在机器学习中减少基于模型的偏差

Alfa Yohannis 约
克大学 英国约克
alfa.yohannis@york.ac.uk

迪米特里斯-科洛
沃斯 英国约克大学
约克分校
dimitris.kolovos@york.ac.uk

摘要

机器学习生成的模型不能保证没有偏差，尤其是在使用歧视性环境下生成的数据进行训练和测试时。偏差可能是不道德的，主要是当数据包含性别、种族、年龄等敏感属性时。一些方法通过提供偏差度量和缓解算法，为缓解此类偏差做出了贡献。面临的挑战是，用户必须用通用/统计编程语言来实现他们的代码，这对缺乏编程和机器学习公平性经验的用户来说要求很高。我们提出了 FairML，这是一种基于模型的方法，可在减少软件开发工作量的同时促进偏差测量和缓解。我们的评估结果表明，FairML 只需要较少行代码，就能生成与基线代码生成的测量值相当的测量值。

综合传播战略概念

• 计算理论 → *学习模型*； - **应用计算** → **法律、社会和行为科学**； - **软件及其工程** → **源代码生成**；**特定领域语言**。

关键词

模型驱动工程、生成式编程、偏差缓解、偏差度量、机器学习

ACM 参考格式：

Alfa Yohannis 和 Dimitris Kolovos。2022.机器学习中基于模型的偏差缓解（Towards Model-based Bias Miti- gation in Machine Learning）。In *ACM/IEEE 25th International Conference on Model Driven Engineering Languages and Systems (MODELS '22), October 23-28, 2022, Montreal, QC, Canada.*ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3550355.3552401

1 引言

如今，机器学习的应用无处不在，从个人日常活动，如用于身份验证的人脸识别和与智能助手（如 Alexa、Siri）交谈时的语音处理，到执行敏感的道德任务，如预测犯罪活动，再到对犯罪活动的侦查，机器学习都能发挥重要作用。

MODELS '22，2022 年 10 月 23-28 日，加拿大不列颠哥伦比亚省蒙特利尔市

978-1-4503-9466-6/22/10.. . $15.00
https://doi.org/10.1145/3550355.3552401

虽然机器学习技术确实能提高效率，但机器学习产生的模型并不能保证不带偏见，尤其是在使用歧视性环境下产生的数据进行训练和测试时。这可能是不可接受的，主要是当它触及性别、种族、年龄等敏感属性时，会放大不公平。

2016 年，一种预测累犯的算法 COMPASS 被发现对白人产生较高的假阴性率，对黑人产生较高的假阳性率[5]。一些商业人脸识别服务对深肤色女性的准确率也明显较低[11]。此外，一个求职平台发现，合格的女性求职者的排名远远低于合格的男性求职者，即使他们具有相似的属性[33]。这些事例表明，机器学习中的偏见会造成不公平。

一些方法通过提供偏差度量和去毛刺算法（详见第 2.2 节和第 2.3 节）来减少此类偏差。不同的工具包已经实现了这些指标和算法。不过，它们的方法和功能各不相同，用户需要深入了解后才能决定哪种工具包最适合特定场景[34]。

数据科学家通常利用直觉来缩小算法、参数和其他因素的组合数量，从而为给定目标、数据集和领域找到最佳模型[37]。之后，通过大量的实验和试错 [12]，他们必须对所有缩小范围的组合进行分析，并对生成的模型进行测试，以确定哪些模型是最好的。此外，无论机器学习库是否可用，数据科学家都必须使用通用/统计编程语言（如 Python、R）从头开始设计搜索过程。

模型驱动软件开发（MDSE）通过隐藏可自动化实现的技术细节，充分利用了软件开发的弊端[10]。因此，用户可以通过使用更简单的建模语言专注于重要方面，目标实现可以自动生成，这反过来又提高了生产率[47]。MDSE 的使用将使数据科学家受益，因为它允许他们在更高的抽象层次上为给定案例搜索最佳的偏差缓解方法。他们也不必用通用/统计编程语言来编码搜索过程的实现，因为它可以
自动生成，然后再进行微调。

在本文中，我们介绍了 FairML[1]，它是一种实现 MDSE 方法的工具，用于机器学习中偏差测量和缓解的建模和自动化。

[1] FairML 原型见 https://doi.org/10.5281/zenodo.7007839 和 https://github.com/York-and-Maastricht-Data-Science-Group/fairml

(1) FairML 提高了偏差测量和减缓的抽象程度，因此用户可以用 YAML（YAML 不是标记语言）这种对人类友好的声明式语言 [19]来配置他们的偏差减缓模型，而无需使用通用/统计编程语言进行编码。
(2) FairML 支持一定程度的表达能力，允许用户尝试不同类型的偏差度量、偏差缓解算法、数据集、分类器及其参数，以找到既能减少偏差又能保证可接受准确度的最佳组合。
(3) 支持工具会自动生成 Python 和 Jupyter Notebook 文件，用户可以执行这些文件来测量和减轻给定数据集上的偏差。所有生成的文件都可以修改和扩展，以便进行微调和进一步开发。

本文的结构如下。首先，我们在第 2 节讨论机器学习中的公平性。该部分涵盖了一些相关术语的定义、实际示例、偏差度量和偏差缓解。在第 3 节中，我们概述了我们的开发情况，以 及 我们在机器学习领域实施去除法技术以自动减缓偏差的方法。本节还介绍了用户如何使用现有工具包来表达减轻分类偏差的解决方案。我们还演示了 FairML 如何以更易读、更简洁的方式表达解决方案，并自动生成与人工创建的解决方案具有相似正确性的偏差测量和缓解代码。在第 4 节中，我们以偏差缓解工具包的现有代码为基准，对 FairML 的表达能力、正确性、生成时间和执行时间进行了评估。第 4.1 节介绍并讨论了评估结果。我们还在第 5 节中反思了在研究和开发 FairML 的过程中获得的一些经验教训。第 6 节讨论相关工作，第 7 节介绍本研究的结论和未来工作。

2 机器学习中的偏见

本节简要讨论了机器学习中的公平性，包括一些相关术语的定义、实际案例、偏差度量和偏差缓解。

2.1 定义和示例

公平的定义是 "对个人或群体没有任何基于其固有或后天特征的偏见或偏袒"[36]。偏差是指由于数据收集和处理、研究设计、分析和解释中的缺陷而导致的系统性错误或对实际情况的扭曲[38]。当偏见使享有特权的群体与未享有特权的群体相比处于有利地位时，就会产生不公平[7]。如果基于敏感属性（种族、年龄、性别、性别差异）的刻板印象和偏见是有意或无意的，那么有害的歧视也会造成偏差、但群体公平并不一定表示个人公平。此外，由知名厂商提供的公开商业人脸识别在线服务被发现对肤色较深的女性的准确率要低得多[11]。机器学习中存在偏见的例子还有很多，但这三个例子已经说明，在实践中，机器学习并不总是公平的，而且这种不公平会给特定群体带来不利。

2.2 偏差度量

目前已开发出一些指标来检测和测量机器学习中的偏差。每个指标都有其计算偏差的方法，因此在特定情况下优于其他指标。IBM AI Fairness 360 [26, 35] 和 Aequitas2 提供了选择指导，如图 1 所示。因此，如果偏差测量需要包含群体和个人公平性的指标，那么Theil 指数[7, 16]和广义熵指数[44]更可取。它们可以作为统一的指标来衡量群体和个人利益分配的不平等程度[26, 35]。较低的分数反映了较强的公平性，而较高的分数则显示了相反的情况。0 表示完全公平 [25]。

欧氏距离（Euclidean Distance）、曼哈顿距离（Manhattan Distance）和马哈罗诺比距离（Mahalanobis Dis- tance）是衡量个体公平性的指标。这些指标用于测量同一个体在原始数据集和去偏数据集中的距离[7]。由于在预处理中应用偏差缓解时，数据转换可以实现群体公平，但可能会对个人造成不公平，因此需要这些指标来确保一致性。在文献[13]中，这些指标被用作限制其偏差缓解失真的约束条件，从而同时实现群体公平和个体公平。

关于群体公平性，世界上主要有两种观点：平等公平和亲和公平[26, 35]。平等公平是指代表性不足的群体在预测结果方面应享有与其他群体相似的机会，通常适用于被认为存在结构性歧视的系统。例如，SAT 就被怀疑存在结构性歧视 [26，35]。统计均等差异 [18, 26, 35] 和差异影响 [20, 26, 35] 是衡量平等公平性的常用指标。

统计均等度差的计算方法是特权组和非特权组之间被贴上有利标签的概率之差（等式 1）[4, 7, 18]。

$$SPD = Pr(\hat{\square} = 1 | D = \text{unpriv}) - Pr(\hat{\square} = 1 | D = \text{priv}) \quad (1)$$

该指标的理想值为 0。负值意味着标签有利于特权群体，而正值则意味着相反[4, 7]。差异影响[4, 7, 20]将公平性计算为非特权群体和特权群体之间被贴上有利标签的概率之比（等式 2）。

$$\frac{Pr(\hat{\square} = 1 | D = \text{无特权})}{}$$

性别等）[15, 36]。

例如，2006 年，研究人员发现，一种算法用于在预测累犯时，黑人的假阳性率远远高于白人 [5]。另一个例子是一个求职平台，人们发现该平台将资质较低的男性求职者排在了前面高于更合格的女性候选人[33]。该论文指出实现了男女群体之间的公平、

$$\square\square = Pr(\square\hat{} = 1 | D = \text{特权}) \quad (2)$$

约哈尼斯和科洛沃斯

该指标的理想值为 1.0。值 **<** 1 表示标签有利于特权群体，值 **>** 1 则不利于非特权群体 [4，7]。

[2] http://aequitas.dssg.io/static/images/metrictree.png

偏差度量树

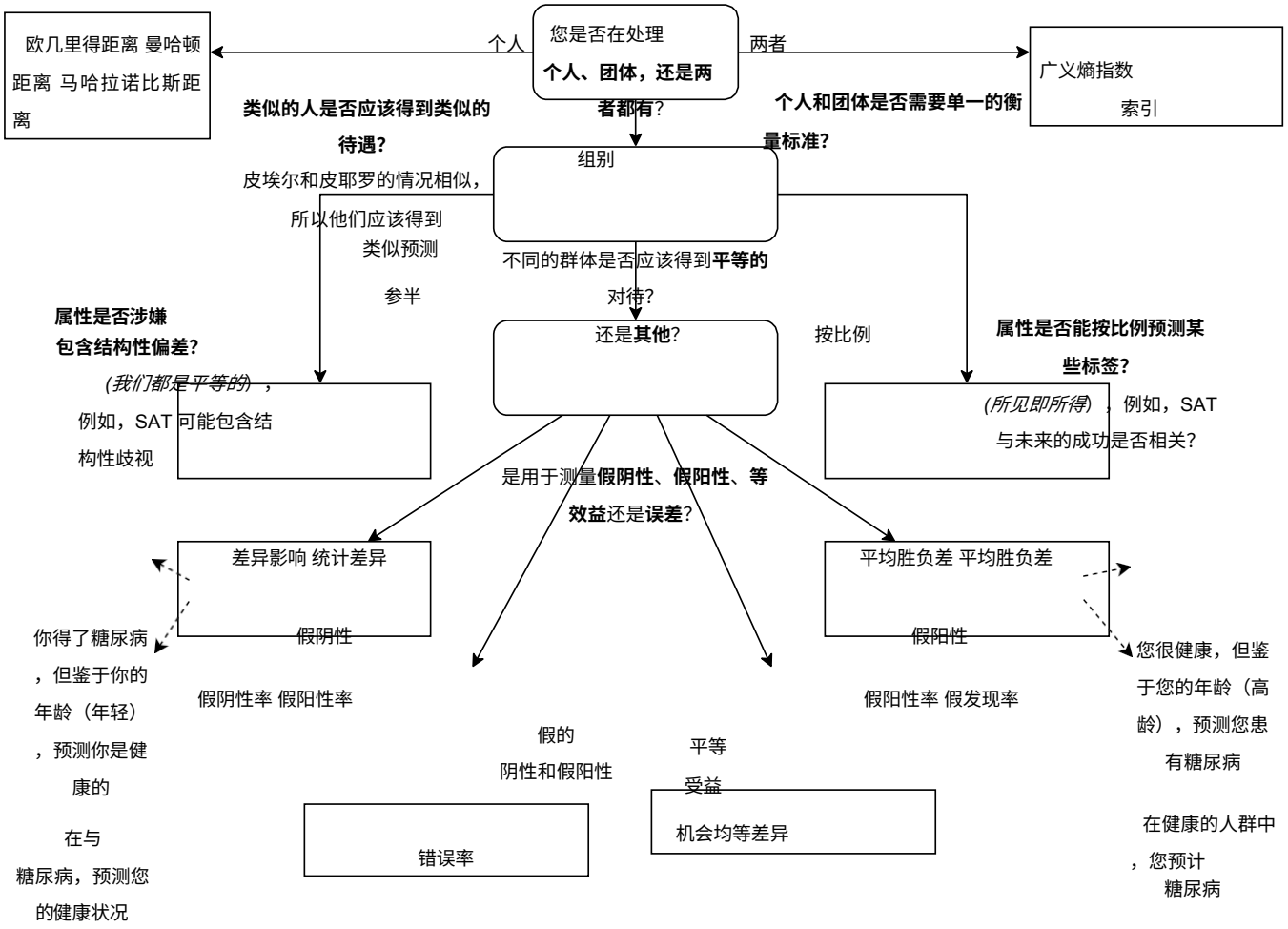


图 1：针对特定情况自动选择最合适偏差指标的决策树。

与此相反，比例公平（又称所见即所得（WYSIWYG））认为某些特征与某些标签相关，因此这些特征可用来预测标签。例如，SAT 分数与未来收入相关[26, 35]。平均赔率差是衡量

比例公平性。该指标将偏差计算为假阳性率（□ □ □ = □/□，假阳性/所有阴性）和真阳性率（□ □ □ = □ □/□，真阳性/所有阳性）（等式 3) [4, 7]。

$$AOD = \frac{1}{2} * ((F_{\square\square\square=\text{unprivileged}} - F_{\square\square\square=\text{privileged}}) + (TPRD=\text{unprivileged} T_{\square\square\square=\text{privileged}})) \quad (3)$$

理想的度量值为 0。

特权群体比非特权群体更有利，而正值则表示相反[4, 7]。

真阳性率是给定组别的真阳性结果与所有阳性结果总数的比率（□ □□ = □ □/□）。理想值为 0，负值表示偏好程度较高而正则相反 [4, 7]。该领域的其他指标包括假阴性率比值和差值、假遗漏率比值和差值、错误率、假阳性率比值和差值以及假发现率比值和差值 [26，35]。

2.3 减少偏差

目前已开发出去锯齿算法，以减少测量数据中的偏差。

它们可以根据机器学习管道中的应用阶段进行分类：预处理

、

还有一些指标介于平等公平和有利公平之间。例如，

机会均等差值[4, 7]是一种将偏差计算为无特权群体与有特权群体之间真实阳性率之差的指标（等式 4）。

$$EOD = \square\square\square_{=\text{unprivileged}} - \square\square\square_{=\text{privileged}} \quad (4)$$

处理中和处理后。文献[26, 35]提供了针对具体情况选择最合适的去毛刺算法的指导。图 2 总结了这些指导。

参考文献[35]建议尽早减少不公平现象，尤其是在预处理阶段，因为在该阶段减少的偏差是数据集固有的。遗憾的是，有些情况下是禁止修改数据集的。因此，如果不允许修改数据集，重权重[27]不失为一种选择

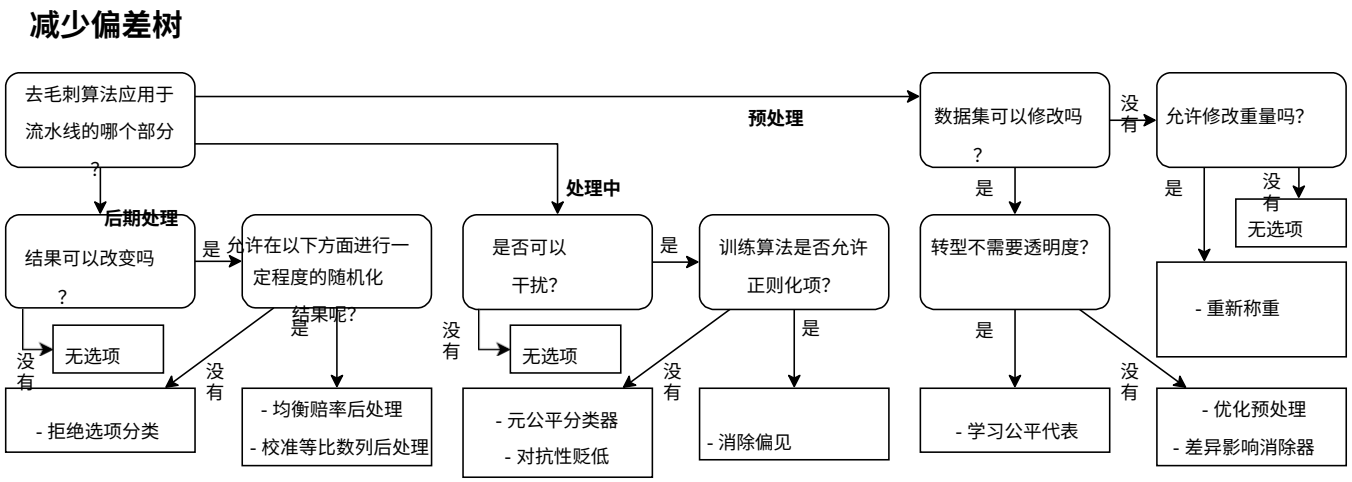


图 2：针对特定情况选择最合适去毛刺算法的决策树。

因为它不会改变特征值，而只是通过改变每个标签/预测的权重来获得公平性。如果允许对数据集进行修改，优化预处理法[13]、差异影响去除法[20]和学习公平表示法（LFR）[50]都是合适的选择，但 LFR 由于将数据编码到潜空间中，因此不提供反转性，因此在需要数据集转换的透明度时，前两种方法更可取。优化预处理可学习一种概率变换，这种变换可修改数据集的标签和特征，但在个体失真、数据保真度和群体公平性方面有目标和约束[13, 35]。差异影响移除器（Disparate Impact Remover）会改变特征值，以提高分组的公平性，同时保留组内的排序[20, 35]。LFR 通过将数据编码到潜空间来预处理数据集，并混淆受保护属性的值[35, 50]。

在处理过程中，也可以通过干扰训练算法来减少偏差。在处理过程中应用的去偏算法主要通过惩罚偏差来实现，通常使用公平性约束或目标来限制或规范分类器中的偏差[35]。例如，Prejudice Remover [29] 在学习目标中添加了一个歧视意识正则化项。不过，它仅限于支持正则化项的学习算法 [26, 35]。由于存在这种限制，用户可以考虑采用其他的内处理算法，例如对抗去偏差算法[26, 35]，这些算法允许使用更通用的学习算法集。Adversarial Debiasing [51] 通过学习分类器来最大限度地提高其预测准确性，同时降低对手从预测中确定受保护属性的能力。这种方法使预测结果无法携带敌方可以利用的任何群体歧视信息，因此可以获得公平的分类结果[26]。其他内处理除杂算法有元公平分类法 [14]、格里公平分类法 [30，31]、幂级数公平分类法 [27]、幂级数公平分类法 [28]、幂级数公平分类法 [29]。
梯度缩减法 [2] 和网格搜索缩减法 [2, 3]。

只有在预测结果可以改变的情况下，才能进行后处理偏差缓解。剔除选项分类[28]有利于非特权群体的结果，而在决策周围的置信区间内则不利于特权群体的结果。

不确定性最高的边界。均衡赔率后处理[23, 39]和校准均衡赔率后处理[39]试图找到改变输出标签的概率，以优化均衡赔率目标。前者通过求解线性程序来实现，后者则对校准分类器分数进行优化。这两种均衡赔率算法都有随机成分，因此都是首选。否则，剔除选项算法就是备选方案，因为它是确定性的。

3 基于模型的偏差缓解

本节介绍我们在机器学习中提供基于模型的偏差缓解方法。

3.1 工具包选择

MDSE 主要依靠代码生成来生成目标软件的代码。生成的代码通常使用现有工具或软件库来实现相应领域的某些功能。机器学习也不例外；可以利用现有的工具来减少机器学习中的 道德偏见。

Lee等人[34]分析了算法伦理学开源工具包的现状，尤其是公平机器学习方面。工具包的标准是：1）应该是开源的；2）有可能被从业者使用；3）正在实施与公平相关的方法论，通过数据科学家的焦点小组讨论，他们选出了六个最好的公平机器学习工具包，并在分析中进一步使用了这些工具包。这些工具包是 Aequitas[41]、Google What-If [48]、Scikit-fairnet/scikit-lego[42, 43]、Fairlearn[9]、和 IBM AI Fairness 360 [7]。

我们对这六个工具包进行了评估，以确定它们是否适合作为我们基于模型的方法的基础结构。选择的主要标准是：1）它应是开源的；2）它应支持偏差缓解；3）它应提供应用编程接口（API），以便可编程并集成到基于模型的方法中；4）它应支持各种偏差测量和偏差缓解。

减小偏差算法，使用户能够找到最佳的减小偏差策略。

所有工具包都符合第一条标准，因为它们都是开源项目。我们将 Aequitas 排除在外，因为它无法进行偏差缓解（第二项标准）；它只支持偏差测量，而不支持偏差缓解。我们还删除了谷歌 What-If，因为它没有 API3 与其他应用程序集成（第三条标准）；它的定制功能仅限于定制预测功能4。我们还删除了 Google What-If，因为它没有 API3 可与其他应用程序集成（第三条标准）；其定制功能仅限于定制预测功能[4]。IBM AI Fairness 360 的功能比其他产品更多。它声称自己提供了至少 10 种最先进的偏差缓解算法和 77 个偏差度量指标[5]，而 Fairlearn 和 Scikit-fairness/lego 位居第二和第三[34]。它们都符合第四条标准。不过，由于 IBM AI Fairness 360 比其他公司拥有更多的功能，因此我们选择它作为我们基于模型的解决方案的基础架构。

3.2 结构和工作流程

要对机器学习中的偏见缓解进行建模，特定领域语言应该能够让用户表达该领域的基本结构和工作流程[47]。文献[7]中的工作记录了使用 IBM AI Fairness 360 工具包实现偏差缓解的重要结构和典型工作流程/管道，我们在此对其进行简要讨论。

设置数据集（T01）。减轻偏差通常从设置*数据集*开始。在这项任务中，用户要定义每个数据集的*来源*，通常是 *CSV 文件*。他们还要确定哪个属性是*预测属性*，包括at- tribute 中的*有利类别*、*敏感属性*及其*特权*和*非特权类别*。数据集也可以从某些工具包提供的*内置数据集中*加载。*训练*、*验证*和*测试数据集*也在此设置，无论它们是全部来自不同的数据集，还是同一数据集但按不同比例*分割*。他们还可以选择将应用于数据集、*模型*和*预测的偏差度量*和*缓解算法*。

预测前测量原始偏差（T02）。在这项任务中，用户用不同的指标测量数据集的*原始偏差*，如准确率、平均差等。这些结果随后将作为基准与预测（T04）或偏差缓解（T06）后的测量结果进行比较。

训练和预测（T03）。该任务通常使用任务 T01 中定义的训练数据集上的某些*分类器/分类算法来训练模型*。如果需要，训练可以扩展到*验证*，以调整超参数，从而为模型获得最佳*参数*。之后，用户使用模型对任务 T01 中定义的测试数据集进行*预测*。如果不需要预测，可以跳过这个任务。例如，我们只想测量和减轻原始数据集的偏差（预处理）。

预测后测量原始偏差（T04）。在该任务中，用户在完成任务 T03 中的预测后测量原始准确度和偏差。这些结果将作为以后与偏差缓解后的偏差进行比较的基准。如果未执行预测或任务 T03，则跳过此操作。

减少偏差（T05）。在这项任务中，用户可以选择去偏差算法的类型--预处理、处理中或处理后

[3] https://groups.google.com/g/what-if-tool/c/zw4Rk5kxPIM
[4] https://pair-code.github.io/what-if-tool/get-started/
[5] https://aif360.mybluemix.net/

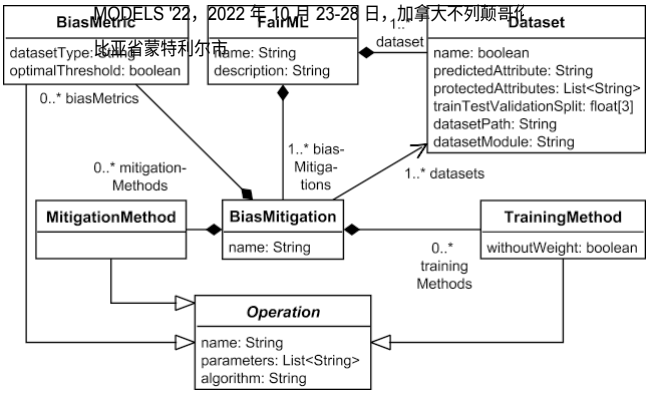


图 3： FairML 的元模型。

- 如果需要对去毛刺算法进行超参数调整，也可以进行验证。如果需要进行超参数调整，以获得去毛刺算法的最佳参数，也可以进行*验证*。

测量去毛刺后的偏差（T06）。在该任务中，用户使用测试数据集测量任务 T05 中应用去毛刺算法后的准确度和偏差。

总结（T07）。在该任务中，将去偏差后的偏差与任务 T02 和 T04 中获得的原始偏差进行比较。然后，用户分析结果，找出数据集、分类器和去偏差算法的最佳组合，使其偏差和准确性符合自己的目的。比较结果可以用数字、表格和图表的形式呈现，以方便用户分析数据集、分类器和除杂算法的效果。

3.3 元模型

上述构造已被纳入 FairML 的元模型，以便用户在定义减轻偏差实验时可以表达它们。图 3 显示了元模型的简化版本。FairML 类代表了机器学习中的偏差缓解项目。它可以包含一个或多个数据集和偏差缓解。

数据集类定义数据集。在该类中，我们指定了数据集的名称、预测属性、保护属性、训练、测试和验证之间的分割比例，以及 CSV 文件的路径。

BiasMitigation 类指定了要在 FairML 项目中执行的偏差缓解措施。*name* 属性用作偏差缓解的标识符。该类由另外三个重要的类组成：TrainingMethod（训练方法）、MitigationMethod（缓解方法）和 Bias- Metric（偏差度量）。所有这些类都是从操作抽象类派生出来的，因为它们可以被执行、具有参数并返回结果。属性 *name* 和 *parameters* 分别用作标识符和定义操作的参数。操作类的算法属性定义了每个派生类所执行的选定算法。TrainingMethod 类负责定义用于训练和预测的分类器，而 MitigationMethod 类则指定用于减少偏差的去毛刺算法。BiasMetric 类确定用于衡量公平性的指标。

3.4 FairML 示例

清单 1 显示了一个 FairML 模型，它符合图 3 中的元模型，用 YAML 表示，是 IBM AI Fairness 360 示例 demo[6] 的 FairML 版本。该示例使用*成人*[7] 数据集来训练和测试预测模型。该示例没有减少数据集中的偏差（预处理偏差缓解），而是在机器学习管道的内处理阶段应用幂梯度降低技术来减少偏差。最后，该示例比较了使用和未使用处理内偏差缓解技术所生成模型的准确性和公平性。原始示例有 93 行 Python 代码，不包括注释行和空行，几乎是清单 1 的三倍。

该模型有一个名称和描述，分别为 Demo 和使用 EGR 的去偏差（第 3-4 行）。模型有一个数据集（第 7 行）和一个偏差缓解（第 15 行）。数据集名为 Adult，从 csv 文件 load_preproc_data_adult.csv 加载。我们还将性别和种族设置为受保护属性，将二进制作为预测属性，并将原始数据集按 7:3:0 的比例分割为训练数据集、测试数据集和验证数据集（第 7-12 行）。

第 15-32 行定义了在演示模型中执行的偏差缓解的定义。偏差缓和被命名为指数梯度降低，并使用第 7-12 行定义的成人数据集作为其数据集。

当预测仅使用 Logistic 回归分类器而未去重时（第 19-21 行），以及当预测使用幂梯度还原法去重时（第 23-25 行），偏差缓解使用了三种不同的指标来分析准确性和公平性之间的权衡（平均差异、平均几率差异）（第 27-32 行）。前者只使用了一个参数；"lbfgs "被设为求解器，而后者有三个参数；同一分类器被用作估计器，"EqualizedOdds "被用作约束条件，并且没有放弃受保护的属性。

清单 1：使用 YAML 表示的 Demo Exponentiated Gradient Reduction 减少偏差。

```
1  nsuri：fairml
2  fairml：
3  - 名称： 演示
4  - 描述： 使用 EGR 去毛刺
5
6  # set the dataset
7  - 数据集：
8      - 名称：成人
9      - 预测属性：收入二进制
10     - 受保护属性：性别、种族
11     - train TestValidation Split：7，3
12     - 数据集路径：load_preproc_data_adult.csv
13
14 # define the bias mitigation
15 - 减少偏差：
16     - 名称： 指数梯度削减法
17     - 数据集：成人
18
19 - 培训方法：
```

[6] https://github.com/Trusted-AI/AIF360/blob/master/examples/demo_exponentiated_gradient_reduction.ipynb.
[7] https://archive.ics.uci.edu/ml/datasets/Adult

```
    - 算法： 逻辑回归
    - 参数： 求解器='lbfgs'

  - 缓解方法：
    - 算法： 幂级数递减法
    - parameters：estimator= LogisticRegression(solver='lbfgs'), constraints='EqualizedOdds', drop_prot_attr=False

  - 偏差度量：
    - 名称： 精确度
  - 偏差度量：
    - 名称： 平均差
  - 偏差度量：
    - 名称： 平均赔率差值
```

3.5 向导

虽然 FairML 提供了指定偏差缓和的方法，但用户仍需要了解所有支持的除杂算法和偏差度量，才能正确使用它们。使用向导以自然语言提问的方式帮助用户构建 FairML 模型非常重要。该向导遵循图 1 和图 2 中的决策树，两者分别在第 2.2 和 2.3 节中讨论。

完成向导后，FairML 会自动生成一个 FairML 模型 n Flexmi（见下文）。如果用户以后想添加其他算法或修改现有的除杂算法和偏差指标，可以修改该文件。向导还会自动生成生成模型的 Python 和 Jupyter 笔记本文件。

清单 2：FairML 向导中提出的一些问题，以帮助用户选择最佳除杂算法和偏差指标。

```
...
---- 缓解算法 ----
# 1.预处理
在预处理中应用偏差缓解 （默认值： true）：
数据集的权重可以修改 （默认值： true）：
偏差缓解允许潜在空间 （默认值： false）：
...
...
---- Bias Metric ----
衡量组的公平性 （默认值： true）：
衡量个体公平性 （默认值： false）：
对个人和组使用单一指标 （默认值： false）：
衡量公平性 （默认值： false）：
...
```

3.6 一代人

图 4 描述了消耗 FairML 模型并生成 Python 和 Jupyter Notebook 文件的转换过程。用户定义他们的

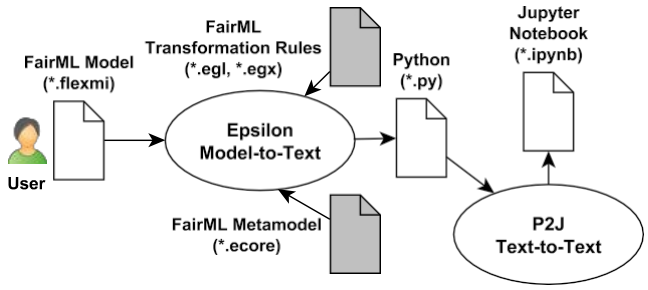


图 4：fairML 中模型到文本的转换。

Flexmi[8] [32]中的偏差缓解模型。Flexmi 是 EMF 模型的反映文本语法[45]，它使用模糊匹配将 YAML/XML 文档中的标记和属性映射到目标元模型中的 Ecore 类/特征名[45]。除了支持默认的基于 XML 的格式外，Flexmi 还能读取以 YAML 格式表达的模型。一个用 Epsilon 生成语言（EGL）[40] 编写的转换器可以读取符合 FairML 元模型的模型，并生成 Python 文件。注意到同样的转换可以用任何模型到文本语言来实现。然后使用 P2J[9] 引擎将生成的 Python 文件转换为 Jupyter 笔记本文件。

3.7 生成文件

根据清单 1 中定义的模型，FairML 会生成一个 Jupyter 笔记本文件，其中包含执行偏差缓解的代码、对所测偏差的解释以及对所选分类器和去重算法的引用，以帮助用户理解所应用的偏差缓解和所测指标。它还包含一个带颜色编码的表格，比较了不同数据集、分类器和去噪算法测得的偏差指标，以帮助用户选择适合其情境的最佳组合。

例如，图 5 显示了清单 1 生成的偏差测量汇总表，它是生成的笔记本截图的一部分。有三种预测组合。在第一种组合中，我们只测量平均差、

即原始数据集中的统计奇偶校验差。在使用逻辑回归进行预测后，第二个组合在没有任何偏差缓解的情况下测量了一些指标。从表中可以看出，平均差异略微恶化为-0.206，远离 0，有利于特权组（负号）。在使用指数梯度降低算法作为消除偏差的算法后，预测准确率略微降低到 0.79，但公平性有了显著提高；平均差和平均几率差都比消除偏差前的值更接近零。

在有些情况下，用户会遇到许多不同的数据集、分类器、去偏算法和偏差度量值组合。在这些选项中，用户通常需要选择最适合其偏差缓解策略的组合。因此，我们通过将每个指标的最佳测量值以粗体格式化，并用颜色标记来帮助用户。偏差度量值

[8] https://www.eclipse.org/epsilon/doc/flexmi
[9] https://pypi.org/project/p2j/

编码等级从白色到黄色再到绿色；每个等级对应各自指标的最差、中等和最佳测量值。在图 5 中，用户可以毫不费力地发现第二种组合的准确率最高。用户还可以发现，第三个选项是偏差最小的组合，但准确率仍然很高。

4 评估

我们评估了 FairML 的表现力、正确性、简洁性和执行时间。在表现力评估方面，我们的目标是回答 "*FairML 能否支持真实世界中的偏见缓解用例？* 为此，我们使用 IBM Fairness AI 360 提供的示例作为评估 FairML 表达能力的基准。该工具包附带了一些示例（表 1），展示了该工具包在现实世界中的用例。我们使用了工具包 3.0 版中的 12 个示例。我们确定了每个示例中使用的所有元素--偏差度量、除杂算法、分类器和数据集--以衡量其复杂性。示例涵盖的元素越多，FairML 能表达的情况就越复杂。

我们将原始示例中测得的偏差度量值（测量值）与生成代码中测得的值进行了比较，以评估 FairML 生成的偏差度量值的正确性。

代码。两者应产生相同或相似的值，误差在 ± 0.1 范围内。
在简洁性评估方面，我们比较了书写的
对比 FairML 生成的代码与示例中的代码。虽然由于多种因素（见第 4.2 节），这一指标并不总能保证生产率，但用户编写的行数却大大减少了。在测量中，空白行和注释行没有计算在内。

我们还测量了 FairML 的执行时间。首先，我们测量了*生成时间*--FairML 生成偏差检测和缓解代码所需的时间。其次，我们比较了生成的目标代码与示例代码完成操作所需的时间（*生成执行时间与原始执行时间*）。评估是在装有 Windows 10 64 位操作系统、第 11 代英特尔（R）酷睿（TM）i9-11900H @ 2.50GHz 8 核处理器、32.0 GB DDR4 内存、Open- JDK Runtime Environment 18.9（build 11.0.14.1+1）和 Python 3.9.7 的机器上进行的。

4.1 结果与讨论

在本节中，我们将介绍和讨论 FairML 的评估结果[10]。

*4.1.1 表现力和正确性。*FairML 能够再现表 1 中的所有 12 个示例。这些场景总共包括 6 个独特的数据集（成人、德语、Compas、MEPSDataset19、MEPSDataset20、MEPSDataset21）、6 个分类器（逻辑回归、线性回归、线性 SVR、决策树、核

岭、Ran- dom Forest）、11 个偏差缓解算法（Adversarial Debiasing、校准均衡赔率、差异影响消除、指数化梯度降低、Gerryfair、学习公平表征、重权、元公平分类、优化预处理、Re- ject 选项分类、偏见消除），以及 13 个偏差度量指标

---

[10] 评估数据见 https://github.com/York-and-Maastricht-Data-Science-Group/fairml/blob/main/data/evaluation.xlsx

最差 = 白色，最佳 = 绿色，中等 = 黄色

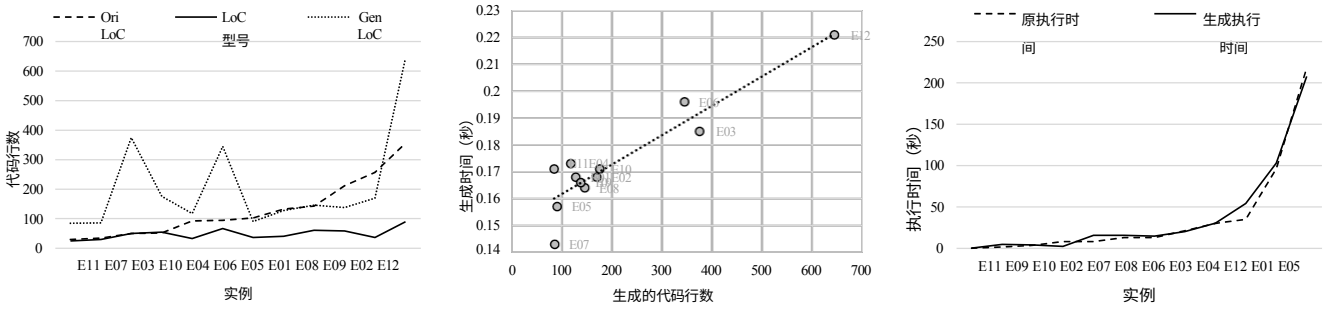
	缓解	数据	集分类器准确率平均值_差值平均值_差值		
1	原始成人（7.0:3.0:0.0）		无	-0.198048	无
2	原始成人（7.0:3.0:0.0）	逻辑回归求解器='lbfgs'	**0.804204**	-0.205572	-0.272736
3	ExponentiatedGradientReduction estimator=LogisticRegression(solver='lbfgs') 、 constraints='EqualizedOdds' 、 drop_prot_attr=False	成人（7.0:3.0:0.0）	ExponentiatedGradientReduction estimator=LogisticRegression(solver='lbfgs') 、 constraints='EqualizedOdds', drop_prot_attr=False	0.787552	**-0.052739**	**0.010994**

图 5：根据清单 1 中定义的模型生成的 Juypter 笔记本文件中显示的汇总表。

表 1：IBM AI Fairness 360 示例以及每个示例中的独特数据集、分类器、去重算法、偏差指标和测量值的数量。

代码	示例/文件名 (*.ipynb)	数据集	分类器	去重算法	偏差度量	措施
	E01演示对抗性贬损	成人	不适用		准确性，平衡准确性、差异影响、平均几率、统计均等、机会均等、Theil 指数	28
E02	演示校准方程奇数后处理	成人，德语，Compas	逻辑回归	校准等价几率后处理	平均差，假阳性率、假阴性率、均衡准确率、机会均等	13
	E03演示差异影响消除器	成人	逻辑	回归差异影响清除剂	差异影响	11
	E04演示幂级数降低	成人	逻辑回归	指数化减少梯度	精度，平均差、平均赔率	8
	E05演示 LFR	成人	逻辑	回归学习公平代表性差异	均衡准确性，平均值，差别影响	5
	E06演示元分类器 13	成人		N/AMeta 公平分类器	准确率，平衡准确率、差异影响，虚假发现率	
	E07演示优化预编程成人	成人	不适用	优化预处理	平均差	2
	E08演示剔除选项分类	成人，德语，Compas	逻辑回归	剔除选项分类	均衡准确性，差异 26 影响、平均几率、统计均等、机会均等、Theil 指数	
E09	演示重称预处理	成人，德语，Compas	逻辑回归	重重权衡	均衡精度，差异 16 影响、平均几率统计均等、机会均等、Theil 指数	

E10	演示简短的格里非尔测试	成人	线性回归、线性 SVR、决策树、核岭		GerryfairGamma 差异	15
E11	信用评分教程	德国	不适用	重新称重	平均差异	2
E12	医学教程支出	MEPS-数据集 19、MEPS-数据集 20、MEPS-数据集21	随机森林逻辑回归	重新称重、偏见清除剂	平衡精度，差异影响，平均赔率、统计均等，平等机会，Theil 指数	90



(a) 原始代码行与模型代码行 **(b) 生成时间与代码行数** **(c) 原始代码与生成代码的执行时间对比**

图 6：FairML 在简洁性、生成时间和执行时间方面的评估。

(准确度、平衡准确度、平均差异、统计均等差异、差异影响、机会均等差异、Theil 指数、伽马差异、平均赔率差异、平均绝对误差、假阳性率、假阴性率、假发现率）。此外，FairML 还支持从外部 CSV 文件加载数据。这样，用户就可以通过数据集、分类器、去偏算法和偏差指标的不同组合来表达他们的偏差缓解策略。

在正确性方面，由于机器学习过程中的随机性，生成的代码中并非所有值都与示例中的相应值完全相等。但是
值仍在其 ± 0.1 的误差范围内。
各自的。

*4.1.2 简洁性。*图 6a 显示了原始示例、FairML 模型和生成代码之间的代码行数（LoC）比较。从图中可以看出，模型 LoC（实线）往往少于原始 LoC（虚线），这表明用户只需编写较少的代码就能产生与示例相同的结果（正确性见第 4.1.1 节）。此外，随着原始 LoC 数量的增加，效率也会越来越高。从示例 E11（最左侧）到 E12（最右侧），原始 LoC 从 30 个增加到 356 个，我们可以发现这样一个规律：用户只需编写各自原始示例中 LoC 的 83% 到 25%，效率也随之提高。

我们预计 FairML 产生的 LoC（虚线）会比原始示例多，因为它们包含了原始示例中没有的特征，例如 (1) 用于解释分类器、偏差缓解算法和偏差度量的代码，以及 (2) 以汇总表和图表形式显示测量结果的不同方法。

*4.1.3 生成和执行时间。*图 6b 显示了 FairML 生成代码所需的持续时间（生成时间）。在生成时间方面，FairML 可以在 0.2 秒内生成原始示例的每个生成代码版本，基于示例 E12，它可以达到 2,927 行/秒（644 行/0.22 秒）的速度。

图 6c 显示了生成代码的执行时间与原始示例的性能对比（以秒为单位）。一般来说，生成代码（实线）的执行时间略长于原始示例（虚线）。我们预计

正如第 4.1.2 节所述，生成代码版本的功能更多，代码行数更多。

4.2 对有效性的威胁

虽然我们已经根据 IBM AI Fairness 360 的文档和示例开发并测试了 FairML，但我们还没有进行用户评估，以获得该领域经验丰富的用户的反馈。这需要一些在偏差识别和缓解方面有经验的数据科学家的参与，而这是非常稀缺的资源[34]。尽管如此，我们转载的文档和示例都是由专家撰写的，其操作的数据也是来自不同领域的真实世界数据。

5 经验教训

我们从这项工作中汲取了一些经验教训：

- 针对不同类型的指标，有许多广为人知的算法来支持偏差识别和缓解工作

  和偏差。因此，它们可以为 FairML 等边界清晰的 DSL 提供坚实的基础。
- 识别和减少偏见的过程容易出错
  (例如，以错误的顺序应用算法）。表达它们

  在 DSL 中使用模型到文本的转换会对生成的代码施加一定的结构，而在手动编写代码时，这些结构不会自动执行。
- 直接从模型生成 Jupyter 笔记本
  模型到文本的转换或模型到 JSON 的转换

  的技术要求很高。不过，我们可以利用第三方引擎（如 P2J）来避免这种情况，它可以将文本 Python 程序转化为 Jupyter 笔记本。

6 相关工作

有些工具箱是为测量机器学习中的偏差而开发的。另一个由 [1] 开发的工具箱 FairML，通过利用四种输入排序算法和模型压缩，计算不同输入对预测结果的相对影响，从而审核预测模型的公平性。FairTest [46] 计算敏感属性与预测标签之间的关联，以检查数据集中的偏差。它还提供了一种方法来识别输入空间中算法可能产生异常高错误率的区域。Themis [22]，一个偏差工具箱、

它允许自动生成测试，以测量预测系统决策中的歧视。Fairness Measures [49] 支持对不同偏见指标的测量，如差异影响、平均几率比例和平均差异。Aequitas [41] 是一个审计工具包，可测量不同指标的公平性。它还提供了一个决策树，指导用户针对特定情况选择最合适的指标。

其他一些工具包也能减少偏差，不仅用于衡量公平性。它们是 ThemisML [6], Fairness Comparison [21], Aequitas [41], Google What-If [48], Scikit-fairnet/scikit-lego [42，43]、Fairlearn [9] 和 IBM AI Fairness 360 [7]。

Rapidminer [24]、Knime [8]和 Orange [17]等平台使用低代码、基于模型的方法，允许用户通过组装基于组件的程序对机器学习管道进行可视化编程。它们支持数据探索、转换、可视化以及机器学习和数据挖掘的不同算法。此外，它们都具有可扩展性，这意味着用户可以添加新的模块或脚本。不过，据我们所知，目前还没有用于测量和减轻偏差的内置模块。

与 FairML 最接近的现有作品是 Arbiter [52]，这是一种专为道德机器学习设计的特定领域语言。它是一种类似于 SQL 的声明式语言，用于定义如何训练机器学习模型，其中有四个部分用于描述其道德实践：透明度、公平性、责任性和可重复性。不
遗憾的是，其实施仅限于特定的指标和分类器[11]。

7 结论和今后的工作

本文介绍了 FairML 工具包，该工具包实现了一种基于模型的偏差缓解自动化方法。利用 FairML，用户可以用简洁的声明式方法生成偏差缓解代码，并据此生成可执行的 Python 代码。此外，就正确性而言，生成的代码产生的偏差度量值与原始示例中测量的值相似。

在未来的工作中，FairML 生成的代码行数仍可进一步减少，方法是将可重复的行合并到函数中，并删除首先通过静态分析确定的不必要代码。作为连带效应，删除不必要的代码可以优化生成代码的执行时间。

8 致谢

这项工作由约克-马斯特里赫特合作组织的 "负责任的设计数据科学 "计划（https://www.york.ac.uk/maastricht）。我们感谢马斯特里赫特团队的所有宝贵贡献。

参考文献

[1] Julius A Adebayo 等人，2016 年。*FairML: ToolBox for diagnosing bias in predictive modeling*.博士论文。麻省理工学院。

[2] Alekh Agarwal 、Alina Beygelzimer 、Miroslav Dudik 、John Langford 和 Hanna Wallach。2018.公平分类的还原方法》。*第 35 届机器学习国际会议论文集》（《机器学习研究论文集》，第 80 卷）*，Jennifer Dy 和 Andreas Krause （编辑）。PMLR, 60-69. https://proceedings.mlr.press/v80/agarwal18a.html

[3] Alekh Agarwal、Miroslav Dudik 和 Zhiwei Steven Wu.2019.公平回归：定量定义与基于还原的算法。在

[11] https://github.com/julian-zucker/arbiter

第 36 届机器学习国际会议（《机器学习研究论文集》第 97 卷），Kamalika Chaudhuri 和 Ruslan Salakhutdinov（编辑）。
PMLR，120-129。https://proceedings.mlr.press/v97/agarwal19d.html

[4] 人工智能公平 360 (AIF360) 作者。2022.AI Fairness 360 documentation. https://aif360.readthedocs.io/en/stable/ 访问日期：2022-01-30。

[5] Julia Angwin、Jeff Larson、Surya Mattu 和 Lauren Kirchner。2016.机器偏见：全国各地都在使用预测未来罪犯的软件。它对黑人有偏见。ProPublica (2016).https://www.propublica. org/article/machine-bias-risk-assessments-in-criminal-sentencing Accessed: 2022-01-18.

[6] 尼尔斯-班蒂兰2018.Themis-ml：用于端到端歧视发现与缓解的公平感知机器学习界面》。*人类服务技术期刊》* 36，1（2018），15-30。https://doi.org/10.1080/15228835.2017.1416512 arXiv:https://doi.org/10.1080/15228835.2017.1416512

[7] Rachel K. E. Bellamy、Kuntal Dey、Michael Hind、Samuel C. Hoffman、Stephanie Houde、Kalapriya Kannan、Pranay Lohia、Jacquelyn Martino、Sameep Mehta、Aleksandra Mojsilovic、Seema Nagar、Karthikeyan Natesan Ramamurthy、John Richards、Diptikalyan Saha、Prasanna Sattigeri、Moninder Singh、Kush R. Varshney 和张云峰。2018. AI Fairness 360：用于检测、理解和缓解不必要算法偏见的可扩展工具包。arXiv:1810.01943 [cs.AI] https://arxiv.org/abs/1810.01943

[8] Michael R. Berthold, Nicolas Cebron, Fabian Dill, Thomas R. Gabriel, Tobias Kötter、Thorsten Meinl、Peter Ohl、Christoph Sieb、Kilian Thiel 和 Bernd Wiswedel。2008.KNIME: The Konstanz Information Miner.*数据分析、机器学习与应用》*，Christine Preisach、Hans Burkhardt、Lars Schmidt-Thieme 和 Reinhold Decker（编辑）。Springer Berlin Heidelberg, Berlin, Heidelberg, 319- 326。

[9] Sarah Bird、Miro Dudík、Richard Edgar、Brandon Horn、Roman Lutz、Vanessa Milan、Mehrnoosh Sameki、Hanna Wallach 和 Kathleen Walker。2020.*Fairlearn：评估和改进人工智能公平性的工具包*。技术报告 MSR-TR-2020-32.Microsoft. https://www.microsoft.com/en-us/research/publication/fairlearn- a-toolkit-for-assessing-and-improving-fairness-in-ai/

[10] M.Brambilla, J. Cabot, and M. Wimmer.2017.*模型驱动软件工程实践》*。Morgan & Claypool Publishers. https://books.google.co.uk/books? id=dHUuswEACAAJ

[11] Joy Buolamwini 和 Timnit Gebru.2018.性别阴影：商业性别分类中的 跨部门准确性差异。*第一届公平、责任和透明度会议论文集》（《机器学习研究论文集》，第 81 卷）*，Sorelle A. Friedler 和 Christo Wilson（编辑）。PMLR, 77-91. https://proceedings.mlr.press/v81/buolamwini18a.html

[12] C.Byrne.2017. *数据科学家的开发工作流》*。O'Reilly Media. https://books.google.co.uk/books?id=84HgwQEACAAJ

[13] Flavio Calmon、Dennis Wei、Bhanukiran Vinzamuri、Karthikeyan Natesan Ramamurthy 和 Kush R Varshney。2017.防止歧视的优化预处理。*神经信息处理系统进展》*，I. Guyon、U. V. Luxburg、S. Bengio、H. Wallach、R. Fergus、S. Vishwanathan 和 R. Garnett（编辑），第 30 卷。https://proceedings.neurips.cc/paper/2017/file/9a49a25d845a483fae4be7e341368e36-Paper.pdf

[14] L.Elisa Celis、Lingxiao Huang、Vijay Keswani 和 Nisheeth K. Vishnoi。2019.带有公平性约束的分类：具有可证明保证的元算法。In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*.美国计算机协会，纽约、
NY, USA, 319-328. https://doi.org/10.1145/3287560.3287586

[15] 陈家豪、内森-卡卢斯、毛晓杰、杰弗里-斯瓦查和马德琳-乌德尔。2019.Unawareness：评估未观察到受保护阶 级 时的不平等。*公平、问责与透明会议论文集*（美国佐治亚州亚特兰大市）*（FAT* '19）*。Association for Computing Machin- ery, New York, NY, USA, 339-348. https://doi.org/10.1145/3287560.3287594

[16] Pedro Conceição 和 Pedro Ferreira.2000.1 Theil Index 的年轻人指南：Suggesting Intuitive Interpretations and Exploring Analytical Appli- cations.

[17] Janez Demšar、Tomaž Curk、Aleš Erjavec、Črt Gorup、Tomaž Hočevar、Mitar Milutinovič、Martin Možina、Matija Polajnar、Marko Toplak、Anže Starič、Miha Štajdohar、Lan Umek、Lan Žagar、Jure Žbontar、Marinka Žitnik 和 Blaž Zupan。2013.橙色：Python中的数据挖掘工具箱。*机器学习杂志 Research* 14 (2013)，2349-2353。http://jmlr.org/papers/v14/demsar13a.html

[18] Cynthia Dwork、Moritz Hardt、Toniann Pitassi、Omer Reingold 和 Richard Zemel。2012.通过认知实现公平。*第三届理论计算机科学创新会议论文集*（马萨诸塞州剑桥）*（ITCS '12）*。美国计算机协会，纽约州纽约市，214-226。https://doi.org/10.1145/2090236.2090255

[19] Clark Evans, O Ben-Kiki, and I döt Net.2017.YAML Ain't Markup Language (YAML™) Version 1.2. https://yaml.org/spec/1.2.2 Accessed: 2022-01-19.

[20] Michael Feldman、Sorelle A. Friedler、John Moeller、Carlos Scheidegger 和 Suresh Venkatasubramanian.2015.认证和消除差异影响。In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Sydney, NSW, Australia) *(KDD '15)*.协会

Computing Machinery, New York, NY, USA, 259-268. https://doi.org/10.1145/2783258.2783311

[21] Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth.2019.机器学习中增强公平性干预的比较研究》。*公平性、问责制和透明度会议论文集*（美国佐治亚州亚特兰大市）*（FAT* '19）*。Association for Computing Machinery, New York, NY, USA, 329-338. https://doi.org/10.1145/3287560.3287589

[22] Sainyam Galhotra、Yuriy Brun 和 Alexandra Meliou.2017.公平测试：测试软件是否存在歧视。In *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering* (Paderborn, Germany) *(ESEC/FSE 2017)*.美国计算机协会，纽约州纽约市，498-510。https：//doi.org/10.1145/3106237.3106277

[23] 莫里茨-哈特、埃里克-普莱斯、埃里克-普莱斯和纳蒂-斯雷布洛。2016.监督学习中的机会均等》。In *Advances in Neural Information Process- ing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (Eds.), Vol. 29.https://proceedings.neurips.cc/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf

[24] M.Hofmann 和 R. Klinkenberg。2016.*RapidMiner：数据挖掘用例和商业分析应用》*。CRC Press. https://books.google.co.id/books? id=Y_wYCwAAQBAJ

[25] IBM AI Research.2022. 欢迎访问 LALE 的 API 文档！ https://lale.readthedocs.io/en/latest/modules/lale.lib.aif360.util.html#lale.lib.aif360.util.theil_index Accessed: 2022-01-30.

[26] IBM Research Trusted AI.2022.选择衡量标准和缓解措施指南。https://aif360.mybluemix.net/resources#guidance 访问时间：2022-01-30。

[27] Faisal Kamiran and Toon Calders.2011.无差别分类的数据预处理技术。*Knowl.Inf.Knowl.*33, 1 (2011), 1-33. https：//doi.org/10.1007/s10115-011-0463-8

[28] Faisal Kamiran, Asim Karim, and Xiangliang Zhang.2012.识别感知分类的决策理论》。*2012 IEEE 第 12 届数据挖掘国际会议*。924–929. https://doi.org/10.1109/ICDM.2012.45

[29] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma.2012.带有消除偏见正则的公平感知分类器。*数据库中的机器学习和知识发现》*，Peter A. Flach、Tijl De Bie 和 Nello Cristianini（Eds.）Springer Berlin Heidelberg, Berlin, Heidelberg, 35-50.

[30] Michael Kearns、Seth Neel、Aaron Roth 和 Zhiwei Steven Wu.2018.Preventing Fairness Gerrymandering：子群公平性的审计与学习》。*第 35 届机器学习国际会议论文集》（《机器学习研究论文集》，第 80 卷）*，Jennifer Dy 和 Andreas Krause（编辑）。PMLR, 2564-2572. https://proceedings.mlr.press/v80/kearns18a.html

[31] 迈克尔-卡恩斯（Michael Kearns）、塞斯-尼尔（Seth Neel）、亚伦-罗斯（Aaron Roth）和吴志伟（Zhiwei Steven Wu）。2019.机器学习丰富子群公平性的实证研究。In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*.Association for Computing Machinery, New York, NY, USA, 100-109. https://doi.org/10.1145/3287560.3287592

[32] Dimitrios S. Kolovos、Nicholas Matragkas 和 Antonio García-Domínguez.2016.迈向灵活解析结构化文本模型表征。In *Pro- ceedings of the 2nd Workshop on Flexible Model Driven Engineering co-located with ACM/IEEE 19th International Conference on Model Driven Engineering Languages & Systems (MoDELS 2016), Saint-Malo, France, October 2, 2016 (CEUR Workshop Pro- ceedings, Vol. 1694)*, Davide Di Ruscio, Juan de Lara, and Alfonso Pierantonio (Eds.). CEUR-WS.org, 22-31. http://ceur-ws.org/Vol-1694/FlexMDE2016_paper_3.pdf

[33] Preethi Lahoti、Krishna P.Gummadi, and Gerhard Weikum.2019. iFair：为算法决策学习独立公平的数据表示。In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*.1334–1345. https://doi.org/10.1109/ICDE.2019.00121

[34] Michelle Seng Ah Lee 和 Jat Singh。2021.*开源公平性工具包的格局与差距》*。美国计算机协会，纽约州纽约市、USA. https://doi.org/10.1145/3411764.3445261

[35] T.T. Mahoney、K.R. Varshney、M. Hind 和 O'Reilly Media Company Sa- fari。2020.*人工智能的公平性：如何衡量和减少机器学习中不必要的偏见》*。https://books.google.co.id/books?id= uSbfzQEACAAJ

[36] Ninareh Mehrabi、Fred Morstatter、Nripsuta Saxena、Kristina Lerman 和 Aram Galstyan。2021.机器学习中的偏见与公平性调查》。*ACM Comput. Surv.*54, 6, Article 115 (jul 2021), 35 pages. https://doi.org/10.1145/3457607

[37] A.C. Müller and S. Guido.2016.*Python 机器学习入门：数据科学家指南》*。https://books.google.co.uk/books?id= vbQlDQAAQBAJ

[38] 牛津参考》。2022. 偏见。https://www.oxfordreference.com/view/10.1093/oi/authority.20110803095504939 访问日期：2022-01-16。

[39] Geoff Pleiss、Manish Raghavan、Felix Wu、Jon Kleinberg 和 Kilian Q Weinberger。2017. 论公平与校准》。*神经信息处理系统进展》*，I. Guyon、U. V. Luxburg、S. Bengio、H.Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30.https://proceedings.neurips.cc/paper/2017/file/b8b9c74ac526fffbeb2d39ab038d1cd7-Paper.pdf

[40] Louis M. Rose, Richard F. Paige, Dimitrios S. Kolovos, and Fiona A. C. Polack.2008.Epsilon 生成语言。In *Model Driven Architecture - Foundations and Applications*, Ina Schieferdecker and Alan Hartman (Eds.).Springer Berlin Heidelberg, Berlin, Heidelberg, 1-16.

[41] Pedro Saleiro、Benedict Kuester、Loren Hinkson、Jesse London、Abby Stevens、Ari Anisfeld、Kit T．Rodolfa, and Rayid Ghani.2019.Aequitas：A Bias and Fairness Audit Toolkit. arXiv:1811.05577 [cs.LG] https://arxiv.org/abs/1811.05577

[42] scikit-fairness.2022. scikit-fairness. https://scikit-fairness.netlify.app/ Accessed: 2022-01-30.

[43] scikit-lego.2022. scikit-lego. https://scikit-lego.readthedocs.io/en/latest/index.html 访问日期：2022-01-30。

[44] Till Speicher、Hoda Heidari、Nina Grgic-Hlaca、Krishna P. Gummadi、Adish Singla、Adrian Weller 和 Muhammad Bilal Zafar。Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar.2018.量化算法不公平的统一方法》：通过不平等指数衡量个人和群体的不公平。In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (London, United Kingdom) *(KDD '18)*.美国计算机协会，纽约州纽约市，2239-2248。https://doi.org/10.1145/3219819.3220046

[45] D.Steinberg, F. Budinsky, and E. Merks.2009.*EMF：Eclipse 建模框架*。Addison-Wesley. https://books.google.co.id/books?id=oAYcAAAACAAJ

[46] Florian Tramèr、Vaggelis Atlidakis、Roxana Geambasu、Daniel Hsu、Jean-Pierre Hubaux、Mathias Humbert、Ari Juels 和 Huang Lin。2017.FairTest：在数据驱动型应用中发现不必要的关联。In *2017 IEEE European Symposium on Security and Privacy (EuroS P)*.401-416. https://doi.org/10.1109/ EuroSP.2017.29

[47] M.Völter, T. Stahl, J. Bettin, A. Haase, S. Helsen, K. Czarnecki, and B. von Stockfleth.2013. *模型驱动软件开发》：技术、工程、管理agement*.https://books.google.co.uk/books?id=9ww_D9fAKncC

[48] James Wexler、Mahima Pushkarna、Tolga Bolukbasi、Martin Wattenberg、Fernanda Viégas 和 Jimbo Wilson。2020.What-If 工具：机器学习模型的交互式探测。*IEEE Visualization and Computer Graphics* 26, 1 (2020), 56-65. https://doi.org/10.1109/TVCG.2019.2934619

[49] "Meike Zehlike、Carlos Castillo、Francesco Bonchi、Ricardo Baeza-Yates、Sara Hajian 和 Mohamed Megahed"。2017.公平措施：A Platform for Data Collection and Benchmarking in discrimination-aware ML. https:// fairnessmeasures.github.io. https://fairnessmeasures.github.io

[50] Rich Zemel、Yu Wu、Kevin Swersky、Toni Pitassi 和 Cynthia Dwork。2013.学习公平表征。*第 30 届机器学习国际会议论文集》（《机器学习研究论文集》，第 28 卷）*，Sanjoy Dasgupta 和 David McAllester（编辑）。PMLR, Atlanta, Georgia, USA, 325-333. https://proceedings.mlr.press/v28/zemel13.html

[51] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell.2018.用对抗学习减轻不想要的偏见。In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (New Orleans, LA, USA) *(AIES '18)*.As- sociation for Computing Machinery, New York, NY, USA, 335-340. https://doi.org/10.1145/3278721.3278779

[52] Julian Zucker 和 Myraeka d'Leeuwen.2020.*Arbiter：道德机器学习的特定领域语言》*。美国计算机协会，纽约州纽约市，421-425。https://doi.org/10.1145/3375627.3375858。