DOI:10.7544/issn1000-1239.2021.20200758
58(2):264-280,2021

面向深度学习的公平性研究综述

陈晋音[1,2] 陈奕芃[2] 陈一鸣[2] 郑海斌[2] 纪守领[3] 时 杰[4] 程 瑶[4]
[1](浙江工业大学网络空间安全研究院 杭州 310023)
[2](浙江工业大学信息工程学院 杭州 310023)
[3](浙江大学计算机科学与技术学院 杭州 310058)
[4](华为国际有限公司新加坡研究院 新加坡 138589)
(chenjinyin@zjut.edu.cn)

Fairness Research on Deep Learning

Chen Jinyin[1,2], Chen Yipeng[2], Chen Yiming[2], Zheng Haibin[2], Ji Shouling[3], Shi Jie[4], and Cheng Yao[4]
[1](*Institute of Cyberspace Security, Zhejiang University of Technology, Hangzhou* 310023)
[2](*College of Information Engineering, Zhejiang University of Technology, Hangzhou* 310023)
[3](*College of Computer Science and Technology, Zhejiang University, Hangzhou* 310058)
[4](*Huawei International Pte Ltd, Singapore* 138589)

Abstract Deep learning is an important field of machine learning research, which is widely used in industry for its powerful feature extraction capabilities and advanced performance in many applications. However, due to the bias in training data labeling and model design, research shows that deep learning may aggravate human bias and discrimination in some applications, which results in unfairness during the decision-making process, thereby will cause negative impact to both individuals and socials. To improve the reliability of deep learning and promote its development in the field of fairness, we review the sources of bias in deep learning, debiasing methods for different types biases, fairness measure metrics for measuring the effect of debiasing, and current popular debiasing platforms, based on the existing research work. In the end we explore the open issues in existing fairness research field and future development trends.

Key words deep learning; algorithm fairness; debiasing method; fairness metric; machine learning

摘 要 深度学习是机器学习研究中的一个重要领域,它具有强大的特征提取能力,且在许多应用中表现出先进的性能,因此在工业界中被广泛应用.然而,由于训练数据标注和模型设计存在偏见,现有的研究表明深度学习在某些应用中可能会强化人类的偏见和歧视,导致决策过程中的不公平现象产生,从而对个人和社会产生潜在的负面影响.为提高深度学习的应用可靠性、推动其在公平领域的发展,针对已有的研究工作,从数据和模型2方面出发,综述了深度学习应用中的偏见来源、针对不同类型偏见的去偏方法、评估去偏效果的公平性评价指标、以及目前主流的去偏平台,最后总结现有公平性研究领域存在的开放问题以及未来的发展趋势.

关键词 深度学习;算法公平性;去偏方法;公平性指标;机器学习

中图法分类号 TP391

收稿日期:2020-09-15;**修回日期**:2020-10-28

基金项目:国家自然科学基金项目(62072406);浙江省自然科学基金项目(LY19F020025);宁波市"科技创新2025"重大专项(2018B10063)
This work was supported by the National Natural Science Foundation of China (62072406), the Natural Science Foundation of Zhejiang Province (LY19F020025), and the Major Special Funding for "Science and Technology Innovation 2025" in Ningbo (2018B10063).

目前,深度学习算法已经取得了巨大的进步,并且越来越多地用于影响个人生活的决策应用中,包括图像分类[1]、欺诈检测[2]、情绪分析[3]、面部识别[4]、语音理解[5]、自动驾驶[6]、医学诊断[7]等,深度学习在这些复杂任务上的性能已经达到甚至超过了人类决策的水平,能够实现比机器学习更高的准确率.然而,深度学习在基于种族、年龄、性别等敏感属性上的应用仍然具有不公平性,这种基于数据的学习方法会过度关联敏感属性,可能会对受保护群体表现出歧视行为,从而对个人和社会产生潜在的负面影响.例如,美国法院使用 COMPAS 作为刑事司法系统中的风险评估工具,用来衡量每一个被告再次犯罪的概率.然而,对此工具的调查发现 COMPAS 对于种族这一敏感属性存在不公平性,非裔美国人被告再次犯罪的风险估计平均高于白人被告[8].在医学领域,年龄作为一种潜在的敏感属性,会影响基于深度学习诊断系统的评估结果.例如,来自 UCI 机器学习知识库的 Heart Dataset 包含了 906 名不同年龄段患者的 14 个处理过的特征[9].这个数据集的目标是准确地预测一个人是否患有心脏病,而研究发现系统对年龄的偏见可能会导致不必要的医疗护理.在某些简历筛选工具中,存在对性别这一敏感属性产生歧视性行为的现象,导致男性在应聘过程中比女性更有优势.深度学习在应用过程中存在的不公平现象引起了业界和学术界的广泛关注,Du[10]和 Ross 等人[11]使用局部解释对深度模型进行正则化训练从而实现模型的公平;Elazar[12]和 Zhang 等人[13]使用对抗性训练从模型的隐层表示中去除敏感属性的信息,从而得到一个公平的分类器.

与机器学习方法相同,深度学习存在的偏见也是来自于数据和模型.一方面,深度学习是基于数据驱动的学习范式,它使模型能够自动从数据中学习有用的表示.但是这些数据在标注过程中会引入偏见,这些数据偏见被深度模型复制甚至放大.另一方面,深度模型的结构是基于经验设计的,其训练是一个黑盒过程,因此很难确定训练好的模型是基于正确的理由做出的决定,还是受偏见影响做出的不公平判断,这也使得模型去偏成为极具挑战性的任务.

目前,面向深度学习的公平性研究领域还有很大的发展空间,针对来自数据、模型的偏见问题已经成为重点关注对象,仍需要不断的探索.同时由于深度学习在高风险领域中的应用,对数据偏见的预处理去偏、对模型偏见的中处理去偏、以及后验性去偏方法,正在引起业界和学术界的关注.

为了更好地探究深度学习的公平性与未来的发展方向,本文将综述深度学习偏见的不同来源并分类,对预处理去偏方法、深度模型的公平性训练方法以及后验去偏方法进行介绍,并列举目前主流的面向深度学习的去偏平台及去偏方法的公平性评估指标,同时对未来可能的研究方向作出展望.

1 偏见的来源

由于训练数据标注和深度模型结构设计本身存在偏见,会导致深度学习任务的预测结果存在不公平现象.根据偏见的来源不同,我们将偏见类型分为数据偏见和模型偏见.

1.1 数据偏见

训练数据中可能存在由历史社会原因产生的偏见,在有偏见的数据上学习的模型可能会导致预测结果的不公平性.数据的偏见会以多种形式存在,Suresh 等人[14]讨论了数据偏见的不同来源,以及这些偏见的产生方式;Olteanu 等人[15]准备了一份完整的不同类型偏见的列表,并对由于数据偏见而产生的后果进行分析;Mehrabi 等人[16]总结了以上 2 篇论文中引入的一些最普遍数据偏见的来源,但是缺少对偏见来源的细粒度分类.

在本文中,我们将介绍这些数据偏见的定义并进行详细说明,此外还将按照发生的原因对这些数据偏见进行细粒度的分类.我们将其分为时间偏见、空间偏见、行为偏见、群体偏见、先验偏见、后验偏见.

1.1.1 时间偏见

时间偏见是指由于时间维度的差异引起的偏见.例如,在 Twitter 上可以观察到一个例子,人们谈论一个特定的话题时开始使用标签来吸引注意力,然后不使用标签继续讨论该事件[15,17],这是由不同时期人群和行为的差异产生的[15].另一个典型的时间偏见是纵向数据偏见,观察性研究经常把横断面数据当作纵向的.例如,对大量 Reddit 数据的分析显示,评论长度会随着时间的推移而减少[18].

然而,大量的数据代表的是人口的横截面快照,实际上包含了不同年份加入 Reddit 的不同群体.当数据按队列分列时,发现每个队列中的评论长度随时间增加[18].时间偏见可能会导致数据缺失,对后续的分析统计带来困难.

1.1.2 空间偏见

空间偏见主要指的是由数据空间维度产生的偏见,也就是常说的维数灾难.Verleysen 等人[19]指出

基于学习原理的数据分析工具可从学习样本中推断出知识或信息.显然,通过学习建立的模型仅在可获得学习数据的空间范围内有效.模型不可能对与所有学习点都不相同的数据进行概括.

因此,成功开发学习算法的关键要素之一就是要有足够的数据进行学习,以便它们可以填充模型必须包含的空间.在保持其他所有约束不变的情况下,学习数据的数量应随维度呈指数增长,例如,学习二维数据需要 100 个具有相同平滑度的模型;对于 3 维模型,则需 1 000 个.指数级增长是维数灾难后果,这些数据通常会对算法的行为和性能产生不利影响.对于这类偏见,我们通常采用降维的方法进行偏见的减轻.

1.1.3 行为偏见

行为偏见可以分为社会行为偏见[20-21]和用户行为偏见[21].其中社会行为偏见是由社会历史固有的偏见或者他人的行为引起的偏见,可分为社会偏见[20]、紧急偏见[22]、历史偏见[14]、资助偏见[16].社会偏见[18]的产生是由于他人的行为可能会影响我们的判断,例如,用户想要评价或回顾一个得分较低的项目,但当受到其他高评分的影响时,用户可能认为自己太过苛刻,从而会改变自己的评分[20-21].紧急偏见[22]的发生由于人口、文化价值观或社会知识的变化而产生的,这种偏见更可能在用户界面中被观察到,因为通过设计,界面倾向于反映未来用户的能力、特征和习惯.历史偏见[14]是指世界上已经存在的偏见和社会技术问题,即使给定一个完美的采样和特征选择,也会渗透到数据生成过程中.资助偏见[16]是指当公司为了满足资助机构的要求而进行虚假报告,从而出现人为的偏见.例如,当公司的员工为了让资助机构满意而在他们的数据和统计中报告进行杜撰,使报告结果产生偏见.

用户行为偏见[23]源于跨平台、上下文或不同数据集的不同用户行为.这类偏见的典型例子可在 Miller 等人[24]的研究中观察到,其中作者展示了不同平台之间的表情符号表达的差异如何导致人们的不同反应和行为,有时甚至导致交流错误.用户行为偏见可以分为用户交互偏见[23]、内容产生偏见[23]和流行偏见[23].用户交互偏见[23]不仅可以在 Web 上观察到,而且可以从 2 个来源触发——用户界面和通过用户自己选择的偏见行为[16].这种偏见可能会受到其他类型和子类型的影响,比如呈现偏见[20]和排名偏见[20].呈现偏见[20]是信息如何呈现的结果,例如,在 Web 上,用户只能单击他们看到的内容,因此其他内容不会被单击,也可能是用户没有看到 Web 上的所有信息.排名偏见[20]是由于人们认为排名靠前的搜索结果是最相关、最重要的,这种想法会吸引更多的点击量.这种偏见影响了搜索引擎[20]和众包应用程序[25].内容产生偏见[15]源于用户生成的内容在结构、词汇、语义和句法上的差异.例如,Nguyen 等人[26]讨论了不同性别和年龄群体在使用语言方面的差异.流行偏见[27-28]是由于越受欢迎的物品越容易被曝光.这种偏见可以在搜索引擎或推荐系统中看到,在这些系统中,受欢迎的对象会更多地呈现给公众.行为偏见会使用户在决策过程中受到其他外界因素的影响,导致获得的信息不足或者带有偏见,从而产生歧视性行为.

1.1.4 群体偏见

群体偏见[15]产生于数据集或平台中所表示的用户群体中的统计数据、代表数据和用户特征与原始目标群体不同的时候.典型的例子是对于不同社交平台上不同用户的统计数据,女性更倾向于使用 Pinterest、Facebook、Instagram 等社交平台,而男性在 Reddit 或 Twitter 等在线论坛上更活跃.Huang 等人[29]调查了根据性别、种族、民族和父母教育背景划分的年轻人使用社交媒体的例子和数据.

群体偏见可分为聚集偏见[14]和 Simpson 悖论[30].聚集偏见[14]是由于人们观察其他不同的子群体得出错误结论时或者对一个群体的错误假设影响模型的结果和定义时产生的.例如,在临床辅助工具中用于糖尿病诊断和监测的糖化血红蛋白水平在不同性别和种族之间存在复杂的差异.由于这些因素以及它们在不同的子群体中的不同意义和重要性,单一的模型很可能不适合一个群体中的所有群体[14].

Simpson 悖论[30]可能会对由不同行为的子群体或个体组成的异构性数据的分析产生偏见.这类悖论的一个比较著名的例子是对加州大学伯克利分校的性别歧视诉讼[31].在分析了研究生院的招生数据后,可以发现与男性相比,女性被录取为研究生的比例更小.然而,当对各个院系的招生数据进行分析后发现女性申请者具有平等的地位,在某些情况下甚至比男性小有优势.Simpson 悖论在许多领域都得到了观察,包括生物学[32]、心理学[33]、天文学[34]和计算社会科学[35].群体偏见会导致用户得到错误的数据,从而得到错误的结论.

1.1.5 先验偏见

先验偏见发生在我们选择、利用和测量特定特征的方式上.先验偏见可以分为抽样偏见[16]、自我

选择偏见[16]、链接偏见[15]和遗漏变量偏见[16].抽样偏见[16]是由于子组的非随机抽样而产生的,结果是对一个种群估计的趋势可能不能推广到从一个新种群收集的数据.自我选择偏见[16]是抽样偏见的一种亚型,它是指研究对象在这种调查研究中选择自己.例如,在一项关于成功学生的调查研究中,一些不那么成功的学生可能会认为他们是成功的,这就会影响分析的结果.

链接偏见[15]是指当从用户连接、活动或交互中获得的网络属性不同并歪曲了用户的真实行为的现象.Mehrabi 等人[36]指出,仅考虑网络中的链接,而不考虑网络中用户的内容和行为,社交网络会偏向低度节点.Wilson 等人[37]也表明,用户交互与基于特征的社交链接模式有显著差异.网络中的差异可能是许多因素造成的,如网络采样,它可以改变网络度量,导致不同类型的问题[38-39].

遗漏变量偏见[16]发生于当一个或多个重要的变量被排除在模型之外的时候.例如,当公司设计模型来预测老客户继续订阅他们服务的占比,然而很快发现,多数的用户会取消订阅并不遵从设计模型.取消订阅的原因可能是市场上出现了一个新的强有力的竞争对手,它提供同样的解决方案,但价格减半.然而预测模型并没有考虑到竞争者的出现,因此,它被认为是一个被忽略的变量.

1.1.6 后验偏见

后验偏见主要是由于研究人员或观察者行为导致的偏见,可以分为评估偏见[14]、因果偏见[16]和观察者偏见[16].评估偏见[14]发生在研究人员评估过程中,例如,在评价诸如 Adience 和 IJB-A 等应用时,使用不适当的基准,从而造成偏见.因果偏见[16]是由于观察者认为相关性意味着因果关系这一谬论的结果.

例如,公司的数据分析师想要分析顾客的忠诚度有多成功,这位分析师认为,参加了忠诚度计划的顾客比没有参加的顾客在该公司的商店里花更多的钱.这是有问题的,参加忠诚度计划的顾客与计划在此商店花更多钱这一相关性并不意味着它们之间的因果关系.观察者偏见[16]一般发生在研究人员下意识地将他们的期望投射到研究中的时候.当研究人员在采访和调查中无意地影响参与者,或者当他们挑选对他们的研究有利的参与者或统计数据时,这种类型的偏见就会发生.由于观察者的异常或者错误行为会导致后验偏见,从而得到有歧视性的决策结果.

1.2 模型偏见

深度学习算法本身工作方式上存在细微差别,这些差别可能导致深度模型做出不公平的决策.Du 等人[40]从计算的角度将深度模型的不公平性分为预测结果歧视和预测质量差异 2 类.

1.2.1 预测结果歧视的偏见

歧视[41]是指由于某些群体的成员身份,深度模型对这些群体成员产生不利决策结果的现象.深度学习是基于数据驱动的学习范式,它使模型能够自动从数据中学习有用的表示.这些数据中有可能包含偏见,这会导致深度模型复制、甚至放大数据中存在的偏见.更糟糕的是,深度模型不仅依赖这些数据中的偏见来做决策,还会做出毫无根据的联想,放大对某些敏感属性的刻板印象[42-43],这最终会产生具有算法歧视的训练模型.预测结果歧视[40]可以进一步分为输入歧视和表征歧视 2 类.Du 等人[40]对这 2 个子类别进行了详细的描述.

输入歧视是尽管深度模型没有明确地将种族、性别、年龄等敏感属性作为输入,但仍可能导致预测结果的歧视[44].大多数深度模型直接使用原始数据作为输入,因此在输入数据中没有对敏感属性进行分类处理.虽然没有明确敏感属性,但深度模型仍可能表现出无意的歧视,主要是由于存在一些与类成员高度相关的特征[40].例如,邮政编码和姓氏可以用来表示种族,文本输入中的许多单词可以用来推断被预测成员的性别,模型预测过程可能与受保护群体高度相关.最终,模型可能对某些受保护的群体产生不公平的决策.例如,在就业系统中,简历筛选工具认为男性更有优势,对女性存在偏见;贷款批准制度对属于特定邮政编码的人给予负面评价,导致对特定地域的歧视;在刑事司法领域,再犯预测系统预测将黑人囚犯归类为“高风险”的可能性是白人囚犯的 3 倍.

有的时候预测结果歧视需要从表征的角度进行诊断和减轻[40].在某些情况下,将偏见归因于输入几乎是不可能的,例如在图像输入领域,卷积神经网络可以通过视网膜图像识别患者自我报告的性别,并有可能基于性别产生歧视.此外,在某些应用场景中如果输入维度太大,那么查找输入的敏感属性就很困难[43].在这些情况下,某些受保护属性的类成员关系可以在深度模型中表示,模型将根据这些信息做出决策,并产生歧视[40].例如在信用评分中,使用原始文本作为输入,作者的人口统计信息被编码在基于深度模型中间表示的信用评分分类器中.

1.2.2 预测质量差异的偏见

预测质量差异[40]的偏见是指不同受保护群体模型的预测质量差异较大.与其他群体相比,深度模型对某些群体的预测质量较低.预测结果歧视主要涉及高风险领域的应用,而预测质量差异涉及一般领域的应用.例如,在计算机视觉领域[45],对肤色较深的女性面部识别的表现较差;在自然语言处理中[46],语言识别系统在处理某些种族的人产生的文本时表现明显较差;在医疗保健领域[47],重症监护病房死亡率和精神病 30 天再入院模型预测准确度在性别和保险类型之间存在显著差异.这通常是由于训练数据代表性不足导致的问题,在这种情况下,用户对人口的某些方面收集的数据可能不够充足或不够可靠.因为深度模型训练的典型目标是将总体误差最小化,也就是说模型如果不能同时适合群体中的所有个体,它将以适合群体中的大多数个体为目标.虽然这可以最大限度地提高整体模型预测的准确性,但它可能因为缺乏代表性数据从而导致对少数类群体的预测表现出不公平性.

综上所述,我们根据偏见的来源将其分为数据偏见和模型偏见,并进一步将数据偏见分为时间偏见、空间偏见、行为偏见等 6 个子类,将模型偏见分为预测结果歧视和预测质量差异 2 个子类,并且对这些偏见进行了详细的介绍.在表 1 中我们对数据偏见和模型偏见进行列举,如表 1 中的"偏见类型";并表示出它们的子类型,见第 2 列的"子类型";以及这些子类型的组成,见第 3 列的"组成".这些偏见可能发生在不同的阶段,例如,数据本身存在的偏见(表 1 中用"数据阶段"表示)、由于用户行为导致的偏见(表 1 中用"用户行为"表示)以及由于算法细微的差别产生的偏见(表 1 中用"算法阶段"表示),在表中用"√"表示偏见所发生的阶段.此外,在表 1 最后一列"去偏方法"中介绍了上述偏见的常用去偏处理方法,去偏方法的具体内容将在第 2~4 节进行详细介绍.

Table 1 Classification of Bias's Sources

表 1 偏见来源的分类

偏见类型	子类型	组成	数据阶段	用户行为	算法阶段	去偏方法
数据偏见	时间偏见(见 1.1.1 节)	纵向数据偏见[18]	√			数据重采样[48] 类别均衡采样[49] SMOTE[50] 数据增强[51] 代价敏感方法[52] PCA[53] One-hot 编码[54] 对抗性训练[12-13] 模型正则化[10-11]
	空间偏见(见 1.1.2 节)	维数灾难[19]	√			
	行为偏见(见 1.1.3 节)	社会行为偏见[20-21]	√	√		
		用户行为偏见[21]		√		
	群体偏见(见 1.1.4 节)	聚集偏见[14]	√			
		Simpson 悖论[30]	√			
	先验偏见(见 1.1.5 节)	抽样偏见[16]		√		
		自我选择偏见[16]	√	√		
		链接偏见[15]		√		
		遗漏变量偏见[16]		√		
	后验偏见(见 1.1.6 节)	评估偏见[14]		√	√	
		因果偏见[16]		√		
		观察者偏见[16]		√	√	
模型偏见	预测结果歧视(见 1.2.1 节)	输入歧视[40]	√		√	对抗性训练[12-13] 模型正则化[10-11] ADF[55] THEMIS[56] AEQUITAS[57] 符号生成[58]
		表征歧视[40]			√	
	预测质量差异(见 1.2.2 节)	缺乏代表性数据[40]			√	
	预测质量差异(见 1.2.2 节)	表征歧视[40]			√	
		缺乏代表性数据[40]			√	

2 基于预处理的数据去偏方法

预处理技术通过对数据进行处理,以减轻预测模型潜在的歧视.如果允许算法修改训练数据,则可以使用预处理技术[59].例如,可以通过获取更多数据来扩充数据集,对于代表性不强的数据集,更多的数据往往能得到更多的分布信息.数据预处理去偏

可以分为数据层面处理方法和算法层面处理方法.

2.1 数据层面去偏处理方法

数据层面处理方法多借助数据采样法等方法使整体训练集样本趋于平衡,从而达到去偏效果.常用的方法有数据重采样、类别均衡采样、数据合成、数据增强等.

2.1.1 基于数据重采样的去偏方法

对数据集重新采样,即向数据集中添加新元素或删除现有元素[48].Burnaev 等人[48]通过实验研究了重采样对分类准确性的影响,比较了重采样方法并突出了重采样的重点和难点.对于数据集 S 中类的不平衡,作者引入不平衡率 $IR(S)=\frac{|C_0(S)|}{|C_1(S)|}$ 来衡量,其中,$C_0(S)=\{(X_i,y_i)\in S\mid y_i=0\}$,$C_1(S)=\{(X_i,y_i)\in S\mid y_i=1\}$.若不平衡率 $IR(S)\geqslant 1$ 或更高,则数据集 S 越不平衡.数据重采样的方法分 2 步完成:首先,使用重采样方法 r 对数据集 S 进行重采样,即丢弃 S 中的某些观测值或向 S 添加一些新的合成观测值.然后,在 $r(S)$ 上学习一些标准分类模型 h,从而得出分类器 $h_{r(S)}:\mathbb{R}\rightarrow\{0,1\}$.

Burnaev 等人[48]的实验结果表明重采样对数据集质量的影响在很大程度上取决于重采样乘数,并且重采样方法的性能取决于所使用的分类器,此方法在人工数据集上的去偏效果要好于真实数据集.如果正确选择了方法,那么在大多数情况下,重采样可以改善不平衡数据集的分类,从而达到去偏的效果.但是通过重采样来对数据集进行去偏并不是总能达到预期效果,在某些情况下,数据重采样可能会引入大量重复样本,会减慢训练速度,使模型在过采样时容易过拟合,或者丢弃重要的重要示例.

2.1.2 基于类别均衡采样的去偏方法

样本类别分布不均衡也是导致深度模型不公平的一个原因,类别均衡采样是解决这类问题一个方法.常用的类别均衡方法就是根据每个类别的观察次数重新采样和重新加权.Cui 等人[49]认为随着样本数量的增加,新添加的数据点带来的好处将减少.他们提出了一种新颖的理论框架,通过将每个样本与其较小的邻域相关联来测量数据.有效样本数通过简单公式 $(1-\beta^n)/(1-\beta)$ 来计算,其中 n 是样本数,$\beta\in[0,1)$ 是超参数.Cui 等人[49]设计了一种重新加权方案,该方案使用每个类的有效样本数来重新平衡损失,从而产生类平衡的损失.

类别均衡采样方法可以使不平衡样本分布均衡,从而达到数据去偏的效果.但是,这种方法可能会破坏原属性的线性关系,改变原样本的某些特征值.此外,Shrivastava[60]等人提出了 OHEM 方法对样本不平衡的问题进行处理.

2.1.3 基于合成数据的去偏方法

Chawla 等人[50]提出了一种叫做 Synthetic Minority Over-sampling Technique(SMOTE)的合成数据的方法.SMOTE 通过创建“综合”示例而不是通过替换来对少数群体进行过采样.通过获取每个少数种群样本以及基于距离度量选择类别下 2 个或者更多的相似样本引入综合示例,对少数种群进行过采样.

合成数据是通过以下方式生成的:取所考虑的特征向量(样本)与其最近邻域之间的差,将该差乘以 0 到 1 之间的一个随机数,并将其添加到所考虑的特征向量中.这将导致沿着 2 个特定特征之间的线段选择一个随机点,这样就构造了许多新数据.Chawla 等人[50]的实验结果表明 SMOTE 方法可以提高少数群体分类器的准确性.SMOTE 不仅提供了一种新的过采样方法,并且 SMOTE 和欠采样的组合比纯欠采样性能更好.合成数据这一去偏方法对数据量较少数据集的去偏效果较好,同时还能提高分类器的准确性,但是合成数据可能会引入重复样本.

2.1.4 基于增强数据的去偏方法

数据增强[51](data augmentation)针对有限数据问题的数据空间提供解决方案,包含一套技术可用于加强深度学习所使用的数据集的大小和质量,从而给用户提供更好的深度学习研究条件.

使用数据增强技术可以构建模型.例如,当输入数据集是图像时,可以应用图像数据增强图像方法.该增强方法包括几何变换、色彩空间增强、抖动、混合图像、随机擦除、特征空间增强、对抗训练、生成对抗网络、神经样式转换和元学习等算法.数据增强旨在增加样本数量,当数据量以及多样性很少的情况下是非常有效的,但它无法克服小型数据集存在的所有偏差,例如,在犬种分类任务中,如果只有斗牛犬并且没有金毛寻回犬,则数据增强方法不会创建金毛寻回犬.但是,使用数据增强可以避免或至少可以大大减少偏差的几种形式,例如照明、遮挡、缩放等.数据增强的不足之处是可能引入重复样本.

2.2 算法层面去偏处理方法

算法层面处理不平衡样本问题的方法有代价敏感、主成分分析、One-hot 编码等.

2.2.1 基于代价敏感的去偏方法

基于代价敏感[52]的去偏方法是使用代价来调整分类器的权重,代价敏感的特性能够在分类器上得到满足.若某个训练集存在 N 个样本,形如$[x_n, y_n]_{n=1}^N$,所谓的代价敏感方法是指利用 $K\times K$ 的矩阵 $\boldsymbol{C}$ 对不同样本类别施加权重[52].$C(y_i, y_j)\in[0,\infty)$表示类别 y_i 错分为类别 y_j 的惩罚[52].训练目标被施加代价后将会变为:训练得到某分类器 g 使得期望之和 $\sum_{n=1}^{N}C(y_n, g(x_n))$ 最小[52].

为了能够对少数类样本进行比较准确的识别,可采用基于代价敏感学习的方法,将少数类视为重要类别,并令其错分代价大于多数类的错分代价.

2.2.2 基于主成分分析的去偏方法

主成分分析[53](principal component analysis, PCA)是一种线性、无监督、生成和全局特征学习方法,可以对空间偏见进行减轻.它是通过创建新的不相关变量来实现的,从而连续地最大化方差.查找主成分变量的过程可以简化为求解特征值以及特征向量的问题,并且新变量是通过现有的数据集定义的,而不是先验的,因此 PCA 是自适应的数据去偏分析技术.从另一种意义上说,它也是自适应的,因为已经开发了针对各种不同数据类型和结构量身定制的技术变体.

2.2.3 One-hot 编码

One-hot 编码的操作十分简单,从业人员经常将其用作更复杂技术的第一步.One-hot 编码[54]定义如下:令 x 为具有 n 个不同值 $x_1, x_2, \cdots, x_n$ 的某个离散类别随机变量.然后,特定值 x_i 的 One-hot 编码是向量 $\boldsymbol{v}$,其中 $\boldsymbol{v}$ 中第 i 个分量值为 1,其余每个分量均为零.例如,假设我们有一些随机变量 x 取自设置 $S=\{a,b,c\}$.令 $x_1=a$, $x_2=b$ 和 $x_3=c$.x 的一次编码为:(1,0,0),(0,1,0)和(0,0,1).由于分类变量级别的 One-hot 编码仅取决于级别的数量,因此 One-hot 编码属于确定的用于编码分类变量的技术,可以用于神经网络.对数据进行预处理去偏时,通常要确定 2 个相似个体特定特征之间的度量距离,One-hot 编码能更加合理的计算特征之间的距离,从而达到去偏的效果.Ruoss 等人[61]使用 One-hot 编码对数据进行预处理.

除了以上介绍的一般数据预处理的方法外,各类文献中也提出了各种方法.为了减轻数据偏见对最终决策带来的影响,Benjamin[62],Gebru[63]等人将数据表作为数据的支持文件来报告数据集创建方法、其特征、动机及其偏见.Holland 等人[64]提出了标签,就像食品上的营养标签一样,以便更好地对每个任务的每个数据进行分类.除了这些一般的技术,一些工作还针对更具体类型的偏见.例如,Alipourfard[65],Zhang 等人[66]提出了自动发现数据中 Simpson 悖论的方法.在一些工作中,因果模型和图表也被用于检测数据中的直接歧视,以及对数据进行修改的预防技术,以使预测不存在直接歧视.Hajian 等人[67]还致力于防止数据挖掘中的歧视,针对直接歧视、间接歧视和同时产生的影响.生成式对抗网络可用于为少数生成合成数据,这可以提高少数群体的预测质量,同时又不影响未受保护群体的预测性能,从而避免对这些群体的歧视.

3 深度学习模型去偏

在介绍了数据预处理方法之后,我们在本节中将介绍深度学习模型去偏方法,确保深度学习模型的公平性.模型去偏方法通常可以分为模型正则化和对抗性训练 2 类.前者通过在总体目标函数中添加辅助正则化项来实现,显式或隐式地对某些公平性度量施加约束,后者可以从深度模型的中间表示中去除敏感属性的信息,从而得到一个公平的分类器.

3.1 基于模型正则化的去偏方法

正则化是模型去偏的一种方法,具体来说,使用局部解释对模型训练进行正则化训练[10-11].对于整个输入 x,除了真值 y 之外,这种正则化还需要特性方面的注释 r,指定输入中的每个特性是否与受保护的属性相关,r 可以进一步融入到训练过程中,目的是使深度模型更加公平.正则化的总损失函数如式(1)所示:

$$L(\theta,x,y,r)=d_1(y,\hat{y})+\lambda_1 d_2(f_{\text{loc}}(x),r)+\lambda_2 R(\theta), \tag{1}$$

其中,d_1 为正态分类损失函数,$R(\theta)$为正则化项.函数 $f_{\text{loc}}(x)$是局部解释方法,d_2 是距离度量函数.这 3 个术语分别用于指导深度模型进行正确的预测,超参数 λ_1 和 λ_2 用于平衡这 3 个术语.

例如,Du 等人[10]采用一种名为 CREX(CRedible EXplanation)的方法对深度模型进行正则化训练,使用的损失函数如式(2)所示:

$$L(\theta,x,y,r)=L_{\text{sup}\,v}+\lambda_1 L_{\text{rationale}}+\lambda_2 L_{\text{sparse}}, \tag{2}$$

其中,作者使用的正态分类损失函数为交叉式损失 $L_{\text{sup}\,v}$.CREX 的核心思想是深度模型应该依靠合理的证据来做出决定.CREX 的示意图如图 1 所示.

在图 1 中,黑色实线表示向前的路径,两端带箭头的虚线是损失,一侧带有箭头的虚线表示坡度流.$\boldsymbol{x}_n$,$\boldsymbol{r}_n$,$\boldsymbol{y}_n$ 三个向量从左到右分别是输入、解释和基本原理.

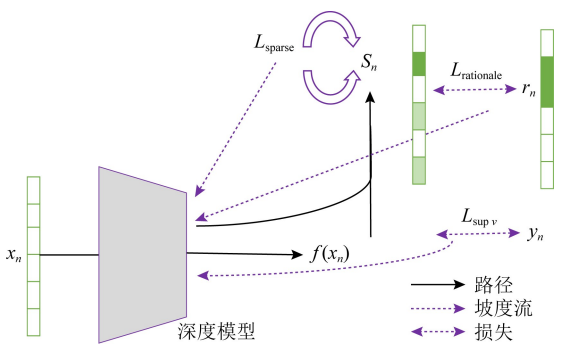


Fig. 1 Schematic of CREX[10]

图 1 CREX 示意图[10]

另一个典型的例子是对式(1)中的 $f_{loc}(x)$ 使用输入梯度解释方法$\frac{\partial \hat{y}}{\partial x}$,$L_2$ 范数作为距离度量 d_2[11].直观上是缩小模型对与敏感属性高度相关的特征的关注,从而减少模型对这些公平敏感特征的依赖.由此产生的公平模型更多地依赖于与任务相关的整体信息,同时较少地依赖于敏感属性.

3.2 基于对抗性训练的去偏方法

从模型训练的角度来看,对抗性训练是一种典型的解决方案,可以从深度模型的中间表示中去除敏感属性的信息,从而得到一个公平的分类器[12-13,68].其目标是学习一种高级输入表征,该表征对主要预测任务具有最大信息量,同时对受保护属性具有最小预测性.对抗性训练过程可以表示为式(3):

$$\arg\min_{g} L(g(h(x),z)),$$
$$\arg\min_{h,c} L(c(h(x)),y)-\lambda L(g(h(x))). \quad (3)$$

深度模型可以记为 $f(x)=c(h(x))$,其中,$h(x)$ 是输入 x 的中间表示,$c(\cdot)$负责将中间表示映射到最终的模型预测.$f(x)$可以是通过反向传播学习的任意深度模型.要检查的受保护属性使用 z 表示,主任务 $f(x)=c(h(x))$本身并没有与受保护的属性 z 进行排序.构造了一个对抗性分类器 $g(h(x))$,从表示 $h(x)$中预测受保护属性 z.训练是在 $f(x)$和对抗性分类器 $g(h(x))$之间迭代进行的.经过一定的迭代次数,我们可以得到去偏的深度模型.

图 2 为对抗性训练的示意图,利用对抗性训练,通过表示减少歧视.直觉上是通过加强深度表示来最大限度地预测主要任务标签,同时最小限度地预测敏感属性.

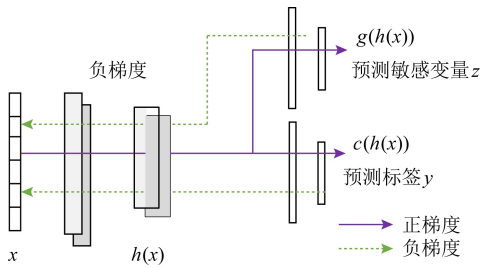


Fig. 2 Schematic of adversarial training[40]

图 2 对抗性训练示意图[40]

对抗训练广泛适用于不同的深度模型架构和不同的输入格式,包括带有图像数据的卷积神经网络[69]、带有文本数据的循环神经网络[70]、以及带有分类数据的多层神经网络[71].Zhang 等人[13]提出了一个模型用来减轻从关联的数据中学习到的模型中的偏见.在这个模型中,他们试图最大限度地提高 y 的预测精度,同时最小化对手预测受保护或敏感变量的能力.

3.3 其他模型去偏方法

除深度模型外,在传统的机器学习中也有很多模型去偏的方法.Bolukbasi 等人[41]注意到在词类比测试中使用最新的词嵌入时,“男人”将被映射为“计算机程序员”,而“女人”将被映射为“家庭主妇”.针对这种对女性的偏见,作者提出一种词嵌入的方法,该方法通过遵循以下步骤对性别中性词进行嵌入:首先识别性别子空间,然后确定捕获偏见的嵌入方向[41],最后将性别子空间与性别中性词分开,并确保在性别子空间中将所有性别中性词都删除并归零[41].

Zhao 等人[43]研究了语义角色标签模型和数据集 imSitu,并发现在 imSitu 中烹饪图像中只有 33% 的代理角色是男人,其余 67%的烹饪图像中有女性代理角色.除了数据集中现有的偏见外,该模型还会放大偏见.因此,他们提出了一种称为 RBA(reducing bias amplification)的校准算法.RBA 是一种通过在结构化预测中校准预测来消除模型偏见的技术,其思想是确保模型预测在训练数据中遵循相同的分布.

在人工智能领域中已经提出了各种方法来消除偏见的影响,这些方法大多数都试图避免敏感或受保护属性对决策过程的影响.Lipton 等人[72]提出通过在训练阶段允许使用受保护属性,但避免在预测期间使用受保护属性,他们认为通过这种方法可以减轻偏见.Louizos 等人[73]使用变分编码器对表示

学习进行去偏.Mehrabi 等人[36]提出了一种新的社区检测方法,以减轻模型对在线社交社区中处境不利群体的损害.这些对其他模型去偏方法目前还没有广泛应用于深度模型.

4 深度学习的后验去偏

深度学习模型后验性去偏使用可解释技术作为一种有效的工具,用户可以利用可解释技术生成特征重要度向量,然后对特征重要度向量进行分析,从而达到去偏的效果.在本节中,我们先介绍深度模型可解释性的 2 种分类,然后对后验性去偏方法进行介绍.

4.1 深度模型的可解释性

可解释性可以作为一种有效的调试工具,对模型进行分析,最终提高模型的透明度,保证模型的公平性.深度模型可解释性一般可分为 2 类:局部解释和全局解释,这取决于目标是局部理解特定的预测,还是全局理解的预测[43].

4.1.1 深度模型的局部解释技术

局部解释可以说明模型是如何对特定输入进行某种预测的(图 3(a)).它是通过对模型的输入特征进行属性预测来实现的,最后以特征重要度可视化的形式进行说明.以贷款预测为例,该模型的输入是一个包含分类特征的向量,其中得分较高的特征表示与分类任务的相关度较高.局部解释方法大致可分为 4 类:基于局部逼近的[41]、基于扰动的[41]、基于反向传播的[70]以及基于分解的[70]方法,这些方法都可以被用来生成一个输入的特征重要向量.

4.1.2 深度模型的全局解释技术

全局解释的目标是提供一个关于预先训练的 DNN 所捕获的知识,并以一种直观的方式向人类说明所学的表示(图 3(b)).解释可以看作是一个函数 $f_{\text{global}}: E_h \rightarrow E_m$,从中间表示 E_h 映射到人们可以理解的概念 E_m[74].在本例中,E_h 是由特定层上的特定通道派生的表示,多个神经元的组合可以代表更抽象的概念[75].这里的 E_h 对应着不同通道甚至不同层的组合,特别是那些受保护的概念,它们通常基于多个基本的低级概念.例如,在人脸图像识别应用中,可以通过多个局部线索来显示性别和种族概念.因此,与单个神经元学习的概念相比,由多个神经元组合产生的概念与深度模型的公平性更相关.

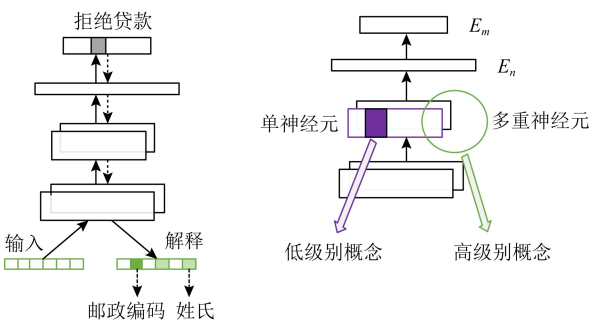


Fig. 3 Illustration of DNN local interpretation as well as global interpretation[40]

图 3 DNN 局部解释和全局解释的示意图[40]

4.2 模型的后验去偏方法

深度学习模型后验性去偏方法主要是搜索模型中的歧视实例,通过检测偏见来进行模型再训练,以减少歧视,达到模型的公平性.第 1 种方法采用自顶向下的方法,利用局部解释生成特征重要度向量,然后对特征重要度向量进行分析.第 2 种解决方案以自底向上的方式实现.人们首先预先选择他们怀疑与受保护属性相关联的特性,然后分析已识别的特性的重要性[76].这些对公平性敏感的特征被干扰,通过特征被直接删除或特征被替代来实现.然后将扰动输入到深度模型中,观察模型预测的差异.如果这些被怀疑为公平敏感特征的扰动最终导致模型预测发生显著变化,则可以断言深度模型捕获了偏见,并根据受保护的属性进行决策.第 3 种方法利用全局解释,首先,利用全局解释来分析深度模型对受保护属性相关概念的学习程度.这通常是通过指向深度模型中间层激活空间的一个方向来实现的[74-75,77].其次,在确认一个深度模型已经学习了一个受保护概念后,我们将进一步测试该概念对模型最终预测的贡献.可以采用不同的策略来量化概念敏感度,包括自上而下计算深度模型预测对概念向量的方向导数[74],自下而上将该概念向量添加到不同输入的中间激活中,观察模型预测[78]的变化.最后,使用数值分数来描述受保护属性的表示偏见水平.在 2 种方式中,数值敏感性得分越高,该概念对深度模型预测的贡献越显著.

Zhang 等人[55]提出了一个基于梯度的可扩展的算法,称为 ADF(adversarial discrimination finder),用于生成个体歧视实例,它是专门为深度模型设计的.ADF 的概述如图 4 所示.

ADF 由 2 部分组成,即全局生成(左边的部分)和本地生成(右边的部分).在全局生成过程中,对原始数据集中的样本进行聚类,并以循环方式从每个聚类中选择种子实例.全局生成的目标是增加所生成

的个体歧视实例的多样性.在全局生成中使用梯度通过最大化 2 个相似实例的深度模型输出之间的差异来指导个体歧视实例的生成.如果成功生成了一定数量的个别歧视性实例或超时,则全局生成将停止.识别出的个别歧视实例然后作为本地生成的输入.其思想是搜索个体歧视性实例的邻域以寻找更多的歧视性实例.梯度在本地生成中以不同的方式使用作为引导,即我们利用代表每个属性重要性的梯度的绝对值来识别与种子差异最小的个体歧视性实例,同时保持它们的模型预测[55].

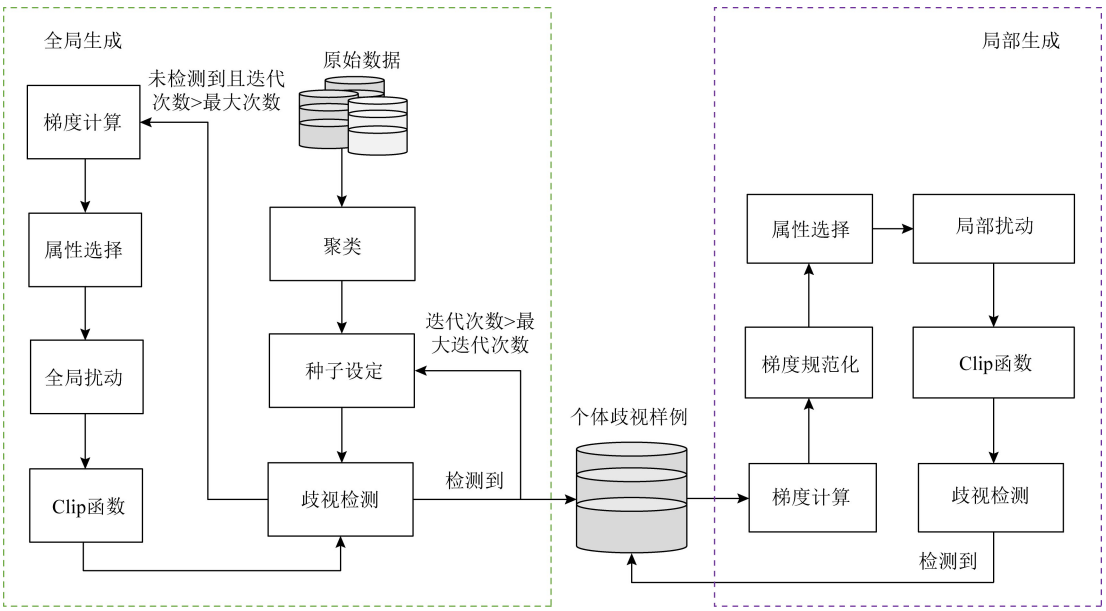


Fig. 4 An overview of ADF[55]

图 4 ADF 概述图[55]

除此之外,Galhotra 等人[56]提出了 THEMIS,通过在其域内随机采样每个属性并识别出那些有偏见的实例来衡量歧视的发生频率.Udeshi 等人[57]开发了 AEQUITAS,它包括一个全局搜索和一个本地搜索,即 AEQUITAS 首先搜索输入空间的随机抽样(又名全局搜索),然后基于全局搜索的结果来进行本地搜索,通过将已识别的个体歧视实例与选定的属性沿随机方向进行干扰,以识别尽可能多的歧视的实例.Agarwal 等人[58]提出了一种符号生成的方法,该方法首先使用现有的方法生成一个局部解释决策树来近似模型决策,然后根据决策树进行符号执行来生成测试用例[79].与 AEQUITAS 一样,它还将基于决策树的全局搜索与局部搜索相结合,前者的目标是最大化路径覆盖,后者的目标是最大化歧视性实例的数量.

对第 2~4 节介绍的基于数据预处理的去偏方法、模型去偏方法以及后验去偏方法及其相关原理整理在表 2 中.表格的第 1 列"类型"为 3 类不同的去偏方法,分别为基于数据预处理的去偏方法、深度学习模型去偏方法以及模型的后验去偏方法,根据这些去偏方法应用的阶段不同,我们在表格中将它们表示为"数据预处理去偏、模型去偏、后验去偏".表格的第 2 列"方法"列举了相应的算法,其中列举了 8 种预处理方法、2 类模型去偏方法以及 4 种后验性去偏方法,并且在第 3 列"原理"表示出对应算法的原理,在第 4 列"贡献"和第 5 列"不足"分别列举出它们的贡献和不足.

5 去偏实验平台和公平性指标

在本节中,我们列举常用的去偏实验平台和公平性指标,方便之后的研究.近年来,随着人工智能技术的快速发展,许多科技公司推出了对应的深度学习去偏实验平台.其中 Microsoft 的 Fairlearn、IBM 的 AI Fairness 360 以及 Google 的 ML-fairness-gym 具有功能较全、丰富的演示代码以及持续迭代更新等显著特点.本节将详细介绍这 3 个去偏实验平台.

Table 2 List of Debiasing Methods for Different Sources of Bias

表 2 针对不同偏见来源的去偏方法列表

类型	方法	原理	贡献	不足
数据预处理去偏	数据重采样[48](见 2.1.1 节)	向数据集中添加新元素或删除现有元素	改善不平衡数据集的分类	减慢训练速度,在过采样时容易过拟合
	类别均衡采样[49](见 2.1.2 节)	把样本按类别分组	保证每个类别参与训练的机会均等	可能会破坏原属性的线性关系
	合成数据[50](见 2.1.3 节)	通过创建"综合"示例来对少数群体进行过采样	增加样本数量	可能会引入重复样本
	数据增强[51](见 2.1.4 节)	通过翻转、旋转、平移、缩放等方式构造数据.	增加样本数量	可能会引入重复样本
	代价敏感方法[52](见 2.2.1 节)	将代价用于权重的调整	准确识别少数类样本	对多数类样本效果不佳
	PCA[53](见 2.2.2 节)	将特征的主要成分找出,并去掉无关的成分	达到降维的目的	对非线性数据效果不佳
	One-hot 编码[54](见 2.2.3 节)	使用 N 位状态寄存器来对 N 个状态进行编码	使分类器能够更好处理属性数据	使数据变得稀疏
模型去偏	模型正则化[10,11](见 3.1 节)	使用局部解释对模型训练进行正则化训练	减少模型对公平敏感特征的依赖	很难校准
	对抗性训练[12-13](见 3.2 节)	构造对抗性分类器	对深度模型进行去偏	可能降低准确性
后验去偏	ADF[55](见 4.2 节)	基于梯度的可扩展的算法	用于生成个体歧视实例	步长参数依赖于训练数据集
	THEMIS[56](见 4.2 节)	在其域内随机采样每个属性	通过识别出偏差实例来衡量歧视的发生频率	识别出的歧视实例较少
	AEQUITAS[57](见 4.2 节)	基于全局搜索的结果来进行本地搜索	识别歧视实例	搜索空间较小
	符号生成[58](见 4.2 节)	生成一个局部解释决策树来近似模型决策	识别歧视实例	搜索时间较长

5.1 数据集

本文提到的去偏实验平台应用了 Adult 数据集、Bank marketing 数据集、Boston 房价数据集、COMPAS 数据集、Greman credit 数据集、医疗支出小组调查(Medical expenditure panel survey, MEPS)等 6 个数据集.

Adult 数据集包括 48 842 个连续或者离散的实例,其中训练集实例 32 561 个,测试集实例 16 281 个,该数据集可用于预测一个人的年收入是否多于 5 万美元;该数据集包括年龄、工种、学历、职业、性别、种族等 14 个特征,其中 6 个连续变量,8 个名词属性变量,其中性别和种族是敏感属性.

Bank marketing 数据集与葡萄牙的银行有直接关联,根据相关信息进行电话推销,与该数据集对应的任务是分类任务,目的是用于预测客户是否会认购定期存款;该数据集包括年龄、工作类型、婚姻状况、受教育背景、信用情况、个人贷款、最后联系月份、最后一次接触距离上一次接触的时间,以前的活动中联系的次数等一共 50 个特征以及 41 188 个实例,其中年龄和信用情况是敏感属性.

Boston 房价数据集包含美国人口普查局收集的美国 Boston 住房价格的有关信息,这个数据集的每一行数据都是对波士顿周边或者城镇的房价的描述数据统计于 1978 年,数据中包含 14 个特征,506 个案例.特征例如城镇人均犯罪率、住宅所占比例、城镇中黑人比例、低收入人群数等,其中城镇中黑人比例是敏感属性.

COMPAS 数据集使用一种算法来评估刑事被告再次犯罪可能性,开发者为一般累犯和暴力累犯以及审前不当行为制定了风险量表,这种风险量表是根据"与累犯和犯罪职业高度相关的"行为和心理结构设计的,目前已经在美国纽约州、威斯康星州、加利福尼亚州等地投入使用,其敏感属性为种族属性.

Greman credit 数据集通过一组属性描述将申请人员分类为良好或不良信用风险,该数据集是根据个人的银行贷款信息和申请客户贷款逾期发生情况来预测客户贷款违约情况,数据集包含 24 个维度的共 1 000 条数据.该数据集包括年龄、借款持续时间、现有的信贷数量等特征,其中现有的信贷数量和年龄是敏感属性.

MEPS 数据集始于 1996 年,其收集内容包括对家庭和个人、医疗提供者和雇主的大规模调查,并提供有关受访者使用的医疗服务、服务的成本和频率、人口统计等数据.

5.2 Microsoft 公司的 Fairlearn 去偏实验平台

Microsoft 推出了 Fairlearn[①](version 0.4.6)工具包,它能够评定和改正人工智能技术系统软件的公平性.可让人工智能系统的开发人员评估其系统的公平性并减轻任何客观存在但是不明显的不公平问题.全世界四大会计师会计师事务所之一的安永,在用于全自动评定借款管理决策的机器学习模型中,运用 Fairlearn 工具包来减少与性别有关的不合理结果,其剖析数据显示最初男士借款的成功率比女士高 15.3 个百分点.根据正模型,安永的开发设计工作组改善了计划方案的精确度,将性别导致的差别降至了 0.43 个百分点.

Fairlearn 去偏实验平台涵盖阈值优化器[80]、网格搜索[81]以及幂梯度[82]等去偏算法.其中,阈值优化器算法基于监督学习中机会均等原理,将现有分类器和敏感特征作为输入,对深度模型进行后处理去偏.

该平台还包含群体总结、平均预测以及选择率等 16 种度量指标,用来衡量深度模型去偏效果.其中平均预测度量指标用于计算(加权)平均预测结果,选择率计算与输出"良好"结果相匹配的预测标签的比例.

5.3 IBM 公司的 AI Fairness 360 去偏实验平台

IBM 公司发布的 AI Fairness 360 工具包[②](version 0.3.0)是一种可扩展的开放源代码库,可帮助检测和减轻整个 AI 应用程序生命周期中机器学习模型的偏见.

AI Fairness 360 去偏实验平台涵盖一共 11 种去偏算法,例如优化预处理[83]、不同影响消除[84]、均等赔率后处理[80]、校准后的均等赔率后处理[85]、学习公平表示[86]、对抗性去偏[13]、公平分类的元算法[87]、重新加权[88]、基于拒绝选项的分类[89]、正则化去偏[90]等去偏算法.对抗性去偏是一种过程中去偏的技术,学习分类器通过对抗生成的方式以最大程度地提高预测准确性,同时降低对手根据预测确定受保护属性的能力,因为预测结果不可以携带任何敌手可以利用的会造成歧视的信息,因此保证了公平性.

该平台还包含超过 30 种公平性度量指标,所有的度量指标可以根据选择率和错误率分为如下 4 类:全面的群体公平性度量标准、全面的样本失真指标集、广义熵指数[91]以及差异公平和偏见放大[92]等.在这 4 类度量中具有代表性的有群体总结、平均预测以及选择率等度量指标,这些度量指标在 5.2 节已经介绍.

5.4 Google 公司的 ML-fairness-gym 去偏实验平台

Google 提出了 ML-fairness-gym[③](version 0.1.0),用于评估机器学习系统的公平性以及评估静态数据集上针对系统的各种输入的误差度量的差异.ML-fairness-gym 是用于构建简单模拟的一组组件,这些模拟探索了在社会环境中部署基于机器学习的决策系统的潜在长期影响.随着机器学习公平性的重要性变得越来越明显,最近的研究集中在最初在静态环境中定义的执行公平性度量的潜在的令人惊讶的长期行为.

ML-fairness-gym 去偏实验平台包括注意力分配[93]去偏算法以及长期公平[94]去偏算法,注意力分配算法通过对深度模型动态分配不同的注意力权重以避免包含偏见较大的部分参与总体决策,从而实现公平性.该平台也包括错误率指标、借贷指标以及价值追踪指标等公平性度量指标.

我们在表 3 中对以上 3 种去偏实验平台所使用的数据集、度量标准以及平台所支持的去偏算法进行分类整理.因篇幅有限,在这里仅列举出部分具有代表性的度量标准.

6 未来研究方向

本文面向深度学习的公平性进行了尽可能全面的调研,对去偏实验平台以及公平性指标进行了介绍.本节我们针对深度学习中的公平性,探讨其在未来的研究发展方向,从不同角度分析之后可发展的研究内容.

6.1 公平性的度量标准

我们在第 5 节中对国际主流去偏实验平台中的公平性度量标准进行了介绍,但是目前关于公平的度量方法仍然没有形成共识.在某些情况下,一些度量可能与其他度量相冲突.一个模型可能在某一指标上是公平的,但可能导致其他类型的不公平,所以探讨公平性的度量标准是有必要的.

① https://github.com/fairlearn/fairlearn

② https://github.com/Trusted-AI/AIF360

③ https://github.com/google/ml- fairness-gym

Table 3 Debiasing Experiment Platform

表 3 去偏实验平台

公司	平台	数据集	算法	#度量指标
Microsoft	Fairlearn	Adult Bank marketing Boston 房价	阈值优化器[80] 网格搜索[81] 幂梯度[82]	16(例如群体总结、平均预测、选择率等)
IBM	AIF 360	Adult Bank marketing COMPAS German credit 医疗支出小组调查(MEPS)	优化预处理[83] 不同影响消除[84] 均等赔率后处理[80] 校准后的均等赔率后处理[85] 学习公平表示[86] 对抗性去偏[13] 公平分类的元算法[87] 重新加权[88] 基于拒绝选项的分类[89] 正则化去偏[90] 子群体公平[79]	30+(例如样本失真、广义熵指数[91]、偏见放大[92]等)
Google	ML-fairness-gym	Adult	注意力分配[93] 长期公平[94]	3(错误率、借贷指标、价值追踪指标)

6.2 联邦学习公平性问题

联邦学习是一个机器学习框架,能有效帮助多个机构在满足用户隐私保护、数据安全和政府法规的要求下,进行数据使用和机器学习建模[95].在联邦成员共享加密的模型参数或者中间计算结果的同时,也会共享各自存在的偏见,甚至是偏见叠加.对于联邦学习的公平性,我们可以在联邦环境下进行边缘端偏见检测,首先分析不同的联邦成员在上传加密的模型参数或者中间结果时,对其中携带的偏见信息进行检测;然后,分析云端在下发共享参数信息时,检测云端训练的模型从成员中学到的叠加偏见.

6.3 迁移学习的公平性问题

迁移学习[96]是一种机器学习方法,就是把为任务 A 开发的模型作为初始点,重新使用在为任务 B 开发模型的过程中.迁移学习让 AI 系统获得"举一反三"能力,但是从源域到目标域的迁移过程中,极大可能存在偏见的转移.针对迁移学习中的公平性问题,可以从数据偏见转移、算法偏见转移、迁移的新增偏见 3 方面展开研究.在基于实例和基于特征的迁移中,研究数据的偏见转移,对目标域的公平性影响.首先检测源域中数据集存在的偏见,检测目标域中数据集存在的偏见;对源域和目标域中的数据进行偏见对齐,得到偏见分布相似的数据集;使用偏见对齐的目标域中的数据进行迁移训练,检测目标模型与源模型的偏见差异,若偏见评价结果相等或更小,则有效消除数据偏见.

6.4 元学习的公平性问题

元学习利用以往的知识经验来指导新任务的学习,具有学会学习的能力[97].在基于记忆的元学习中,网络的输入把上一次的 y 也作为输入,并且添加了外部记忆存储上一次的 x 输入,这使得下一次输入后进行反向传播时,可以让 y 和 x 建立联系,使得之后的 x 能够通过外部记忆获取相关图像进行比对来实现更好的预测[97].因此在历史记忆中存在的偏见可能会不断积累,对该偏见的消除十分重要.对于元学习的公平性,我们可以对记忆单元设计不同的权重分配策略,减弱历史偏见的积累.

参 考 文 献

[1] Haralick R M, Shanmugam K, Dinstein I. Textural features for image classification [J]. IEEE Transactions on Systems, Man, and Cybernetics, 1973, SMC-3(6): 610-621

[2] Fu Kang, Cheng Dawei, Tu Yi, et al. Credit card fraud detection using [C] //Proc of the 23rd Neural Information Processing Int Conf. Berlin: Springer, 2016: 483-490

[3] Zhang Ying, Wang Chao, Guo Wenya, et al. Multi-source news comment sentiment prediction based on two-way hierarchical semantic model [J]. Journal of Computer Research and Development, 2018, 55(5): 933-944 (in Chinese)
(张莹, 王超, 郭文雅, 等. 基于双向分层语义模型的多源新闻评论情绪预测[J]. 计算机研究与发展, 2018, 55(5): 933-944)

[4] Schroff F, Kalenichenko D, Philbin J. Facenet: A unified embedding for face recognition and clustering [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 815-823

[5] Xin Yu, Yang Jing, Tang Chuheng, et al. An overlapping semantic community detection algorithm based on local semantic cluster [J]. Journal of Computer Research and Development, 2015, 52(7): 1510-1521 (in Chinese)
(辛宇, 杨静, 汤楚蘅, 等. 基于局部语义聚类的语义重叠社区发现算法[J]. 计算机研究与发展, 2015, 52(7): 1510-1521)

[6] Bojarski M, Del Testa D, Dworakowski D, et al. End to end learning for self-driving cars [J]. arXiv preprint, arXiv: 1604.07316, 2016

[7] Vieira S, Pinaya W H L, Mechelli A. Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications [J]. Neuroscience & Biobehavioral Reviews, 2017, 74: 58-75

[8] Wadsworth C, Vera F, Piech C. Achieving fairness through adversarial learning: An application to recidivism prediction [J]. arXiv preprint, arXiv: 1807.00199, 2018

[9] Lichman M. UCI machine learning repository [OL]. 2013 [2020-09-15]. http://archive.ics.uci.edu/ml

[10] Du Mengnan, Liu Ninghao, Yang Fan, et al. Learning credible deep neural networks with rationale regularization [C] //Proc of the IEEE Int Conf on Data Mining. Piscataway, NJ: IEEE, 2019: 150-159

[11] Ross A S, Hughes M C, Doshi-Velez F. Right for the right reasons: Training differentiable models by constraining their explanations [J]. arXiv preprint, arXiv: 1703.03717, 2017

[12] Elazar Y, Goldberg Y. Adversarial removal of demographic attributes from text data [J]. arXiv preprint, arXiv: 1808.06640, 2018

[13] Zhang B H, Lemoine B, Mitchell M. Mitigating unwanted biases with adversarial learning [C] //Proc of the 2018 AAAI/ACM Conf on AI, Ethics, and Society. New York: ACM, 2018: 335-340

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

[15] Olteanu A, Kiciman E, Castillo C. A critical review of online social data: Biases, methodological pitfalls, and ethical boundaries [C] //Proc of the 11th ACM Int Conf on Web Search and Data Mining. New York: ACM, 2018: 785-786

[16] Mehrabi N, Morstatter F, Saxena N, et al. A survey on bias and fairness in machine learning [J]. arXiv preprint, arXiv: 1908.09635, 2019

[17] Tufekci Z. Big questions for social media big data: Representativeness, validity and other methodological pitfalls [J]. arXiv preprint, arXiv: 1403.7400, 2014

[18] Barbosa S, Cosley D, Sharma A, et al. Averaging gone wrong: Using time-aware analyses to better understand behavior [C] //Proc of the 25th Int Conf on World Wide Web. New York: ACM, 2016: 827-839

[19] Verleysen M, François D. The curse of dimensionality in data mining and time series prediction [C] //Proc of Int Work-Conf on Artificial Neural Networks. Berlin: Springer, 2005: 758-770

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

[21] Wang Ting, Wang Dashun. Why Amazon's ratings might mislead you: The story of herding effects [J]. Big Data, 2014, 2(4): 196-204

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

[23] Berk R, Heidari H, Jabbari S, et al. Fairness in criminal justice risk assessments: The state of the art [J]. Sociological Methods & Research, 2018 (1): 1-42

[24] Miller H, Thebault-Spieker J, Chang S, et al. "blissfully happy" or "ready to fight": Varying interpretations of emoji [C] //Proc of the 10th Int Conf on Web and Social Media. Menlo Park, CA: AAAI, 2016: 259-268

[25] Lerman K, Hogg T. Leveraging position bias to improve peer recommendation [J]. PloS One, 2014, 9(6): e98914

[26] Nguyen D, Gravel R, Trieschnigg D, et al. "How old do you think I am?": A study of language and age in Twitter [C] //Proc of the 7th Int AAAI Conf on Weblogs and Social Media. Menlo Park, CA: AAAI, 2013: 439-448

[27] Ciampaglia G L, Nematzadeh A, Menczer F, et al. How algorithmic popularity bias hinders or promotes quality [J]. Scientific Reports, 2018, 8(1): 1-7

[28] Introna L, Nissenbaum H. Defining the Web: The politics of search engines [J]. Computer, 2000, 33(1): 54-62

[29] Huang L, Vishnoi N K. Stable and fair classification [J]. arXiv preprint, arXiv: 1902.07823, 2019

[30] Blyth C R. On Simpson's paradox and the sure-thing principle [J]. Journal of the American Statistical Association, 1972, 67(338): 364-366

[31] Bickel P J, Hammel E A, O'Connell J W. Sex bias in graduate admissions: Data from Berkeley [J]. Science, 1975, 187(4175): 398-404

[32] Chuang J S, Rivoire O, Leibler S. Simpson's paradox in a synthetic microbial system [J]. Science, 2009, 323(5911): 272-275

[33] Kievit R, Frankenhuis W E, Waldorp L, et al. Simpson's paradox in psychological science: A practical guide [J]. Frontiers in Psychology, 2013, 4: No.513

[34] Minchev I, Matijevic G, Hogg D W, et al. Yule-Simpson's paradox in Galactic Archaeology [J]. Monthly Notices of the Royal Astronomical Society, 2019, 487(3): 3946-3957

[35] Lerman K. Computational social scientist beware: Simpson's paradox in behavioral data [J]. Journal of Computational Social Science, 2018, 1(1): 49-58

[36] Mehrabi N, Morstatter F, Peng N, et al. Debiasing community detection: the importance of lowly connected nodes [C] //Proc of Int Conf on Advances in Social Networks Analysis and Mining. Piscataway, NJ: IEEE, 2019: 509-512

[37] Wilson C, Boe B, Sala A, et al. User interactions in social networks and their implications [C] //Proc of the 4th ACM European Conf on Computer Systems. New York: ACM, 2009: 205-218

[38] González-Bailón S, Wang N, Rivero A, et al. Assessing the bias in samples of large online networks [J]. Social Networks, 2014, 38: 16-27

[39] Morstatter F, Pfeffer J, Liu H, et al. Is the sample good enough? Comparing data from Twitter's streaming api with Twitter's firehose [J]. arXiv preprint, arXiv: 1306.5204, 2013

[40] Du Mengnan, Yang Fan, Zou Na, et al. Fairness in deep learning: A computational perspective [J]. arXiv preprint, arXiv: 1908.08843, 2020

[41] Bolukbasi T, Chang K W, Zou J Y, et al. Man is to computer programmer as woman is to homemaker? debiasing word embeddings [C] //Proc of the 29th Advances in Neural Information Processing Systems. New York: Curran Associates, 2016: 4349-4357

[42] Zhao Jieyu, Wang Tianlu, Yatskar M, et al. Men also like shopping: Reducing gender bias amplification using corpus-level constraints [J]. arXiv preprint, arXiv: 1707.09457, 2017

[43] Du Mengnan, Liu Ninghao, Hu Xia. Techniques for interpretable machine learning [J]. Communications of the ACM, 2019, 63(1): 68-77

[44] Kallus N, Mao X, Zhou A. Assessing algorithmic fairness with unobserved protected class using data combination [J]. arXiv preprint, arXiv: 1906.00285, 2019

[45] Blodgett S L, Green L, O'Connor B. Demographic dialectal variation in social media: A case study of African-American English [J]. arXiv preprint, arXiv: 1608.08868, 2016

[46] Jurgens D, Tsvetkov Y, Jurafsky D. Incorporating dialectal variability for socially equitable language identification [C] // Proc of the 55th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2017: 51-57

[47] Chen I Y, Szolovits P, Ghassemi M. Can AI help reduce disparities in general medical and mental health care? [J]. AMA Journal of Ethics, 2019, 21(2): 167-179

[48] Burnaev E, Erofeev P, Papanov A. Influence of resampling on accuracy of imbalanced classification [C] //Proc of the 8th Int Conf on Machine Vision. International Society for Optics and Photonics, Bellingham, WA: SPIE, 2015: No.987521

[49] Cui Yin, Jia Menglin, Lin T Y, et al. Class-balanced loss based on effective number of samples [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 9268-9277

[50] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: Synthetic minority over-sampling technique [J]. Journal of Artificial Intelligence Research, 2002, 16(1): 321-357

[51] Shorten C, Khoshgoftaar T M. A survey on image data augmentation for deep learning [J]. Journal of Big Data, 2019, 6(1): 1-48

[52] Weiss G M, McCarthy K, Zabar B. Cost-sensitive learning vs. sampling: Which is best for handling unbalanced classes with unequal error costs? [C] //Proc of Int Conf on Data Mining. Las Vegas, NV: DMIN, 2007: 35-41

[53] Abdi H, Williams L J. Principal component analysis [J]. Wiley Interdisciplinary Reviews: Computational Statistics, 2010, 2(4): 433-459

[54] Hancock J T, Khoshgoftaar T M. Survey on categorical data for neural networks [J]. Journal of Big Data, 2020, 7(1): 1-41

[55] Zhang Peixin, Wang Jingyi, Sun Jun, et al. White-box fairness testing through adversarial sampling [C] //Proc of the 42nd Int Conf on Software Engineering. New York: ACM, 2020: 23-29

[56] Galhotra S, Brun Y, Meliou A. Fairness testing: Testing software for discrimination [C] //Proc of the 11th Joint Meeting on Foundations of Software Engineering. New York: ACM, 2017: 498-510

[57] Udeshi S, Arora P, Chattopadhyay S. Automated directed fairness testing [C] //Proc of the 33rd ACM/IEEE Int Conf on Automated Software Engineering. New York: ACM, 2018: 98-108

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

[59] D'Alessandro B, O'Neil C, LaGatta T. Conscientious classification: A data scientist's guide to discrimination-aware classification [J]. Big Data, 2017, 5(2): 120-134

[60] Shrivastava A, Gupta A, Girshick R. Training region-based object detectors with online hard example mining [C] //Proc of the IEEE conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 761-769

[61] Ruoss A, Balunović M, Fischer M, et al. Learning Certified Individually Fair Representations [J]. arXiv preprint, arXiv: 2002.10312, 2020

[62] Benjamin M, Gagnon P, Rostamzadeh N, et al. Towards Standardization of Data Licenses: The Montreal Data License [J]. arXiv preprint, arXiv: 1903.12262, 2019

[63] Gebru T, Morgenstern J, Vecchione B, et al. Datasheets for datasets [J]. arXiv preprint, arXiv: 1803.09010, 2018

[64] Holland S, Hosny A, Newman S, et al. The dataset nutrition label: A framework to drive higher data quality standards [J]. arXiv preprint, arXiv: 1805.03677, 2018

[65] Alipourfard N, Fennell P G, Lerman K. Can you Trust the Trend? Discovering Simpson's Paradoxes in Social Data [C] //Proc of the 8th ACM Int Conf on Web Search and Data Mining. New York: ACM, 2018: 19-27

[66] Zhang Lu, Wu Yongkai, Wu Xintao. Achieving non-discrimination in data release [C] //Proc of the 23rd ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2017: 1335-1344

[67] Hajian S, Domingo-Ferrer J. A methodology for direct and indirect discrimination prevention in data mining [J]. IEEE Transactions on Knowledge and Data Engineering, 2012, 25(7): 1445-1459

[68] Wang Tianlu, Zhao Jieyu, Yatskar M, et al. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations [C] //Proc of the 2019 IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 5310-5319

[69] Du Mengnan, Liu Ninghao, Song Qingquan, et al. Towards explanation of dnn-based prediction with guided feature inversion [C] //Proc of the 24th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2018: 1358-1367

[70] Du Mengnan, Liu Ninghao, Yang Fan, et al. On attribution of recurrent neural network predictions via additive decomposition [C] //Proc of the World Wide Web Conf. New York: ACM, 2019: 383-393

[71] Ribeiro M T, Singh S, Guestrin C. "Why should I trust you?" Explaining the predictions of any classifier [C] //Proc of the 22nd ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2016: 1135-1144

[72] Lipton Z C, Chouldechova A, McAuley J. Does mitigating ml's disparate impact require disparate treatment [J]. arXiv preprint, arXiv: 1711.07076, 2017

[73] Louizos C, Swersky K, Li Y, et al. The variational fair autoencoder [J]. arXiv preprint, arXiv: 1511.00830, 2015

[74] Kim B, Wattenberg M, Gilmer J, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors [C] //Proc of Int Conf on Machine Learning. Stockholm, Sweden: PMLR, 2018: 2668-2677

[75] Fong R, Vedaldi A. Net2vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition, Piscataway, NJ: IEEE, 2018: 8730-8738

[76] Kiritchenko S, Mohammad S M. Examining gender and race bias in two hundred sentiment analysis systems [J]. arXiv preprint, arXiv: 1805.04508, 2018

[77] Zhou B, Sun Y, Bau D, et al. Interpretable basis decomposition for visual explanation [C] //Proc of the 15th European Conf on Computer Vision. Berlin: Springer, 2018: 119-134

[78] Upchurch P, Gardner J, Pleiss G, et al. Deep feature interpolation for image content changes [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 7064-7073

[79] Kearns M, Neel S, Roth A, et al. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness [C] //Proc of Int Conf on Machine Learning. Stockholm, Sweden: PMLR, 2018: 2564-2572

[80] Hardt M, Price E, Srebro N. Equality of opportunity in supervised learning [C] //Proc of Advances in Neural Information Processing Systems. New York: ACM, 2016: 3315-3323

[81] Agarwal A, Dudík M, Wu Z S. Fair regression: Quantitative definitions and reduction-based algorithms [J]. arXiv preprint, arXiv: 1905.12843, 2019

[82] Agarwal A, Beygelzimer A, Dudík M, et al. A reductions approach to fair classification [J]. arXiv preprint, arXiv: 1803.02453, 2018

[83] Calmon F, Wei D, Vinzamuri B, et al. Optimized preprocessing for discrimination prevention [J]. IEEE Journal of Selected Topics in Signal Processing, 2018, 12(5): 1106-1119

[84] Feldman M, Friedler S A, Moeller J, et al. Certifying and removing disparate impact [C] //Proc of the 21st ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2015: 259-268

[85] Pleiss G, Raghavan M, Wu F, et al. On fairness and calibration [C] //Proc of Advances in Neural Information Processing Systems. New York: ACM, 2017: 5680-5689

[86] Zemel R, Wu Y, Swersky K, et al. Learning fair representations [C] //Proc of Int Conf on Machine Learning. Stockholm, Sweden: PMLR, 2013: 325-333

[87] Celis L E, Huang L, Keswani V, et al. Classification with fairness constraints: A meta-algorithm with provable guarantees [C] //Proc of the Conf on Fairness, Accountability, and Transparency. New York: ACM, 2019: 319-328

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

[89] Kamiran F, Karim A, Zhang X. Decision theory for discrimination-aware classification [C] //Proc of the 12th Int Conf on Data Mining. Piscataway, NJ: IEEE, 2012: 924-929

[90] Kamishima T, Akaho S, Asoh H, et al. Fairness-aware classifier with prejudice remover regularizer [C] //Proc of Joint European Conf on Machine Learning and Knowledge Discovery in Databases. Berlin: Springer, 2012: 35-50

[91] Speicher T, Heidari H, Grgic-Hlaca N, et al. A unified approach to quantifying algorithmic unfairness: Measuring individual & group unfairness via inequality indices [C] //Proc of the 24th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2018: 2239-2248

[92] Foulds J R, Islam R, Keya K N, et al. An intersectional definition of fairness [C] //Proc of the 36th Int Conf on Data Engineering. Piscataway, NJ: IEEE, 2020: 1918-1921

[93] Atwood J, Srinivasan H, Halpern Y, et al. Fair treatment allocations in social networks [J]. arXiv preprint, arXiv: 1911.05489, 2019

[94] Alexander D, Hansa S, James A, et al. Fairness is not static: Deeper understanding of long term fairness via simulation studies [C] //Proc of the 2020 Conf on Fairness, Accountability, and Transparency. New York: ACM, 2020: 525-534

[95] Li T, Sanjabi M, Beirami A, et al. Fair resource allocation in federated learning [J]. arXiv preprint, arXiv: 1905.10497, 2019

[96] Pan S J, Tsang I W, Kwok J T, et al. Domain adaptation via transfer component analysis [J]. IEEE Transactions on Neural Networks, 2010, 22(2): 199-210

[97] Hospedales T, Antoniou A, Micaelli P, et al. Meta-learning in neural networks: A survey [J]. arXiv preprint, arXiv: 2004.05439, 2020

Chen Jinyin, born in 1982. PhD. Associate professor with Zhejiang University of Technology. Her main research interests include artificial intelligence security, graph data mining and evolutionary computing.

陈晋音,1982 年生.博士,副教授.主要研究方向为人工智能安全、图数据挖掘和进化计算.

Chen Yipeng, born in 1996. Master candidate at Zhejiang University of Technology. Her main research interests are deep learning and artificial intelligence.

陈奕芃,1996 年生.硕士研究生.主要研究方向为深度学习和人工智能.

Chen Yiming, born in 1996. Master candidate at Zhejiang University of Technology. His main research interests are deep learning and artificial intelligence.

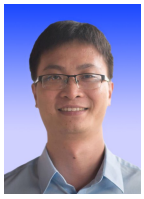
陈一鸣,1996 年生.硕士研究生.主要研究方向为深度学习和人工智能.

Zheng Haibin, born in 1995. PhD candidate at Zhejiang University of Technology. His main research interests include deep learning, artificial intelligence security, and fairness algorithm.

郑海斌,1995 年生.博士研究生.主要研究方向为深度学习、人工智能安全和公平性算法.

Ji Shouling, born in 1986. PhD, professor in the College of Computer Science and Technology at Zhejiang University. His current research interests include data-driven security and privacy, AI security and big data analytics.

纪守领,1986 年生.博士,研究员.主要研究方向为数据驱动安全和隐私、AI 隐私和大数据分析.



Shi Jie, born in 1983, PhD, security expert in Huawei Singapore Research Center. His main research interests include trustworthy AI, machine learning security and privacy, data security and privacy and applied cryptography.

时　杰,1983 年生.博士,华为新加坡研究院安全专家.主要研究方向为可信 AI、机器学习安全和隐私、数据安全和隐私以及应用的加密技术.

Cheng Yao, born in 1987. PhD, senior researcher in Huawei Singapore Research Center. Her main research interests include security and privacy in deep learning systems, blockchain technology applications, Android framework vulnerability analysis.

程　瑶,1987 年生.博士,华为新加坡研究院高级研究员.主要研究方向为深度学习系统安全和隐私、区块链技术应用和安卓框架脆弱性分析.