语

预测结果	真实情况	
	真实-负($Y=0$)	真实-正($Y=1$)
预测-正	TP $PPV=TP/(TP+FP)$ $TPR=TP/(TP+FN)$	FP $FDR=FP/(TP+FP)$ $FPR=FP/(FP+TN)$
预测-负	FN $FOR=FN/(TN+FN)$ $FNR=FN/(TP+FN)$	TN $NPV=TN/(TN+FN)$ $TNR=TN/(FP+TN)$

真阳性　真实正类被预测为正类，用 TP(True Positive)表示其对应样例的数目；

真阴性　真实负类被预测为负类，用 TN(True Negative)表示其对应样例的数目；

假阳性　真实负类被预测为正类，用 FP(False Positive)表示其对应样例的数目；

假阴性　真实正类被预测为负类，用 FN(False Negative)表示其对应样例的数目；

阳性预测率　PPV(Positive Predictive Value)，正类预测结果中真实正类的占比，$PPV=TP/(TP+FP)$，也称为准确率或查准率，它表示正类预测中真实正类的概率 $P(Y=1|d=1)$(注：$P(u|v)$ 表示条件 v 下 u 的概率或者 u 在条件 v 下的条件概率)；

《计算机学报》

阴性预测率 NPV(Negative Predictive Value),负类预测结果中真实负类的占比,$NPV=TN/(TN+FN)$,它表示负类预测中真实负类的概率 $P(Y=0|d=0)$;

假发现率 FDR(False Discovery Rate),正类预测结果中被错误预测为正类的占比,$FDR=FP/(TP+FP)$,它表示正类预测中被误预测为正类的概率 $P(Y=0|d=1)$;

假漏报率 FOR(False Omission Rate),负类预测结果中被错误预测为负类的占比,$FOR=FN/(TN+FN)$,它表示负类预测中被误预测为负类的概率 $P(Y=1|d=0)$;

真阳性率 TPR(True Positive Rate),真实正类中被预测为正类的占比,$TPR=TP/(TP+FN)$,也称为召回率或查全率,它表示真实正类被预测为正类的出现概率 $P(d=1|Y=1)$;

真阴性率 TNR(True Negative Rate),真实负类中被预测为负类的占比,$TNR=TN/(FP+TN)$,它表示真实负类被预测为负类的出现概率 $P(d=0|Y=0)$;

假阳性率 FPR(False Positive Rate),真实负类中被误预测为正类的占比,$FPR=FP/(FP+TN)$,它表示真实负类中被误预测为正类的出现概率 $P(d=1|Y=0)$;

假阴性率 FNR(False Negative Rate),真实正类中被误预测为负类的占比,$FNR=FN/(TP+FN)$,它表示真实正类中被误预测为负类的出现概率 $P(d=0|Y=1)$;

总体精准度 OR(Overall Accuracy),样例中得到正确预测的占比,$OR=(TP+TN)/(TP+FP+TN+FN)$;

基础率 BR(Base Rate),样例中预测为正类的占比,$BR=(TP+FN)/(TP+NP+TN+FN)$.

2.2.1 统计度量与定义

公平性的统计定义和度量大致可分为基于预测结果、基于预测和真实结果、基于预测概率和真实结果共三类[40],下面分别进行介绍.

(1)基于预测结果

定义 1. 统计公平. 如果受保护群体和非受保护群体具有相同的正类预测概率,或者预测与敏感属性无关,则称为统计公平(Statistical Parity)[41],又称为群体公平(Group Fairness),或者接受率平等(Equal Acceptance Rate)[42].

定义 2. 条件统计公平. 在属性 $L\subseteq X$ 下,如果受保护群体和非受保护群体具有相同的正类预测概率,则称为条件统计公平(Conditional Statistical Parity)[43].

(2)基于预测和真实结果

定义 3. 预测公平. 如果受保护群体和非受保护群体的 PPV 等值,则称为预测公平(Predictive Parity)[44],又称为结果检验(Outcome Test)[45].

定义 4. 假阳性率平衡. 如果受保护群体和非受保护群体的 FPR 等值,则称为假阳性率平衡(False Positive Error Rate Balance)[44],又称为预测平等(Predictive Equality)[43].

定义 5. 假阴性率平衡. 如果受保护群体和非受保护群体的 FNR 等值,则称为假阴性率平衡(False Negative Error Rate Balance)[44],又称为平等机会(Equal Opportunity)[46].

定义 6. 条件过程精准度平等. 如果受保护群体和非受保护群体的 TPR 等值、FPR 也等值,则称为条件过程精准度平等(Conditional Procedure Accuracy Equality)[47],又称为均衡几率(Equalized Odds)[46]. 从数理逻辑角度,定义 6 是定义 4 和定义 5 中两个条件的合取.

定义 7. 条件使用精准度平等. 如果受保护群体和非受保护群体的 PPV 等值、NPV 也等值,则称为条件使用精准度平等(Conditional Use Accuracy Equality)[47].

定义 8. 总体精准度平等. 如果受保护群体和非受保护群体具有相等的总体精准度 OR,则称为总体精准度平等(Overall Accuracy Equality)[47].

定义 9. 处置平等. 如果受保护群体和非受保护群体具有相等的 FPR 和 FNR,则称为处置平等(Treatment Equality)[47]. 该定义关注的是错误预测的比率,而不是精准度.

(3)基于预测概率和真实结果

定义 10. 检验公平. 对于预测概率 S,受保护群体和非受保护群体中属于真实正类的概率相等,则称为检验公平(Test-fairness)[45],又称为校准(Calibration)[45].

定义 11. 良态校准. 对于预测概率 S,受保护群体和非受保护群体中属于真实正类的概率都为 S,则称为良态校准(Well-calibration)[48].

定义 12. 正类平衡. 如果受保护群体和非受保护群体中的正类具有相等的平均预测概率 S,则

称为正类平衡(Balance for Positive Class)[48].

定义 13. 负类平衡. 如果受保护群体和非受保护群体中的负类具有相等的平均预测概率 S,则称为负类平衡(Balance for Negative Class)[48].

为了便于直观比照,表 2 列出了公平性的 13 种统计度量与定义方式.

表 2　统计度量与定义的对比

分类依据	名称	定义	数学表示
基于预测结果	统计公平	受保护群体和非受保护群体具有相同的正类预测概率.	$P(d=1\|G=m)=P(d=1\|G=f)$
	条件统计公平	在属性 $L\subseteq X$ 下,受保护群体和非受保护群体具有相同的正类预测概率.	$P(d=1\|L,G=m)=P(d=1\|L,G=f)$
基于预测和真实结果	预测公平	受保护群体和非受保护群体的 PPV 等值.	$P(Y=1\|d=1,G=m)=P(Y=1\|d=1,G=f)$
	假阳性率平衡	受保护群体和非受保护群体的 FPR 等值.	$P(d=1\|Y=0,G=m)=P(d=1\|Y=0,G=f)$
	假阴性率平衡	受保护群体和非受保护群体的 FNR 等值.	$P(d=0\|Y=1,G=m)=P(d=0\|Y=1,G=f)$
	条件过程精准度平等	受保护群体和非受保护群体的 TPR、FPR 等值.	$P(d=1\|Y=i,G=m)=P(d=1\|Y=i,G=f)$, $i\in\{0,1\}$
	条件使用精准度平等	受保护群体和非受保护群体的 PPV、NPV 等值	$(P(Y=1\|d=1,G=m)=P(Y=1\|d=1,G=f))\wedge$ $(P(Y=0\|d=0,G=m)=P(Y=0\|d=0,G=f))$
	总体精准度平等	受保护群体和非受保护群体具有相等的总体精准度.	$P(d=Y,G=m)=P(d=Y,G=f)$
	处置平等	受保护群体和非受保护群体具有相等的 FPR 和 FNR.	$P(d=1\|Y=0,G=m)=P(d=1\|Y=0,G=f)\wedge$ $P(d=0\|Y=1,G=m)=P(d=0\|Y=1,G=f)$
基于预测概率和真实结果	检验公平	对于预测概率 S,受保护群体和非受保护群体中属于真实正类的概率相等.	$P(Y=1\|S=s,G=m)=P(Y=1\|S=s,G=f)$
	良态校准	对于预测概率 S,受保护群体和非受保护群体中属于真实正类的概率都为 S.	$P(Y=1\|S=s,G=m)=P(Y=1\|S=s,G=f)=s$
	正类平衡	受保护群体和非受保护群体中的正类具有相等的平均预测概率 S.	$E(S\|Y=1,G=m)=E(S\|Y=1,G=f)$
	负类平衡	受保护群体和非受保护群体中的负类具有相等的平均预测概率 S.	$E(S\|Y=0,G=m)=E(S\|Y=0,G=f)$

2.2.2　相似性度量与定义

统计度量和定义考虑了敏感(受保护)属性,而忽视了其他属性,这样可能隐藏不公平. 相似性度量与定义考虑了非敏感属性,克服了这一局限.

定义 14. 因果歧视. 针对具有相同属性 X 的个体,预测也相同,称为不具有因果歧视,否则,称为因果歧视(Causal Discrimination)[49].

定义 15. 无意识公平. 决策过程中没有显式地使用敏感(受保护)属性,称为无意识公平(Fairness Through Unawareness)[50].

定义 16. 有意识公平. 相似的个体具有相似的预测,称为有意识公平(Fairness Through Awareness)[51]. 在这里,个体之间的相似性通过距离来度量,为了满足公平性,个体预测之间的距离最多应该是个体之间的距离.

2.2.3　因果推理定义

无论是统计度量和定义,还是相似度量和定义,都是从结果观察角度来建立公平性的度量和定义,缺乏对过程细节所导致的不公平的刻画和描述. 因果推理从数据生成的因果过程视角,对受保护属性相关的因果关系所引发的公平性进行了定义,揭示了前两类定义所忽略之处[40,50,52].

因果图是一个有向无环图,其中节点表示个体的属性,有向边表示属性之间的关系[53]. 在因果图中,代理(Proxy)属性是一种能够派生出另一个属性的值的属性,解析(Resolving)属性是一种受保护属性以非歧视性方式影响的属性. 例如,图 2 是一个用于贷款申请评估的因果图,由受保护属性 G、授信额度属性、雇佣长度属性和信用记录属性,以及预测结果 d 等组成. 在该图中,雇佣长度属性是 G 的代理属性:从雇佣长度属性可得出申请者的性别;G 对于授信额度属性是无歧视的,亦即,不同的 G 对应不同的授信额度并不认为是歧视的,授信额度属性是 G 的一个解析属性.

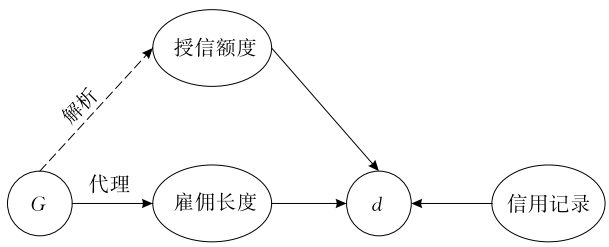


图 2　因果图简例

定义 17. 反事实公平. 如果个体的预测在不同反事实场景中保持不变,称为反事实公平(Counterfactual Fairness)[40,50]. 在因果图中,如果预测结果

d 不依赖于受保护属性 G 的后代，则因果图是反事实公平的. 例如，图 2 中，d 依赖于信用记录、授信额度和雇佣长度，雇佣长度是 G 的直接后代，该因果图模型不具有反事实公平性.

定义 18. 无非解析歧视. 如果因果图中不存在从保护属性 G 到预测结果 d 的路径，或者存在通过解析属性的路径，则称该因果图是无非解析歧视(No Unresolved Discrimination)[40,52]. 例如，图 2 中，通过授信额度从 G 到 d 的路径是非歧视性的，因为授信额度是一个解析属性，通过雇佣长度从 G 到 d 的路径是歧视性的，该图不是无非解析歧视.

定义 19. 无代理歧视. 如果因果图中不存在从受保护属性 G 到预测结果 d 被代理属性阻塞的路径，则称该因果图是无代理歧视(No Proxy Discrimination)[40,52]. 例如，在图 2 所示的因果图中，存在一条通过代理属性雇佣长度从 G 到 d 的路径，该因果图不是无代理歧视.

定义 20. (推论公平) 如果因果图中不存在从受保护属性 G 到预测结果 d 的不合法路径，则称该因果图是推论公平(Fair Inference)[40,54]. 例如，在做出信贷相关的决定时考虑雇佣长度可能是有意义的. 由此，在图 2 所示的因果图中，即使雇佣长度是 G 的代理属性，通过雇佣长度从 G 到 d 的路径也被认为是合法的. 但是，通过授信额度从 G 到 d 的路径是不合法的，这个因果图不是推论公平.

各种公平性度量和定义满足了一定场景的需求，具有一定的优势也存在某些局限. 公平的统计度量和定义需要使用经过验证的真实结果数据，也要求这些数据遵循一定的分布. 相似性度量和定义需要建立个体之间的距离度量，不仅距离度量有一定的难度，而且可能嵌入设计人员的偏见，预测也会由此产生新的偏见. 因果推理的定义，有赖于因果图的建立和搜索，对于复杂和大规模问题，搜索空间可能非常大，导致有些问题难解、甚至不可解. 表 3 对 3 种主要公平性的度量和定义进行了对比.

表 3 公平性度量/定义间的对比

类别	典型特点
统计度量和定义	(1) 从结果观察角度建立公平性的度量和定义，都缺乏对过程细节所导致的不公平的刻画和描述. 仅考虑了敏感(受保护)属性，而忽视了其他属性，可能存在隐藏不公平的现象. (2) 需要使用经过验证的真实结果数据，也要求这些数据遵循一定的分布，其实际应用的场景受到了一定的限制.
相似性度量和定义	(1) 从结果观察角度建立公平性的度量和定义，都缺乏对过程细节所导致的不公平的刻画和描述. 考虑了非敏感属性，解决了隐藏不公平的现象. (2) 需要建立个体之间的距离度量，不仅距离度量有一定的难度，而且可能嵌入设计人员的偏见，预测也会由此产生新的偏见.
因果推理定义	(1) 从数据生成的因果过程视角，对受保护属性相关的因果关系所引发的公平性进行了定义，揭示了过程细节. (2) 可以对过程细节进行形式化规范描述. (3) 有赖于因果图的建立和搜索，对于复杂和大规模问题，搜索空间可能非常大，导致有些问题难解、甚至不可解.

除上述度量和定义外，研究人员还提出了其他公平性相关的定义和度量[6,9,31,33]. 这些公平性度量和定义各有侧重和特点，存在如下方面的主要困难：(1) 不同文化倾向于用不同的方式来看待公平，关于公平性的定义及度量仍然缺乏共识；(2) 难以选择适合具体应用的公平性度量和定义. Makhlouf 等人给出的选择公平性度量和定义的一系列定性准则是解决此问题的有益探索[9]；(3) 某些定义和度量之间存在的矛盾和冲突. 例如，良态校准、正类平衡和负类平衡之间是互不相容的[48]. 如何克服或折衷处理这些冲突也是需要研究的问题[44,47-48].

3 机器学习的公平性分析

机器学习各个阶段所涉及的数据、技术和算法都可能存在导致模型预测不公平的偏见. 偏见是引发歧视和导致不公平的主要来源. 本节对机器学习中可能存在的各种形式的偏见以及不公平性的发现技术等进行介绍和讨论.

3.1 偏见的类型

Friedman 和 Nissenbaum 首先开展了偏见相关方面的研究，给出了计算机系统中偏见分析的一个框架，并结合应用案例进行了阐释[8]. Baeza-Yates 从数据、算法和用户交互等方面对网页生态系统中的相关偏见进行了定义和剖析[55]. Olteanu 等人分析了数据平台及其相关技术特征，从产生的来源和表现的形式，阐述总结了社交数据相关的偏见[56]. Sures 和 Guttag 从数据生成、模型开发及部署两阶段出发，定义和分析了机器学习中可能存在的五种偏见[18]. 本文从机器学习生命周期中数据管理、模型训练、模型评测、模型部署等阶段出发[57]，对各个阶段中存在的偏见进行了梳理(参见图 3). 下面从数据、算法和人机交互三个方面(图 4)，对机器学习中的主要偏见进行分类介绍和讨论.

《计算机学报》

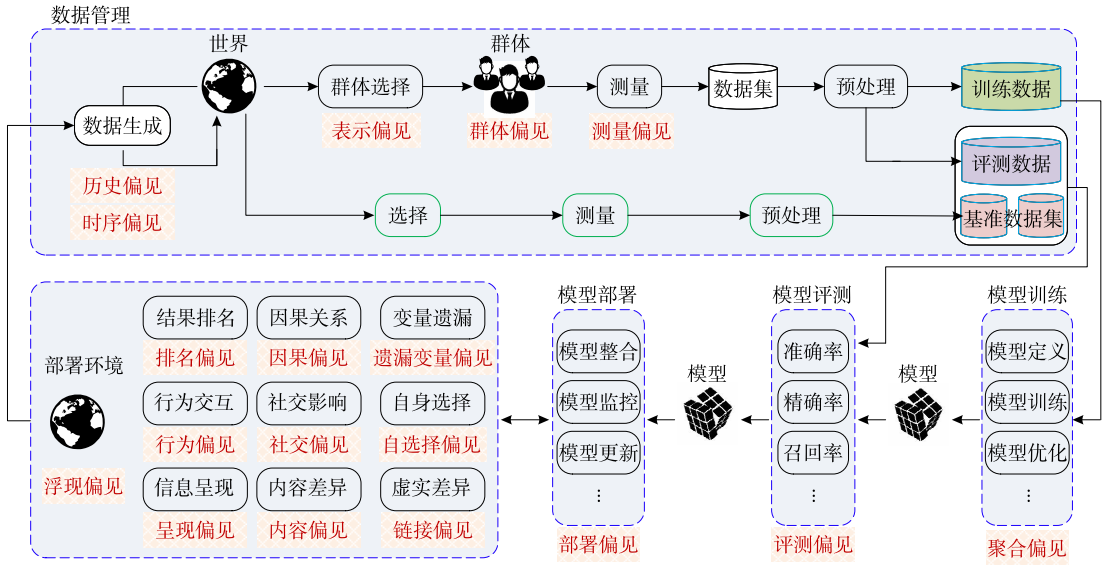


图 3　机器学习的生命周期及其偏见

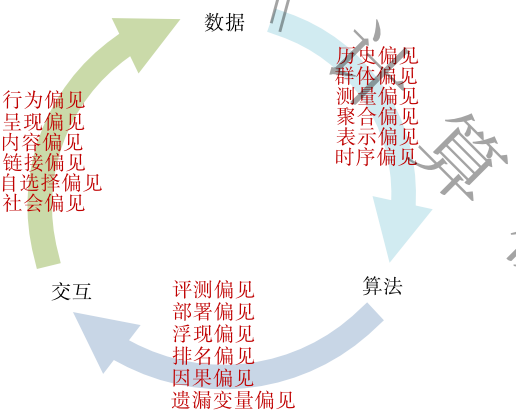


图 4　机器学习的主要偏见类型

3.1.1　数据类偏见

(1) 历史偏见. 社会、文化和习俗等方面的成见渗透到数据中产生的偏见，称为历史偏见(Historical Bias)[18]. 无论选择如何完美的采样和特征匹配技术，这种偏见都是难以避免的. 例如，化妆品、香烟广告集中投送不同的性别对象，存在性别方面的历史偏见.

(2) 群体偏见. 数据所表示群体的统计特征和属性特征与应用目标群体的不同，由此引起的偏见称为群体偏见(Population Bias)[56]. 例如，不同社交平台上群体性别的代表性存在差异[58-59]，女性较多使用 Facebook 和 Instagram 等，而男性则在 Reddit 或 Twitter 等论坛上更为活跃.

(3) 测量偏见. 在选择、收集或计算用于预测的特征和标签时，可用或能测量的数据往往是所感兴趣的特征和标签的有噪声代理，在选择了要测量的代理后，测量过程本身又增加了噪声，由此产生了测量偏见(Measurement Bias)[18]. 例如，逮捕率通常被用来代替犯罪率，可获得测量的逮捕率作为代理导致了测量偏见.

(4) 聚合偏见. 群体成员可能具有不同的背景和文化等，一个给定的变量对于不同群体中个体也可能意味着不同的东西，不同群体采用单一的通用模型，会产生聚合偏见(Aggregation Bias)[18]. 例如，糖尿病诊断和监测的糖化血红蛋白水平在不同性别和族群之间存在较大的差异，预测并发症的模型不可能适合所有人群.

(5) 表示偏见. 训练数据没有充分覆盖或代表所有预测空间，某些样本空间没有得到足够的表示，这种由代表性不足产生的偏见称为表示偏见(Representation Bias)[18]. 例如，ImageNet 包含约 1400 余万张图像，其中约 45% 来自美国，仅有 1% 和 2.1% 分别来自中国和印度. ImageNet 训练代表性不足国家的图像效果明显差于北美国家[60].

(6) 时序偏见. 不同时期的群体行为会有所漂移，不同时间点采集数据的场景可能有所不同，时间的粒度会影响观测的长期效应，数据会随时间而发生变化，这些因时间变化所导致的偏见，称为时序偏见(Temporal Bias)[56]. 例如，2013 年约 3 亿用户将 1 周内发布推文的 2.4% 删除[61]，用户关注特定话题的时间和时长受到当前热点的影响[62].

3.1.2　算法类偏见

(1) 评测偏见. 算法评测的测试数据或基准数据使用不当，导致不能完全代表目标群体，由此产生的偏见称为评测偏见(Evaluation Bias). 例如，人脸

识别算法用于性别判别和微笑检测中，对于深色皮肤的女性的表现明显较差.一方面是训练数据中的表示偏见造成的，另一方面是模型训练的基准数据并没有发现和纠正这一点[63].

(2)部署偏见.如果模型是为某一特定任务而建立的，而该任务并不是或不匹配部署后实际执行的任务，就不能保证获得所评测的良好性能，由此产生的偏见称为部署偏见(Deployment Bias)[18].例如，罪犯风险评估工具是设计用来对未来可能犯罪的风险进行预测评估的模型，通过该工具来确定罪犯的刑期，就会产生部署偏见.

(3)浮现偏见.模型在设计完成并运行一段时间后出现的偏见，称为浮现偏见(Emergent Bias)[8].浮现偏见一般在系统部署的后续应用中出现.例如，自动航班预订系统初始是为提供国内航线的国内航空公司所设计，该航班预订系统在选择国内航线的承运公司上会对国际航空公司产生系统性的不公平.

(4)排名偏见.推荐系统和信息检索会产生多个结果，这些结果具有相关性或重要性的优先次序，排名靠前的会吸引更多的用户关注并得到更多的用户点击，由此产生的偏见称为排名偏见(Ranking Bias)[18].例如，在美食推荐应用中，美团外卖会对推荐给用户的商家进行排名，用户选择排名靠前的商家的可能性就很大.

(5)因果偏见.将关联关系误认为因果关系所导致的偏见，称为因果偏见(Cause-effect Bias)[6].例如，在学校接受辅导的同学的考试成绩比没有接受辅导的同学的考试成绩差，接受辅导并不是考试成绩差的原因.再如，客户忠诚度测试中客户在某电商平台的消费成倍多于其他客户，这未必成功，因为客户可能在其他地方也有成倍的消费.

(6)遗漏变量偏见.如果模型设计中遗漏了一个或多个重要的变量，模型预测就会产生偏见，并称之为遗漏变量偏见(Omitted Variable Bias)[6].例如，市场分析预测模型突然发现大量的用户取消了某项服务的订阅，用户取消订阅的原因是，新的竞争对手提供价格减半的同样服务，预测模型并没有考虑竞争者，竞争者是一个被遗漏的变量.

3.1.3 人机交互类偏见

(1)行为偏见.不同平台和应用场景下个体或群体的行为以及不同个体或群体之间交互的行为，所产生的偏见称为行为偏见(Behavioral Bias)[56].例如，具有社会联系用户的交互明显高于没有联系的用户，20%具有社会联系的用户占取了80%交互量.Twitter的查询侧重于即时信息和人物，网页查询则侧重于用户对所关注主题的学习和了解[56].

(2)社交偏见.某些个体或群体的行为或判断影响到其他个体或群体的行为或判断，由此所产生的偏见称为社交偏见(Social Bias)[6,56].例如，Twitter上最受欢迎的0.05%的用户吸引了近50%的参与者[64]，也就是说，少数的Twitter用户影响了约有一半的其他用户.Facebook 2009年将近4万名活跃用户中7%的用户发布了50%的帖子[56].

(3)自选择偏见.研究对象或受试者自身进行选择所带来的偏见，称为自选择偏见(Self-selection Bias)[6].例如，关于成功企业家行为的问卷调查研究中，99%的答卷并不是出自成功企业家.再如，一个知识库在线产品，通过对是否阅读过其中的资料的不同人群进行比较来了解产品的效果，阅读过资料的人群比不阅读的人群活跃50%.

(4)呈现偏见.信息呈现给用户的方式影响着交互的效果，由此产生的偏见称为呈现偏见(Presentation Bias)[56].例如，Web用户只能点击所看到的内容，而其没有看到的内容就不会被点击，此外，图片附近的内容更有可能被点击.在视频流媒体服务中，用户仅仅浏览数百条推荐信息，这些都会带来呈现偏见.

(5)内容偏见.生成内容的结构、词法、语法和语义的差异所引发的偏见，称为内容偏见(Content Bias)[56].例如，不同国家、甚至不同地区的群体使用的语言各有不同.公众和专家生成的内容有别于常规用户生成的内容，Twitter的专家用户专注于生成专家主题的内容.这些都会对用户分类、热点分析、内容过滤等带来内容偏见.

(6)链接偏见.从用户的交互及活动构建出的社会网络特征与真实特征之间存在一定的差异，由此带来的偏见称为链接偏见(Linking Bias)[56].用户之间的交互或联系方式会随时间或场景的变化而不同.例如，个体或群体的地理位置与在线社会网络的特征有关[65-66]，线下社会关系影响着用户创建在线社会连接和在线交互.

偏见必然导致不公平，厘清机器学习中可能出现的各种偏见，有助于不公平的发现.尽管机器学习中的偏见可粗略分为上述三类，但这种分类是不够

《计算机学报》

严谨的，事实上，这些偏见之间还难以找到严格的界限. 除上述介绍的主要偏见外，研究人员还提出了机器学习中可能出现的其他偏见[6,8,18,55-56].

3.2　不公平的发现

发现机器学习的不公平是纠正偏见和消除歧视的前提. 歧视类型和偏见类别的概念定义为发现不公平提供了不同视角下的可能技术路径. 机器学习中不公平发现的主要技术包括关联规则挖掘、k 最邻近分类、概率因果网络、隐私攻击和基于深度学习的方法等. 表 4 对不公平发现技术进行了对比.

表 4　不公平发现技术的对比

技术方法	基本思想	优点/用途	缺点/不足
关联规则挖掘方法	采用关联规则挖掘技术抽取隐藏在历史决策记录数据中的决策规则，依据潜在歧视项引起的置信度增益揭示该决策规则是否存在歧视.	(1) 克服了传统统计分析方法对大量历史数据搜索的能力局限 (2) 详尽发现隐藏在历史决策中分类规则的歧视 (3) 可发现直接性/间接性歧视	(1) 不能用于发现因果关系导致的不公平、回归问题中的不公平以及隐私保护数集的不公平等
k 最邻近分类方法	借助相似性度量函数在 k 最邻近范围内搜索具备相似特征的测试者，并使用给定决策进行测试，根据结果揭示给定决策是否存在歧视.	(1) 克服了关联规则挖掘依赖于规范属性以及局部关联规则缺乏整体性等不足 (2) 能够实施具有区间取值属性相关的歧视发现	(1) 相似性度量考虑了所有属性，难以区分具体属性对歧视的影响 (2) 不能用于隐私保护数据集的不公平发现
概率因果网络方法	基于概率因果理论和有向无环图，综合考虑并清晰刻画各种属性之间的关系及其对决策的影响.	(1) 能够完善关联规则挖掘方法、k 最邻近分类方法中存在的因果关系刻画、在等方面的不足 (2) 具备发现直接性歧视、间接性歧视、可解释性歧视、个体不公平、群体不公平等的能力	(1) 实施效果有赖于方法中定理所假设条件的满足 (2) 不能发现回归问题的不公平、隐私保护数据中的不公平
隐私攻击方法	利用最小攻击、推理控制算法等得到受保护/不受保护属性等内容，进一步基于背景知识、借助于费雷歇边界定理发现歧视.	(1) 为隐私保护数据的歧视发现探索了可行的技术途径 (2) 具备间接性歧视发现、隐私保护数据的歧视发现等功能	(1) 难以发现可解释歧视，不能发现回归问题的不公平

3.2.1　关联规则挖掘方法

关联规则挖掘发现歧视(不公平)的基本思想在于[67]：历史决策记录中决策规则可视为历史决策记录数据的分类规则，该规则具有与之对应的置信度，置信度表示了在给定前提(前件)下得出决策(后件)的概率. 决策规则中使用的事实(项)包含有(潜在)歧视项和(潜在)非歧视项，前者表示了法律条规、政策文件和社会习俗等限定的受保护属性，后者表示了决策场景相关的特征. 通过采用关联规则挖掘技术，提取历史决策记录数据中特定形式的分类规则(频繁项集)，就可以获得隐藏在数据集中的决策规则. 依据所抽取的决策规则的前件中潜在歧视项所引起的置信度增益来揭示该决策规则的歧视与否(参见图 5).

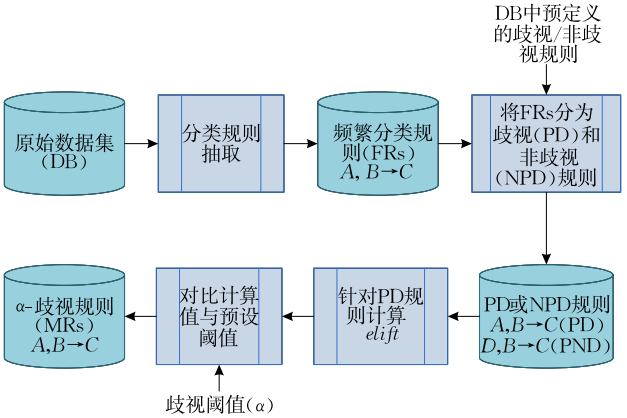


图 5　关联规则挖掘发现歧视

Pedreschi 等人首先将歧视引入到分类规则[68-69]，定义了潜在歧视规则：对于数据集 DB、歧视项集 A、非歧视项集 B、分类项 C、分类规则 $(A,B\rightarrow C)$，如果 A 非空，则该分类规则称为潜在歧视规则，否则，称为非潜在歧视规则.

潜在歧视规则和歧视行为具有不同的含义，为了度量直接性歧视行为，Pedreschi 等人定义了(直接性)α 歧视规则[68-69]：对于潜在歧视规则 $(A,B\rightarrow C)$ 和非潜在歧视规则 $(B\rightarrow C)$，如果 $elift(A,B\rightarrow C)=Conf(A,B\rightarrow C)/Conf(B\rightarrow C)\geq\alpha$($\alpha\geq 0$ 为预设的阈值)，则称潜在歧视规则 $(A,B\rightarrow C)$ 是(直接性)α 歧视的，或者是(直接性)α 歧视规则，其中，$Conf(A,B\rightarrow C)$ 和 $Conf(B\rightarrow C)$ 分别是规则 $(A,B\rightarrow C)$ 和规则 $(B\rightarrow C)$ 的置信度. 如果 $elift(A,B\rightarrow C)<\alpha$，则称规则 $(A,B\rightarrow C)$ 是 α 防护的，或者 α 防护规则.

为了度量间接性歧视行为，Pedreschi 等人定义了(间接性)α 歧视规则[68-69]：对于非潜在歧视规则 $(D,B\rightarrow C)$ 和 $(B\rightarrow C)$、潜在歧视规则 $(A,B\rightarrow C)$、歧视项集 A、$Conf(A,B\rightarrow D)\geq\beta_1$、$Conf(D,B\rightarrow A)\geq\beta_2>0$，如果 $elb(Conf(D,B\rightarrow C),Conf(B\rightarrow C))\geq\alpha$，则称潜在歧视规则 $(A,B\rightarrow C)$ 是间接性 α 歧视的，或者间接性 α 歧视规则，并称非潜在歧视规则 $(D,B\rightarrow C)$ 是红线(Redlining)规则. 红线规则 $(D,B\rightarrow C)$ 以及背景知识规则 $(A,B\rightarrow D)$ 和 $(D,B\rightarrow A)$ 导致了间接性 α 歧视规则 $(A,B\rightarrow C)$. 其中，$\alpha\geq 0$ 为

预设的阈值，$\gamma=Conf(D,B\rightarrow C)$，$\delta=Conf(D,B\rightarrow C)$，$f(x)=\beta_1/\beta_2=(\beta_2+x-1)$，$elb(x,y)=f(x)=\begin{cases}f(x)/y, & f(x)>0\\0, & \text{其他}\end{cases}$.

如果非潜在歧视规则$(D,B\rightarrow C)$以及背景知识规则$(A,B\rightarrow D)$和$(D,B\rightarrow A)$不能导致任何 α 歧视规则$(A,B\rightarrow C)$，则称非潜在歧视规则$(D,B\rightarrow C)$为非红线规则. Pedreschi 等人还讨论了不同歧视度量标准 $slift$、$glift$、$clift$、$olift$ 下分类规则的歧视发现问题[70]. 此后，为了提高歧视发现的效率，Pedreschi 等人提出在各种歧视度量标准下对排列前 k 个规则(Top-k)进行分析来发现歧视[71].

关联规则挖掘克服了传统统计分析方法对大量历史数据搜索的能力局限[72]，能够更详尽地发现隐藏在历史决策记录中分类规则的歧视. 但是，目前还不能用于发现因果关系导致的不公平、回归问题中的不公平以及隐私保护数据集的不公平等.

3.2.2 k 最邻近分类方法

Luong 等人基于司法领域的情景测试[73]给出了不公平发现的 k 最邻近(k-Nearest Neighbor，k-NN)分类方法[74]，其主要思想在于：在给定历史决策记录数据下，对于决策结果为否定的受保护组的每一成员，寻找具有合法相似特征的测试者(受保护组或不受保护组)，如果受保护组测试者和不受保护组测试者的决策结果明显不同，由此就可推断出该否定决策对受保护组具有偏见，即存在不公平性. 相似性通过距离函数来度量，在 k 最邻近范围内搜索相似特征的测试者，不公平性度量可选择 $diff(\boldsymbol{r})=p_1-p_2$、$slift(\boldsymbol{r})=p_1/p_2$ 或 $olift(\boldsymbol{r})=(p_1(1-p_2))/(p_2(1-p_1))=(ad)/(cb)$ 的其中之一(表达式中各符号的含义参见表 5).

表 5 k-NN 中的不公平性度量

群体情况	预测结果			
	Negative	Positive	总量	比值
受保护组	a	b	n_1	$p_1=a/n_1$
不受保护组	c	d	n_2	$p_2=c/n_2$
总量	m_1	m_2	n	$p=m_1/n$

Luong 等人定义了距离函数：对于具有 n 元属性的多元组 $\boldsymbol{r}$ 和 $\boldsymbol{s}$，二者之间的相似性距离为 $d(\boldsymbol{r},\boldsymbol{s})=\sum_{i=1}^{n}d_i(\boldsymbol{r}_i,\boldsymbol{s}_i)/n$. 其中，$d_i(\boldsymbol{r}_i,\boldsymbol{s}_i)$ 依据属性的取值类型(区间数值、规范型、排序编号)以不同方式计算[73-74]. 由此，对于一个多元组 $\boldsymbol{r}$ 和数据集 $R(R=PR\cup UR$，PR 是受保护组，UR 是非受保护组)，可以得出多元组中 i 个属性 $\boldsymbol{r}^i$ 的排序 $rank_R(\boldsymbol{r},\boldsymbol{r}^i)=|\{j|d(\boldsymbol{r},\boldsymbol{r}^i)<d(\boldsymbol{r},\boldsymbol{r}^i)\vee d(\boldsymbol{r},\boldsymbol{r}^i)=d(\boldsymbol{r},\boldsymbol{r}^i)\wedge j\leq i\}|$. 多元组 $\boldsymbol{r}$ 的 k 最邻近集为

$kset_R(\boldsymbol{r},k)=\{\boldsymbol{r}^i\in R|rank_R(\boldsymbol{r},\boldsymbol{r}^i)\leq k\}$ 或者

$kset_R(\boldsymbol{r},k,d)\{\boldsymbol{r}^i\in R|rank_R(\boldsymbol{r},\boldsymbol{r}^i)\leq k\wedge d(\boldsymbol{r},\boldsymbol{r}^i)\leq d\}$

并定义

$$p_1=|\{\boldsymbol{r}'\in kset_{PR\setminus\{\boldsymbol{r}\}}(\boldsymbol{r},k)|dec(\boldsymbol{r}')=dec(\boldsymbol{r})\}|/k,$$

$$p_2=|\{\boldsymbol{r}'\in kset_{UR}(\boldsymbol{r},k)|dec(\boldsymbol{r}')=dec(\boldsymbol{r})\}|/k,$$

$$diff(\boldsymbol{r})=p_1-p_2,$$

其中，$dec(\boldsymbol{r})$ 和 $dec(\boldsymbol{r}')$ 分别表示 $\boldsymbol{r}$ 和 $\boldsymbol{r}'$ 的决策. 对于 $\boldsymbol{r}\in PR$ 和阈值 $t\in[0,1]$，如果 $dec(\boldsymbol{r})$ 为阴性(负类)且 $diff(\boldsymbol{r})\geq t$，则称 $\boldsymbol{r}$ 是 t 歧视的.

Romei 等人将 k 最邻近分类方法应用于科技项目申报评审过程中的不公平发现[75].

k 最邻近分类克服了关联规则挖掘依赖于规范属性以及局部关联规则的整体性缺乏的不足，并能实施具有区间取值属性相关的歧视发现，但是，该方法中相似性的距离度量考虑了所有属性，难以区分具体属性对歧视的影响，也不能用于隐私保护数据集的不公平发现.

3.2.3 概率因果网络方法

Mancuhan 和 Clifton 提出了以贝叶斯网络作为决策过程模型来发现直接性歧视和/或间接性歧视的贝叶斯估计方法[76]，其基本思想在于：关联数据挖掘中规则的置信度 $Conf(A,B\rightarrow C)$ 可视为条件概率 $P(C|A,B)$ 的估计，贝叶斯网络是实现概率 $P(A,B,C)$ 和条件概率 $P(C|A,B)$ 估计的有效技术. 在贝叶斯网络方法中，关联规则挖掘中的 $elift$ 扩展为 $belift$(Bayesian $elift$)：对于受保护属性集 A、非受保护属性集 B 和关联受保护属性集 R，$belift=P(C|A,B,R)/P(C|B)$，满足 $P(C|A,B,R)>t>P(C|B)$，其中 t 是二值决策选取的临界值.

Zhang 等人通过贝叶斯网络中属性与决策的因果关系定义相似性的距离度量[34,77]，对 k 最邻近分类方法中相似特征个体的搜索进行了改进.

Bonchi 等人定义了萨普斯-贝叶斯关系网络(Suppes-Bayes Causal Network，SBCN)用以刻画数据中各种类型歧视的因果关系，并给出了基于 SBCN 模型发现歧视的随机游走方法[78]. Choi 等人给出了朴素贝叶斯分类器中歧视模式度量的上下界以及歧视模式的出现概率，进而提出了朴素贝叶斯分类器中不满足 δ-公平的歧视模式发现的分支定界方法[79].

Wu 等人将排序数据集中的排序位置映射到连续的评分，建立了属性(离散变量)和评分(连续变量)的因果网络模型，给出了条件高斯分布评分下路

径影响度量和歧视度量的理论结果，并应用于排序数据的歧视发现[80].

概率因果网络方法基于概率因果理论和有向无环图，可以综合考虑并清晰刻画各种属性之间的关系及其对决策的影响，从而发现直接性歧视、间接性歧视、可解释性歧视、个体不公平、群体不公平. 但是，此类方法的效果有赖于方法中定理所假设条件的满足.

3.2.4　隐私攻击方法

受隐私攻击和不公平/歧视发现的相似性的启发，Ruggieri 等人给出了间接性歧视发现、隐私保护数据的歧视发现以及歧视数据恢复等隐私攻击方法[81]. 隐私攻击发现间接性歧视的基本思想在于：对于不含有受保护属性的数据，基于弗雷歇边界(Frechet Bounds)定理[82]，利用背景知识(如属性的相关性)，从不含有受保护属性的数据中获取 $diff(c)$ 的下界，并由此来判定是否存在间接性歧视.

隐私攻击发现隐私保护数据中歧视的基本思想在于：数据中含有受保护属性，但是数据经过隐私保护方法进行过加工(干扰了受保护属性)，利用背景知识(如受保护属性)和桶技术[83]将数据集进行受保护属性和不受保护属性的分组，对分组后的数据和背景知识采用弗雷歇边界定理，获得 a、p、p_1 和 p_2 的边界，进而基于 $diff(\boldsymbol{r})=p_1-p_2$ 的下界来发现歧视(a、p、p_1 和 p_2 的含义参见表 5).

隐私攻击实现隐私保护数据恢复的基本思想在于：数据经过隐私保护方法进行过加工隐藏了歧视性决策，最小攻击[84]是对最优匿名隐私数据恢复的有效策略，对于数据的非保护属性、隐私策略和匿名算法已知的情况下，利用背景知识(如隐藏歧视的数量)和推理控制算法重构数据集，进而采用常规方法发现歧视.

隐私攻击方法将数据隐私保护的方法和算法应用于歧视发现，为隐私保护数据的歧视发现探索了可行的技术途径. 但是，该方法难以发现可解释歧视，也不能发现回归中的不公平问题.

3.2.5　深度学习方法

通过将个体公平性测试生成问题表述为深度强化学习问题，并将被测试的机器学习模型(Model Under Test，MUT)视为强化学习环境的一部分，Xie 等人提出了针对机器学习模型的黑盒公平性测试技术[85]. 如图 6 所示，强化学习代理(Agent)通过对环境采取行动生成针对 MUT 的输入，然后通过观察环境状态并获得来自环境的奖励. 通过这种交互式迭代，代理学习到一种在无需访问 MUT 内部动态的情况下(即黑盒)便可高效生成个体歧视性输入的最优策略. 训练完成后的深度强化学习模型可以有效地探索和利用输入空间，并在短时间内检测到更多的个体歧视性输入.

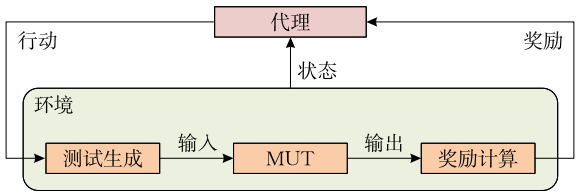


图 6　针对黑盒公平性测试的强化学习框架

从机器学习全生命周期中的不同阶段出发，研究人员对偏见/不公平的类型进行了系统分析. 同时，为应对不同场景需求，提出了多种类型的不公平发现技术，这些研究为公平机器学习模型和算法设计提供了良好的支持.

4　公平机器学习的设计

为了开发公平机器学习系统或者确保机器学习系统的公平性，人们建立了一系列公平机器学习的设计方法. 这些设计方法可以从三个维度来粗略划分：(1)面向特定的机器学习任务，如自然语言处理、人脸识别、推荐系统、分类问题、回归问题、聚类问题等；(2)针对专门的机器学习技术或算法，如深度学习、强化学习、决策树学习、集成学习、表示学习、对抗学习等；(3)依据机器学习的生命周期，分为预处理、中间处理和后处理[86-87]. 下面从机器学习的生命周期维度，介绍和讨论公平机器学习的设计.

4.1　预处理

预处理也称为训练数据预处理，通过发现训练数据中的偏见或歧视，并对数据进行预先修改或重新表示，以消除训练数据中的不公平. 预处理的方法大致可分为修改训练数据和公平表示学习两类.

4.1.1　修改训练数据方法

Kamiran 和 Calders 给出了预处理的数据篡改、数据加权和数据采样方法[88]，以保证训练数据对敏感属性群体的决策具有统计公平性. 在人口普查收入数据集(Census Income Dataset)上的实验表明，在保证高准确度的前提下，以上方法可将歧视率从 17.93%降为 0.11%. 但该研究仅针对二值属性及二分类问题.

数据篡改、数据加权和数据采样方法[89-91]只能依据单一度量标准，在原始数据中发现某一歧视项的歧视，但是歧视性行为可能由多个歧视项或者其组合所引发，这就不能保证处理后的数据集是真正

无歧视的.

Hajian 等人给出了规则防护和规则泛化的数据变换方法[86,92-93]. 该方法将数据集 DB 中的频繁模式规则 FR(Frequent Classification Rules)划分为给定歧视项下的潜在歧视规则集 PD(Potentially Discriminatory)和潜在非歧视规则集 PND(Potentially Nondiscriminatory),并分别基于直接性歧视度量 $elift$ 和间接性歧视度量 elb 获得 PD 中的 α 歧视规则集合 MR 和 PND 中的红线规则及引发的歧视规则的集合 RR.

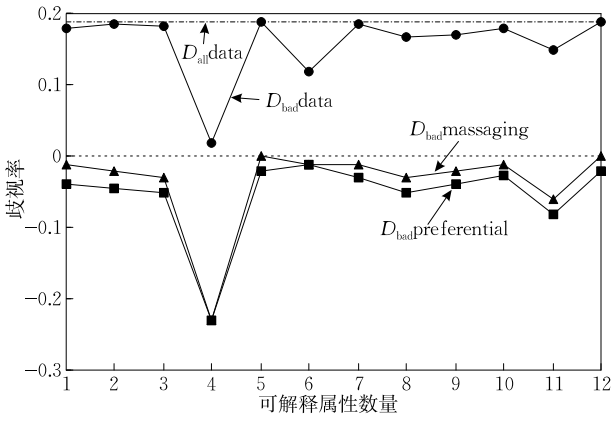
直接性歧视规则的规则防护数据变换方法的基本思想在于:(1)将规则$(\neg A,B\rightarrow\neg C)$的支撑数据集 $DB_C\subseteq DB$ 中记录的 $\neg C$ 变换为 C,实现规则$(\neg A,B\rightarrow\neg C)$到规则$(\neg A,B\rightarrow C)$的变换(更改数据记录中的分类项);或者(2)将支撑数据集 DB_C 中记录中的 $\neg A$ 变换为 A,实现规则$(\neg A,B\rightarrow\neg C)$到规则$(A,B\rightarrow\neg C)$的变换(更改数据记录中的歧视项集). 进而,使得潜在歧视规则$(A,B\rightarrow C)$满足 $elift(A,B\rightarrow C)<\alpha$,规则集合 MR 中的 α 歧视规则就变换为 α 防护规则(其中,$\neg X$ 表示项集 X 的非,是指与 X 具有相同属性,但属性取值为 X 属性取值之外的其它可能取值的项集). 对于 MR 中的直接性歧视规则$(A,B\rightarrow C)$,如果存在至少一个非红线规则$(D,B\rightarrow C)$且满足 $Conf(A,B\rightarrow D)\geq p$($p$ 为 1 或接近于 1 的数值),则可以使用规则泛化数据变换方法来实施直接性歧视规则的预处理.

直接性歧视规则的规则泛化数据变换方法的基本思想在于:通过将规则$(A,B,\neg D\rightarrow C)$的支撑数据集 $DB_C\subseteq DB$ 中记录的 C 变换为 $\neg C$,使得潜在歧视规则$(A,B\rightarrow C)$是非红线规则$(D,B\rightarrow C)$的一个实例[86],从而使规则集合 MR 中的 α 歧视规则引发的直接性歧视得到防护. 规则泛化数据变换方法并不能处理 MR 中的所有直接性歧视规则,直接性歧视的预处理需要将规则泛化数据变换方法和规则防护数据变换方法结合使用:MR 中满足规则泛化数据变换方法的直接性歧视规则使用规则泛化数据变换进行预处理,MR 中其它不满足规则泛化数据变换方法的直接性歧视规则使用规则防护数据变换进行预处理.

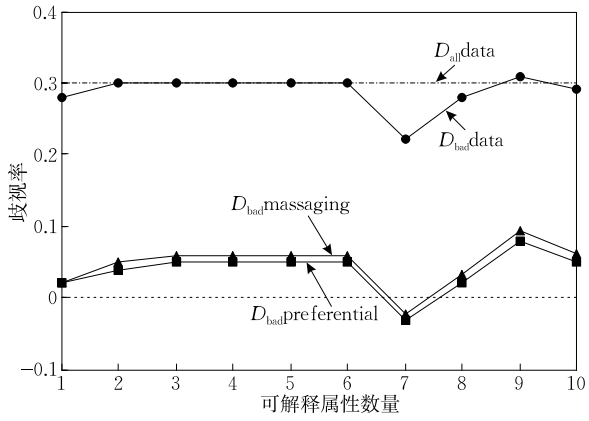
对于 RR 集中的红线规则$(D,B\rightarrow C)$及其间接性 α 歧视规则$(A,B\rightarrow C)$,间接规则防护数据变换方法进行预处理的基本思想在于:(1)通过将规则$(\neg A,B,\neg D\rightarrow\neg C)$的支撑数据集 $DB_C\subseteq DB$ 中记录的$\neg C$ 变换为 C,实现规则$(\neg A,B,\neg D\rightarrow\neg C)$到规则$(\neg A,B,\neg D\rightarrow C)$的变换(更改数据记录中的分类);或者(2)将支撑数据集 $DB_C\subseteq DB$ 中记录的$\neg A$ 变换为 A,实现规则$(\neg A,B,\neg D\rightarrow\neg C)$到规则$(A,B,\neg D\rightarrow\neg C)$的变换(更改数据记录中的歧视项集). 进而,使得红线规则$(D,B\rightarrow C)$满足 $elb(Conf(D,B\rightarrow C),Conf(B\rightarrow C))<\alpha$,从而使规则集合 RR 中间接 α 歧视规则变换为间接 α 防护规则.

规则防护和规则泛化的数据变换方法能够对含有直接性歧视和间接性歧视的数据进行预处理,但是要求数据集满足规定的条件,同时,在预处理实施中需要进行额外的频繁模式挖掘和规则分类计算.

Zliobaitė 等人对 Kamiran 和 Calders 的数据篡改和数据采样方法[88]进行了改进和完善,给出了局部篡改和局部优先采样的预处理方法[94],该方法仅仅剔除不可解释歧视相关的数据,保留了训练数据中可解释歧视相关的部分,能够克服歧视的过度消除. 在 Adult 和 Dutch Census 数据集上的实验表明,以上方法能够显著降低歧视率,图 7、图 8 所示分别为全局、局部预处理技术的歧视消除效果,其中

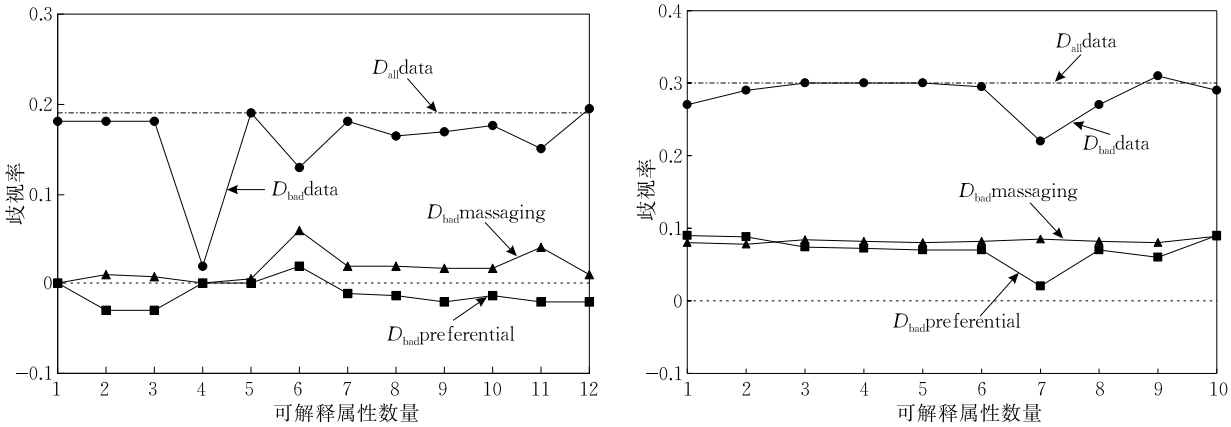


(a) Adult数据集



(b) Dutch Census数据集

图 7 全局预处理方法的歧视率



(a) Adult数据集　　(b) Dutch Census数据集

图 8　局部预处理方法的歧视率

D_{all}data 表示数据集中存在的歧视率，D_{bad} data 表示不可解释属性导致的歧视率，D_{bad} massaging 表示采用篡改技术处理后的歧视率，D_{bad} preferential 表示采用采样技术处理后的歧视率.

Luong 等人提出了将训练数据集中 t 歧视的 $\boldsymbol{r}$ 决策 $dec(\boldsymbol{r})$ 从负类改为正类的预处理方法[74]. Feldman 等人给出了不公平性和受保护属性泄露之间的对应关系，并提出了通过更改数据中的非受保护属性，使得数据集中的受保护属性不能得到预测的预处理方法[95].

Jiang 和 Nachum 假定数据集被无偏见的真实(Ground Truth)标记函数所标记，数据获取智能体的偏见导致数据产生了观测偏见，由此提出了对数据进行加权以消除不公平的预处理方法，并给出了经过加权预处理的数据能够确保分类器不会产生歧视的理论证明[96].

Calmon 等人给出了通过数据概率变换实施数据预处理来减轻歧视概率的优化模型[97]，该模型在概率分布上定义歧视和效用、在采样的基础上控制数据失真，并限制个体数据变换影响来确保个体公平性，从而使得数据预处理中的歧视控制、数据效用和个体数据失真得到折衷平衡.

这些方法一定程度上克服或减轻了训练数据中的偏见或歧视，但是只能对原始训练数据进行预处理加工，不具有对未知数据进行加工的泛化能力. 通过修改训练数据克服或减轻偏见的方法仅从统计公平角度出发解决问题，往往只能针对单一敏感属性.

4.1.2　公平表示学习方法

Zemel 等人首先提出了数据预处理的表示学习方法[98]：将原始数据聚类为 K 个公平聚类空间(Proto-types)，通过个体数据到聚类空间的概率分布的映射，实现数据属性信息的混淆，并尽可能多地编码原始数据的信息，以使得从 K 个公平聚类中难以获取个体受保护属性的信息，从而获得既符合群体公平又符合个体公平的数据表示.

Louizos 等人给出了公平变分自编码器模型[99]，该模型由解码器 $p_\theta(x|z,s)$ 和编码器 $q_\phi(z|x,s)$ 组成，将敏感属性变量 s 和隐变量 z 通过 $p(s)p(z)$ 尽可能分离(隐变量 z 不含有敏感属性信息). 对于隐变量仍存在敏感属性信息的训练数据，他们引入了最大均值差(Maximum Mean Discrepancy，MMD)作为正则项以保持边际后验 $q(z|s=k)$ 和 $q(z|s=k')$ 的一致性.

Sattigeri 等人基于生成对抗网络(Generative Adversarial Network，GAN)构造了用于提升统计公平和机会均等的公平生成对抗网络 Fairness GAN[100]. Fairness GAN 的目标是利用真实数据集构造无偏数据集，并保证两者在样本特征与决策结果间的联合分布(以受保护属性为条件)相近，但无偏数据集能够保证公平性. 图像数据集 CelebA 以及 Soccer 的测试结果表明：Fairness GAN 在构建无偏数据集、保证公平性方面具有良好性能. 但该方法仅能用于二分类任务，无法适用于更复杂任务.

Edwards 和 Storkey 提出了基于对抗的公平表示学习方法[101]，实现原始数据 X(敏感属性为 S、标签为 Y)转换为符合统计公平性的隐式表示 Z，其概念模型由两个深度神经网络(Deep Neural Networks，DNN)组成：一个神经网络生成符合统计公平性的表示 Z，另一个对抗神经网络预测表示中敏感属性 S 以评判生成表示的效果.

Xie 等人提出了基于三方极大极小博弈的不变特征表示学习[102]：编码器将训练数据映射至特征空间，区分器识别希望剔除的特征，预测器借助不变特征对模型进行评估. 在该方法中，可将敏感属性视

作为干扰变量,通过尽量降低或剔除它们对不变特征的影响,来实现数据集的公平表示.

Madras 等人讨论了考虑统计公平、机会均等以及均衡几率等公平性指标下的对抗学习公平表示[103],所给出方法可以生成迁移到新的任务的公平表示,即适用于公平迁移学习.

Oneto 等人提出了通过对基于低秩矩阵分解多任务学习的表示矩阵施加公平性约束,获得满足统计公平性的数据表示的多任务学习方法[104],该公平表示可用于不同任务的多种模型训练,能够对新的任务提供良好的泛化. Tan 等人将统计公平、机会均等、均衡几率等公平性描述为充分降维(Sufficient Dimension Reduction)问题,给出了适用于核模型的公平表示学习方法[105],该方法可以获得数据的公平高斯过程表示.

Gong 等人给出了一种消偏对抗网络[106],该网络由一个身份识别分类器和三个人类特征(性别、年龄和种族)分类器,每一个任务分类器对抗监督其它任务,联合学习出身份以及性别、年龄和种族等特征的无偏见表示.

Lahoti 等人给出了用于排序推荐任务的个体公平表示学习方法 ifair[107]:对数据进行聚类,将个体公平定义为聚类后数据的距离函数,并作为正则项. 该方法基于相似个体具有相似排序的策略,可处理多敏感属性的综合效果,也无需预知敏感属性.

公平表示方法无需改变已有的训练数据,能够保持数据的完整性,可以基于已发布的数据,也可以集成到数据发布中,使用起来比较灵活方便. 但是,对于隐私保护数据、保护属性未知的数据还难以实施. 此外,受偏见、歧视或不公平的多样性影响,此类方法并不能确保经过预处理后的数据中无歧视或者完全公平,对机器学习的精准度的影响程度也难以估计.

4.2 中间处理

中间处理也称为学习模型和算法的处理,是对学习模型或算法的调整、修正和完善(如改变目标函数或施加约束等). 中间处理无需对训练数据进行任何加工,但能消除因训练数据导致的不公平,同时保持模型训练过程不存在不公平. 根据任务不同,公平机器学习大体分为公平分类和公平回归;根据应用场景不同,分为自然语言处理、人脸识别、推荐系统等场景下的公平机器学习. 下面将基于以上分类对公平机器学习的中间处理方法进行讨论和介绍,其中公平分类和公平回归介绍一般意义下的方法和技术.

4.2.1 公平分类

Kamiran 等人通过拆分准则和剪枝策略将非歧视/公平约束深度嵌入决策树,给出了公平分类的决策树学习方法[108]. 该方法在对树节点进行拆分时,不仅要评估拆分对精准率的贡献,还要评估拆分对歧视/不公平性所带来的影响;在叶节点重标记中,标签不仅取决于这个节点训练集的元组的多数类,而且还要满足较少损失精准率下降低歧视. Raff 等人对 CART 决策树(Classification and Regression Tree)的 Gini 纯度进行了修正,定义了适应于多值类别或连续数值两种形式属性的公平信息增益,给出了公平随机森林的构建方法[109],该方法只需对 CART 决策树中的 Gini 纯度和信息增益的计算进行调整,可以方便地实施公平分类或回归.

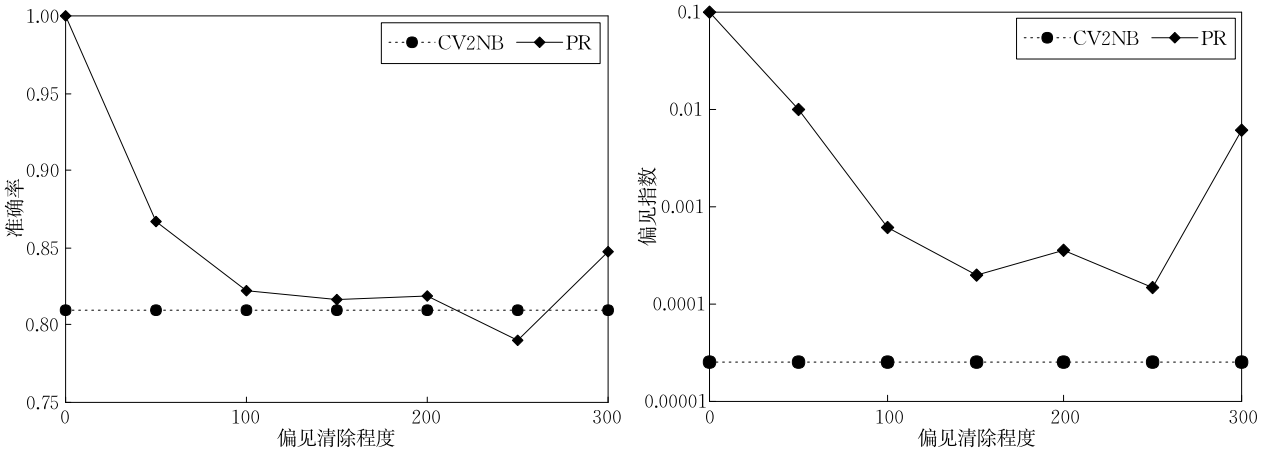
Calders 和 Verwer 给出了朴素贝叶斯分类器的三种改进方法使其适用于公平分类[110]:(1)将贝叶斯模型中关于敏感属性 S、类型变量 C 和其它属性 $A_1,\cdots,A_n$ 的联合概率分布 $P(C,S,A_1,\cdots,A_n)=P(C)\ P(S|C)\ P(A_1|C)\cdots P(A_n|C)$ 中的 $P(S|C)$ 替换为 $P(C|S)$,并调整 $P(C|S)$ 直到新模型中的标签无歧视;(2)基于敏感属性划分训练数据,借助于划分后的训练数据集分别训练多个模型,并对训练出的模型进行平衡;(3)在贝叶斯模型中增加用以表示无偏标签的隐变量,并通过最大期望来优化模型参数.

Choi 等人将朴素贝叶斯学习描述为约束多项式规划问题[70],并通过对该多项式规划问题附加相应形式的 δ-公平约束,学习获取满足 δ-公平的朴素贝叶斯网络的极大似然参数,进而实现公平朴素贝叶斯分类(δ-公平方法). 在三个数据集上的实验表明,δ-公平方法的分类性能均优于 Calders 和 Verwer 的第二种改进方法 CV2NB(Calders and Verwer's 2-Naive-Bayes),对比结果如表 6 所示.

表 6 CV2NB 与 δ-公平方法间的分类准确率对比

分类方法	COMPAS	Adult	German
CV2NB	0.875	0.759	0.679
δ-公平方法	**0.879**	**0.827**	**0.696**

Kamishima 等人将不公平产生的原因总结为[111]:敏感属性影响、训练数据量不足、训练样本存在不公平采样或标签等三个方面,提出了通过互信息正则项来消除或降低敏感属性带来的不公平,并用于对数几率回归,以实现公平分类. 文中所提出的 PR 方法(Prejudice Remover)与 CV2NB[111] 间的性能对比如图 9 所示,其中横坐标表示偏见清除程



(a) 准确率　　(b) 偏见指数

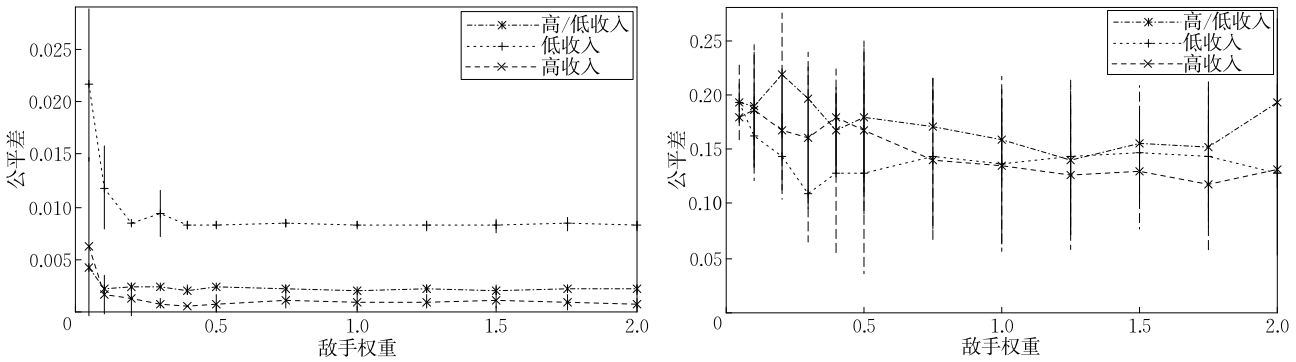
图 9　朴素贝叶斯分类器性能比较

度，(a)中纵坐标表示算法准确率，(b)中纵坐标表示算法偏见指数，取值越少表明算法越公平. PR 在算法准确率方面具有优势，而 CV2NB 在偏见消除方面优势明显.

Krasanakis 等人假定训练样本存在能够生成公平分类的标签数据，进而给出了一种自适应敏感加权机制 ASR(Adaptive Sensitive Reweighting)和权重估计模型，通过对原始数据和数据标签的同时训练，实现了对数几率回归在不平等对待和不平等影响下的公平分类[112]. 实验表明，通过避免不平等对待，ASR 在准确性和偏见间的权衡方面能够获得与协方差方法相似的表现；如果不避免不平等对待，而是在评估数据中提供有关敏感群体的信息，ASR 能够以较小的准确度损失，在权衡不平等对待和影响消除方面获得比协方差方法更好的表现.

Zafar 等人将不平等对待、不平等影响和机会均等分别描述为用户敏感属性和用户特征向量到分类器决策边界的符号距离的协方差，给出了这些公平性约束下的对数几率回归、(非)线性支持向量机模型，以实现公平边际(Fair Margin-based)分类[113]. Jiang 等人给出了含有 Wasserstein-1 距离惩罚项的对数几率回归[114]，以保证在统计公平意义下模型分类决策独立于敏感属性.

Beutel 等人给出了公平分类的多头(Multi-head)深度神经网络模型[115]：一个头用于分类并防止另一个头对敏感属性的预测. 该模型通过对抗学习剔除表示中的敏感信息，模型的训练和使用无需获得敏感属性或者数据的敏感属性标签，可用于隐私保护数据的公平学习分类. Adult 数据集下敏感属性分布对公平性的影响结果如图 10 所示，其中横坐标为与隐私保护相关的敌手权重，纵坐标为公平差(Parity Gap)，其定义为 $|TP-TN|$. 图 10(a)表明，当敏感属性服从均匀分布时，针对不同收入群体的公平差浮动较小，反之如图 10(b)所示浮动较大，两者相差一个数量级.



(a) 敏感属性服从均匀分布　　(b) 敏感属性不服从均匀分布

图 10　敏感属性分布对公平性的影响

《计算机学报》

Agarwal 等人给出了公平性(如统计公平、均衡几率等)的条件矩(Conditional Moments)线性不等式描述,将此类条件矩线性不等式约束下的二分类任务转换为代价敏感分类问题[116],在无需敏感属性信息下,实现公平学习分类.

Liu 和 Vicente 提出了不平等影响和不平等对待下精准率和公平性折衷的随机多目标优化框架[117],给出了求解 Pareto 前沿面(Fronts)的随机多梯度方法,可用于流式数据的公平机器学习.

Dong 等人研究了图神经网络(Graph Neural Networks,GNN)的公平性问题[118],并提出了基于排名的个体公平 GNN 框架 REDRESS(Ranking basEd inDividual faiRnESS).该框架能够实现端到端的训练,并且具备即插即用效果,可推广到任何 GNN 架构.

因果推理能够从分类特征中剔除嵌入的敏感信息,提升分类的公平性,使反事实生成具备对比相反敏感属性假设案例的能力.Kim 等人认为这种方法不能区分由干预(敏感变量)引起的信息和与干预相关的信息[119],因此通过将外部不确定性分解为独立于干预的变量及与干预相关但没有因果关系的变量来应对这一限制,并提出了解纠缠因果效应变分自动编码器(Disentangled Causal Effect Variational Autoencoder,DCEVAE),在没有完整因果图情况下估计总效应和反事实效应.通过添加公平性正则化,DCEVAE 能够生成反事实的公平数据集,同时保留了更多原始信息.

针对几率回归、支持向量机等不同机器学习中的公平分类问题,提出了多种解决方案,但此类研究大多仅适用于单一机器学习模型,且仅基于少数公平性指标进行了性能分析或验证.

4.2.2 公平回归

Berk 等人引入了适用于回归模型的一系列群体和个体公平性度量指标,将其以正则项的形式应用于线性或对数几率回归的损失函数,保持了原损失函数的凸性[120],尤其是,该策略可通过改变公平性正则项的权重,计算出精准率和公平性折衷的前沿面或完全 Pareto 曲线,从而实现高效优化计算的公平回归.

Agarwal 等人定义了连续变量类型受保护属性的统计公平和有界群体损失的公平性度量,从理论上证明了:对于任何 Lipschitz 连续的损失函数(如最小二乘回归、对数几率回归、分位数回归等),模型能够确保最优性和公平性.他们将公平回归描述为统计公平或有界群组损失的约束优化问题,该约束优化问题可以通过由最小经验风险、最小平方损失、代价敏感分类组成的监督学习 Oracles 中的一种方式来求解[121].

Oneto 等人给出了类别型和连续数值型敏感属性的 ϵ-general 公平性,以及一般公平最小经验风险优化模型,理论证明了 ϵ-general 公平性和最小经验风险一致需要满足的条件,并将其用于核方法,以实现类别型和连续数值型敏感属性的公平回归[122].

Fitzsimons 等人定义了期望群体公平[123],并基于此,将约束核回归用于公平决策树回归、公平随机森林回归、公平极端(Extra)树回归、公平提升(Boosted)树回归等模型.

Aghaei 等人给出了公平决策树设计的混合整数线性规划统一框架[124],该框架将不平等对待、不平等影响等公平性度量纳入优化模型,可以设计线性分支(分支的评测取决于多个特征的线性函数)和线性叶节点(叶节点的评分为多个特征的线性函数)的公平回归决策树以及具有线性叶节点的公平分类决策树.同时,也能够通过选择树结构(如树的高度)、分支规则类型(如单一或多特征变量评分)、叶节点的类型(如线性和常量)等来调整决策树的可解释性.从而,支持类别型和连续/有序型变量的公平、可解释决策树生成.

Zhao 和 Chen 讨论了公平多任务回归学习[125]:通过基于排序的非参数统计检验 Mann Whitney U 来衡量目标变量和受保护变量的相关性,将问题描述为群组的排序函数定义的非凸约束的非凸优化问题,给出了一个基于非凸交替方向乘子法的高效模型训练算法.该方法对训练数据的分布没有任何限制要求,能够将公平性作为优化的显式约束以严格控制多任务上的公平性.

4.2.3 公平自然语言处理

自然语言处理是机器理解和解释人类自然语言文本以及实现人机交互的重要技术.人类自然语言的发展和演化经历了漫长的时间,自然语言的形成也带有不同程度的性别、民族、地域和文化等特征.这些特征在一定场合下具有敏感性,不恰当的使用会带来偏见歧视.公平自然语言处理是不具有基于敏感属性的偏见歧视的自然语言处理.

Bolukbasi 等人通过职业相关词向量和词对类比讨论了词向量中的性别偏见[126],利用含有正则项的支持向量机模型对预选的性别专用词进行学习,获取 w2vNEWS 中的中性词,进而给出了对中性词进行硬消偏或软消偏的词向量性别偏见消除方法.该方法借助于奇异值分解进行特征分解,进而度

量单词之间的空间距离(关联)，方法简单但容易产生误差，对后续过程影响较大.

Zhao 等人给出了性别敏感属性的词向量学习方法 GN-GloVe(Gender-Neutral Global Vector)[127]，将词向量 $\boldsymbol{w}$ 表示为中性部分 $w^{(a)}$ 和性别部分(受保护属性)$w^{(g)}$，记为 $\boldsymbol{w}=[w^{(a)};w^{(g)}]$. $w^{(a)}$ 中的编码信息独立于 $w^{(g)}$，从而不受性别部分影响. 该方法能够在学习词向量的同时识别出中性词，无需单独的分类器来识别中性词，从而解决了错误传播问题. 此外，该方法可以进一步与其他词嵌入模型结合.

Brunet 等人剖析了词向量的偏见在训练过程中的产生机理[128]，指出对训练语料库施加扰动会影响词向量的偏见，追溯词向量偏见的原始训练文本可以识别出能够最大程度减少偏见的文本子集. 并给出了基于差分偏见、共现扰动和偏见梯度等概念的差分偏见的近似计算，用于 GloVe 词向量的性别偏见消除.

Kaneko 和 Bollegala 将词分为阳性词、阴性词、中性词和偏见词共四类，给出了消除词向量性别偏见的降噪自编码器方法[129]，该方法的性能优于硬消偏和 GN-GloVe 方法，而且可以对 GN-GloVe 的词向量实施进一步的性别偏见消除.

Shin 等人给出了词向量性别偏见消除的反事实方法[130]，该方法将 Siamese 自编码结构用于性别词对以实现语义和性别隐描述的解纠缠，使用梯度反转层禁止从语义信息中推断性别信息；将词向量通过反事实性别词向量变换为中性词向量，将词向量消偏导致的偏移通过核函数对齐保持在不损害词向量语义的方向上. 该方法在消除词向量性别偏见的同时能够很好地保持词向量的语义信息，并且具有比同类算法更好的性能表现.

Yang 和 Feng 给出了词向量关系中性别偏见消除的因果推理方法[131]，不同于传统通过减少带有性别偏见的词向量和性别方向之间的关系来减轻性别偏见的方法，该方法基于统计关联图和 Half-Sibling 回归，利用带有性别偏见的词向量和性别定义的词向量之间的统计依赖关系学习和抽取虚假性别信息，并将所学虚假性别信息从带有性别偏见的词向量中减去，以实现偏差消除，克服了在词向量的性别方向上消除偏差的局限.

Zhao 等人讨论了上下文相关词向量的性别偏见问题[132]：ELMo(Embeddings from Language Models)的训练数据集中男性实体明显多于女性实体；通过主分量分析发现上下文相关词向量 ELMo 系统性地编码了性别信息；上下文相关词向量 ELMo 对男性实体和女性实体的编码信息存在性别偏见；ELMo中存在的性别偏见会转移到基于 ELMo 实现的下游任务中. 针对以上问题，他们提出了两种解决方法：训练时数据增强技术和测试时嵌入中和技术，并证明了数据增强技术总体表现优于嵌入中和技术.

Basta 等人对上下文相关词向量和标准词向量的性别偏见进行了比较分析，前者具有更低程度的性别偏见[133]，尤其在性别空间、直接偏见、男/女聚类等方面表现突出，但在刻板印象词与隐性性别词组合等方面表现不够理想，甚至放大了性别偏见.

Liang 等人给出了句子表示的性别偏见消除 SENT-DEBIAS 方法[134]：定义具有偏见属性的词；将这些词语境化为带有偏见属性的句子，并通过 BERT(Bidirectional Encoder Representation from Transformers)或 ELMo 获得句子表示；估计句子表示的偏见子空间；通过去除偏见子空间上的投影来消除句子的偏见，即实现硬消偏.

Rudinger 等人借助于 Winograd 模式对基于规则的系统、特征驱动的统计学习系统和神经网络等三类共指系统中的代词消解进行实证分析[135]，指出了共指消解系统中存在不同程度的职业相关词的性别偏见.

Zhao 等人开发了共指消解系统中性别偏见测试的语料库 WinoBias[28]，将其应用于典型共指系统的分析测试，揭示出性别偏见产生的两方面原因：训练数据(OntoNotes 5.0)和辅助资源(词向量). 给出了采用交换性别实体构建附加训练语料库来消除偏见，通过平衡名词短语中男、女性别的出现频次以消除偏见的方法. 并进一步研究了 ELMo 用于共指消解系统中的性别偏见问题，给出了消除性别偏见的数据增强和词向量中性化方法[132].

Lu 等人定义了用于量化自然语言处理任务中性别偏见的通用基准，借助于实证研究指出了机器学习模型中普遍存在针对职业的性别偏见，进一步给出了共指消解中性别偏见消除的反事实数据增强方法 CDA(Counterfactual Data Augmentation)[136]：通过交换语料库中的性别词对、因果干预构造匹配的性别词对来增强语料库，并去除性别词和中性词之间的关联. 并指出，使用梯度下降法进行训练时，性别偏见会随着损失的减少而增加，表明此类优化方法扩大了偏见，但 CDA 能够缓解偏见的发生.

Prates 等人开展了机器翻译中的性别偏见研究[137]，构建了工作职位和性别专用句子的测试集，将英语作为目标语言，同时将 12 种性别中性语言

《计算机学报》

(没有明显的主体性别信息的语言,如匈牙利语、约鲁巴语、中文等)作为源语言,对测试集中的句子执行 Google 翻译,发现 Google 翻译结果表现出强烈的男性倾向,特别是在与不平衡性别分布或定型观念有关的领域,如 STEM(Science, Technology, Engineering, and Mathematics)领域.

Vanmassenhove 等人对性别信息对机器翻译的影响进行了研究,首先对平行语料库 Europarl 加注性别等相关信息,进一步对英语-西班牙语、英语-法语、英语-意大利语等 10 种语言对的翻译效果进行了分析比较,结果表明:增加性别信息可以明显提升某些机器翻译系统的互译质量[138],如表 7 所示,其中 BLEU(Bilingual Evaluation Understudy)评分越高说明翻译效果越好.

表 7 不同机器翻译系统的 BLEU 得分

机器翻译系统	未加注性别信息	加注性别信息
法语	37.82	**39.26**
西班牙语	42.47	42.28
希腊语	31.38	**31.54**
意大利语	31.46	**31.75**
葡萄牙语	36.11	**36.33**
丹麦语	36.69	**37.00**
德语	28.28	28.05
芬兰语	21.82	21.35
瑞典语	35.42	35.19
荷兰语	28.35	28.22

Font 和 Costa-jussà 借助于 Transformer,通过在编码器和/或解码器中使用预训练词嵌入,给出了机器翻译中性别偏见检测和评价的实验框架,以及通过词向量纠偏来消除机器翻译系统中性别偏见的方法[139],在包含 3000 个句子的 Newstest2013 测试集上的实验表明,该方法提高了英语-西班牙语机器翻译平台的 BLEU 评分.

Kiritchenko 和 Mohammad 精心挑选了 8640 个英文句子,并建立了一个涉及性别和种族偏见公平评价语料库(Equity Evaluation Corpus,EEC),并对 219 个情感分析系统进行了评估,发现其中超过 75%的系统存在明显的性别和种族偏见[140],且种族偏见比性别偏见更为普遍,不同情感维度的偏见程度也存在不同. Prabhakaran 等人[141]建立了无意偏见的扰动敏感分析方法,并将其应用于四种不同类型的情感和毒句分析,结果表明:实体名字扰动能够揭示机器学习模型中的偏见.

通过最小化词嵌入关联测试(Word Embeddings Association Test,WEAT)的得分,Popović 等人给出了词向量多种偏见(如性别、种族、宗教等)的联合消除方法[142]:HardWEAT 完全消除多种偏见;SoftWEAT 消除维持原始词关系下的偏见,并将该方法应用于影评文本的情感分析,结果表明偏见能够被部分甚至全部消除,词嵌入向量间有意义的联系也得以保持.

Bordia 和 Bowman 给出了一种能够消除性别偏见的词级语言模型的建立方法[143]:定义性别子空间,并将词向量在性别子空间上的投影作为惩罚正则项,以达到量化和消除词级语言模型中性别偏差的目的. 文本语料库 PTB、WikiText-2、CNN/Daily Mail 等和递推神经网络语言模型上的测试表明了该语言模型对性别偏见消除的有效性.

Qian 等人将男性和女性词的输出概率均衡作为损失函数,给出了能够降低性别偏见的 LSTM(Long Short Term Memory)语言模型[144],有效地减少了语言模型在训练过程中的性别偏见. 与现有的数据增强、词嵌入去偏等去偏策略相比,该方法在减少职业词的性别偏见方面表现突出.

Huang 等人使用反事实评价研究了语言模型如何受到敏感属性(如国家、职业、性别)的影响而产生的情感偏见[145]. 将个体公平性指标和群组公平性指标用于反事实情感偏见的度量,并在新闻文章和维基百科两个语料库上进行了模型训练,表明了情感偏见的存在. 进一步给出了词向量公平性正则项和情感公平性正则项两种模式的语言模型,能够在保持文本生成语义的同时较好地消除情感偏见.

表 8 列出了公平自然语言处理研究所关注的偏见类型及研究过程中所采用的实验数据集.

表 8 公平自然语言处理研究对比

文献	偏见类型	实验数据集
[126]	性别	w2vNEWS
[127]	性别	SemBias
[128]	性别	Wikipedia, New York Times
[129]	性别	WS, RG, MTurk, RW, MEN, SimLex
[130]	性别	Sembias
[131]	性别	SemBias, OntoNotes 5.0, WinoBias
[132]	性别	WinoBias, OntoNotes
[133]	性别	WMT18
[134]	性别	SST-2, CoLA, QNLI
[135]	性别	MTurk, B&L, BLS
[136]	性别	CoNLL-2012
[137]	性别	Google Translate Female, BLS Female Participation
[138]	性别	原创数据集
[139]	性别	WMT newstest2013
[140]	性别/种族	EEC
[141]	性别/种族	FB-Pol, FB-Pub, Reddit, Fitocracyc
[142]	性别/种族	Conservapedia, Rationalwiki, Wikipedia
[143]	性别	Penn Treebank, WikiText-2, CNN/Daily Mail
[144]	性别	Daily Mail stories
[145]	情感	WikiText-103, WMT-19

自然语言处理中的词向量、共指消解、机器翻译、情感分析、语言模型、对话生成等都存在一定程度的偏见和不公平性问题. 目前研究大多研究仅针对单一偏见(如性别偏见)展开，针对其他偏见或者同时消除多种偏见的研究成果较为缺乏.

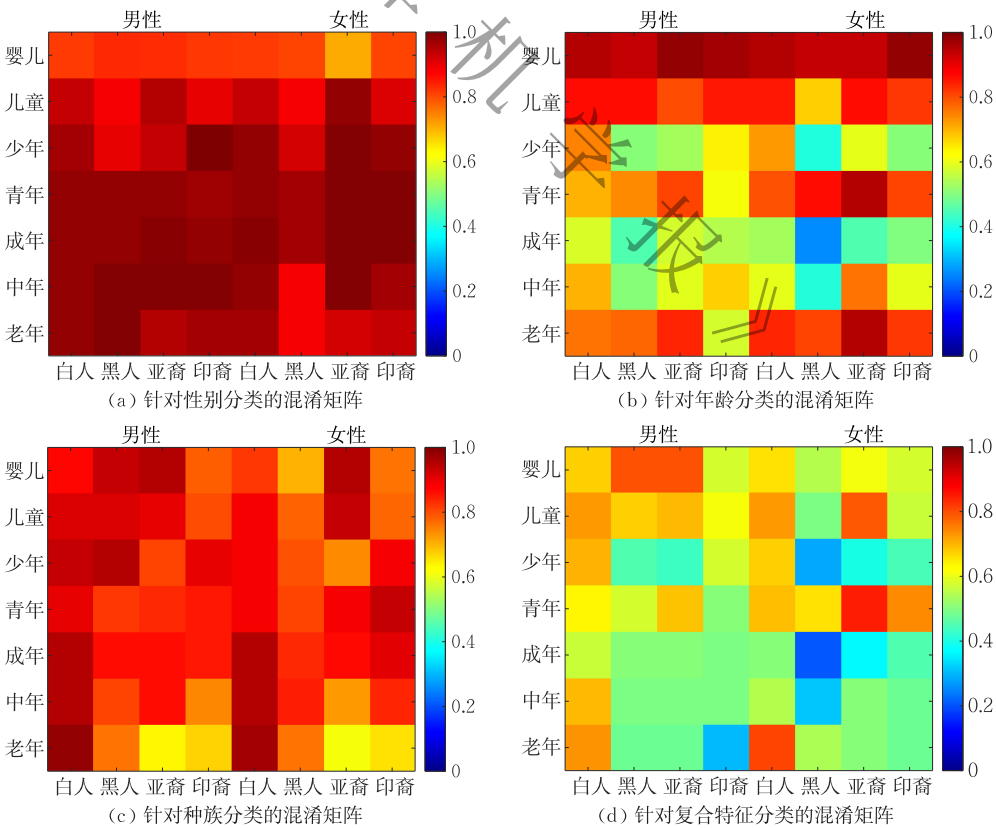
4.2.4　公平人脸识别

人脸识别需要通过人脸图像的生物特征实现图像或视频中人脸的检测、分析和比对. 性别、民族、地域、肤色等生物特征具有一定的敏感性，这些特征在训练数据和模型训练过程中很容易导致偏见歧视. 公平人脸识别就是确保机器学习的模型预测不存在敏感属性偏见的人脸识别.

Buolamwini 和 Gebru 通过建立的具有 Fitzpatrick 皮肤分型标记的人脸数据集，对商用人脸识别系统进行分析，发现人脸识别中存在明显的性别偏见[63]：深色皮肤女性的面部识别错误率高达 34.7%，而浅色皮肤男性的面部识别错误率最高仅为 0.8%.

Terhörst 等人给出了基于个体公平性的公平评分正规化方法[146]：依据性别、年龄和种族等特征对样本进行聚类以使得具有相似特征的样本属于相同的类；分别计算出各类的最优局部阈值以确保个体公平. 该方法能够在改善公平性的同时提高人脸识别系统性能，并可以集成到现有人脸识别系统中.

Das 等人给出了年龄、性别和种族等多特征人脸识别及(类内)偏见消除的多任务卷积神经网络(Multi-Task Convolutional Neural Network，MTCNN)方法[147]：利用特征之间的不相交性，分别增加对应于年龄(婴儿、儿童、少年、青年、成年、中年、老年)、性别(男性、女性)和种族(白人、黑人、亚裔、印裔)的卷积神经网络层；训练中通过联合动态加权机制自动调整每项任务的损失函数权重. MTCNN 能够较好地改善以上特征的公平性，并保证分类准确率. 图 11 借助于混淆矩阵进行了说明. 图 11(a)、(c)表明，MTCNN 针对性别和种族的分类效果较好，但针对婴儿(尤其是亚裔女性婴儿)的性别分类、针对老年人(尤其是亚裔和印裔老年人)的种族分类相对较差；图 11(b)表明，在多数种族和性别下，针对成年人(尤其是黑人女性)的分类效果较差；图 11(d)表明，当考虑复合特征时，性能有所下降，尤其是针对少年和成年黑人女性的分类效果下降较为明显.



(a) 针对性别分类的混淆矩阵　(b) 针对年龄分类的混淆矩阵

(c) 针对种族分类的混淆矩阵　(d) 针对复合特征分类的混淆矩阵

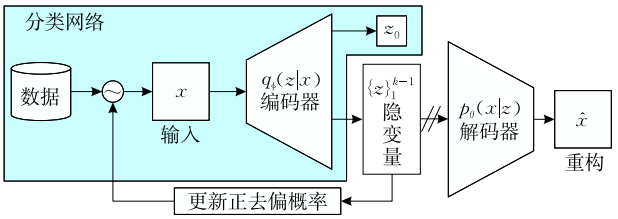
图 11　人脸识别的混淆矩阵

Smith 和 Ricanek 给出了通过数据增强策略实现性别和年龄偏见消除的深度卷积神经网络方法[148]，对不同群组学习出不同的增强策略，达到不同性别和年龄组群人脸识别率的公平.

Amini 等人给出了基于变分自编码器(Variational Autoencoder，VAE)的人脸识别中性别和种

《计算机学报》

族偏见消除的 VAE(Debiasing-VAE,DB-VAE)方法[149],如图 12 所示,VAE 的编码器学习数据中隐变量的近似分布 $q_\phi(z|x)$(隐变量 $z\in\mathbf{R}^k$);解码器重构 $p_\theta(x|z)$;损失函数由交叉熵、输入输出重构损失和隐变量 KL 散度等三部分组成.



注:// 表示当y=o时阻塞梯度

图 12 DB-VAE 框架

Wang 等人分别用数据集泄漏和模型泄露来度量数据集和训练模型关于受保护变量(性别)的编码偏见,指出平衡数据集无法完全保证公平性,并进一步给出了人脸识别中从卷积神经网络的中间表示中消除保护变量相关联特征的对抗方法[150],能够在基本保持原有人脸识别模型精准度下有效降低性别偏见.

Dhar 等人指出男性和女性数目相当的平衡数据集也未必能排除人脸识别的性别偏见,给出了分阶段学习的对抗性别消偏(Adversarial Gender De-biasing,AGD)方法[14]:人脸特征为输入生成性别消偏表示;性别消偏表示再输入到分类器和性别预测集成学习模型;损失函数由分类交叉熵损失和消偏损失两部分组成.该方法可用于不平衡数据集的公平人脸识别.

Wang 和 Deng 给出了基于强化学习的种族平衡网络(Reinforcement Learning Based Race Balance Network,RL-RBN)[151],将种族裕度(Margin)的优化求解描述为马尔可夫决策过程,通过深度 Q-learning 训练智能体选择逼近 Q-值函数的合理裕度的策略.此外,提供了两个用于研究种族偏见问题的训练数据集 BUPT Globalface 和 BUPT Balancedface.实验结果表明,有效降低了人脸识别的种族偏见.

表 9 列出了公平人脸识别研究所关注的偏见类型、所使用的机器学习模型及研究过程中所采用的实验数据集.

表 9 公平人脸识别研究对比

文献	偏见类型	机器学习模型	数据集
[106]	性别/年龄/种族	深度神经网络	CACD, IMDB, UTKFace, AgeDB, AFAD, AAF, FG-NET, RFW, IMFDB-CVIT, Asian-DeepGlint, PCSO, MS-Celeb-1M, LFW, IJB-A, IJB-C
[146]	性别/年龄/种族	Facenet	Adience, ColorFeret, Morph
[147]	性别/年龄/种族	卷积神经网络	UTKFace, BEFA
[148]	性别/年龄	ResNet,卷积神经网络	IMDB, Wiki, MORPH-II, PPB
[149]	性别/年龄	DB-VAE	CelebA, ImageNet, PPB
[150]	性别	ResNet,条件随机场	COCO, imSitu,
[151]	种族	深度强化学习,卷积神经网络	BUPT-Globalface, BUPT-Balancedface, RFW

公平人脸识别对于身份识别和认证、情感分析、疾病诊治等机器学习应用具有重要的意义.公平人脸识别和个人隐私保护具有密切的关联性,如何从受隐私保护的人脸数据中实现公平的人脸识别?公平人脸识别不仅能够降低不同群组人脸识别的偏见,是否也能够提高人脸识别在某种意义上的精准率?这些都是公平人脸识别值得进一步研究的问题.

4.2.5 公平推荐

推荐系统是依据用户的兴趣爱好和行为特征提供商品、新闻、娱乐、就业等信息服务的系统.推荐系统基于观测到的用户历史数据,利用机器学习进行偏好预测,难免会继承已有数据的已有偏见或固有成见,导致偏见、不公平问题的出现.公平推荐系统是能够提供无偏见的公平推荐服务的推荐系统.

Kamishima 等人将推荐结果不受敏感属性影响定义为推荐独立,并给出了增强推荐独立的正则方法和生成模型方法[152].前者是他们提出的公平分类的正则方法在推荐系统中的推广[111],为了满足评分的连续值需要,采用了概率矩阵分解形式的正则项;后者采用了协同过滤的隐语义模型,模型中评分变量独立于受保护属性变量.这些方法成功应用于年代和性别为敏感属性的电影评分的公平推荐.

Yang 和 Stoyanovich 给出了公平排序结果的归一化折扣差、归一化折扣 KL 散度、归一化折扣率的度量[153],揭示了实际数据案例中的不公平性问题,并对 Zemel 等人提出的公平表示学习方法[98]的目标函数进行修正,使其能够在满足统计公平性的同时保持分类的精准度,以实现公平排序推荐.

Yao 和 Huang 指出了协同过滤推荐系统中统计公平性度量的不足,讨论了推荐系统中不公平的四种情形:均衡用户群组和均衡观测概率、均衡用户群组和偏见观测概率、偏见用户群组和均衡观测概

率、偏见用户群组和偏见观测概率等.并给出了将这些不公平度量作为增强矩阵分解目标函数的正则项的学习算法,针对真实和合成数据在基本不影响排序性能的同时降低了不公平性[154].Burke 等人通过对损失函数增加邻近平衡正则项,给出了协同过滤推荐系统中公平排序推荐的改进稀疏线性推荐 SLIM(Sparse Linear Method)算法[155],并将其用于性别为保护属性的公平电影推荐.

Bobadilla 等人借助于原始评级和人口统计信息、用户和项目的少数索引、精准预测和公平推荐等四个抽象层次的深度学习模型训练,提出了基于深度学习的协同过滤推荐系统 DeepFair.该系统能够获得公平性和准确性之间的最佳平衡,并且无需获取用户的人口统计信息(如性别、年龄等)便可实现公平推荐,有助于保护用户隐私[156].

Singh 和 Joachims 讨论了排序推荐中群组公平的约束线性规划框架,给出了统计公平、不公平对待、不公平影响的约束描述[157].

表 10 列出了公平推荐研究所关注的偏见类型、所使用的推荐算法及研究过程中所采用的实验数据集.

表 10　公平推荐研究对比

文献	偏见类型	推荐算法	实验数据集
[152]	年份/性别/年龄/食物	基于规则的推荐	Movielens 1M, Flixster, Sushi
[153]	性别/种族	基于效用的推荐	ProPublica, German Credit
[154]	性别	协同过滤推荐	Movielens Million
[155]	性别	协同过滤推荐	MovieLens 1M
[156]	性别/年龄	协同过滤推荐	MovieLens 1M
[157]	性别/地域	基于效用的推荐	Yow news

推荐系统对于解决信息超载问题,为人们提供高质量的信息服务具有较大意义,但也存在大数据杀熟、信息茧房等偏见歧视问题.虽然国内外学者针对公平推荐系统展开了研究,但绝大多数工作仍局限于协同过滤推荐及基于效用推荐.

4.3　后处理

后处理是对训练后的模型或模型预测的数据进行处理,以消除训练数据和/或训练过程中残余的不公平.对于黑盒方式的机器学习,无法修改任何训练数据或学习算法,那么后处理是保证此类机器学习公平性所不得不采用的技术路径.

Hardt 等人给出了基于真实结果、受保护属性和其它属性分布的贝叶斯最优机会均等预测器[46],对已有预测器进行后续修正或者对已有模型预测进行后处理,该方法无需改变原始训练模型,能够保持所有受保护属性在机会均等下的公平性.

Jiang 等人给出了模型分类精度最小变化的强统计公平(Strong Demographic Parity)后处理方法[114]:模型信任变量分布的传输所导致的预测分类变换的期望,等值于模型信任变量分布的 WASSERSTEIN-1 距离;强统计公平下模型预测变化最小的后处理,对应于 WASSERSTEIN-1 最优传输映射.该方法能够在最小修改模型预测下实现公平性.

在排序推荐任务中,Zehlike 等人定义了公平 Top-k[158]:从数目超过 k 个的候选集中,选择效用最大的前 k 个最佳候选,Top-k 中每一候选的所有前序候选中敏感属性候选需要超过某一阈值.并给出了对排序推荐进行后处理求解公平 Top-k 的 FA*IR 算法.Wu 等人将排序推荐中候选项的位次映射为连续评分变量,建立了离散特征和连续评分因果图[80],并将路径影响推广至混合变量因果图,给出了基于路径影响的直接性和间接性歧视的度量,将消除歧视和重构排序描述为有约束二次规划问题,该方法实现 FRank(Fair Ranking)算法时的实验效果优于 FA*IR 算法.

Karako 和 Manggala 将欧氏距离相似度引入最大边缘相关算法 MMR(Maximal Marginal Relevance),给出了用于图片公平排序推荐的 FMMR(Fair MMR)算法[159],对 k-最邻近算法结果的后处理实验表明:FMMR 较基线 MMR 具有更高的精度和更好的公平性.

后处理是公平机器学习的一种事后补救措施,虽然可以消除数据预处理和模型算法中间处理中残留的偏见和不公平性,但是,某些训练数据或模型算法的固有偏见是本质上难以事后纠正或消除的.也正是后面的可能原因,除排序推荐这一特殊任务外,后处理的研究工作还比较有限.

5　公平性与隐私保护

隐私是个人或群体不愿意泄露的敏感信息,包括身份、属性及其相关的数据.隐私保护是通过适当的政策法规和技术手段来保障个人或群体的隐私不被泄露.公平性是确保个人或群体都有平等的机会获得一些利益的行为的性质.不公平行为源于基于个人或群体的敏感属性的带有偏见或歧视的决策.对个人或群体敏感属性/数据进行隐私保护,可以防止歧视者获得并利用敏感属性/数据采取带有偏见或歧视的决策.显然,公平性和隐私保护有着

计算机学报

一定的联系.

Hajian 和 Domingo-Ferrer 讨论了数据匿名保护与数据歧视防护的联系[160]:整体泛化的数据 k 匿名保护或多或少有助于数据的歧视防护(α 防护),单元泛化的数据 k 匿名保护或多或少有助于数据的歧视防护(α 防护);拟制操作的数据 k 匿名保护对数据歧视防护(α 防护)的影响,取决于拟制类型中记录的多少、拟制属性的值.

Ruggieri 研究了数据的 t-closeness 匿名保护和数据歧视的 α 防护之间的联系[161],给出了数据匿名保护和歧视防护的多维泛化算法 dMondrain 和桶分组算法 dSabre,实现了数据隐私保护和偏见歧视防护的一体化处理. Hajian 等人讨论了数据发布中隐私入侵和偏见歧视的潜在风险,以及与之相对应的隐私保护和歧视防护的技术相似性,给出了既能够对原始数据进行歧视防护(α 防护)又能够实现原始数据隐私保护(k 匿名)的泛化方法[162].

Kilbertus 等人将多方安全计算用于加密属性数据的公平机器学习[163]:监管方认证学习模型、签发公平模型,并验证机器学习模型是经过认证的公平模型;机器学习方通过用户的加密属性数据和其它数据训练出公平模型;用户提供加密的属性数据和其它数据.

Hu 等人给出了一个分布式隐私保护公平学习框架[164],该框架将数据分布在数据中心和第三方,前者拥有模型学习的非隐私数据,后者拥有能够与数据中心私密通信来辅助模型训练的隐私属性数据. 基于该框架,他们建立了隐私保护公平学习模型的设计策略:数据中心构造一个基于第三方私密通信的随机且公平的假设空间,数据中心在假设空间中通过标准学习方法训练精准模型. 基于该策略设计了分布式公平岭回归、分布式公平核岭回归、分布式公平对数几率回归、分布式公平主分量分析等.

Xu 等人给出了对数几率回归中同时实现差分隐私保护和公平性的两种方法[165]. (1)将决策边界公平约束作为惩罚项引入对数几率回归的目标函数,对数几率回归的决策边界公平性约束定义为受保护属性和非保护属性向量到决策边界的符号距离的协方差,进一步描述为受保护群体和非保护群体重心之间的符号距离. 继而,采取函数机制对约束目标函数的多项式参数增加零均值 Laplace 扰动实现差分隐私;(2)基于差分隐私保护和公平性之间的联系,公平性约束和函数机制都对原有目标函数的多项式参数产生扰动,对数几率回归的决策边界公平性约束可处理为受保护群体和非保护群体重心之间的符号距离引起的多项式参数的偏移. 由此,公平约束不再作为惩罚项,二者可以合并、转换为非零均值 Laplace 分布的噪声.

Jagielski 等人给出了 Agarwal 等人的公平学习算法[116]的差分隐私后处理算法和中间处理算法[166]. Bagdasaryan 等人分析讨论了差分隐私随机梯度下降(Differentially Private Stochastic Gradient Descent,DPSGD)训练的神经网络中的公平性问题[167]:性别分类中黑色面部远远低于白色面部的识别精度,差分隐私模型的识别精度的差别高于非差分隐私模型,亦即,差分隐私模型会导致更大的不公平性.

Yaghini 等人讨论了机器学习模型在成员推理攻击(Membership Inference Attacks,MIA)下不同群组(如性别、年龄等)脆弱性的公平性[168],并指出差分隐私学习有助于防止这类不公平. Chang 等人研究了公平机器学习在投毒攻击下的鲁棒性[169],给出了一种可以改变均衡几率公平机器学习模型中群组数据采样的投毒攻击算法,攻击明显降低了模型的精准度、并加大了不公平性.

Lyu 等人提出了具有公平性和隐私保护功能的公平差分隐私分散式深度学习(Fair and Differentially Private Decentralized Deep Learning,FDPDDL)框架[170-171],该框架基于区块链构建 FDPDDL 的分散化结构,通过建立的信誉系统来保障各参与方的公平性,在初始化和更新阶段分别采用差分隐私生成对抗网络(Differentially Private Generative Adversarial Network,DPGAN)和差分隐私随机梯度下降以防护隐私泄露以及生成对抗网络攻击. 实验结果表明,FDPDDL 能够保持较好的公平性且具有与集中式和分布式深度学习相当的精准度,具有优于独立方式深度学习的精准度.

为了确保隐私敏感机器学习应用中分类结果的公平性,Zhang 等人提出了公平联邦学习框架 FairFL(Fair Federated Learning)[172]. FairFL 的核心是多智能体强化学习(Multi-Agent Reinforcement Learning,MARL)和安全聚合协议. FairFL 允许用户单独制定本地更新策略,但 MARL 通过设定奖励和状态函数,引导用户协同做出本地更新决策,从而优化全局模型的公平性和准确性,安全聚合协议则能够确保学习过程中不侵犯用户隐私. 此外,两者协作,还具备应对受限信息和受限协调挑战的能力. 数据集 Adult 和 COMPAS 等的实验结果表明:FairFL 不仅可以

显著提高模型的公平性(针对 Adult、COMPAS 数据集的公平性最高可分别提升 68.2%、69.4%),而且还实现了更好的分类精度.

表 11 列出了公平性与隐私保护相关研究中所采用的隐私保护技术及实验数据集.

表 11　公平性与隐私保护研究对比

文献	隐私保护技术	实验数据集
[160]	数据匿名	/
[161]	数据匿名	German credit, Adult
[162]	数据匿名	German credit, Adult
[163]	多方安全计算	German credit, Adult, Bank
[164]	/	Community Crime, COMPAS
[165]	差分隐私	Adult, Dutch Census
[166]	差分隐私	Community Crime
[167]	差分隐私	DiF, UTKFace
[168]	差分隐私	Adult, COMPAS, UTKFace
[169]	/	COMPAS
[170]	差分隐私	MNIST, SVHN, Adult, Hospital
[171]	同态加密/差分隐私	MNIST, SVHN
[172]	安全聚合协议	Adult, COMPAS

机器学习的公平性和安全隐私的已有研究大致可分为:匿名保护和公平性、安全多方计算与公平性、差分隐私与公平性、公平机器学习的安全攻击及防护、区块链技术与公平性等,但差分隐私与公平性相关研究占据较大比例.具有安全防护和隐私保护的公平机器学习是机器学习的发展方向.隐私保护是否会对原始数据带来偏见?安全攻击防护是否也能防止歧视?能否开发出同时保护隐私和敏感属性的公平隐私保护技术?这些都是值得进一步研究的问题.

6　公平性与可解释性

解释是对概念或行为提供可理解的术语说明.机器学习的可解释性是指以人类或用户可以理解的方式对其行为和结果进行说明的能力[173-174].一方面,可解释性对于公平机器学习的应用部署具有重要的意义,另一方面,可解释性能够对机器学习的公平性满足与否进行说明和判定,有助于改善机器学习的公平性.

Abdollahi 和 Nasraoui 讨论了推荐系统中可解释和公平性之间的联系和影响[175],可解释以模型预测可解释和模型训练可解释两种方式提升推荐系统的透明性(图 13),进而通过向用户分别提供预测结果和模型训练过程说明的方式改善推荐系统的公平性.

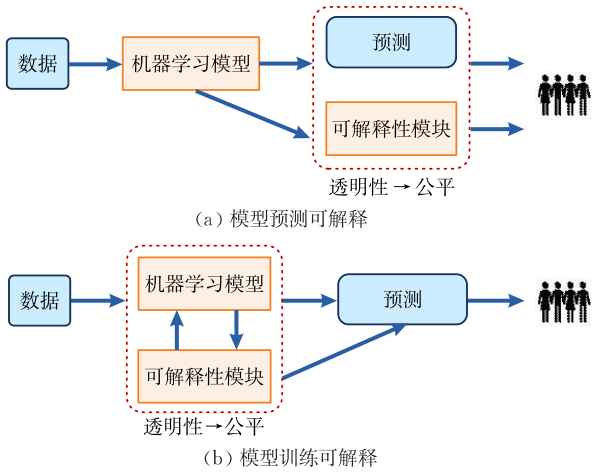


图 13　可解释性与公平性

Dodge 等人开展了解释对机器学习系统公平性评判的影响的实证研究[176],在实际数据集训练的机器学习模型上自动生成四种类型的解释(输入影响解释、统计解释、敏感属性解释、事例解释),探讨解释在揭示两类公平性问题(数据偏见引发的模型范围的公平性,特征空间的不同导致的公平性)的有效性.基于 Mechanical Turk 众包平台的用户调研结果表明:公平性的评判不仅取决于解释设计,还取决于个人对算法公平性的态度,包括公众对基于机器学习系统决策支持的信任和个人对使用某特定特征的立场,并提供了改善公平性评判的解释设计指导方针.

Du 等人讨论了可解释用于理解和发现深度学习的不公平性的可能途径[177].对于输入引发的不公平性,可以采用自顶向下和自底向上两种方法:前者利用局部解释生成特征重要性向量,在得到所有输入特征的特征重要性后,对重要性评分较高的特征进一步分析,识别出对公平性敏感的特征;后者则预先选择出可能与受保护属性相关联的特征,分析这些特征的特征重要性,对重要性评分较高的特征子集进行扰动以产生新的数据样本(即反事实),然后将反事实数据输入模型来观察预测结果,如果这些被怀疑为公平性敏感特征的扰动导致预测结果发生显著变化,则可以断言模型基于受保护属性做出了不公平的决策.对于表示导致的不公平性,首先,利用全局解释分析模型是否学习了一个敏感属性,然后,对于学习了敏感属性的模型继续测试敏感属性对模型最终预测的贡献度:自上而下计算模型预测对敏感属性向量的导数,或者自下而上将敏感属性向量添加到不同输入中观察模型预测的变化.使用数值评分来刻画敏感属性表示的不公平程度,数

《计算机学报》

值敏感性评分越高，该敏感属性对模型预测的不公平贡献越显著.

He 等人给出了可解释公平表示的几何方法[178]，该方法通过选择合理的超参数，从理论上保证去偏后的特征独立于敏感属性特征、高度相关于原有特征，并且去偏后的特征具有与原始特征一样的可解释性.该方法可视作为训练数据的预处理技术，所得到的表示可用于线性回归、随机森林、支持向量机和神经网络等机器学习.

Wang 等人建立了将先验知识引入已有公平学习的可解释公平表示学习方法[179]，例如，人脸识别中用人类难以判别种族的“模糊图片”作为先验知识，这一先验知识表征了数据拥有者所理解的可解释公平表示.该方法由先验知识学习和强制公平约束两个阶段组成：第一阶段利用先验知识训练编码器和判别器以使得编码器生成可解释表示，第一阶段对前阶段训练所得编码器添加到已有的公平表示学习以进一步强制执行期望的公平性约束.合成数据的实验结果表明，该方法的精度、公平性和可解释性均优于已有的基线结果.

为了研究 DNN 模型在图像处理任务中做出决策时所关注的关键区分性区域，Zhou 等人提出了名为类激活图(Class Activation Maps，CAM)的局部 DNN 可解释方法[180].Nagpal 等人利用 CAM 方法，从种族、年龄等方面研究了用于人脸识别的深度学习模型的公平性[181]，并基于 VGG-Face2、MS-Celeb-1M、CMU Multi-PIE、Craniofacial Longitudinal Morphological、Album-2、Racial Faces in the Wild、Adience dataset、Cross-Age CelebrityDataset 等数据集在 LightCNN-9、Light CNN-29、ResNet50、SENet50 等深度学习模型上进行了验证.

Hickey 等人在公平性和可解性的统一框架下定义了可解释公平性(Fairness by Explicability)[182]：如果不存在能够对特定模型公平性进行说明或解释的外部替代模型，那么该特定模型就可以被认为是可解释公平的，或者具有可解释公平性.并通过替代对抗模型的 SHAP(Shapley Additive exPlanations)值将公平性约束引入到可解释模型方法，SHAP 值的使用采取了两种方式：(1)构造可微公平正则项；(2)修改 AdaBoost 算法使其权重更新中包含对抗属性值.SHAP 值提供了解释模型预测的一个统一框架[183]，SHAP 值能够捕获模型特征中的敏感属性的统计不公平性，由此，可解释性和公平性得以有机融合.

表 12 列出了公平性与可解释性相关研究中所关注的机器学习模型及实验数据集.

表 12 公平性与可解释性研究对比

文献	机器学习模型	实验数据集
[176]	逻辑回归	COMPAS
[177]	深度神经网络	ImageNet，Adult，COMPAS
[178]	AdaBoost，支持向量机，随机森林，线性回归	German，Adult，COMPAS
[179]	卷积神经网络	Dsprite，ColorMNIST
[181]	深度神经网络	GG-Face2，MS-Celeb-1M，CMU Multi-PIE，Craniofacial Longitudinal Morphological，Album-2，Racial Faces in the Wild，Adience dataset，Cross-Age Celebrity-Dataset
[182]	XGBoost	Synthetic Data，Adult，Private Credit Risk
[183]	深度神经网络，卷积神经网络	MNIST

机器学习的公平性和可解释性的结合研究还非常有限，但是可解释性是机器学习应用部署必须解决的问题，尤其是，安全关键或生命攸关领域的机器学习应用.机器学习公平性和可解释性的一体化定义和度量、可解释公平机器学习技术与方法、隐私保护数据的可解释公平机器学习等都是值得关注的研究问题.

7 进一步工作展望

机器学习已经获得长足发展，基于机器学习的预测/决策已逐渐渗透到人类社会的各个方面，在自然语言处理、图像处理、个性化推荐、语音识别以及自动驾驶等领域获得广泛应用，机器学习预测的公平性直接影响着个人或群体的日常生活，影响着用户对机器学习应用部署的信心和接受程度.虽然公平机器学习逐渐受到了关注，但是总体而言，相关研究尚处于起步阶段，仍存在许多亟待解决的问题和挑战，如下是一些值得关注的研究：

(1)公平性定义及其度量

歧视和公平是道德、政治、哲学、法学等人文社科领域关注的热点问题，并且在多个方面仍然存在争议[32]：公平应该是确保每个人都有平等的机会获得一些利益，还是应该把对弱势群体的伤害降到最低？是否可以通过参照某些特定的非歧视模式来确定不公平性？隶属于自然科学技术领域的机器学习的公平或非歧视又意味着什么？该领域研究人员建立的 20 余种公平性的概念定义及度量是否已经足够用来解决机器学习中的公平性问题？现有的一些

公平性概念定义及度量是不能被同时满足的[48]，如何处理这些冲突和不相容？如何从最大限度降低弱势群体伤害、特定的非歧视模式等视角，建立机器学习公平性的概念定义及度量？这些视角下的不公平性定义及度量与现有公平性概念定义及度量是否存在不一致，如何协调和统一？符合群组公平性的群组内个体是否存在不公平？如何针对性地选择适合具体机器学习任务的公平性度量？这些都是机器学习的公平性概念定义及度量亟待解决的问题.

(2) 公平机器学习的评测

评测机器学习的不公平性是机器学习应用开发和部署的必要环节[184-185]，对于提升机器学习的可信性具有重要的意义[186]. FairTest 能够通过模型输出结果和受保护群体之间的无根据关联(Unwarranted Associations)，发现诱发不公平影响的关联缺陷，测试可疑的缺陷，并帮助开发人员调试降低不公平影响[187]. Themis 能够自动完成机器学习模型的群组公平性测试及歧视因果分析，能通过随机测试生成技术对不公平性进行定量评估[188]. AEQUITAS 能够通过对输入训练数据随机采样，发现导致个体不公平性的歧视性输入，检测出个体不公平性漏洞，还能对机器学习模型再训练以降低模型决策的不公平[189]. Agarwal 等人使用符号执行和局部可解释组合技术来生成个体公平性黑盒测试的测试用例，其数量高达 Themis 的 3.72 倍[190]. 此外，IBM、Microsoft 和 Google 等公司分别开发出了 AI Fairness 360、FairLearn 和 ML-fairness-gym 等公平性综合工具平台[191-192]，用于评测和消除机器学习的不公平性. 尽管如此，公平性测试研究还相当有限. 一方面，需要扩展已有机器学习模型的测试技术[184-185]，使之能应用于公平性测试. 另一方面，在软件测试领域已有了成熟的技术和方法[193]，机器学习的公平性测试可以从中得到借鉴，如变异测试[49,189]、蜕变测试[194-195]、白盒测试[196]等.

(3) 公平机器学习新模式

公平机器学习的研究大多集中于决策树、朴素贝叶斯、神经网络等模型. 强化学习依据奖励函数来选择行为策略，学习过程的公平性体现于[197]：算法选择差行为的概率不会高于选择好行为；算法不会偏好低质量的行为. 公平强化学习需要研究符合公平性的奖励函数和行为策略设计算法. 主分量分析会因群体的不同导致不同的重构误差[198]，公平主分量分析要求能够保持不同群体具有相当的数据保真度(Fidelity)以实现平衡的重构误差. 动态公平性是为了适应群体的时间演化特征而提出的[199-200]，动态公平机器学习需要构建群体动态模型和决策反馈影响机制的学习算法. 迁移学习将某一任务的训练模型用于另一任务，弥补了机器学习中训练数据的不足，公平迁移学习需要克服从源域到目标域迁移过程中引发的各种不公平[201]. 联邦学习是一种分布式机器学习范式，可以让成员在不共享数据的基础上联合建模. 联邦成员在共享加密的模型参数和中间计算结果的同时，也会共享各自存在的不公平，甚至叠加不公平. 公平联邦学习需要有效的机制来避免这些不公平性[202-203]. 元学习利用以往的经验知识来指导新任务的学习，具有“学会如何学习”的能力，在学会如何学习的同时，难免会积累历史的不公平. 公平元学习需要研究消除不公平累积的学习策略和记忆机制[204-205]. 研究集公平性为一体的新型公平机器学习模式值得关注.

(4) 符合伦理的机器学习

机器学习的公平性、隐私保护和可解释性是人工智能伦理范畴的属性. 欧盟委员发布的《可信赖 AI 的伦理指导原则》指出，可信赖 AI 应满足七个方面的条件要求：受人类监管、技术的稳健性和安全性、隐私和数据管理、透明度、非歧视性和公平性、社会和环境福祉、问责制等[206]. 国际电气电子工程师协会发布了《符合伦理设计：人工智能和自主系统促进人类福祉的远景》[207]，对于人工智能和自主系统的伦理设计提供了指导性建议. 人工智能发展的新机遇得益于机器学习的成功，机器学习作为人工智能的一个重要分支，在人类决策中发挥着愈来愈重要的作用，机器学习应用的推广有赖于人们对其信任程度的提高，符合伦理(Ethically Aligned)的机器学习是必然的发展方向[208-210]. 一方面，机器学习的目标是模型预测的精准度，公平性、隐私保护、可解释性等要求势必带来精准度的损失，这一矛盾是可信赖 AI 需要折衷考虑的问题，以寻求不同要求间的最优平衡；另一方面，当前针对可信赖 AI 的研究大多是从公平性、隐私保护或可解释性等单一维度进行的，而这些维度之间既存在某种程度上的一致性，也存在相互制约的情形，因此，集成公平性、可解释性、隐私保护的伦理机器学习的一体化机制、算法、模式和框架是值得开展的研究[170-171].

参 考 文 献

[1] Jordan M I, Mitchell T M. Machine learning: Trends, perspectives, and prospects. Science, 2015, 349(6245): 255-260

[2] Mitchell T M. Machine Learning. New York, USA: McGraw-Hill, 1997

[3] Zhou Zhi-Hua. Machine Learning. Beijing: Tsinghua University Press, 2016(in Chinese)
(周志华. 机器学习. 北京：清华大学出版社，2016)

[4] Chouldechova A, Roth A. A snapshot of the frontiers of fairness in machine learning. Communications of the ACM, 2020, 63(5): 82-89

[5] Barocas S, Hardt M, Narayanan A. Fairness and machine learning: Limitations and opportunities. http://www.fairmlbook.org, 2019

[6] Mehrabi N, Morstatter F, Saxena N, et al. A survey on bias and fairness in machine learning. ACM Computing Surveys, 2022, 54(6): 1-35

[7] Shifrin C A. Justice will weigh suit challenging airlines' computer reservations. Aviation Week & Space Technology, 1985, 122(12): 105-111

[8] Friedman B, Nissenbaum H. Bias in computer systems. ACM Transactions on Information Systems, 1996, 14(3): 330-347

[9] Makhlouf K, Zhioua S, Palamidessi C. On the applicability of machine learning fairness notions. ACM SIGKDD Explorations Newsletter, 2021, 23(1): 14-23

[10] Lambrecht A, Tucker C. Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. Management Science, 2019, 65(7): 2966-2981

[11] Datta A, Tschantz M C, Datta A. Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. Proceedings on Privacy Enhancing Technologies, 2015, 2015(1): 92-112

[12] Morik M, Singh A, Hong J, et al. Controlling fairness and bias in dynamic learning-to-rank//Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence. Montreal, Canada, 2021: 4804-4808

[13] Albiero V, Krishnapriya K S, Vangara K, et al. Analysis of gender inequality in face recognition accuracy//Proceedings of the IEEE Winter Conference on Applications of Computer Vision Workshops. Colorado, USA, 2020: 81-89

[14] Dhar P, Gleason J, Souri H, et al. An adversarial learning algorithm for mitigating gender bias in face recognition. arXiv:2006.07845, 2020

[15] Osoba O A, Welser IV W. An Intelligence in Our Image: The Risks of Bias and Errors in Artificial Intelligence. Santa Monica, USA: Rand Corporation, 2017

[16] Spanakis E K, Golden S H. Race/ethnic difference in diabetes and diabetic complications. Current Diabetes Reports, 2013, 13(6): 814-823

[17] Obermeyer Z, Powers B, Vogeli C, et al. Dissecting racial bias in an algorithm used to manage the health of populations. Science, 2019, 366(6464): 447-453

[18] Suresh H, Guttag J V. A framework for understanding unintended consequences of machine learning. arXiv:1901.10002, 2019

[19] Fang B, Jiang M, Cheng P, et al. Achieving outcome fairness in machine learning models for social decision problems//Proceedings of the 29th International Joint Conference on Artificial Intelligence. Yokohama, Japan, 2020: 444-450

[20] O'neil C. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Portland, USA: Broadway Books, 2016

[21] Waters A, Miikkulainen R. GRADE: Machine learning support for graduate admissions. AI Magazine, 2014, 35(1): 64

[22] Friedler S A, Scheidegger C, Venkatasubramanian S. On the (im) possibility of fairness. arXiv:1609.07236, 2016

[23] Santelices M V, Wilson M. Unfair treatment? The case of Freedle, the SAT, and the standardization approach to differential item functioning. Harvard Educational Review, 2010, 80(1): 106-134

[24] Brennan T, Dieterich W, Ehret B. Evaluating the predictive validity of the COMPAS risk and needs assessment system. Criminal Justice and Behavior, 2009, 36(1): 21-40

[25] Meyers J R, Schmidt F. Predictive validity of the Structured Assessment for Violence Risk in Youth (SAVRY) with juvenile offenders. Criminal Justice and Behavior, 2008, 35(3): 344-355

[26] Vaithianathan R, Maloney T, Putnam-Hornstein E, et al. Children in the public benefit system at risk of maltreatment: Identification via predictive modeling. American Journal of Preventive Medicine, 2013, 45(3): 354-359

[27] Aseervatham V, Lex C, Spindler M. How do unisex rating regulations affect gender differences in insurance premiums? The Geneva Papers on Risk and Insurance-Issues and Practice, 2016, 41(1): 128-160

[28] Zhao J, Wang T, Yatskar M, et al. Gender bias in coreference resolution: Evaluation and debiasing methods//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. New Orleans, USA, 2018: 15-20

[29] Ross K, Carter C. Women and news: A long and winding road. Media, Culture & Society, 2011, 33(8): 1148-1165

[30] Zhao J, Wang T, Yatskar M, et al. Men also like shopping: Reducing gender bias amplification using corpus-level constraints//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark, 2017: 2979-2989

[31] Hutchinson B, Mitchell M. 50 years of test (un) fairness: Lessons for machine learning//Proceedings of the Conference on Fairness, Accountability, and Transparency. New York, USA, 2019: 49-58

[32] Binns R. Fairness in machine learning: Lessons from political philosophy//Proceedings of the Conference on Fairness, Accountability and Transparency. New York, USA, 2018: 149-159

[33] Žliobaitė I. Measuring discrimination in algorithmic decision making. Data Mining and Knowledge Discovery, 2017, 31(4): 1060-1089

《计算机学报》

[34] Zhang L, Wu Y, Wu X. A causal framework for discovering and removing direct and indirect discrimination//Proceedings of the 26th International Joint Conference on Artificial Intelligence. Melbourne, Australia, 2017: 3929-3935

[35] Chen J, Kallus N, Mao X, et al. Fairness under unawareness: Assessing disparity when protected class is unobserved//Proceedings of the Conference on Fairness, Accountability, and Transparency. Atlanta, USA, 2019: 339-348

[36] Kamiran F, Žliobaitė I. Explainable and non-explainable discrimination in classification//Custers B, Calders T, Schermer B, et al, eds. Discrimination and Privacy in the Information Society. New York, USA: Springer, 2013: 155-170

[37] Rivera L A. Hiring as cultural matching: The case of elite professional service firms. American Sociological Review, 2012, 77(6): 999-1022

[38] Phelps E S. The statistical theory of racism and sexism. The American Economic Review, 1972, 62(4): 659-661

[39] Cleary T A. Test bias: Prediction of grades of Negro and white students in integrated colleges. Journal of Educational Measurement, 1968, 5(2): 115-124

[40] Verma S, Rubin J. Fairness definitions explained//Proceedings of the 2018 IEEE/ACM International Workshop on Software Fairness. Gothenburg, Sweden, 2018: 1-7

[41] Berk R. A primer on fairness in criminal justice risk assessments. The Criminologist, 2016, 41(6): 6-9

[42] Žliobaitė I. On the relation between accuracy and fairness in binary classification//Proceedings of the 2nd Workshop on Fairness, Accountability, and Transparency in Machine Learning. Lille, France, 2015: 1-5

[43] Corbett-Davies S, Pierson E, Feller A, et al. Algorithmic decision making and the cost of fairness//Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Halifax, Canada, 2017: 797-806

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

[45] Simoiu C, Corbett-Davies S, Goel S. The problem of infra-marginality in outcome tests for discrimination. The Annals of Applied Statistics, 2017, 11(3): 1193-1216

[46] Hardt M, Price E, Srebro N. Equality of opportunity in supervised learning//Proceedings of the 30th International Conference on Neural Information Processing Systems December. Barcelona, Spain, 2016: 3315-3323

[47] Berk R, Heidari H, Jabbari S, et al. Fairness in criminal justice risk assessments: The state of the art. Sociological Methods & Research, 2018, 50(1): 3-44

[48] Kleinberg J, Mullainathan S, Raghavan M. Inherent trade-offs in the fair determination of risk scores. arXiv: 1609.05807, 2016

[49] Galhotra S, Brun Y, Meliou A. Fairness testing: Testing software for discrimination//Proceedings of the 11th Joint Meeting on Foundations of Software Engineering. Paderborn, Germany, 2017: 498-510

[50] Kusner M J, Loftus J, Russell C, et al. Counterfactual fairness//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA, 2017: 4069-4079

[51] Dwork C, Hardt M, Pitassi T, et al. Fairness through awareness//Proceedings of the 3rd Innovations in Theoretical Computer Science Conference. Cambridge, USA, 2012: 214-226

[52] Kilbertus N, Carulla M R, Parascandolo G, et al. Avoiding discrimination through causal reasoning//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA, 2017: 656-666

[53] Pearl J. Causality. Cambridge, UK: Cambridge University Press, 2009

[54] Nabi R, Shpitser I. Fair inference on outcomes//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. Orleans, USA, 2018: 1931-1940

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

[56] Olteanu A, Castillo C, Diaz F, et al. Social data: Biases, methodological pitfalls, and ethical boundaries. Frontiers in Big Data, 2019, 2(13): 1-33

[57] Ashmore R, Calinescu R, Paterson C. Assuring the machine learning lifecycle: Desiderata, methods, and challenges. ACM Computing Surveys, 2021, 54(5): 1-39

[58] Anderson M. Men catch up with women on overall social media use. Washington, USA: Pew Research Center, Technical Report, 2015

[59] Hong L, Convertino G, Chi E H. Language matters in twitter: A large scale study//Proceedings of the 5th International AAAI Conference on Weblogs and Social Media. Barcelona, Spain, 2011: 518-521

[60] Shankar S, Halpern Y, Breck E, et al. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. arXiv: 1711.08536, 2017

[61] Almuhimedi H, Wilson S, Liu B, et al. Tweets are forever: A large-scale quantitative analysis of deleted tweets//Proceedings of the 2013 Conference on Computer Supported Cooperative Work. San Antonio, USA, 2013: 897-908

[62] Resnick P, Garrett R K, Kriplean T, et al. Bursting your (filter) bubble: Strategies for promoting diverse exposure//Proceedings of the 2013 Conference on Computer Supported Cooperative Work Companion. San Antonio, USA, 2013: 95-100

[63] Buolamwini J, Gebru T. Gender shades: Intersectional accuracy disparities in commercial gender classification//Proceedings of the Conference on Fairness, Accountability and Transparency. New York, USA, 2018: 77-91

[64] Wu S, Hofman J M, Mason W A, et al. Who says what to whom on twitter//Proceedings of the 20th International Conference on World Wide Web. Hyderabad, India, 2011: 705-714

[65] Mehrabi N, Morstatter F, Peng N, et al. Debiasing community detection: The importance of lowly connected nodes//Proceedings

《计算机学报》

of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Vancouver, Canada, 2019: 509-512

[66] Scellato S, Noulas A, Mascolo C. Exploiting place features in link prediction on location-based social networks//Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Diego, USA, 2011: 1046-1054

[67] Pedreschi D, Ruggieri S, Turini F. The discovery of discrimination//Custers B H M, Calders T, Schermer B W, et al, eds. Discrimination and Privacy in the Information Society. New York, USA, 2013: 91-108

[68] Pedreschi D, Ruggieri S, Turini F. Discrimination-aware data mining//Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Las Vegas, USA, 2008: 560-568

[69] Ruggieri S, Pedreschi D, Turini F. Data mining for discrimination discovery. ACM Transactions on Knowledge Discovery from Data, 2010, 4(2): 1-40

[70] Pedreschi D, Ruggieri S, Turini F. Measuring discrimination in socially sensitive decision records//Proceedings of the 2009 SIAM International Conference on Data Mining. Sparks, USA, 2009: 581-592

[71] Pedreschi D, Ruggieri S, Turini F. A study of top-*k* measures for discrimination discovery//Proceedings of the 27th Annual ACM Symposium on Applied Computing. Trento, Italy, 2012: 126-131

[72] Finkelstein M O, Levin B. Statistics for Lawyers. New York, USA: Springer, 2001

[73] Bendick M. Situation testing for employment discrimination in the United States of America. Horizons Stratégiques, 2007, 3(N5): 17-39

[74] Luong B T, Ruggieri S, Turini F. *k*-NN as an implementation of situation testing for discrimination discovery and prevention //Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Diego, USA, 2011: 502-510

[75] Romei A, Ruggieri S, Turini F. Discrimination discovery in scientific project evaluation: A case study. Expert Systems with Applications, 2013, 40(15): 6064-6079

[76] Mancuhan K, Clifton C. Combating discrimination using Bayesian networks. Artificial Intelligence and Law, 2014, 22(2): 211-238

[77] Zhang L, Wu Y, Wu X. Situation testing-based discrimination discovery: A causal inference approach//Proceedings of the 25th International Joint Conference on Artificial Intelligence. New York, USA, 2016: 2718-2724

[78] Bonchi F, Hajian S, Mishra B, et al. Exposing the probabilistic causal structure of discrimination. International Journal of Data Science and Analytics, 2017, 3(1): 1-21

[79] Choi Y J, Farnadi G, Babaki B, et al. Learning fair naive Bayes classifiers by discovering and eliminating discrimination patterns//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA, 2020: 10077-10084

[80] Wu Y, Zhang L, Wu X. On discrimination discovery and removal in ranked data using causal graph//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. London, UK, 2018: 2536-2544

[81] Ruggieri S, Hajian S, Kamiran F, et al. Anti-discrimination analysis using privacy attack strategies//Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Nancy, France, 2014: 694-710

[82] Dobra A, Fienberg S E. Bounds for cell entries in contingency tables given marginal totals and decomposable graphs. Proceedings of the National Academy of Sciences, 2000, 97(22): 11885-11892

[83] Xiao X, Tao Y. Anatomy: simple and effective privacy preservation//Proceedings of the 32nd International Conference on Very Large Data Bases. Seoul, Korea, 2006: 139-150

[84] Wong R C W, Fu A W C, Wang K, et al. Minimality attack in privacy preserving data publishing//Proceedings of the 33rd International Conference on Very Large Data Bases. Vienna, Austria, 2007: 543-554

[85] Xie W, Wu P. Fairness testing of machine learning models using deep reinforcement learning//Proceedings of the 19th International Conference on Trust, Security and Privacy in Computing and Communications. Guangzhou, China, 2020: 121-128

[86] Hajian S, Domingo-Ferrer J. A methodology for direct and indirect discrimination prevention in data mining. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(7): 1445-1459

[87] Kashid A, Kulkarni V, Patankar R. Discrimination-aware data mining: A survey. International Journal of Data Science, 2017, 2(1): 70-84

[88] Kamiran F, Calders T. Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems, 2012, 33(1): 1-33

[89] Kamiran F, Calders T. Classifying without discriminating//Proceedings of the 2nd International Conference on Computer, Control and Communication. Karachi, Pakistan, 2009: 1-6

[90] Calders T, Kamiran F, Pechenizkiy M. Building classifiers with independency constraints//Proceedings of the 2009 IEEE International Conference on Data Mining Workshops. Miami, USA, 2009: 13-18

[91] Kamiran F, Calders T. Classification with no discrimination by preferential sampling//Proceedings of the 19th Annual Machine Learning Conference of Belgium and The Netherlands. Leuven, Belgium, 2010: 1-6

[92] Hajian S, Domingo-Ferrer J, Martinez-Balleste A. Rule protection for indirect discrimination prevention in data mining //Proceedings of the International Conference on Modeling Decisions for Artificial Intelligence. Changsha, China, 2011: 211-222

[93] Hajian S, Domingo-Ferrer J, Martinez-Balleste A. Discrimination prevention in data mining for intrusion and crime detection//Proceedings of the 2011 IEEE Symposium on Computational Intelligence in Cyber Security. Paris, France, 2011: 47-54

[94] Žliobaitė I, Kamiran F, Calders T. Handling conditional discrimination//Proceedings of the IEEE 11th International Conference on Data Mining. Vancouver, Canada, 2011: 992-1001

[95] Feldman M, Friedler S A, Moeller J, et al. Certifying and removing disparate impact//Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Sydney, Australia, 2015: 259-268

[96] Jiang H, Nachum O. Identifying and correcting label bias in machine learning//Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics. Palermo, Italy, 2020: 702-712

[97] Calmon F, Wei D, Vinzamuri B, et al. Optimized pre-processing for discrimination prevention//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA, 2017: 3995-4004

[98] Zemel R, Wu Y, Swersky K, et al. Learning fair representations//Proceedings of the 30th International Conference on Machine Learning-Volume. Atlanta, USA, 2013: 325-333

[99] Louizos C, Swersky K, Li Y, et al. The variational fair autoencoder//Proceedings of the 4th International Conference on Learning Representations. San Juan, Puerto Rico, 2016

[100] Sattigeri P, Hoffman S C, Chenthamarakshan V, et al. Fairness GAN: Generating datasets with fairness properties using a generative adversarial network. IBM Journal of Research and Development, 2019, 63(4/5): 3:1-3:9

[101] Edwards H, Storkey A. Censoring representations with an adversary//Proceedings of the 4th International Conference on Learning Representations. San Juan, Puerto Rico, 2016: 1-14

[102] Xie Q, Dai Z, Du Y, et al. Controllable invariance through adversarial feature learning//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA, 2017: 585-596

[103] Madras D, Creager E, Pitassi T et al. Learning adversarially fair and transferable representations//Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden, 2018: 3384-3393

[104] Oneto L, Donini M, Pontil M, et al. Learning fair and transferable representations with theoretical guarantees//Proceedings of the 2020 IEEE 7th International Conference on Data Science and Advanced Analytics. Sydney, Australia, 2020: 30-39

[105] Tan Z, Yeom S, Fredrikson M, et al. Learning fair representations for kernel models//Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics. Palermo, Italy, 2020: 155-166

[106] Gong S, Liu X, Jain A K. Jointly de-biasing face recognition and demographic attribute estimation//Proceedings of the European Conference on Computer Vision. Glasgow, UK, 2020: 330-347

[107] Lahoti P, Gummadi K P, Weikum G. ifair: Learning individually fair data representations for algorithmic decision making//Proceedings of the IEEE 35th International Conference on Data Engineering. Macao, China, 2019: 1334-1345

[108] Kamiran F, Calders T, Pechenizkiy M. Discrimination aware decision tree learning//Proceedings of the 2010 IEEE International Conference on Data Mining. Sydney, Australia, 2010: 869-874

[109] Raff E, Sylvester J, Mills S. Fair forests: Regularized tree induction to minimize model bias//Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. Honolulu, USA, 2018: 243-250

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

[111] Kamishima T, Akaho S, Asoh H, et al. Fairness-aware classifier with prejudice remover regularizer//Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Bristol, UK, 2012: 35-50

[112] Krasanakis E, Spyromitros-Xioufis E, Papadopoulos S, et al. Adaptive sensitive reweighting to mitigate bias in fairness-aware classification//Proceedings of the 2018 World Wide Web Conference. Lyon, France, 2018: 853-862

[113] Zafar M B, Valera I, Gomez-Rodriguez M, Gummadi K P. Fairness constraints: A flexible approach for fair classification. Journal of Machine Learning Research, 2019, 20(75): 1-42

[114] Jiang R, Pacchiano A, Stepleton T, et al. Wasserstein fair classification//Proceedings of the 35th Uncertainty in Artificial Intelligence Conference. Tel Aviv, Israel, 2020: 862-872

[115] Beutel A, Chen J, Zhao Z, et al. Data decisions and theoretical implications when adversarially learning fair representations//Proceedings of the 2017 Workshop on Fairness, Accountability, and Transparency in Machine Learning. Halifax, Canada, 2017

[116] Agarwal A, Beygelzimer A, Dudík M, et al. A reductions approach to fair classification//Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden, 2018: 60-69

[117] Liu S, Vicente L N. Accuracy and fairness trade-offs in machine learning: A stochastic multi-objective approach. arXiv:2008.01132, 2020

[118] Dong Y, Kang J, Tong H, et al. Individual fairness for graph neural networks: A ranking based approach//Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. Singapore, 2020: 300-310

[119] Kim H, Shin S, Jang J H, et al. Counterfactual fairness with disentangled causal effect variational autoencoder//Proceedings of the 35th AAAI Conference on Artificial Intelligence. Palo Alto, USA, 2021, 8128-8136

[120] Berk R, Heidari H, Jabbari S, et al. A convex framework for fair regression. arXiv:1706.02409, 2017

[121] Agarwal A, Dudík M, Wu Z S. Fair regression: Quantitative definitions and reduction-based algorithms//Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA, 2019: 166-183

计算机学报

[122] Oneto L, Donini M, Pontil M. General fair empirical risk minimization//Proceedings of the 2020 International Joint Conference on Neural Networks. Glasgow, UK, 2020: 1-8

[123] Fitzsimons J, Al Ali A, Osborne M, Roberts S. A general framework for fair regression. Entropy, 2019, 21(8): 741-756

[124] Aghaei S, Azizi M J, Vayanos P. Learning optimal and fair decision trees for non-discriminative decision-making//Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu, USA, 2019: 1418-1426

[125] Zhao C, Chen F. Rank-based multi-task learning for fair regression//Proceedings of the 2019 IEEE International Conference on Data Mining. Beijing, China, 2019: 916-925

[126] Bolukbasi T, Chang K W, Zou J, et al. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings//Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona, Spain, 2016: 4356-4364

[127] Zhao J, Zhou Y, Li Z, et al. Learning gender-neutral word embeddings//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium, 2018: 4847-4853

[128] Brunet M E, Alkalay-Houlihan C, Anderson A, et al. Understanding the origins of bias in word embeddings//Proceedings of the 36th International Conference on Machine Learning. California, USA, 2019: 803-811

[129] Kaneko M, Bollegala D. Gender-preserving debiasing for pre-trained word embeddings//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy, 2019: 1641-1650

[130] Shin S, Song K, Jang J H, et al. Neutralizing gender bias in word embeddings with latent disentanglement and counterfactual generation//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings. Punta Cana, Dominican, 2020: 3126-3140

[131] Yang Z, Feng J. A causal inference method for reducing gender bias in word embedding relations//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA, 2020: 9434-9441

[132] Zhao J, Wang T, Yatskar M, et al. Gender bias in contextualized word embeddings//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, USA, 2019: 629-634

[133] Basta C, Costa-jussà M R, Casas N. Evaluating the underlying gender bias in contextualized word embeddings//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, USA, 2019: 33-39

[134] Liang P P, Li I M, Zheng E, et al. Towards debiasing sentence representations//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Washington, USA, 2020: 5502-5515

[135] Rudinger R, Naradowsky J, Leonard B, et al. Gender bias in coreference resolution//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. New Orleans, USA, 2018: 8-14

[136] Lu K, Mardziel P, Wu F, et al. Gender bias in neural natural language processing//Nigam V, Ban Kirigin T, Talcott C, et al, eds. Logic, Language, and Security. New York, USA, 2020: 189-202

[137] Prates M O R, Avelar P H, Lamb L C. Assessing gender bias in machine translation: A case study with Google translate. Neural Computing and Applications, 2019, 32: 6363-6381

[138] Vanmassenhove E, Hardmeier C, Way A. Getting gender right in neural machine translation//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium, 2018: 3003-3008

[139] Font J E, Costa-jussà M R. Equalizing gender bias in neural machine translation with word embeddings techniques//Proceedings of the 1st Workshop on Gender Bias in Natural Language Processing. Florence, Italy, 2019: 147-154

[140] Kiritchenko S, Mohammad S. Examining gender and race bias in two hundred sentiment analysis systems//Proceedings of the 7th Joint Conference on Lexical and Computational Semantics. New Orleans, USA, 2018: 43-53

[141] Prabhakaran V, Hutchinson B, Mitchell M. Perturbation sensitivity analysis to detect unintended model biases//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong, China, 2019: 5740-5745

[142] Popović R, Lemmerich F, Strohmaier M. Joint multiclass debiasing of word embeddings//Proceedings of the 25th International Symposium on Methodologies for Intelligent Systems. Graz, Austria, 2020: 79-89

[143] Bordia S, Bowman S R. Identifying and reducing gender bias in word-level language models//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop. Minneapolis, USA, 2019: 7-15

[144] Qian Y, Muaz U, Zhang B, et al. Reducing gender bias in word-level language models with a gender-equalizing loss function//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop. Florence, Italy, 2019: 223-228

[145] Huang P S, Zhang H, Jiang R, et al. Reducing sentiment bias in language models via counterfactual evaluation//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings. Punta Cana, Dominican, 2020: 65-83

[146] Terhörst P, Kolf J N, Damer N, et al. Post-comparison mitigation of demographic bias in face recognition using fair score normalization. Pattern Recognition Letters, 2020, 140: 332-338

《计算机学报》

[147] Das A, Dantcheva A, Bremond F. Mitigating bias in gender, age and ethnicity classification: A multi-task convolution neural network approach//Proceedings of the European Conference on Computer Vision Workshops. Munich, Germany, 2018: 573-585

[148] Smith P, Ricanek K. Mitigating algorithmic bias: Evolving an augmentation policy that is non-biasing//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops. Colorado, USA, 2020: 90-97

[149] Amini A, Soleimany A P, Schwarting W, et al. Uncovering and mitigating algorithmic bias through learned latent structure//Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. Honolulu, USA, 2019: 289-295

[150] Wang T, Zhao J, Yatskar M, et al. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea, 2019: 5310-5319

[151] Wang M, Deng W. Mitigating bias in face recognition using skewness-aware reinforcement learning//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 9322-9331

[152] Kamishima T, Akaho S, Asoh H, et al. Model-based approaches for independence-enhanced recommendation//Proceedings of the IEEE 16th International Conference on Data Mining Workshops. Barcelona, Spain, 2016: 860-867

[153] Yang K, Stoyanovich J. Measuring fairness in ranked outputs//Proceedings of the 29th International Conference on Scientific and Statistical Database Management. Chicago, USA, 2017: 1-6

[154] Yao S, Huang B. Beyond parity: Fairness objectives for collaborative filtering//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA, 2017: 2925-2934

[155] Burke R, Sonboli N, Mansoury M, et al. Balanced neighborhoods for fairness-aware collaborative recommendation//Proceedings of the FATREC Workshop on Responsible Recommendation Proceedings. Como, Italy, 2017: 5

[156] Bobadilla J, Lara-Cabrera R, N González-Prieto, et al. DeepFair: Deep learning for improving fairness in recommender systems. arXiv:2006.05255, 2020

[157] Singh A, Joachims T. Fairness of exposure in rankings//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. London, UK, 2018: 2219-2228

[158] Zehlike M, Bonchi F, Castillo C, et al. FA*IR: A fair top-k ranking algorithm//Proceedings of the 2017 ACM Conference on Information and Knowledge Management. Singapore, 2017: 1569-1578

[159] Karako C, Manggala P. Using image fairness representations in diversity-based re-ranking for recommendations. arXiv: 1809.03577, 2018

[160] Hajian S, Domingo-Ferrer J. A study on the impact of data anonymization on anti-discrimination//Proceedings of the IEEE 12th International Conference on Data Mining Workshops. Brussels, Belgium, 2012: 352-359

[161] Ruggieri S. Using t-closeness anonymity to control for non-discrimination. Transactions on Data Privacy, 2014, 7(2): 99-129

[162] Hajian S, Domingo-Ferrer J, Farràs O. Generalization-based privacy preservation and discrimination prevention in data publishing and mining. Data Mining and Knowledge Discovery, 2014, 28(5): 1158-1188

[163] Kilbertus N, Gascón A, Kusner M, et al. Blind justice: Fairness with encrypted sensitive attributes//Proceedings of the International Conference on Machine Learning. Stockholm, Sweden, 2018: 2630-2639

[164] Hu H, Liu Y, Wang Z, et al. A distributed fair machine learning framework with private demographic data protection//Proceedings of the 2019 IEEE International Conference on Data Mining. Beijing, China, 2019: 1102-1107

[165] Xu D, Yuan S, Wu X. Achieving differential privacy and fairness in logistic regression//Proceedings of the 2019 World Wide Web Conference. San Francisco, USA, 2019: 594-599

[166] Jagielski M, Kearns M, Mao J, et al. Differentially private fair learning//Proceedings of the International Conference on Machine Learning. California, USA, 2019: 3000-3008

[167] Bagdasaryan E, Poursaeed O, Shmatikov V. Differential privacy has disparate impact on model accuracy//Proceedings of the Advances in Neural Information Processing Systems. Montreal, Canada, 2019: 15453-15462

[168] Yaghini M, Kulynych B, Troncoso C. Disparate vulnerability: On the unfairness of privacy attacks against machine learning. arXiv: 1906.00389, 2019

[169] Chang H, Nguyen T D, Murakonda S K, et al. On adversarial bias and the robustness of fair machine learning. arXiv: 2006.08669, 2020

[170] Lyu L, Li Y, Nandakumar K, et al. How to democratise and protect AI: Fair and differentially private decentralised deep learning. IEEE Transactions on Dependable and Secure Computing, 2020, PP(99): 1-1

[171] Lyu L, Yu J, Nandakumar K, et al. Towards fair and privacy-preserving federated deep models. IEEE Transactions on Parallel and Distributed Systems, 2020, 31(11): 2524-2541

[172] Zhang D Y, Kou Z, Wang D. FairFL: A fair federated learning approach to reducing demographic bias in privacy-sensitive classification models//Proceedings of the 2020 IEEE International Conference on Big Data. Atlanta, USA, 2020: 1051-1060

[173] Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. arXiv:1702.08608, 2017

[174] Chen Ke-Rui, Meng Xiao-Feng. Interpretation and understanding in machine learning. Journal of Computer Research and Development, 2020, 57(9): 1971-1986(in Chinese)
(陈珂锐，孟小峰. 机器学习的可解释性. 计算机研究与发展, 2020, 57(9): 1971-1986)

《计算机学报》

[175] Abdollahi B, Nasraoui O. Transparency in fair machine learning: the case of explainable recommender systems//Kao Y F, Venkatachalam R, eds. Human and Machine Learning. New York, USA: Springer, 2018: 21-35

[176] Dodge J, Liao Q V, Zhang Y, et al. Explaining models: An empirical study of how explanations impact fairness judgment//Proceedings of the 24th International Conference on Intelligent User Interfaces. Marina del Ray, USA, 2019: 275-285

[177] Du M, Yang F, Zou N, et al. Fairness in deep learning: A computational perspective. IEEE Intelligent Systems, 2020, 36(4): 25-34

[178] He Y, Burghardt K, Lerman K. A geometric solution to fair representations//Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York, USA, 2020: 279-285

[179] Wang T, Buçinca Z, Ma Z. Learning interpretable fair representations. Harvard University, Cambridge, USA: Technical Report, 2021

[180] Zhou B, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 2921-2929

[181] Nagpal S, Singh M, Singh R, et al. Deep learning for face recognition: Pride or prejudiced? arXiv:1904. 01219, 2019

[182] Hickey J M, Di Stefano P G, Vasileiou V. Fairness by explicability and adversarial SHAP learning//Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Ghent, Belgium, 2021: 174-190

[183] Lundberg S M, Lee S I. A unified approach to interpreting model predictions//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA, 2017: 4768-4777

[184] Zhang J M, Harman M, Ma L, et al. Machine learning testing: Survey, landscapes and horizons. IEEE Transactions on Software Engineering, 2022, 48(1): 1-36

[185] Wang Zan, Yan Ming, Liu Shuang, et al. Survey on testing of deep neural networks. Journal of Software, 2020, 31(5): 1255-1275(in Chinese)
(王赞, 闫明, 刘爽等. 深度神经网络测试研究综述. 软件学报, 2020, 31(5): 1255-1275)

[186] Liu Wen-Yan, Shen Chu-Yun, Wang Xiang-Feng, et al. Survey on Fairness in trustworthy machine learning. Journal of Software, 2021, 32(5): 1404-1426(in Chinese)
(刘文炎, 沈楚云, 王祥丰等. 可信机器学习的公平性综述. 软件学报, 2021, 32(5): 1404-1426)

[187] Tramer F, Atlidakis V, Geambasu R, et al. FairTest: Discovering unwarranted associations in data-driven applications//Proceedings of the 2017 IEEE European Symposium on Security and Privacy. Paris, France, 2017: 401-416

[188] Angell R, Johnson B, Brun Y, et al. Themis: Automatically testing software for discrimination//Proceedings of the 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering. Lake Buena Vista, USA, 2018: 871-875

[189] Udeshi S, Arora P, Chattopadhyay S. Automated directed fairness testing//Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering. Montpellier, France, 2018: 98-108

[190] Agarwal A, Lohia P, Nagar S, et al. Automated test generation to detect individual discrimination in AI models. arXiv:1809.03260, 2018

[191] Bellamy R K E, Dey K, Hind M, et al. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM Journal of Research and Development, 2019, 63(4/5): 4:1-4:15

[192] Chen Jin-Yin, Chen Yi-Peng, Chen Yi-Ming, et al. Fairness research on deep learning. Journal of Computer Research and Development, 2021, 58(2): 264-280(in Chinese)
(陈晋音, 陈奕芃, 陈一鸣等. 面向深度学习的公平性研究综述. 计算机研究与发展, 2021, 58(2): 264-280)

[193] Ammann P, Offutt J. Introduction to Software Testing. Cambridge, UK: Cambridge University Press, 2016

[194] Ma P, Wang S, Liu J. Metamorphic testing and certified mitigation of fairness violations in NLP models//Proceedings of the 29th International Joint Conference on Artificial Intelligence. Yokohama, Japan, 2020: 458-465

[195] Sharma A, Wehrheim H. Testing machine learning algorithms for balanced data usage//Proceedings of the 12th IEEE Conference on Software Testing, Validation and Verification (ICST). Xi'an, China, 2019: 125-135

[196] Soremekun E, Udeshi S, Chattopadhyay S. Astraea: Grammar-based fairness testing. IEEE Transactions on Software Engineering, 2022, PP(99): 1-1

[197] Jabbari S, Joseph M, Kearns M, et al. Fairness in reinforcement learning//Proceedings of the International Conference on Machine Learning. Sydney, Australia, 2017: 1617-1626

[198] Samadi S, Tantipongpipat U, Morgenstern J, et al. The price of fair PCA: One extra dimension//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montreal, Canada, 2018: 10999-11010

[199] Wen M, Bastani O, Topcu U. Fairness with dynamics. arXiv:1901.08568, 2019

[200] Liu L T, Dean S, Rolf E, et al. Delayed impact of fair machine learning//Proceedings of the International Conference on Machine Learning. Stockholm, Sweden, 2018: 3150-3158

[201] Coston A, Ramamurthy K N, Wei D, et al. Fair transfer learning with missing protected attributes//Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. Honolulu, USA, 2019: 91-98

[202] Li T, Sanjabi M, Beirami A, et al. Fair resource allocation in federated learning//Proceedings of the Eighth International Conference on Learning Representations. Addis Ababa, Ethiopia, 2020: 1-27

[203] Zhang J, Li C, Robles-Kelly A, et al. Hierarchically fair federated learning. arXiv:2004.10386, 2020

[204] Slack D, Friedler S, Givental E. Fair meta-learning: Learning

how to learn fairly. arXiv:1911.04336, 2019

[205] Zhao C, Li C, Li J, et al. Fair meta-learning for few-shot classification//Proceedings of the 2020 IEEE International Conference on Knowledge Graph. Nanjing, China, 2020: 275-282

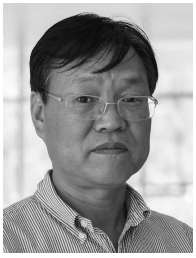
[206] Floridi L. Establishing the rules for building trustworthy AI. Nature Machine Intelligence, 2019, 1(6): 261-262

[207] How J P. Ethically aligned design. IEEE Control Systems Magazine, 2018, 38(3): 3-4

[208] Gu Tian-Long, Li Long. Artificial moral agents and their design methodology: Retrospect and prospect. Chinese Journal of Computers, 2021, 44(3): 632-651(in Chinese)
(古天龙，李龙. 伦理智能体及其设计：现状和展望. 计算机学报, 2021, 44(3): 632-651)

[209] Loi M, Christen M. How to include ethics in machine learning research. ERCIM News, 2019, 116(3): 5-6

[210] Yapo A, Weiss J. Ethical implications of bias in machine learning//Proceedings of the 51st Hawaii International Conference on System Sciences. Hawaii, USA, 2018: 5365-5372



GU Tian-Long, Ph. D., professor. His research interests mainly include formal methods, trustworthy artificial intelligence, ethically aligned machine design, artificial intelligence ethics, and data governance.

LI Long, Ph. D., lecturer. His mainly research interests include artificial intelligence security, fair machine learning and logic programming.

CHANG Liang, Ph. D., professor. His research interests mainly include knowledge graph, know-ledge representation, and reasoning, description logics.

LUO Yi-Qin, Ph. D. candidate. Her mainly research interests include fair machine learning and fair representation learning.

《计算机学报》

Background

This paper is the frontier research in the field of machine learning and trustworthy artificial intelligence. Machine learning is an important branch of artificial intelligence, which is the study of automatically improving the performance of computer systems or algorithms through data or previous experience. In recent years, abundant data and computing power accelerate the development of machine learning, and their applications have been closely related to many respects of public life, such as data mining, computer vision, natural language processing, speech recognition, disease diagnosis, personalized recommendation, credit rating, welfare resource allocation, and evaluation of students' quality, etc. In these applications, machine learning plays an important role in assisting or replacing human to predict and make decisions. Influenced by the nature and technical characteristics of machine learning itself, the prediction and decision-making of machine learning will inevitably produce bias or unfairness, which has gradually attracted the attention of scientific researchers, industrial practitioners and the public. In the decision-making process, fairness refers to the absence of any prejudice, preference, discrimination or injustice based on the inherent or acquired characteristics of individuals or groups. Therefore, an unfair algorithm is one whose decisions are biased against individuals or specific groups, which leads to unfair treatment of the individual or disadvantaged groups and damages the interests of them. How to protect the interests of disadvantaged groups in these applications? How to ensure fair or unbiased decisions in these applications? These issues have important impacts on the society and the public's trust in machine learning, and the public acceptance of artificial intelligence technology and their applications' deployment. In this paper, the fairness or justice in machine learning, the concept of fair machine learning, discrimination discovery in the applications of machine learning, and design methodology of fair machine learning algorithms are introduced and discussed. Meanwhile, fair machine learning via security and privacy, and fair machine learning via interpretability are illustrated. Moreover, the challenges and further research topics regarding fair machine learning are presented and outlooked. Recently, there are several review articles regarding fair machine learning, but they only focus on some aspects, and the scope of review is not comprehensive enough. The work of this paper benefits from the research experiences of the NSFC general projects and key projects hosted by the author in recent years. These projects have carried out a lot of research on formal methods, artificial intelligence, machine learning, knowledge engineering, big data of urban governance, and big data of education. The author has published some works, such as "formal method of software development" and "ordered binary decision graph and application", and some academic papers. Researchers can fully understand the research status of fair machine learning at home and abroad from the work of this paper, and it is helpful to guide interested researchers in this field to realize the state of the art of fair machine learning and to grasp the topics of further research.