arXiv:2304.12573v1 [cs.LG] 25 Apr 2023

真相发现算法中的公平性与偏见：实验分析

Simone Lazier* Saravanan Thirmuruganathan† Hadis Anahideh‡

摘要
基于机器学习（机器学习）的方法正日益应用于众多具有社会影响的应用中。训练机器学习模型通常需要大量的标注数据，而众包是从多名工人处获取标签的主流范式。众包工人有时可能会提供不可靠的标签。为了解决这一问题，会应用多数投票等真实值发现（TD）算法，根据相互冲突的工人响应确定共识标签。然而，需要注意的是，这些共识标签仍可能基于性别、种族或政治倾向等敏感属性而产生偏见。即使不涉及敏感属性，标签也可能由于对毒性等主观方面的不同观点而产生偏差。本文对TD算法的偏差与公平性进行了系统研究。我们使用两个现有的众包标注数据集开展研究，结果表明，有相当一部分工人会提供有偏结果，而使用简单的TD方法并非最优。我们的研究还表明，流行的TD算法并非万能药。此外，我们量化了这些不公平工人对下游机器学习任务的影响，并表明在这一场景中，实现公平性和纠正标签偏差的传统方法并不奏效。最后，我们呼吁设计新颖的、能够感知偏差的真相发现算法，以缓解这些问题。

1 引言

机器学习（ML）日益普及，并已对人们的生活产生了重大影响。这些模型所做的决策往往会以不均衡的方式影响不同的子群体。因此，研究人员纷纷致力于设计能够兼顾公平性的机器学习算法。本文认为，机器学习流程中的一个关键环节——数据标签的收集，也必须得到同等重视。我们的研究立足于两大范式的交汇处：以数据为中心的人工智能和公平机器学习。前者致力于利用高质量数据构建人工智能系统，后者则确保最终系统具备公平性。正如我们即将展示的，当前的标签收集流程过于简单，导致产生有偏差的标签，并对下游机器学习模型的准确率和公平性造成不成比例的影响。尽管设计公平机器学习算法的研究已相当广泛，但开发能够生成正确且公平标签的算法方面，相关研究仍然有限。

数据香肠的制作。数据集对于机器学习社区的发展不可或缺，它们使新模型的设计成为可能，并作为衡量进展的基准。尽管数据集非常重要，机器学习领域绝大多数工作仍致力于开发新算法，而非数据管理实践 [18]。近期，公平性研究领域揭示了知名基准数据集中的许多问题，例如标签错误和偏差 [17, 19]。在这项工作中，我们系统地分析了另一种偏差来源——通过简单汇总存在偏差的众包工人的回答而生成的标签。

机器学习成功的一个关键因素是能够获得大量已标注数据集。然而，聘请领域专家创建准确标签的成本高得令人难以承受。因此，常见的替代方法是借助众包工人生成标签。人们通常通过 Amazon Mechanical Turk 等众包市场雇用工人。与领域专家相比，这些工人的工作质量往往较低，但成本效益显著更高。质量控制通常通过让不同工人为每个数据点提供多个标签，并采用额外的筛选机制来实现，例如，在开始标注主数据集之前，先向工人提供一组答案已知的“黄金任务”进行预测试 [23]。工人给出的标签可能彼此冲突。在这种情况下，会使用聚合方法，更广泛地说，即真值发现算法（TD），来确定最终标签。一种常见方法是采用多数投票。在许多情况下，由此获得的标签相当准确，可用于训练机器学习模型。许多大型

*伊利诺伊大学芝加哥分校。（slazie2@uic.edu）
†卡塔尔计算研究所、哈马德·本·哈利法大学。（sthirumuruganathan@hbku.edu.qa）
‡伊利诺伊大学芝加哥分校。（hadis@uic.edu）

版权 © 2023年 SIAM。未经授权的复制本文受到禁止

基准数据集都是通过这种方法获得的。即使只考虑计算机视觉这一个领域，此类数据集也包括但不限于 ImageNet [8], Places [24], MS COCO[14], ActivityNet [6], Open Street Map（OSM）数据集 [11], YouTube-8M [2], 和 Intrinsic Images in the Wild[5]。众包数据集在其他机器学习领域中也同样普遍。例如，自然语言处理领域发表的论文中，多达 10 % 使用了众包数据集 [21]。

广泛使用众包来获取标签存在问题。使用人工标注者是一把双刃剑。虽然人类可以为具有挑战性和主观性的问题提供答案，但他们的回答也可能受到各种偏差的影响。随着机器学习越来越多地应用于假新闻、毒性和仇恨言论检测等敏感领域，这些领域的标签可能具有很强的主观性，并涉及细微复杂的概念。不同工人可能会根据各自的生活经历，为同一个数据点分配不同的标签，这可能导致主观价值观和偏见被注入数据集。使用包含大量有偏见标签的数据集，可能会导致模型在准确率和公平性方面表现欠佳。

我们的贡献总结。我们对众包数据集中标签的聚合过程进行了系统研究。我们的研究包括两个公开可用的多样化数据集——*Crowd Judgement*和*Twitter Toxicity*，并且研究结果与我们在仇恨言论检测领域的其他多个数据集上进行实验所得的结果一致。我们的研究结果表明，有偏见的工人普遍存在，并且会对数据集标签的质量产生重大影响。令人惊讶的是，即使包含真实标签的数据集本身可能存在偏见，某个工人仍然可能在保持准确的同时具有偏见。因此，这排除了移除有偏见工人等简单解决方案，因为此举可能会对其他项目的准确率造成附带损害。关于真值发现算法，已有大量研究致力于将相互冲突的标签聚合为单一的“真实值”。然而，我们发现，使用此类TD算法并不能有效解决这一问题，尽管它们比多数投票更加复杂。我们使用两种流行算法（DS 和 LFC）进行分析：Dawid Skene（DS） [7] 和 Learning From Crowds（LFC） [20],

这些算法针对众包工人偏见采用了不同的概率生成建模方法。尽管如此，它们的性能并未显著优于多数投票。此外，我们研究了有偏标签对下游机器学习任务的影响，并表明每种真值发现算法都存在类似的缺陷，导致模型准确性更低且有失公平。我们还表明，使用公平机器学习算法并不能消除有偏标签的问题。

我们的主要贡献如下：

1. 我们发现，用于汇总众包工人回答的真值发现算法是众包数据集中偏差的一个主要来源。

2. 我们开展了广泛的实验研究，以考察这种偏差的各个方面，并量化其对下游机器学习任务的影响。

3. 我们为实践者提供切实可行的建议，并指出有前景的研究问题。

论文结构。 本文其余部分安排如下。第 2 节介绍众包、公平性和真值发现算法的相关背景。第 3 节概述本研究所用的数据集和研究方法。我们的实验分析围绕五个主要维度展开，详见第 4 至第 8 节。最后，我们将在第 9 节总结全文并提出一些思考。

2 背景

用于机器学习及其缺陷的众包。 在过去十年中，众包被广泛用于推动机器学习研究。众包工人被用于数据生成、标签生成以及机器学习模型的评估。机器学习社区已采用众包来获取训练热门视觉与语言处理任务所需的图像和语言标注。一个关键障碍在于，工人提供的标签可能存在噪声或不准确。工人可能并不完美、缺乏积极性，或受到各种认知偏差的影响[10]。当这些标签用于训练执行具有社会影响的任务的机器学习模型时，这些偏差可能造成严重后果。一项开创性研究确定了可能影响机器学习算法的各种偏差来源[16]。人们提出了真值发现算法来解决前述问题，同时提出了公平机器学习算法来解决后者。

众包中的真实值发现。 众包中的质量控制涉及从众包工人提供的低质量答案中确定正确答案。遗憾的是，绝大多数研究采用多数投票等简单方法，而这通常是不够的。复杂的方法通常基于概率生成模型，可对影响工人标签生成的各种因素进行建模。这些因素可能与工人相关（准确率、偏差、动机、串通、主题相关专业知识等），也可能与任务相关（难度、

版权 © 2023年 SIAM 未经授权的复制本文是被禁止的

清晰度等）。真值发现算法的分类可参见 [23]。

真实值发现——问题设置。考虑一个任务集合 $\mathcal{T} = \{t_1, \ldots, t_n\}$，其规模为 n。设 $\mathcal{W}$ 为标注者集合，W_i 为由 $\mathcal{W}$ 中的部分标注者收集的任务 t_i 的答案向量。设 m 为标注者总数。每个任务 t_i 都可能由一部分工人进行标注。通过向这些标注者进行询问，我们希望确定任意任务 $t_i \in \mathcal{T}$ 的真实类别标签 y_i。随后，将标注者 j 对任务 t_i 给出的标签记为 l_{ij}。设 $\tilde{y}_i^w$ 为标注者 w 对任务 t_i 给出的答案，y 为~未知的真实值。给定 Y 的答案矩阵，使用真值发现算法，可以将每个任务的真实值推断结果确定为 $\hat{y}_i$。

机器学习中的算法公平性。近年来，算法公平性的研究在机器学习社区中受到了越来越多的关注。这些研究旨在确保人工智能系统不会基于种族、性别或年龄等敏感特征歧视某些群体。有关各个方面的详细描述可见于 [4]。尽管关于机器学习公平性的文献日益丰富，但仍有许多挑战和局限性需要解决。例如，目前尚无广泛接受的公平性定义，这可能导致对人工智能系统公平性的评估缺乏一致性 [16]。此外，许多用于缓解算法偏见的现有方法依赖于对建模流程进行调整，包括预处理 [22]，、处理中 [3]，和后处理[15] 阶段，而不是解决偏差的根本原因（例如数据收集偏差）；在众包环境中，这一问题尤为突出。

3 实验方法

在本文中，我们关注一种会导致不公平标签的特定类型的偏差。例如，工人可能会为涉及特定性别或种族成员的任务提供错误标签。使用多数投票等 TD 算法会将这些有偏差的标签延续到数据集中，进而导致不公平模型。理想情况下，这些不公平标签应在标签聚合过程中得到纠正。

在本节中，我们概述所采用的实验方法，以系统地研究和量化真值发现算法在存在不公平工人时的工作方式。为便于说明，我们将其划分为五个研究假设。在描述这些假设后，我们将进一步介绍实验设置的详细信息，例如代表性真值发现算法和数据集列表。

实验假设。我们开展了广泛的

实验，以研究不同的TD算法如何影响对不完美工作者提交的标签进行聚合。具体而言，我们聚焦于阐释标签聚合中偏差问题的五个关键维度。第4至第8节将详细介绍相应的研究假设。本节提供一个高层次概述，使读者能够全面了解我们的实验。

H1：通过分析准确性与公平性指标量化不公平工人的普遍性。我们首先对单个工人的准确性与公平性指标进行探索性分析。结果表明，不公平工人并非异常情况，而是广泛存在。

H2：不公平工人对真值发现算法中标签翻转的影响。我们表明，不公平工人可能会翻转相当一部分任务的标签。接下来，我们研究不公平工人对TD算法的影响，以表明不公平工人可能会翻转相当一部分任务的标签。具体而言，我们重点研究多数投票这一广泛使用的方法。值得注意的是，这一问题不仅影响多数投票，也影响其他传统真值发现算法。

H3：不公平工人会降低*TD*算法的性能。 接下来，我们研究三种具有代表性的真值发现算法的性能。具体而言，我们将这些方法估计的真实标签与真实值进行比较。我们表明，所有这些算法的准确率和公平性都相近，且均未达到最优。这凸显了开发新的公平性感知真值发现算法的必要性，而不能依赖现有算法。

H4：不公平工人会对下游机器学习任务产生负面影响。下一个假设回应了一个潜在担忧：对真值发现算法性能的研究并非仅具有学术意义。相反，我们证明，使用简单真值发现算法的输出训练机器学习模型，会得到一个准确率更低且公平性更差的模型。

H5：公平机器学习与公平真值发现的互补性。在最后一个假设中，我们探讨了这样一种观点：使用公平机器学习算法无法完全消除传统真值发现算法所引入的偏差。我们还评估了 FairTD（一种在 [13]中提出的公平真实值发现算法）的有效性。

真值发现算法。 人们已经对聚合工人相互冲突的回答的技术进行了广泛研究。本文比较了三种具有代表性的算法。有关其比较性能的详细讨论，请参阅

版权 © 2023年 SIAM 未经授权的复制本文，禁止进行

[23]。多数投票（*MV*）是一种简单且广泛使用的算法，它选择大多数工人提供的回答作为共识。如果工人具备合理的能力且不存在任何串通，MV 可能会有效。MV 的一个关键局限在于，它对每位工人的回答赋予相同的权重。然而，在工人准确率可能存在差异的众包环境中，这种做法可能不是最优的。例如，如果只有一名工人是专家而其他人都是新手，MV 可能会失效。在这种情况下，有必要对工人准确性进行建模，从而计算加权多数投票。Dawid-Skene（DS） [7] 是一种更复杂的模型，它根据工人混淆矩阵的内容为其分配不同的权重。工人准确性通过迭代式期望最大化算法进行估计。从众包中学习（LFC） [20] 是一种概率模型，为标注者提出了双硬币方法。每位工人都有两枚带偏差的硬币，其偏差分别对应灵敏度和特异度。工人根据任务的真实标签选择其中一枚硬币进行投掷，并利用结果确定其回答。这种方法已成功应用于许多领域，例如聚合放射科医生的预测。三种模型都得到了广泛使用，并展现出实证上的成功。本文研究这些模型在存在工作者偏差时的表现。

公平性指标。 关于衡量和缓解数据集及机器学习模型中的偏差，已有大量研究。然而，对于适当的公平性指标，目前尚未形成太多共识，因为不同指标都要求机器学习模型的某些方面在所有子群体中保持可比性。本文采用两种广泛使用的关键指标，用于评估机器学习模型的公平性——人口统计均等和均等赔率。将每位众包工人视为不完美预言机（类似于机器学习模型），我们便可复用这两种指标来衡量工人的公平性。当特定预测（例如贷款发放）的概率不取决于敏感群体归属（例如性别=男性）时，即实现了人口统计均等（DP）。均等赔率（EO）公平性指标要求机器学习模型的预测既不受敏感群体归属影响，也要求真阳性率和假阳性率具有可比性。值得注意的是，EO比DP更严格。为了更有效地评估模型的公平性，我们使用了DP和EO的两种衍生指标——差值和比率。差值为0表示实现了DP/EO，比率为1也表示实现了DP/EO。这些衍生指标很有用，因为它们能将公平性指标表示为单个标量，从而便于比较。

数据集。 我们使用来自不同领域的两个现有的、多样且具有代表性的数据集开展实验，这些数据集来自不同领域。*Crowd Judgement* 数据集 [9]包含某大型众包平台工人对臭名昭著的 COMPAS数据集中1000个再犯案例所作的标签。每组20名工人对同一批50名被告进行了评估。第二个数据集*Jigsaw* 毒性 包含工人对 Twitter 评论是否具有毒性的标签。该数据集收集于 [1] 我们使用了其中14000条评论的子集，这些评论针对毒性、淫秽内容、威胁、侮辱和仇恨进行了标注。

4 H1：通过分析准确率与公平性指标量化不公平工人的普遍性

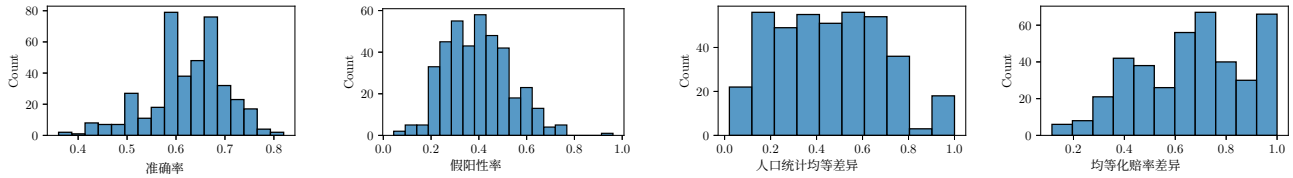
我们的第一组实验对数据集进行探索性分析，以计算工人的准确率和公平性指标。尽管普遍观点认为工人往往并不完美且/或不公平，但量化这些因素对于理解真值发现算法所面临的挑战至关重要。

两个数据集（*Crowd Judgement* 和 *JigsawToxicity*）都包含真实值。这与典型的众包数据集形成对比，后者通常无法获得真实值。这些元数据使我们能够衡量工人的准确性和公平性。具体而言，我们关注两个与准确率相关的指标（准确率和假阳性率）以及两个与公平性相关的指标（人口统计差异和均等化赔率差异）。我们开展了三组相互关联的实验。在第一组实验中，我们生成四个准确率与公平性指标的直方图。其次，我们创建散点图，对比一个准确率指标与一个公平性指标。最后，我们以替代表格格式展示准确率与公平性的分布，因为这种格式能够提供独特的洞见。

工人准确率指标的分布。图 1a 和图 1b 展示了*Crowd Judgement*数据集的工人准确性和假阳性率直方图。可以看到，工人准确性大致介于 0.4 和 0.8 之间。换言之，这些工人既不完美（准确性接近 1.0），也不无能（准确性接近 0.0）。然而，仍有相当数量的工人具有较高的假阳性率，这凸显了真值发现算法所面临的挑战。这些工人的准确性很高，但往往会出现较高的假阳性率。因此，使用多数投票等简单的聚合指标可能不足以解决问题。

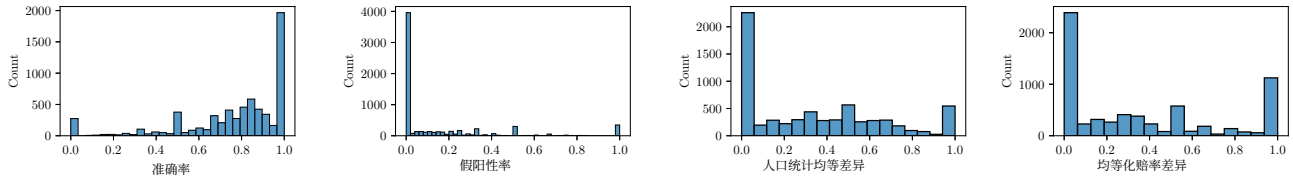
*Jigsaw Toxicity*数据集对应的直方图见图 2a 和图 2b。该数据集的工人准确性分布与 *Crowd Judgement* 数据集不同。对于这个数据集，大多数工人准确率很高，相当一部分工人完全准确，

版权 © 2023年 SIAM 未经授权的复制本文受到禁止



（a）准确率　（b）假阳性率　（c）人口统计均等　（d）均等赔率

图 1：*Crowd Judgement* 数据集中工人的准确性与公平性指标直方图



（a）准确率　（b）假阳性率　（c）人口统计均等　（d）均等赔率

图 2：*Jigsaw Toxicity* 数据集中工人的准确性与公平性指标直方图

且只有少数工人表现出较高的假阳性率。尽管两个数据集中的工人在准确率方面差异很大，我们将证明，两者都包含数量不可忽视的不公平工人。

工人公平性指标的分布。图 1c 和图 1d 分别展示了 *Crowd Judgement* 数据集中工人人口统计均等差异和均等化赔率差异指标的直方图。请注意，值越接近 0，表示工人越公平。这两个指标的直方图均显示，工人的公平性存在很大差异。尽管有一些公平工作者，但绝大多数工人在某种程度上都不公平。图 2c 和图 2d 展示了 *Jigsaw Toxicity* 数据集的直方图，结果表明，与 *Crowd Judgement*数据集中的工人相比，这些工人总体上更加公平。鉴于黄金任务和更明确的评分标准，这并不令人意外。然而，其中仍包含相当数量的不公平工人。我们还可以得出结论，仅仅增加更多工人并不能解决这一问题，因为其中相当一部分工人同样会是不公平的。因此，必须设计一种能够考虑到不公平工人比例较大的真实值发现算法，以便更好地聚合工人标签。

工人准确率—公平性权衡。一个自然会产生的问题是：不准确工人与不公平工人之间是否存在相关性。换言之，不准确的工人是否也不公平？如果这些指标之间存在高度相关性，那么开发单独的公平性感知真值发现算法便没有必要。我们可以通过将准确率指标与公平性指标进行绘图，来研究这些指标之间的关系。图 3a 显示了工人准确率在 X 轴上、工人人口统计学奇偶差异在 Y 轴上的散点图。图 3b 展示了准确率与均等化赔率差异指标的类似图示。我们可以看到，各数值分布在整个图中，没有形成显示准确率与公平性之间关系的紧密聚类。换言之，一个工人可能既准确又不公平。图 3a 和图 3b 中针对*Jigsaw Toxicity* 数据集的相应图示表明，随着准确率提高，不公平工人的数量也会增加。因此，工人可能既准确又不公平；在存在不公平工人的情况下，直接使用以准确率为导向的传统真值发现算法是不充分的。

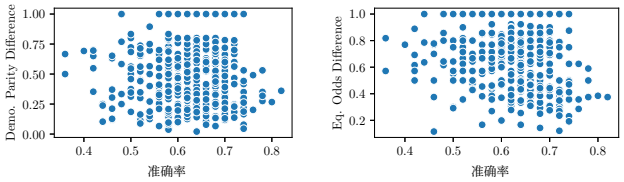
表1通过先根据工人的准确率对其分组，再计算每组中所有工人的公平性指标，展示了图3和图4的细化版本。再次可以看到，这两个指标之间没有明显的相关性。对于 *Crowd Judgement*数据集，两者之间存在弱相关性：工人的准确率越高，其不公平程度越低（即该指标趋近于 0）。然而，在毒性数据集中并未出现这种现象。

5 H2：不公平工人对真相发现算法中标签翻转的影响

在下一组实验中，我们将探究不公平工人的存在如何影响多数投票算法生成的标签。该算法广泛用于聚合大多数机器学习数据集中的众包工人标签。我们从两个不同的角度开展这项研究。

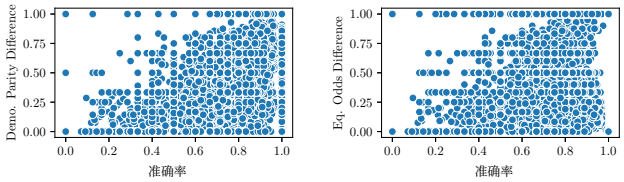
任务中的不公平工人比例。直观地说，

版权 © 2023年 SIAM 未经授权禁止复制本文



（a）准确率与人口统计学的关系（b）准确率与均等化奇偶差异赔率差异的关系

图 3：*Crowd Judgement* 数据集



（a）准确率与人口统计学（b）准确率与均等化奇偶差异赔率差异

图 4：*Twitter Toxicity* 数据集

	Crowd Judgement		Twitter Toxicity	
准确率范围	**演示。奇偶校验。差值**	**相等. 赔率. 差异**	**演示. 奇偶校验. 差异**	**相等. 赔率. 差异**
(0.0, 0.1]	NA	NA	0.05	0.05
(0.1, 0.2]	NA	NA	0.12	0.15
(0.2, 0.3]	NA	NA	0.11	0.22
(0.3, 0.4]	0.62	0.72	0.24	0.35
(0.4, 0.5]	0.49	0.67	0.26	0.42
(0.5, 0.6]	0.45	0.69	0.3	0.5
(0.6, 0.7]	0.46	0.67	0.34	0.53
(0.7, 0.8]	0.4	0.55	0.37	0.5
(0.8, 0.9]	0.36	0.38	0.45	0.43
(0.9, 1.0]	NA	NA	0.32	0.22

表1：两个数据集的准确性与公平性指标分布。

MV 将多数工人提供的回答选为共识。因此，如果分配给某项任务的多数工人最终被判定为不公平工人，那么该任务的输出也将是不公平的。我们的第一组实验研究了根据公平性的定义，分配给一项任务的工人比例如何变化。我们使用人口统计均等和均等化赔率差异指标来确定工人的公平性，其中 0.0 表示不存在偏差。为了将工人指定为公平工作者或不公平工人，我们对这些公平性指标应用阈值。显然，得分为 0.0 的工人是公平工作者，而得分为 1.0 的工人是不公平工人。我们系统地改变公平性的最小阈值（从 0，0。1，.....1。0）。随后，我们统计不公平工人多于公平工作者、从而污染多数投票输出的任务比例。图5使用两个公平性指标绘制了两个数据集中该比例的变化情况。可以看到，即使阈值放宽到 0.2，仍有很大比例的任务由不公平工人占主导。这一观察再次表明，设计公平性感知的TD算法十分重要。

移除不公平工人的影响。一种自然的应对方式可能是剔除不公平工人，然后在剩余工人上运行 TD 算法。在接下来的一组实验中，我们将展示这种方法徒劳无功。回顾图3和图4可知，工人可能总体上准确，却并不公平。图6展示了移除不公平工人的影响。当我们根据上一项实验所使用的公平性指标将工人判定为不公平并予以移除时，可能会导致两个关键参数发生变化。首先，移除不公平工人可能会影响任务的准确率。我们分别使用人口统计学均等准确率和均等机会准确率指标来衡量人口统计均等和均等赔率。其次，有响应的任务数量也可能发生变化。回顾图5可知，不公平工人在许多任务中占据主导地位。因此，某些任务中的所有工人都可能不公平，导致任务被放弃。人口统计学均等任务数和均等机会任务数这两个指标用于衡量这一现象。从图6可以看出，移除不公平工人后，整体准确率有所下降。这是因为其中相当一部分工人是准确的，移除他们会造成准确率下降这一附带损害。此外，随着不公平工人被剔除，被放弃的任务数量也会增加。这说明，对于此类任务，移除不公平工人并不可行。根据这两项观察，我们可以得出结论：仅仅剔除不公平工人是一种次优方案，因为这会降低准确率。要克服这一挑战，关键在于设计具备公平性意识的 TD 算法，在不牺牲准确率的前提下抵消不公平工人的影响。

一个关键的研究挑战在于，真值发现算法本质上是无监督的。需要注意的是，如果已知众包任务的真实值，就不需要工人了。这些算法通常采用EM式迭代方法。它们首先对相关指标（例如工人准确性）进行估计（如随机初始化），然后在每次迭代中改进该估计的

版权 © 2023年 SIAM 未经授权复制本文内容，禁止转载

相关指标，例如工人准确性，并在每次迭代中改进该估计。尽管改进工人准确性估计的方法已广为人知，但设计相应的公平性算法仍是一个有待研究的问题。

6 H3：不公平工人会降低TD算法的性能。

在本节中，我们研究不同TD算法在存在不公平工人时的表现。回顾一下，这些算法通常是为准确率不高的工人设计的。我们的实验如下进行：我们将三种具有代表性的TD算法应用于工人的回答，以获得真实标签的近似估计。由于$Crowd\ Judgment$和$Jigsaw\ Toxicity$都为我们提供了实际的真实标签，因此我们可以将其与TD算法的输出进行比较。这也使我们能够计算每种TD算法的准确率与公平性指标。表2和表3展示了这一实验的结果。

总体而言，我们可以看到，三种算法的准确率大致相当，其中基于生成模型的更复杂TD算法（DS和LFC）提供的估计结果略为准确。在公平性指标方面，我们观察到了类似的模式，没有任何TD算法能够有效处理不公平工人。使用先进的TD算法似乎反而加剧了这一问题，因为某些基础假设会因不公平工人的存在而被违反，例如LFC所采用的双硬币标注者过程。我们可以得出结论：直接使用现有TD算法并非最优选择，因此有必要设计公平性感知的TD算法，明确对工人的不公平性进行建模。

7 H4：不公平工人会对下游机器学习任务产生负面影响。

机器学习算法通常依赖大量数据来学习模式并进行预测。然而，使用从不可靠或存在偏差的来源收集的数据，可能会对下游机器学习任务产生严重影响。偏见数据的一个潜在来源是众包工人，他们可能具有固有偏见或个人信念，从而影响其提供的数据质量。在这种情况下，了解不公平众包工人对下游机器学习任务的影响至关重要。这一点极其重要，因为这些偏见可能会延续并放大现有的社会偏见，导致对某些群体的歧视和不平等对待。

在本节中，我们研究两种具有代表性的分类器——逻辑回归和随机森林所受到的影响。其他分类器也呈现出类似结果，例如支持向量机或神经网络。我们的实验流程如下：使用真实标签训练机器学习模型 $\mathcal{M}_G$。这是 TD 算法达到完美效果时我们本应获得的理想模型。请注意，在现实世界中，真实值往往无法获取。尽管如此，由于我们的两个数据集都提供了真实值，我们得以开展这项实验。接下来，我们根据 TD 算法生成的近似标签训练机器学习模型 $\mathcal{M}_{TD}$。我们将数据集平均划分为训练集和测试集，并重复实验 10 次，以计算平均准确率和公平性指标。通过比较两个模型在同一测试集上的性能，我们可以量化 TD 算法所付出的代价。

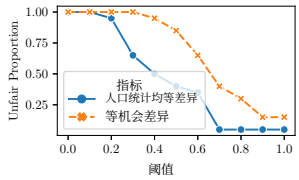
表4和表5展示了两个数据集、两个分类器和三个TD算法的结果。为节省空间，我们仅报告指标之间的差值。具体而言，如果两个模型的准确率分别为 a_G和 a_{TD}，我们报告 $a_G - a_{TD}$。对于公平性指标，我们采用相同的处理方式。如表中所示，使用TD算法会导致模型准确率降低且公平性变差。这并不令人意外，因为TD算法并非专门用于优化公平性。

8 H5：公平机器学习与公平真值发现的互补性。

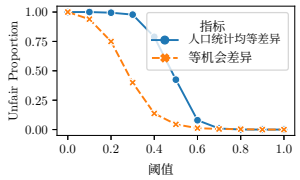
根据前述实验结果可以清楚地看出，在存在不公平工人的情况下，主要的TD算法表现欠佳。人们可能会想，公平机器学习领域的大量既有研究能否减轻TD算法产生的不公平标签所带来的影响。换言之，旨在生成对不同子群体公平的输出的公平机器学习算法，能否克服TD算法产生的错误或不公平标签？本节旨在全面比较公平机器学习和公平真值发现这两种可能的方法，并深入了解每种方法在决策系统中实现公平性方面的优势与局限性。

我们将 FairTD（首个也是唯一一个公平性感知真值发现算法） [13], 的输出与两种公平机器学习算法进行比较：Exponentiated Gradient Reduction[3]和 Prejudice Remover [12]。我们将这些方法分别用于多数投票（MV）和 Dawid-Skene（DS）真值发现策略。实验结果如图7所示，其中比较了所考察方法的准确率随公平性约束变化的情况。对于 FairTD 和 Exponentiated Gradient 方法，公平性约束指允许的公平性违反程度；对于 Prejudice Remover，则指正则化项的倒数。尽管

版权 © 2023年 SIAM 未经授权的复制本文受到禁止

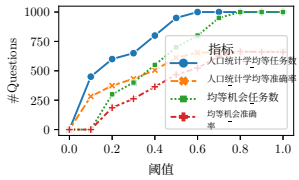


（a） *Crowd Judgement*

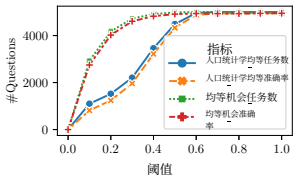


（b）Jigsaw Toxicity

图5：在不同公平性阈值下由不公平工人主导的任务比例



（a） *Crowd Judgement*



(b) *Jigsaw Toxicity*

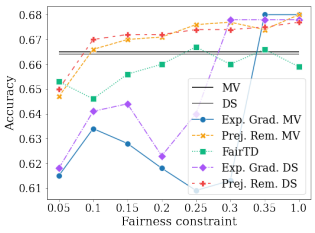
图6：移除不公平工人对整体准确率的影响

TD 算法	准确率	人口统计学均等差异	人口统计均等比率	均等赔率差值	均等赔率比率
多数投票	0.658	0.166	0.706	0.12	0.703
Dawid-Skene	0.663	0.183	0.667	0.136	0.665
从众包中学习	0.664	0.184	0.668	0.142	0.676

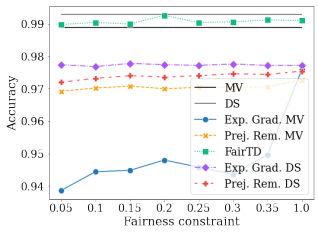
表2：不同TD算法在 *Crowd Judgement* 数据集上的性能

在某些情况下，增大公平性约束违反程度会提高准确率，但这种关系并未在所有受测数据集和方法中一致出现。此外，当公平工作者的准确率始终高于不公平工作者时，这一观察结果可能并不成立。我们可以看到，Prejudice Remover 通常能在保持较高准确率的同时，比 Exponentiated Gradient 更有效地缓解偏差。另一方面，FairTD 似乎是一种稳健的方法，在两种情况下都表现良好；即使约束值严格，它的表现也往往能与 Prejudice Remover 持平或更胜一筹。

根据我们的结果，与仅依赖公平机器学习相比，单独使用 FairTD 或将其与公平机器学习结合使用，似乎是未来研究中更有前景的方法。尽管指数梯度和偏见消除本身都是有价值的技术，但 FairTD 提供了一种稳健且精确的方法来减少偏差。因此，我们认为，未来的研究应集中于改进和优化 FairTD 方法，使其在应对真值发现算法公平性方面更加有效。



（a）公平机器学习与公平真值发现对比 *Crowd Judgement*



（b）公平机器学习与公平真值发现对比 *Jigsaw toxicity*

图7：公平机器学习与公平真值发现

9 结论

在本文中，我们研究了使用众包生成的机器学习数据集中的一个关键偏差来源。大多数此类数据集都使用简单的多数投票来汇总众包工人的冲突响应。然而，在存在不公平工人的情况下，这种方法并非最优，并且会产生较差的结果。我们开展了广泛的实验，以研究不公平工人对各种真值发现算法的影响，并得出结论：这种影响对标签准确率具有显著作用。现有的真值发现算法主要关注准确率，通常无法处理不公平工人。因此，有必要设计公平性感知真值发现算法，以抵消不公平工人对机器学习数据集标签的影响。

参考文献

[1] *Jigsaw unintended bias in toxicity classification*。 https://www.kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification/ data。访问日期：2023年3月20日。

[2] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, *Youtube-8m: A large-scale video classification benchmark*, arXiv 预印本 arXiv: 1609 08675, 。(2016年)。

[3] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach, *A reductions approach to fair classification*, 载于国际机器学习会议，PMLR，2018年，第60–69页。

[4] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning: Limitations and Opportunities*, 2019, fairmlbook.org，。http://www. fairmlbook.org。

版权 © 2023年 SIAM 未经授权的复制本文章，禁止复制

TD 算法	准确率	人口统计学奇偶校验差值	人口统计均等比率	均等赔率差值	均等赔率比率
多数投票	0.92	0.607	0.08	0.016	0.08
Dawid-Skene	0.93	0.636	0.08	0.051	0.05
从众包中学习	0.95	0.64	0.08	0.06	0.06

表 3：不同 TD 算法在 *Jigsaw Toxicity* 数据集上的性能

分类器	TD 算法	准确率变化	人口统计均等差异变化	均等赔率差异变化
随机森林	多数投票	6.04	0.06	0.07
随机森林	Dawid-Skene	7.63	0.04	0.06
随机森林	从众包中学习	7.54	0.04	0.05
逻辑回归	多数投票	2.55	0.04	0.07
逻辑回归	Dawid-Skene	2.77	0.05	0.08
逻辑回归	从众包中学习	2.86	0.05	0.07

表 4：性能　　使用无公平性意识的 TD 算法输出训练的机器学习模型，用于众包判断　　t

分类器	TD 算法	准确率变化	人口统计均等差异变化	均等赔率差异变化
随机森林	多数投票	5.44	0.07	0.09
随机森林	Dawid-Skene	4.63	0.05	0.07
随机森林	从众包中学习	6.23	0.05	0.06
逻辑回归	多数投票	4.32	0.05	0.01
逻辑回归	Dawid-Skene	3.99	0.05	0.08
逻辑回归	从众包中学习	3.96	0.04	0.07

表5：使用无公平性意识的 TD 算法输出训练的机器学习模型在 *Jigsaw Toxicity* 上的性能

[5] S. Bell, K. Bala, and N. Snavely, *Intrinsic images in the wild*, 《ACM 图形学汇刊》（TOG），33（2014年），第1–12页。[6] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles *Activitynet: A large-scale* , *video benchmark for human activity understanding*, 载于计算机视觉与模式识别会议，2015年，第961–970页。[7] A. P. Dawid and A. M. Skene, *Maximum likelihood estimation of observer error-rates using the em algorithm*, 《英国皇家统计学会会刊》，28（1979年），第20–28页。[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei , *Imagenet: A large-scale hierarchical image database* 2009, 248 255 ，载于计算机视觉与模式识别会议，电气电子工程师学会，第–页。[9] J. Dressel and H. Farid, *The accuracy, fairness, and limits of predicting recidivism*, 《科学进展》，4（2018年），第eaao5580页。[10] C. Eickhoff, *Cognitive biases in crowdsourcing*, 载于网页搜索与数据挖掘会议，2018年，第162–170页。[11] M. Haklay and P. Weber, *Openstreetmap: User-generated street maps*, 《IEEE 普适计算》，7（2008年），第12–18页。[12] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, *Fairness-aware classifier with prejudice remover regularizer* 2012, 35 50 ，载于欧洲机器学习与知识发现数据库会议，施普林格，第–页。[13] Y. Li, H. Sun, and W. H. Wang, *Towards fair truth discovery from biased crowdsourced answers*, 载于知识发现与数据挖掘会议，2020年，第599–607页。[14] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona D. Ramanan P. Dollár and C. L. , , , Zitnick, *Microsoft coco: Common objects in context*, 载于欧洲计算机视觉会议，施普林格，2014年，第740–755页。[15] P. K. Lohia, K. N. Ramamurthy, M. Bhide, D. Saha K. R. Varshney and R. Puri , , , *Bias mitigation post-processing for individual and group fairness*, 载于国际声学、语音与信号处理会议，电气电子工程师学会，2019年，第2847–2851页。[16] N. Mehrabi、F. Morstatter N. Saxena K. Lerman and A. Galstyan , , , , *A survey on bias and fairness in machine learning*, ACM 计算调查（CSUR），54（2021），第1–35页。[17] C. G. Northcutt、 A. Athalye、 and J. Mueller、*Pervasive label errors in test sets destabilize machinelearning benchmarks*, arXiv 预印本 arXiv:2103.14749, （2021）。[18] K. Peng、 A. Mathur、and A. Narayanan、 *Mitigating dataset harms requires stewardship: Lessons from 1000 papers* 2108 02922, 2021 ，arXiv 预印本 arXiv:。（）。[19] V. U. Prabhu and A. Birhane、*Large imagedatasets: A pyrrhic win for computer vision?*, arXiv 预印本 arXiv:2006.16923, （2020）。[20] V. C. Raykar、 S. Yu、L. H. Zhao、 G. H. Valadez、C. Florin L. Bogoni and L. Moy *Learning from* , , , *crowds.* 11 2010 ，机器学习研究期刊，（）。[21] B. Shmueli、 J. Fell、 S. Ray、 and L.-W. Ku、*Beyondfair pay: Ethical implications of nlp crowdsourcing*, arXiv 预印本 arXiv:2104.10097, （2021）。[22] B. H. Zhang、 B. Lemoine、and M. Mitchell、*Mitigating unwanted biases with adversarial learning*, 载于2018年 AAAI/ACM 人工智能、伦理与社会会议论文集，2018年，第335–340页。[23] Y. Zheng、 G. Li、 Y. Li、 C. Shan、 and R. Cheng、*Truth inference in crowdsourcing: Is the problem solved?* 10 2017 541 552 ，《VLDB 会刊》基金会，（），第–页。[24] B. Zhou、 A. Lapedriza、 A. Khosla、 and A. Oliva、A. Torralba 、*Places: A 10 million image databasefor scene recognition*, IEEE 模式分析与机器智能汇刊，40（2017），第1452–1464页。

版权 © 2023年 SIAM 未经授权的复制本文被禁止