arXiv:2303.01704v4 [cs.LG] 3 Jun 2024

数据偏差调查中的特征重要性差异

Peter W Chang [1] **Leor Fishman** [2] **Seth Neel** [1]

摘要

人们普遍认为，分类器中下游偏差的成因之一是训练数据中存在偏差。纠正此类偏差可能需要依赖具体情境的干预，例如在不同子群体上分别训练模型、移除在数据收集过程中带有偏差的特征，甚至开展现实世界实验以确定偏差来源。尽管开展此类数据偏差调查十分必要，但鲜有自动化方法能够协助从业者完成这些工作。本文提出了一种此类方法：给定由受保护特征和非受保护特征、结果变量 y以及一个回归器 h 构成的数据集 X，该回归器根据 X预测 y，此方法会输出一个元组（f_j，g），并满足以下性质：g 对应于训练数据集的一个子集（X，y），其中，特征 j^{th} f_j 在子群体 g中的影响力远大于（或小于）其在整个数据集中的影响力；我们将此称为特征重要性差异（FID）。我们在 4 个数据集和机器学习领域普遍关注的 4 种常见特征重要性方法上进行实验，结果表明，即使子群体类别规模呈指数级增长，我们也能高效地找到 FID 值较大的子群体；在实践中，根据标准公平性指标衡量，这些群体对应的子群体可能存在严重的偏差问题。

1. 引言

机器学习正迅速成为我们生活和决策中日益重要、却也愈发不透明的一部分，并被用于再犯风险分析（Angwin 等人，2016）、贷款审批及相关条款（Dastile 等人，2020）和儿童保护服务（Keddell, 2019）等影响日益重大的场景。大规模部署机器学习的一项期待是，这些算法或许能够摆脱人类的偏差与缺陷。遗憾的是，这种期待过于天真。过去十年间，跨学科研究表明，机器学习算法可能存在根深蒂固的偏差，既有隐蔽的，也有直接的（Barocas 等人，2019）；研究人员也一直致力于开发各种技术，以构建更公平的模型（Caton & Haas，2023）。

模型偏差的主要成因之一是训练数据中固有的偏差，而非训练过程本身存在明确的偏见。尽管公平性领域的大多数研究都致力于通过学习数据的更公平表示（Zemel 等人，2013年），或明确约束下游分类器，使其符合特定的公平性概念（Hardt 等人，2016a；Agarwal 等人，2018b），来消除偏差，但研究表明，公平性概念往往很脆弱，且常常彼此矛盾（Dwork 与 Ilvento, 2018年；Kleinberg 等人，2016）。更重要的是，这些方法回避了一个对实践者而言可能更重要的问题：训练数据中的偏差源自何处？数据中哪些子群体受到了影响？

我们将回答这一问题的过程称为数据偏差调查（DBI）。本文提出一种技术，通过帮助分析师（高效且可证明地）识别训练数据中的结构化子集，来辅助开展 DBI，使其能够聚焦于这些子集进行偏差调查。以往的研究通常通过寻找训练数据中最严重违反特定公平性标准的子集来识别此类子集；这通常意味着分类器在该群体上的错误率高于总体错误率。我们采取了截然不同的方法：我们不针对特定的公平性概念进行优化，而是寻找这样的子群体：相较于总体，数据中的某个特定特征在该子群体中的影响显著更大。

为了直观理解为什么特征重要性差异，特征重要性差异 正如我们对其的称呼，在寻找数据集偏差时可能是一个有用的概念，请考虑以下由被广泛视为算法偏差领域先驱的 Cynthia Dwork 在 2015年 接受《纽约时报》采访时给出的例子（Miller，2015）：假设我们有一个少数群体，其中聪明的学生会被引导去学习数学；再假设在多数群体中，聪明的学生则会被引导去学习金融。寻找优秀学生的一个简单方法是寻找正在学习金融的学生；如果少数群体规模较小，这种简单的分类方案可能会找到大多数聪明学生。但这不仅是不公平的，

[1]SAFR AI 实验室、哈佛商学院与 KempnerInstitute，波士顿，马萨诸塞州 [2]哈佛学院，剑桥，马萨诸塞州。通信请联系：Seth Neel < sethneel.ai@gmail.com>。

论文集：41st 国际机器学习会议，奥地利维也纳。PMLR 235，2024年。版权所有 © 2024，作者所有。

对于少数群体中表现优异的学生而言，它的效用也很低。

进一步剖析这个例子可以发现，特征“是否主修金融”有助于在总体中识别优秀学生，但对少数群体却没有预测作用。与此同时，特征“是否主修数学”对少数群体有很强的预测作用，但对多数群体完全没有预测作用。只选择金融专业学生的分类器对少数群体并不公平，正是因为这些特征在两个群体中的重要性不同。确定子群体后，分析师采取的后续行动完全取决于具体情境。这些行动可以包括“简单”的修正措施，例如针对发现的子群体单独训练一个模型，或在训练中排除某个特征；也可以采取更复杂的补救措施，例如调查该群体中特定特征或结果变量的收集方式，以及收集过程中是否存在偏差。

决策树或线性回归等简单模型自带直观的特征重要性度量，但总体而言，复杂模型并没有公认的特征重要性度量。备受关注的方法包括 LIME（Ribeiro 等人，2016）等局部模型无关方法、SHAP（Lundberg 和 Lee，2017）、特定于模型的显著性图（Simonyan 等人，2013），以及基于示例的反事实解释（Molnar，2022）。人们已对最常用的特征重要性度量（包括我们研究的那些度量）的稳定性和稳健性提出质疑（Dai 等，2022；Agarwal 等，2022a；Alvarez-Melis 与 Jaakkola，2018；Bansal 等，2020；Dimanov 等人，2020；Slack 等人，2020）。此外，这些度量往往彼此矛盾，因此没有任何一种可以被视为权威标准（Krishna 等，2022）。尽管存在这些局限，如今这些度量仍在实践中得到广泛应用，并且作为诊断工具仍然有用，正如我们接下来提出的那样。

固定任意一种特征重要性概念后，给定一小组受保护子群体，只需遍历这些子群体和特征，计算每个特征相对于每个子群体的特征重要性差异（FID），然后选出差异最大的特征和子群体即可，这一过程并不复杂。然而，公平性文献也指出，比较少数几个敏感子群体在某项公平性度量上的表现时，分类器看起来可能是公平的；但如果将敏感子群体的概念推广到涵盖敏感特征之间的组合与交互（称为丰富子群体（ Kearns 等，2019）），就可能出现很大的差异。我们在第 5.4 节验证，仅考虑边际子群体时也会出现同样的现象：丰富子群体能够揭示大得多的特征重要性差异。即使对丰富子群体采用简单的定义，例如二元特征的合取，子群体的数量也会随敏感属性的数量呈指数增长，因此无法逐一计算每个敏感子群体的度量并通过穷举找出最大值。这引出了一个显而易见的问题。鉴于前文的讨论，尽管据我们所知，这个问题尚未得到充分研究：当这些特征重要性概念应用于存在偏差问题的分类器和数据集时，它们能否揭示丰富子群体之间显著的特征重要性差异？又能否高效地检测出这些差异？

1.1. 案例研究：COMPAS 数据偏差调查

在深入探讨我们方法的技术细节之前，我们先说明一下，我们的技术在训练随机森林分类器 h 以预测 COMPAS 数据集中的两年内再犯情况时，能够带来哪些有趣的潜在干预措施（Angwin 等人，2016）。利用特征重要性方法 SHAP（Lundberg 和 Lee，2017）和我们的算法 1，我们发现，在一个主要由美洲原住民和非裔美国人男性构成的子群体中，预测两年内再犯情况时，前科次数这一特征的重要性显著更高；该子群体占训练集的 $\sim$9%（图 1）。尽管我们会在附录 A 中进一步讨论，特征重要性方面的差异并不能保证子群体存在公平性差异，但我们在第 5.5 小节中展示了，从经验上看，情况往往如此。在本例中，我们发现，与总体人群相比，该子群体的传统公平性指标略差，准确度低 $\sim$1%，假阳性率（FPR）高 $\sim$1%。一种提高该子群体准确度并可能缩小差异的简单方法，是h_g 为该子群体训练一个单独的模型。我们发现，这样做会使该子群体的假阳性率降低 7.5%，并使准确度提高 7%。此外，使用 h_g计算的前科次数特征平均 SHAP 值，与使用 h 计算的总体人群平均 SHAP 值之间的差异减少了一半以上。另一种直观的技术解决方案是训练一个不包含前科次数特征的新模型h_{-f}。采用这一方案后，公平性差异有所缩小，但模型性能也明显下降，这可能是因为移除了前科次数特征的预测能力。从定性角度看，识别出这一子群体也可能促使人们进一步研究 COMPAS 数据集中的美洲原住民和非裔美国男性是否受到不同的执法对待，这可能导致前科次数特征或两年内再犯结果变量存在测量偏差。

1.2. 结果

虽然前面的示例清楚地展示了我们的方法在宏观层面的实用性，但细节决定成败。本文其余部分将对特征重要性和受保护的丰富子群体进行形式化定义，并介绍我们用于高效检测潜在偏差子群体的方法。

我们最重要的贡献是引入了这一概念

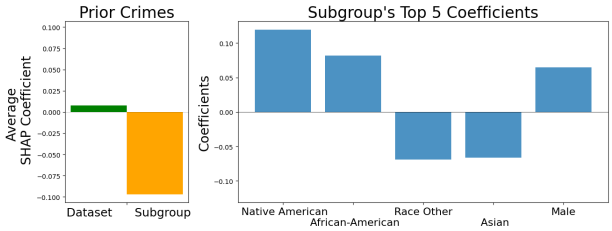


图 1：探索 COMPAS 中具有高特征重要性差异的子群体/特征对。第一张图比较了子群体中 priors-count 的平均 SHAP 特征重要性与整个数据集中的平均 SHAP 特征重要性。第二张图展示了定义该子群体的敏感属性线性函数中绝对值最大的 5 个系数。

：在近期提出的特征重要性度量的背景下，针对丰富子群体（定义 3.1）的特征重要性差异 。我们根据特征重要性度量能否表示为子群体中各点上的求和（定义 3.2），将其归类为可分离的 或不可分离的，并定义了特征重要性差异（FID）的一种变体，即平均特征重要性差异（AVG-FID，定义 3.3）。我们的主要理论贡献是第 4 节中的定理 4.1。该定理大致指出，尽管在最坏情况下，寻找 FID 最大的子群体这一问题是 NP 困难的（附录 G），但只要能够访问关于丰富子群体类 $\mathcal{G}$（定义 F.1）的代价敏感分类 预言机，算法 1 就能针对任意可分离的特征重要性度量，高效地学习 FID 最大的子群体。

在第 5 节中，我们进行了全面的实证评估，使用 LIME、SHAP、显著性图和线性回归系数作为特征重要性概念，对 Student（Cortez & Silva，2008）、COMPAS（Angwin 等人，2016）、Bank（Moro 等人，2014）和 Folktables （ Ding 等人，2021 ）数据集中的高特征重要性差异子群体进行审计。我们的实验得出以下结论：(i) 对于所有（数据集、重要性概念）对，我们都能找到由敏感特征函数定义的子群体，使其相对于给定特征具有较大的特征重要性差异（表 1，图 2、图 3）。(ii) 检查这些子群体的系数，可以深入讨论潜在的数据集偏差（图 1，第 5.3 节）。(iii) 在约一半的情况下，与仅搜索由单一敏感属性定义的子群体相比，丰富子群体具有更高的样本外特征重要性差异，从而证明使用丰富子群体是合理的（第 5.4 节）。(iv) 这些子群体在已被广泛接受的公平性指标（如人口统计均等和校准）上存在差异（表 2）。反之，最大程度违反公平性指标的丰富子群体也表现出较大的特征重要性差异（表 3）。

这些结果在样本外也具有普适性，无论是发现的特征重要性差异（FID）值，还是相应子群体的规模（附录 L）。综合来看，这些理论和实证结果凸显了我们的方法对含有敏感特征的表格数据集中的 DBI 诊断工具包而言，是一项重要补充。

2. 相关工作

已有大量研究探讨机器学习模型及其训练数据中的偏差（Barocas 等人，2019；Caton & Haas，2023）。在较高层面上，我们受到数据集偏差相关研究的启发（Kamiran & Calders，2012；Tommasi 等人，2017；Li & Vasconcelos，2019）。然而，据我们所知，这是首项将丰富子群体中的特征重要性值差异作为公平性诊断进行研究的工作。更多相关工作请参见附录 C。

异常子群发现。在研究方法方面，有两项密切相关的工作（Dai 等，2022；Bal- agopalan 等，2022）将敏感子群体中的公平性问题与模型解释质量联系起来，并通过稳定性和忠实度等属性进行衡量。我们的工作有所不同：我们关注的是解释差异本身的大小，而非其“质量”，并将研究结果扩展到丰富子群设定。我们提出的算法能够在指数规模的子群空间中进行搜索，是该领域一项新颖且必要的补充。另一个研究方向致力于证明所选评分函数满足线性时间子集扫描性质（Neill，2012），随后便可利用该性质在线性时间内搜索子群空间中的分类器偏差（Zhang 和 Neill，2016；Boxer 等，2023）。虽然很难绝对断言这种方法毫无用处，但我们尚不清楚如何让子集扫描方法在丰富子群体上进行优化。

丰富子群体与多重校准。从技术层面来看，与本文最为相关的论文是（Kearns 等人，2018；Hebert-Johnson 等，2018），这些论文在研究针对均等化赔率或校准的分类器学习时，引入了基于敏感特征的丰富子群类 $\mathcal{G}$ 这一概念。我们的算法 1 属于“预言机高效”算法范式，用于求解由 Agarwal 等人（2018a）提出、并由 Kearns 等人（2018；2019；Hebert-Johnson 等，2018）在丰富子群体背景下发展的约束优化问题。近期，人们对学习多重校准预测器表现出浓厚兴趣，因为这与不确定性估计和全能预测器相关（Hu 等，2023；Gopalan 等，2022；Jung 等，2021）。这些研究均未考虑特征重要性差异。

特征重要性概念。关于可解释或可解释机器学习领域，请参阅（Molnar，2022）的综述。最相关的研究涵盖了用于考察给定数据集中某个特征重要性的方法

子集。局部解释方法会为每个点赋予一个特征重要性（x，y），并通过对该子群体内各点的特征重要性求和或取平均，来定义子群体中的重要性度量，正如我们在定义 3.2、3.3 中所做的那样。

3. 预备知识

令 X^n 表示我们的数据集，其中包含 n 个体。这些个体由元组 $((x, x')\ y)$ 定义，其中 $x \in \mathcal{X}_{sense}$ 是受保护特征向量，$x' \in \mathcal{X}_{safe}$ 是非受保护特征向量，$y \in \mathcal{Y}$ 表示标签。令 $X = (x, x') \in \mathcal{X} = \mathcal{X}_{sense} \times \mathcal{X}_{safe} \subset \mathbb{R}^d$ 表示一个联合特征，数据点 (X, y) 独立同分布地取自分布 $\mathcal{R}$。令 $h: \mathcal{X} \to \mathcal{Y}$ 表示一个分类器或回归器，用于根据 X 预测 y。我们将一个丰富的子群体类 $\mathcal{G} = \{g_\alpha\}_{\alpha \in \Omega}$ 定义为函数 $g: \mathcal{X}_{sens} \to [0, 1]$ 的集合，其中 $g(x')$ 表示点 $X = (x, x')$ 属于群体 g 的成员资格。请注意，这与 (Kearns 等人，2018) 中的子群体定义相同，但不再要求 $g(x') \in \{0, 1\}$，从而支持不同程度的群体成员资格。例如，一名混血人士可能以 0.5 的程度属于两个种族群体。令 $f_j, j \in [d]$ 表示 $\mathcal{X} \subset \mathbb{R}^d$ 中的 j^{th} 特征。对于分类器 h 和子群体 $g \in \mathcal{G}$，令 F 为一种特征重要性度量，其中 $F(f_j, g(X^n)\ h)$ 表示 h 赋予子群体 $g(X^n)$ 中特征 j 的重要性，$F(f_j, X^n, h)$ 表示 h 赋予整个数据集中的 f_j 的重要性。我们稍后将给出 F 的更具体实例，但下面将以尽可能一般的形式给出特征重要性差异的定义。

定义 3.1。（特征重要性差异）。给定分类器 h、由 $g \in G$ 定义的 X^n 的一个子群体，以及一个特征 $f_j \in [d]$，再给定特征重要性度量$F(\cdot)$，则相对于 g 的特征重要性差异定义为：

$$\texttt{FID}(f_j, g, h) = \mathbb{E}_{X \sim \mathcal{R}} |F(f_j, g(X^n), h) - F(f_j, X^n, h)|$$

除非需要明确所使用的分类器，否则我们将省略 h，并写作 FID$(j,\ g)$。现在，给定h 和 X^n，我们的目标是找到使 FID$(j,\ g)$ 最大的特征—子群对（$j^*,\ g^*$）$\in [d] \times \mathcal{G}$，即（$j^*,\ g^*$）=最大值点$_{g \in \mathcal{G},\ j \in [d]}FID(j,\ g)$。

现在，我们对特征重要性度量 $F(\cdot)$ 进行更具体的说明。首先，我们定义可分离的 特征重要性度量类：

定义 3.2。（局部分离）。如果特征重要性度量 $F(\cdot)$ 可以分解为局部模型解释值 F' 的逐点和，则称其为局部分离的：

$$F(f_j, X^n, h) = \sum_{X \in X^n} F'(f_j, X, h)$$

由此可知，对于可分离概念，$F(f_j,\ g(X^n)\ h) = \sum_{X \in X^n} g(X) F'(f_j,\ X,\ h)$。给定局部模型解释 F'，我们可以定义一种更具体的特征重要性差异形式，即平均特征重要性差异（AVG-FID），用于比较子群体内的平均特征重要性与整个数据集上的平均重要性。

定义 3.3。（平均情形下的局部分离特征重要性差异）。对于一个 $g \in \mathcal{G}$，令 $|g| = \sum g(X)$。给定一个局部模型解释-$_{X \in X^n}$说明 $F'(\cdot)$，我们定义相应的：

$$\begin{aligned}\texttt{AVG-FID}(f_j, g, h) =& \mathbb{E}_{X^n \sim \mathcal{R}^n} |\frac{1}{|g|} \sum_{X \in X^n} g(X) F'(f_j, X, h) \\ &- \frac{1}{n} \sum_{X \in X^n} F'(f_j, X, h)|\end{aligned}$$

请注意，AVG-FID 不等价于可分离的特征重要性差异，因为我们除以 $|g|$，这会影响求和中的每一项。在第 4 节中，我们将展示如何通过优化带有规模约束的特征重要性差异问题的一个版本来优化 AVG-FID，而借助算法 1 可以高效地完成这一点。

这一可分离性概念对于理解本文其余内容至关重要。在第 4 节中，我们将展示，对于任意可分离的特征重要性差异，算法 1 能够在多项式时间内，以（预言机）高效的方式计算指定规模的最大特征重要性差异子群体。关于“预言机高效”，我们沿用（Agarwal 等人，2018a；Kearns 等人，2018）的定义，指的是能够访问一个可解决（可能是 NP 困难）问题的优化预言机。尽管这听起来是一个很强的假设，但在实践中，我们可以利用能够解决困难非凸优化问题的现代优化算法（例如训练神经网络）。这一框架促成了许多具有坚实理论基础的实用算法的发展（Agarwal 等人，2018a；Kearns 等人，2018；2019；Hebert-Johnson 等，2018）；正如第 5 节所示，它在本文的实际应用中同样表现良好。我们所需的预言机类型称为代价敏感分类（CSC）预言机，其定义见附录 F。

4. 针对 AVG-FID 进行优化

在本节中，我们将展示如何（通过预言机）高效计算使 AVG-FID 最大化的丰富子群体。我们不直接优化 AVG-FID，而是让算法 1 求解一个在群体大小约束下使特征重要性差异最大化的优化问题：

$$\max_{g\in\mathcal{G}} |F(f_j, g(X^n), h) - F(f_j, X^n, h)|$$
$$\text{s.t.} \quad \Phi_L(g) \equiv \alpha_L - \frac{1}{n}\sum_{X\in X^n} g(X) \le 0, \tag{1}$$
$$\Phi_U(g) \equiv \frac{1}{n}\sum_{X\in X^n} g(X) - \alpha_U \le 0$$

其中， Φ_L 和 Φ_U 是给定子群函数 g 的“规模违反”函数。我们用 $g^*_{[\alpha_L,\alpha_U]}$ 表示方程 1 的最优解。我们专注于优化受约束的特征重要性差异，因为以下原语还允许我们高效地优化 AVG-FID：

1. Discretize $[0,1]$ into intervals $(\frac{i-1}{n}, \frac{i}{n}]_{i=1}^n$. Given feature f_j, compute $g^*_{(\frac{i-1}{n},\frac{i}{n}]}$ for $i = 1...n$.

2. 输出 g_{k^*}，其中 $k^* =$最大值点$_k \frac{k}{n}|F(f_j,\ g_k,\ h)|$ 在区间数量 n足够大时，可近似最大化 AVG-FID。

该证明见附录 E。现在我们陈述主要定理，它表明，只需对 CSC 进行多项式次调用$\mathcal{G}$，我们就能求解方程 1 中的受约束的特征重要性差异问题。

定理 4.1。 设 F 为可分离的特征重要性差异概念，固定一个分类器 h、子群类 $\mathcal{G}$和预言机 *CSC*$_\mathcal{G}$。然后，选定精度常数 ν 和界常数 B，并固定一个感兴趣特征 f_j，我们将运行算法 *1* 两次：一次使用由 F给出的特征重要性差异，另一次使用由 $-F$给出的特征重要性差异。令 $\hat{p}^T_\mathcal{G}$ 为算法 *1* 经过 $T = O(\frac{4n^2B^2}{\nu^2})$次迭代后返回的分布，且该分布对应的$\mathbb{E}$[特征重要性差异$(j,g)$]值较大。那么：

$$FID(j, g^*_j) - \mathbb{E}_{g\sim\hat{p}^T_\mathcal{G}}[FID(j,g)] \le \nu$$
$$|\Phi_L(g)|, |\Phi_U(g)| \le \frac{1+2\nu}{B} \tag{2}$$

我们将定理 4.1 的证明推迟至附录 D。简而言之，我们不再针对 $g \in \mathcal{G}$进行优化，而是针对分布 $\Delta(\mathcal{G})$进行优化。这样，我们便可以将方程 1 中的优化问题转化为线性规划，从而构造拉格朗日函数 L。该函数由特征重要性值与规模约束函数构成，后者由对偶变量 λ 加权，然后我们可以应用强对偶性。接下来，我们可以将约束优化转化为求解双人零和博弈的纳什均衡，并应用经典结果 (Freund & Schapire, 1996)：如果两位玩家都采用无遗憾 策略，那么随着两位玩家的平均遗憾趋于零，我们将以相应速率收敛至纳什均衡。算法 1 为最大化玩家实现了指数梯度下降算法 (Kivinen & Warmuth, 1997)，该玩家针对 λ进行优化；对于最小化玩家，则通过 CSC 求解来实现最佳响应。最小化玩家旨在最大化子群差异，以优化丰富子群分布。

我们注意到，我们的算法并非在规模约束下计算使 FID(j,g) 最大化的群组 g，而是输出群组上的一个分布 $\hat{p}^T_\mathcal{G}$，该分布在这些群组上平均而言 满足这一过程。理论上，这似乎不利于可解释性。然而在实践中，我们只需取每轮找到的群组 g_t，并输出其中大小处于适当范围且 FID 值最大的群组。第 5 节的结果验证了，这种启发式选择能够找到既符合约束又具有较大 FID 值的群组。该方法还能够泛化到样本外数据，表明 FID 并未因多重检验而被人为抬高（附录 L）。此外，我们的方法提供了一组候选群组 $(g_t)_{t=1}^T$，可快速评估其 FID 是否较大；这有助于发现最大子群体中不存在的有趣偏差。

5. 实验

这里报告了我们针对 16 种不同数据集/FID 概念配对开展的实证研究结果。这些结果证实，我们的方法能够找到较大的 AVG-FID 值，这些值对应于由受保护属性的简单函数定义的丰富子群体（表 1、图 2、图 3），并且大于仅通过优化边际子群体所找到的值（第 5.4 节）。此外，我们在第 5.5 节发现，高 AVG-FID 子群体往往在传统公平性指标上存在显著差异（表 2）；而最大化 FPR 差异等公平性概念的丰富子群体也表现出较高的 AVG-FID 特征，尽管其数值小于算法 1 找到的结果（表 3）。或许最重要的是，对于旨在协助 DBI 过程的工具而言，所找到的高 AVG-FID 子群体对应于一些暗示潜在数据集偏差的子群体和特征；我们在最初的案例研究中介绍了其中一个例子，并在第 5.3 节进一步给出了更多例子。在附录 J 中，我们构造了两个合成数据集：一个数据集中的特征重要性在所有子群体间都相同，另一个数据集则在特定的丰富子群体中人为引入了差异，以验证我们的方法 i) 能够避免假发现，且 ii) 能够有效识别出高 FID 子群体。我们还给出了相关结果，表明所找到的丰富子群体在样本外就 AVG-FID 和群体规模而言都具有泛化能力 $|g|$（附录 L），对假设类 h 的选择相对稳健（附录 M），并且我们的算法在实践中能够快速收敛（附录 O）。我们实验所使用的代码可在 github.com/safr-ai-lab/xai-disparity 获取。

5.1. 实验细节

数据集：我们使用了四个广受欢迎的数据集进行实验：Student、COMPAS、Bank 和 Folktables。对于每项测试，我们使用了 COMPAS 两次：一次用于预测两年再犯，另一次用于预测十分位风险评分（分别标记为 COMPAS R 和 COMPAS D）。对于每个数据集，我们指定了“敏感”特征，这些特征通常受到平等保护或隐私法的涵盖（例如种族、性别、年龄、健康数据）。附录 I 包含更多详细信息。

计算AVG-FID：我们研究了 3 种基于局部模型解释的可分离特征重要性度量：局部可解释模型无关解释（LIME）（Ribeiro 等人，2016）、Shapley 加性解释（SHAP）（Lundberg 和 Lee，2017）以及显著性图（GRAD）（Simonyan 等人，2013）。对于每种方法和每个数据集，我们都在 α 范围（α_L，α_U）$= \{[.01, .05], [.05, .1], [.1, .15], [.15, .2], [.2, .25]\}$内优化受约束的特征重要性差异。这些较小的范围使我们能够合理地比较表 1 中报告的特征重要性差异值。此外，这些范围涵盖了公平性研究和数据集审计工作中特别值得关注的子群体规模。结果中报告的所有 AVG-FID 值均为样本外值；也就是说，AVG-FID 值是在未用于优化子群体的测试集上计算的。数据集采用 $80-20$训练-测试划分，但 Student 除外；由于其规模较小，该数据集采用 $50-50$ 划分。在所有数据集中，当 FID 使用 LIME 或 SHAP 时，我们将 h 设为随机森林；使用 GRAD 时，则采用逻辑回归，因为它要求分类器的输出对输入可微。正如我们在附录 M 中所述，分类器的具体选择对结果没有明显影响。由于计算资源限制，GRAD 仅在 COMPAS R 数据集上进行了测试。有关我们如何实现重要性概念和算法 1 的细节，请参阅附录 I。

线性特征重要性差异：除了 3 种可分离的特征重要性差异概念外，我们还研究了一种不可分解的重要性概念方法。线性回归（在结果中标记为 LR）是一种流行且本质上具有可解释性的模型；加权最小二乘法（WLS）解的系数代表每个特征的重要性。因此，我们可以将另一种特征重要性差异定义为线性特征重要性差异（LIN-FID），即特征 f_j 在子群体 g 上的 WLS 系数与其在数据集 X^n 上的 WLS 系数之差。由于 LIN-FID 关于 g 可微，我们能够使用非凸优化器找到具有较高 LIN-FID 的局部最优 g；我们使用了 Adam。详情和证明请参见附录 H。

5.2. 实验结果

表 1 总结了实验结果，这些结果以对数比率尺度可视化于图 2中，以便更好地比较不同的重要性度量。在每个数据集和重要性度量下，我们的方法都能够找到具有较高特征重要性差异的子群体，其差异通常相差数个数量级。例如，在使用 LIME 作为重要性度量的 Folktables 数据集中，存在一个子群体，在该子群体中，年龄的平均重要性是其在总体人群中的 225 倍。表 1 还提供了定义特征，即在 g 中具有最大系数的敏感特征。

这项实验引出了一个自然的后续问题：给定一个数据集，特征重要性差异的分布是什么样的？图3展示了银行数据集中 FID 值最高的 10 个特征的分布。我们可以看到，有少数特征能够找到具有较大 FID 的子群体，但这样的情况会显著减少。这种模式在所有数据集和特征重要性概念中均有体现。对于实际应用而言，这是一个积极的结果，因为分析师或领域专家可以专注于少数几个在进行 DBI 时表现差异显著的特征。

在前文第 1.1 节中，我们考察了一个具体案例：我们的方法所发现的 FID 揭示了一个存在偏差的子群体，而其中的公平性差异随后可以通过针对性方法加以缓解。在下一节中，我们针对每个数据集和每种特征重要性度量，都发现了类似的例子，揭示了某种形式的潜在偏差。需要注意的是，并非发现的每个（子群体，特征）配对都必然意味着存在公平性问题。例如，AVG-FID 的部分结果可能由 f_j 与定义 g 的敏感属性之间的相关性所驱动。公平性研究始终面临这样一种情况：在统计上等价的两种情境，在现实世界中可能具有截然不同的公平性影响。因此，我们的方法应被视为协助从业者开展 DBI 的工具，而非偏差的确凿证据。

5.3. 高FID子群体讨论

在图4中，我们为每个数据集选取并展示了一个有趣的（特征、子群体、方法）组合。图4a显示，在Student数据集上，缺勤这一特征在整个数据集中的重要性接近于零，但在一个子群体中，它与学生表现呈强烈负相关；该子群体排名前两位的特征表明，学生的父母是否在一起生活，以及他们是否居住在农村社区。图4b显示，在 COMPAS 数据集上使用 GRAD 方法时，在预测两年内再犯时，被逮捕但未被起诉这一特征通常非常重要。然而，对于一个主要由美洲原住民男性构成的子群体，该特征的重要性显著降低。在 COMPAS 上预测十分位风险评分时，LIME 指出年龄在整个数据集中并不重要；然而，对于非美洲原住民女性少数族裔而言，年龄较大可以用于解释较低的十分位数评分。在 Bank 数据集上使用 LIN-FID 时，我们发现，一个线性回归

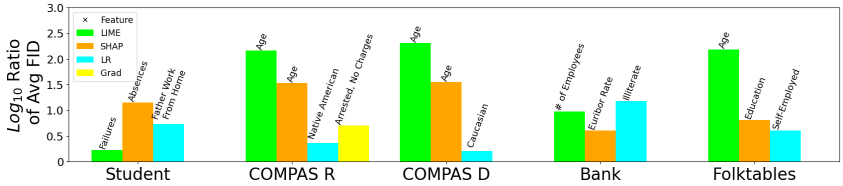


图2：针对每个（数据集、方法）组合，所发现的最高特征重要性差异的摘要。其显示为 $\left|log_{10}(R)\right|$，其中 R 表示在 g^* 中每个数据点的平均重要性与 X 上平均重要性的比值（适用于可分离概念），或者表示 LIN-FID 的系数比值。该尺度支持对不同重要性概念进行比较。与每个 g^* 相关联的特征标注在条形图上方。

表 1：每个实验中 AVG-FID 最高的子群体摘要，以及相应的特征、子群规模和定义特征。实验在多个（α_L、α_U）范围内进行，并显示所发现的最高 AVG-FID。 $\mu(F)$ 是指定群体上的平均特征重要性值。

数据集	概念	特征 f_j	$\mu(F(f_j,\ X))$	$\mu(F(f_j,\ g))$	$\lvert g\rvert$	定义特征
学生	LIME	不及格	$-.006$	$-.011$	.01	酒精使用 ，城市家庭
	SHAP	缺勤	$-.15$	-2.1	.02	父母状况 ，城市家庭
	LR	父亲居家办公	21.7	-4.0	.03	酒精使用 ，健康
COMPAS R	LIME	Age	.0009	$-.14$	.05	美洲原住民
	SHAP	Age	.012	.41	.04	亚裔美国人
	LR	美洲原住民	.5	1.17	.04	亚裔/西裔美国人
	GRAD	逮捕，无指控	.09	.02	.05	美洲原住民
COMPAS D	LIME	Age	$-.0003$	$-.06$	.02	原住民/非裔美国人
	SHAP	Age	.06	2.35	.07	非裔/亚裔美国人
	LR	高加索人	6.7	10.7	.04	美洲原住民
Bank	LIME	员工人数	$-.003$	.03	.03	婚姻状况
	SHAP	欧元银行同业拆借利率	$-.004$	.016	.03	婚姻状况
	LR	不识字	$-.07$	$-.0045$	.01	年龄、婚姻状况
Folktables	LIME	Age	$-.0007$	$-.11$	.21	婚姻状况
	SHAP	教育	.023	.15	.03	亚裔美国人
	LR	自雇	$-.26$	$-.06$	.02	白人美国人

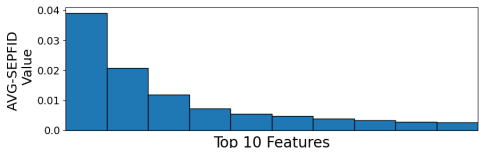
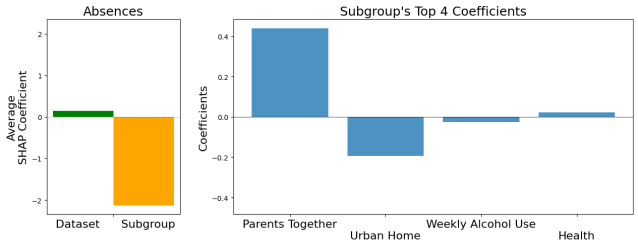


图3：使用 LIME 在 BANK 数据集上计算的排名靠前特征的 AVG-FID 分布。我们看到 AVG-FID 急剧下降。所有数据集和概念中都能观察到这一模式。

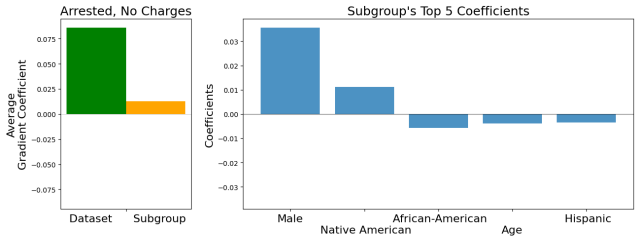
模型在由年龄较大、单身个体构成的子群体样本上训练后，在预测开设账户的可能性时，会更加重视 job=housemaid。最后，在 Folktables 上，我们发现，在预测收入时，与整个数据集相比，LIN-FID 对一个主要由白人和离异人士构成的子群体中的 job=military 特征赋予的权重低得多。这些有趣的示例结合表 1 中报告的结果，凸显了我们的方法在发现子群体方面的实用性；相关分析师或领域专家可以深入研究这些子群体，以确定偏差可能如何体现在数据中，以及如何加以纠正它们。

5.4. 富子群体与边际子群体的 FID 值比较

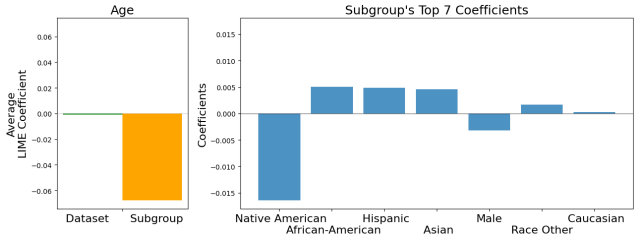
为了更好地量化丰富子群体的优势，我们进行了相同的分析，但仅在边际子群体空间中进行搜索。对于每一对数据集和重要性度量，我们建立了一个有限列表，其中包含由单个敏感特征定义的子群体，并计算了这些子群体各自的特征重要性差异。在图 5 中，我们将图 2 所示的一些最大 AVG-FID 丰富子群体与具有相同特征的最大 AVG-FID 边际子群体进行了比较。在大约一半的情况下，边际子群体的 AVG-FID 与丰富子群体相近。在其他情况下，将子群体类别扩展到包含由敏感属性的线性函数定义的丰富子群体，使我们能够找到 AVG-FID 更高的子群体。例如，在图 5b 中可以看到，在 COMPAS R 数据集上使用 GRAD 作为重要性度量时，「被捕，未被起诉」这一类别的丰富子群体的 AVG-FID 比完整数据集上的低 4 倍。然而，我们未能在边际空间中找到任何一个子



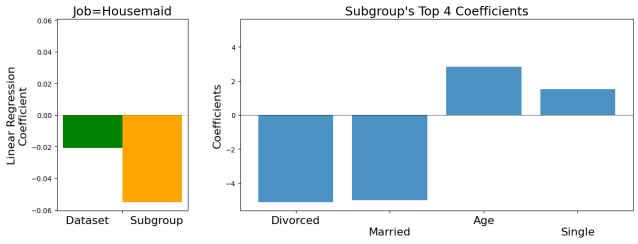
(a) 学生：预测成绩结果



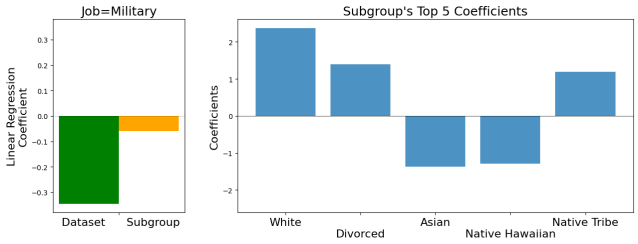
(b) COMPAS：预测两年内再犯



(c) COMPAS：预测十分位风险评分



(d) 银行：预测是否进行了银行存款



(e) Folktables：预测收入 >$50k

图4：对每个数据集所发现的关键子群体/特征对进行探索。第一张图展示了特征重要性从整个数据集到子群体的变化。第二张图展示了用于定义该子群体的主要系数。

群体，使得该特征的重要性差异几乎达到同样大的程度（所有数据集的比较见附录 K）。在某些情况下，边际子群体的表现略优于丰富子群体（图6）。当使用丰富子群体并未比使用边际子群体带来显著优势时，就会出现这种情况；而算法 1 中的经验误差容限会使收敛提前停止。

或许，算法1能够在丰富子群体上进行优化，更重要的实际优势在于，它允许将受保护子群体定义为连续变量的函数。例如，年龄很容易纳入我们的公式，而使用边际子群体来表示年龄，则需要先将年龄划分为年龄组，再对这些组进行独热编码；这种做法在统计、解释和计算方面都有缺陷。如第3节所述，我们的框架还允许个体同时属于多个群体，例如，多种族个体可能更适合被表示为不同种族群体中的分数成员，而不是某个单一离散群体的成员。边际子群体无法捕捉这类数据。

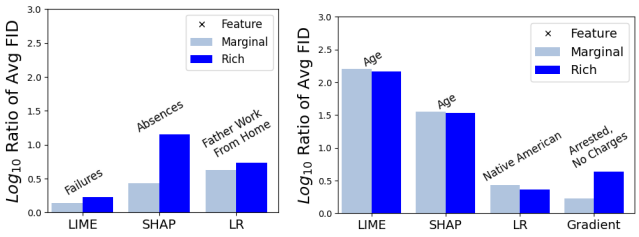
5.5 公平性指标

尽管针对给定特征和重要性度量的较大 AVG-FID 值并不能保证常见公平性指标存在差异（这些指标通常并非根据某个特定参考特征来定义），但自然会让人想知道这些度量之间是否相关：AVG-FID 较大的子群体，其公平性指标的差异是否也较大？公平性指标差异较大的子群体，是否会在某些特征上呈现特别大的 AVG-FID 值？

我们在表 2 中考察了第一个问题。我们发现，这些高 AVG-FID 子群体往往在传统公平性指标上存在显著差异。虽然这些指标在 g 上并非总是更差，但这进一步印证了这样的直觉：对于 AVG-FID 较高的子群体，需要进行更严格的审查。在表3中，我们研究相反的问题：使用 (Kearns 等人，2018) 的 GerryFair 代码，找出最大程度违反 FPR差异的丰富子群体，然后计算这些子群体的 AVG-FID。我们发现，这些子群体也包含 AVG-FID 较高的特征，尽管其数值不及算法 1 找到的特征高，因为算法 1 明确以优化 AVG-FID 为目标。这两项结果凸显了我们的方法在识别潜在高风险子群体方面的作用。

6. 讨论

在本文中，我们确立了特征重要性差异作为协助开展 DBI 的重要工具。我们这项研究的一项优势是，随着特征重要性方法不断进步，这些方法可以应用于我们的算法 1（如果它们是独立的——



（a）学生　　（b）COMPAS R

图 5：将一些特征重要性差异最大的丰富子群体与同一特征上特征重要性差异最大的边际子群体进行比较，使用与图2相同的对数刻度。子群体对应的特征标注在每个条形上方。

表 2：高 AVG-FID 子群体的公平性指标。COMPAS D 和学生被排除在外，因为它们使用非二元y，导致分类指标的可比性较低。我们测量了 (Barocas 等人，2019) 概述的 3 种公平性类型：$P(\hat{Y}=1)$、真阳性率/假阳性率以及期望校准误差。 <指标>$_\Delta$ 是 g 上的指标减去 X上的指标。

数据集	概念	F	定义特征 g	$\hat{Y}_\Delta$	TPR_Δ	FPR_Δ	ECE_Δ
COMPAS R	LIME	Age	美洲原住民	−.16	−.2	−.07	.24
	SHAP	Age	亚裔美国人	.37	.26	.4	−.12
	GRAD	逮捕，无指控	美洲原住民	−.24	−.35	−.12	.37
Bank	LIME	员工人数	婚姻状况	.11	.08	.07	−.15
	SHAP	欧元银行同业拆借利率	婚姻状况	.11	.08	.07	−.17
Folktables	LIME	Age	婚姻状况	−.15	−.09	−.07	.19
	SHAP	教育	亚裔美国人	.17	.2	.05	−.09

表3：比较通过我们的方法（AVG-FID$_{FID}$）发现的 g 排名靠前的特征及其相应的 AVG-FID，与（Kearns 等人，2018）（AVG-FID$_{gerry}$）发现的结果。与表 2 一样，COMPAS D 和学生数据集已排除。

数据集	概念	F_{FID}	AVG-FID$_{FID}$	F_{gerry}	AVG-FID$_{gerry}$
COMPAS R	LIME	Age	.14	Age	.04
	SHAP	Age	.4	Age	.06
	GRAD	逮捕，无指控	.09	Male	.03
Bank	LIME	员工人数	.03	员工人数	.008
	SHAP	欧元银行同业拆借利率	.016	就业方差率	.004
Folktables	LIME	Age	.11	Age	.05
	SHAP	教育	.13	Age	.05

可耕种）。正如我们所讨论的，鉴于观察到的特征/子群体差异，分析师可以采取许多后续措施来消除数据集偏差，包括干预训练或特征选择过程，以及更深入地调查数据的收集方式。这一视角与公平性文献中近期关注理解和记录数据来源的研究相辅相成（Gebru 等，2018年；Fabris 等，2022年）。在算法方面，与开发在模型训练或表示学习过程中消除偏差的方法相比，系统识别数据中偏差来源的研究还很少。本文尝试从特征重要性的角度，开发此类以“数据为中心”的公平性方法。要确定应收集哪些特征子集、如何检测和减轻数据收集过程中的偏差，以及在模型训练前应保留哪些数据或特征子集，仍有大量工作有待开展。

致谢

作者谨向 Cynthia Dwork 致谢，感谢她最初与我们探讨并启发了这一问题方向。

影响声明

本文旨在帮助从业者调查机器学习中常用数据集的潜在偏差来源。因此，这项工作有望推动众多应用领域实现更加公平、更加民主化的算法决策。正如相关文献中常见的观点所强调的，我们认为，没有任何一种方法能够一劳永逸地缓解机器学习系统中的偏差。

参考文献

Agarwal, A., Beygelzimer, A., Dudik, M., Langford, J., and Wallach, H. 一种实现公平分类的归约方法。载于 Dy, J. and Krause, A. (编)，第*35*届国际机器学习会议论文集，第80卷，机器学习研究论文集，第60–69页。PMLR，2018年7月10日至15日。网址：https://proceedings.mlr.press/v80/agarwal18a.html。

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. M. 一种实现公平分类的归约方法。载于 Dy, J. G. and Krause, A.（编），第*35*届国际机器学习大会论文集，*ICML 2018*，*Stockholmsmässan*，斯德哥尔摩，瑞典，*2018*年*7*月*10*日至*15*日，《机器学习研究论文集》第80卷，第60–69页。PMLR，2018b。网址：http://proceedings.mlr.press/ v80/agarwal18a.html。

Agarwal, C., Johnson, N., Pawelczyk, M., Krishna, S., Sax- ena, E., Zitnik, M., andLakkaraju, H. 重新思考基于归因的解释的稳定性，2022年。

Agarwal, C., Krishna, S., Saxena, E., Pawelczyk, M., Johnson, N., Puri, I., Zitnik, M., and Lakkaraju, H. Openxai：迈向对模型解释的透明评估。载于 Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., andOh, A.（编），神经信息处理系统进展，第 35 卷，第 15784–15799 页。Curran Associates, Inc., 2022b。网址：https://proceedings.neurips. cc/paper_files/paper/2022/file/65398a0eba88c9b4a1c38ae405b125ef-Paper-和_基准.pdf。

Alvarez-Melis, D. 和 Jaakkola, T. S. 论可解释性方法的稳健性，2018年。

Angwin, J., Larson, J., Mattu, S., and Kirch-ner, L. 机器偏见：全美各地都在使用一种软件来预测未来罪犯，而它对黑人存在偏差，2016年。网址 https://www.propublica.org/article/machine-bias-risk-assessments-in-crim [访问日期：15.11.2020]。

Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., and Müller, K.-R. 如何解释个体分类决策。机器学习研究期刊，11(61):1803–1831，2010年。网址 http://jmlr. org/papers/v11/baehrens10a.html。

Balagopalan, A., Zhang, H., Hamidieh, K., Hartvigsen, T., Rudzicz, F., 和 Ghassemi, M. 通往可解释性之路 i^s pave d with bias: M_easur ing the fai^rness o f expl 论文集。在*2022*年*ACM*公平性、问责制与透明度会议论文集中，FAccT '22，第1194–1206页，纽约，纽约州，美国，2022年。计算机协会。ISBN 9781450393522。doi：10.1145/3531146.3533179。URL https://doi. org/10.1145/3531146.3533179。

Bansal, N.、Agarwal, C. 和 Nguyen, A. Sam：归因方法对超参数的敏感性，2020年。

Barocas, S.、Hardt, M. 和 Narayanan, A. 公平性与机器学习：局限性与机遇。fairml- book.org，2019年。http://www.fairmlbook.org。Begley, T.、Schwedes, T.、Frye, C. 和 Feige, I. 公平机器学习的可解释性。*CoRR*，abs/2010.07389，2020年。URL https://arxiv.org/abs/2010. 07389。

Bolukbasi, T.、Chang, K.-W.、Zou, J.、Saligrama, V. 和 Kalai, A. 男性之于计算机程序员，正如女性之于家庭主妇？消除词嵌入中的偏差。载于 第 *30* 届国际神经信息处理系统会议论文集，NIPS'16，第 4356–4364 页，美国纽约州雷德胡克，2016 年。Curran Associates Inc. ISBN 9781510838819。

Boxer, K. S.、au2, E. M. I. 和 Neill, D. B. 审计预测模型中的交叉性偏见，2023 年。

Buolamwini, J. 和 Gebru, T.《性别阴影：商业性别分类中的交叉性准确率差异》。载于 Friedler, S. A. 和 Wilson, C.（编），第一届公平性、问责制与透明度会议 数据集，机器学习研究论文集第 81 卷，第 77–91 页。PMLR，2018 年 2 月 23–24 日。网址：https://proceedings.mlr.press/v81/buolamwini18a.html。Bureau, U. C.《2014–2018 年 ACS PUMS 数据词典》，2020 年。网址：https://www2.census.gov/programs-surveys/acs/tech_docs/ pums/data_dict/PUMS_Data_Dictionary_2014-2018.pdf。Caton, S. 和 Haas, C.《机器学习中的公平性：综述》。i nalcomputing ACM 计算机调查，2023 年 8 月。ISSN 0360-0300。doi：10.1145/3616865。网址：https://doi.org/10.1145/3616865。已接受。Chouldechova, A.《具有差别影响的公平预测：再犯预测工具中的偏差研究》。大数据，5(2)：153–163，2017 年。doi：10.1089/big.2016. 0047。网址：https://doi.org/10.1089/big. 2016.0047。PMID：28632438。

Cortez, P. 和 Silva, A.《利用数据挖掘预测中学生的学生表现》。载于第五届未来商业技术会议论文集，第 5–12 页，2008 年。网址：http://www3.dsi.uminho.pt/ pcortez/student.pdf。

Dai, J., Upadhyay, S., Aivodji, U., Bach, S. H. 和 Lakkaraju, H. 通过解释质量实现公平性。载于*2022*年*AAAI/ACM*人工智能、伦理与社会会议论文集。ACM，2022年7月。doi：10.1145/ 3514094.3534159。URL https://doi.org/10. 1145%2F3514094.3534159。

Dastile, X., Celik, T. 和 Potsane, M. 信用评分中的统计学和机器学习模型：系统性文献综述。《应用软计算》，91:106263，2020年。ISSN 1568-4946。doi：https://doi.org/10.1016/j.asoc.2020.106263。URL https://www.sciencedirect.com/science/article/pii/S1568494620302039。

Dimanov, B., Bhatt, U., Jamnik, M. 和 Weller, A. 别相信我：学习能够向多种解释方法掩盖不公平性的模型。发表于 *SafeAI@AAAI*，2020年。URL https: // api.semanticscholar.org/CorpusID: 211838210.

Ding, F., Hardt, M., Miller, J. 和 Schmidt, L. 让成年人退役：用于公平机器学习的新数据集。神经信息处理系统进展 *34*（*NeurIPS*），2021年。URL https://par.nsf.gov/ biblio/10341458.

Dwork, C. 和 Ilvento, C. 组合下的公平性。*CoRR*，abs/1806.06122，2018年。URL http://arxiv.org/abs/1806.06122.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O. 和 Zemel, R. 通过认知实现公平性。发表于 第三届理论计算机科学创新会议论文集，ITCS '12，第214–226页，纽约，纽约州，美国，2012年。计算机协会。ISBN 9781450311151。doi: 10.1145/2090236.2090255。URL https://doi. org/10.1145/2090236.2090255.

Fabris, A.、Messina, S.、Silvello, G. 和 Susto, G. A. 算法公平性数据集：迄今为止的发展历程。*CoRR*，abs/2202.01711，2022年。网址：https://arxiv.org/ abs/2202.01711。

Freund, Y. 和 Schapire, R. E. 博弈论、在线预测与提升。载于第九届计算学习理论年度会议论文集，COLT '96，第 325–332 页，纽约，纽约州，美国，1996年。计算机协会。ISBN 0897918118。doi: 10.1145/238061.238163。URL https://doi.org/10.1145/238061.238163。

Gebru, T.、Morgenstern, J.、Vecchione, B.、Vaughan, J. W.、Wallach, H. M.、III、H. D. 和 Crawford, K. 数据集数据表。*CoRR*，abs/1803.09010，2018年。URL http: // arxiv.org/abs/1803.09010。

Gopalan, P.、Kalai, A. T.、Reingold, O.、Sharan, V. 和 Wieder, U. 全能预测器。载于 Braverman, M.（编），第 *13* 届理论计算机科学创新会议（*ITCS 2022*），第 215 卷，莱布尼茨国际信息学会议论文集（*LIPIcs*），第 79:1–79:21 页，德国 Dagstuhl，2022年。Schloss Dagstuhl – 莱布尼茨信息学中心。ISBN 978-3-95977-217-4。doi：10.4230/LIPIcs.ITCS.2022.79。网址：https://drops.dagstuhl.de/opus/volltexte/2022/15675。

Grabowicz, P. A., Perello, N. 和 Mishra, A. 在监督学习中结合公平性与可解释性。载于*2022* 年 *ACM* 公平性、问责与透明度会议。ACM，2022 年 6 月。doi: 10. 1145/3531146.3533236. URL https://doi.org/ 10.1145%2F3531146.3533236。

Hardt, M., Price, E. 和 Srebro, N. 监督学习中的机会平等。载于 Lee, D. D., Sugiyama, M., von Luxburg, U., Guyon, I. 和 Garnett, R.（编），神经信息处理系统进展第 *29* 卷：*2016* 年神经信息处理系统年度会议，*2016* 年 *12* 月 *5—10* 日，西班牙巴塞罗那，第 3315–3323 页，2016a。URL https://proceedings.neurips. cc/paper_files/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper. pdf。

Hardt, M., Price, E. 和 Srebro, N. 监督学习中的机会平等。载于第 *30* 届神经信息处理系统国际会议论文集，NIPS'16，第 3323–3331 页，美国纽约州雷德胡克，2016b。Curran Associates Inc. ISBN 9781510838819。

Hebert-Johnson, U., Kim, M., Reingold, O. 和 Rothblum, G. 多重校准：针对（可计算识别的）群体的校准。载于 Dy, J. 和 Krause, A.（编），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，第 1939–1948 页。PMLR，2018年7月10日至15日。网址：https://proceedings.mlr.press/v80/hebert-johnson18a.html。

Hu, L., Livni Navon, I. R., Reingold, O. 和 Yang, C. 面向约束优化的全能预测器。载于 Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S. 和 Scarlett, J.（编），第 *40* 届机器学习国际会议论文集，机器学习研究论文集第 202 卷，第

13497–13527 页。PMLR，2023 年 7 月 23–29 日。网址：https:// proceedings.mlr.press/v202/hu23b.html。

Ingram、E.、Gursoy、F. 和 Kakadiaris、I. A. 机器学习犯罪再犯模型的准确度、公平性和可解释性。发表于*2022* 年 *IEEE/ACM*大数据计算、应用与技术国际会议（*BDCAT*），第233–241页，美国加利福尼亚州洛斯阿拉米托斯，2022年12月。IEEE计算机学会。doi：10.1109/BDCAT56447.2022.00040。网址：https://doi.ieeecomputersociety.org/10.1109/BDCAT56447.2022.00040。

Joseph, M.、Kearns, M.、Morgenstern, J.、Neel, S. 和 Roth, A. 面向无限及情境老虎机的精英公平性。载于*2018*年*AAAI/ACM*人工智能、伦理与社会会议论文集，AIES ’18，第158–163页，美国纽约州纽约市，2018年。计算机协会。ISBN 9781450360128。doi：10. 1145/3278721.3278764。URL：https://doi.org/ 10.1145/3278721.3278764。Jung, C.、Lee, C.、Pai, M.、Roth, A. 和 Vohra, R. 不确定性估计的矩多重校准。载于 Belkin, M. 和 Kpotufe, S.（编），第三十四届学习理论会议论文集，第134卷机器学习研究论文集，第2634–2678页。PMLR，2021年8月15日至19日。URL：https://proceedings.mlr.press/v134/jung21a.html。Kamiran, F. 和 Calders, T. 面向无歧视分类的数据预处理技术。知识与信息系统，33(1)：1–33，2012年。doi：https://doi. org/10.1007/s10115-011-0463-8。Kearns, M.、Neel, S.、Roth, A. 和 Wu, Z. S. 防止公平性选区划分操纵：面向子群体公平性的审计与学习。载于 Dy, J. 和 Krause, A.（编），第*35*届国际机器学习会议论文集，第80卷机器学习研究论文集，第2564–2572页。PMLR，2018年7月10日至15日。URL：https://proceedings.mlr.press/v80/kearns18a.html。Kearns, M.、Neel, S.、Roth, A. 和 Wu, Z. S. 面向机器学习的丰富子群体公平性的实证研究。载于公平性、问责制与透明度会议论文集，FAT* ’19，第100–109页，美国纽约州纽约市，2019年。计算机协会。ISBN 9781450361255。doi：10. 1145/3287560.3287592。URL：https://doi.org/ 10.1145/3287560.3287592。Keddell, E. 儿童保护中的算法正义：统计公平性、社会正义及其影响

实践。社会科学，8(10)，2019年。ISSN 2076-0760。doi：10.3390/socsci8100281。URL：https: // www.mdpi.com/2076-0760/8/10/281。Kindler, J. Sion 极小极大定理的一个简单证明。美国数学月刊，112(4)：356–358，2005年。URL：http: // www.jstor.org/stable/30037472。Kingma, D. P. 和 Ba, J. Adam：一种随机优化方法。载于 Bengio, Y. 和 LeCun, Y.（编），第三届国际学习表征会议，*ICLR 2015*，美国加利福尼亚州圣迭戈，*2015*年*5*月*7*日至*9*日，会议论文集，2015年。URL：http: // arxiv.org/abs/1412.6980。Kivinen, J. 和 Warmuth, M. K. 指数梯度与线性预测器的梯度下降。信息与计算，132(1)：1–63，1997年。doi：10.1006/inco.1996.2612。URL：https://doi.org/10.1006/inco.1996. 2612。Kleinberg, J.、Mullainathan, S. 和 Raghavan, M. 风险评分公平判定中的内在权衡。载于 Papadimitriou, C. H.（编），第八届理论计算机科学创新会议（*ITCS 2017*），第67卷莱布尼茨国际信息学会议论文集（*LIPIcs*），第43:1–43:23页，德国 Dagstuhl，2017年。Schloss Dagstuhl–莱布尼茨信息学中心。ISBN 978-3-95977-029-3。doi：10.4230/LIPIcs. ITCS.2017.43。URL：http://drops.dagstuhl. de/opus/volltexte/2017/8156。Kleinberg, J. M.、Mullainathan, S. 和 Raghavan, M. 风险评分公平判定中的内在权衡。载于信息技术融合与服务，2016年。URL：https://api.semanticscholar. org/CorpusID:12845273。Krishna, S.、Han, T.、Gu, A.、Pombra, J.、Jabbari, S.、Wu, S. 和 Lakkaraju, H. 可解释机器学习中的分歧问题：从业者视角。*CoRR*，abs/2202.01602，2022年。URL：https: // arxiv.org/abs/2202.01602。Kusner, M. J.、Loftus, J.、Russell, C. 和 Silva, R. 反事实公平。载于 Guyon, I.、Luxburg, U. V.、Bengio, S.、Wallach, H.、Fergus, R.、Vishwanathan, S. 和 Garnett, R.（编），神经信息处理系统进展，第30卷。Curran Associates, Inc.，2017年。URL：https://proceedings.neurips. cc/paper_files/paper/2017/file/a486cd07e4ac3d270571622f4f316ec5-Paper

pdf. Li, Y. 和 Vasconcelos, N. Repair：通过数据集重采样消除表征偏差。发表于*2019*年 *IEEE/CVF* 计算机视觉与模式识别会议

（计算机视觉与模式识别会议），第 9564–9573 页，2019 年。doi：10.1109/CVPR.2019. 00980。

Liu, J., Zhong, C., Seltzer, M., and Rudin, C. 面向广义线性模型和加性模型的快速稀疏分类。载于 Camps-Valls, G., Ruiz, F. J. R., and Valera, I. (eds.), 第*25*届人工智能与统计国际会议论文集，第 151卷，机器学习研究论文集，第9304–9333页。PMLR，2022年3月28–30日。网址：https://proceedings.mlr.press/ v151/liu22f.html。

Lundberg, S. M. 解释公平性的定量度量。载于公平与负责任人工智能研讨会@ *CHI 2020*，2020年。

Lundberg, S. M. and Lee, S.-I. 一种解释模型预测的统一方法。载于第*31*届神经信息处理系统国际会议论文集，NIPS’17，第4768–4777页，美国纽约州雷德胡克，2017年。Curran Associates Inc. ISBN 9781510860964。

Miller, C. 算法与偏差：与 Cynthia Dwork 的问答，2015年8月。

Molnar, C. 可解释机器学习。第2版，2022年。网址 https://christophm.github.io/ interpretable-ml-book。

Moro, S.、Cortez, P. 和 Rita, P. 一种数据驱动的银行电话营销成功率预测方法。决策支持系统，62：22–31，2014年。ISSN 0167-9236。doi：https://doi.org/10.1016/j.dss.2014.03. 001。网址https://www.sciencedirect.com/ science/article/pii/S016792361400061X。

Neill, D. B. 用于空间模式检测的快速子集扫描。英国皇家统计学会会刊 *B*辑：统计方法学，74(2)：337–360，2012年1月。ISSN 1369-7412。doi：10.1111/j.1467-9868. 2011.01014.x。网址 https://doi.org/10.1111/ j.1467-9868.2011.01014.x。

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. 关于公平性与校准。载于 Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (编)，神经信息处理系统进展，第 30 卷。Curran Associates, Inc.，2017年。URL https://proceedings.neurips. cc/paper_ files/paper/2017/file/ b8b9c74ac526fffbeb2d39ab038d1cd7-Paper pdf.

Quinlan, J. R. 决策树的归纳。机器学习，1(1):81–106，1986年。doi: 10.1007/BF00116251。URL https://doi.org/10.1007/BF00116251。

Ribeiro, M. T., Singh, S., and Guestrin, C. “我为什么应该相信你？”：解释任意分类器的预测。载于第*22*届 *ACM SIGKDD* 国际知识发现与数据挖掘会议论文集，KDD ’16，第 1135–1144 页，纽约，纽约州，美国，2016。计算机协会。ISBN 9781450342322。doi: 10.1145/ 2939672.2939778。URL https://doi.org/10. 1145/2939672.2939778。

Simonyan, K., Vedaldi, A., and Zisserman, A. 卷积网络深处：图像分类模型与显著性图的可视化。*CoRR*，abs/1312.6034，2013年。URL http://dblp.uni-trier.de/db/journals/ corr/corr1312.html#SimonyanVZ13。

Slack, D., Hilgard, S., Jia, E., Singh, S., 和 Lakkaraju, H. 欺骗 LIME 和 SHAP：针对事后解释方法的对抗攻击。载于 *AAAI/ACM*人工智能、伦理与社会会议论文集，AIES ’20，第 180–186 页，美国纽约州纽约市，2020年。计算机协会。ISBN 9781450371100。doi：10. 1145/3375627.3375830。URL https://doi.org/ 10.1145/3375627.3375830。

Sundararajan, M., Taly, A., 和 Yan, Q. 深度网络的公理化归因。载于 Precup, D. 和 Teh, Y. W.（编），第*34*届国际机器学习会议论文集，第 70 卷，机器学习研究论文集，第 3319–3328 页。PMLR，2017年8月6日至11日。URL https://proceedings.mlr. press/v70/sundararajan17a.html。

Tommasi, T., Patricia, N., Caputo, B. 和 Tuytelaars, T. 深入探讨数据集偏差。载于计算机视觉应用中的领域自适应，第 37–55 页。施普林格，2017年。

Zemel, R. S., Wu, Y., Swersky, K., Pitassi, T. 和 Dwork, C. 学习公平表征。载于第*30*届国际机器学习会议论文集，*ICML 2013*，亚特兰大，美国佐治亚州，*2013*年*6*月*16*日至*21*日，第28卷，收录于*JMLR*研讨会与会议论文集，第325–333页。JMLR.org，2013年。网址：http://proceedings.mlr.press/v28/zemel13.html。

Zhang, Z. 和 Neill, D. B. 识别分类器中的显著预测偏差。*arXiv*，abs/1611.08292，2016年。URL https://api.semanticscholar.org/CorpusID:2256765。

.

A. 局限性

重要的是，我们不作任何更广泛的断言，认为较大的特征重要性差异必然意味着关于底层分类模型公平性的数学结论在所有情况下都成立。众所周知，即使是最流行、最自然的公平性指标，也不可能同时满足，因此我们会面临一个问题：如何界定模型公平（Chouldechova，2017；Kleinberg 等人，2017）。通过检测特征重要性方面的异常子群体，我们的方法可以向领域专家提示，特征收集或测量偏差等问题或许值得关注。这将有助于后续检验由此得出的假设，并最终采取干预措施，以解决差异、改善公平性结果。人们已对应用最广泛的特征重要性概念（包括我们研究的概念）的稳定性和稳健性提出担忧（Dai 等，2022；Agarwal 等，2022a；Alvarez-Melis 和 Jaakkola，2018；Bansal 等，2020；Dimanov 等人，2020；Slack 等人，2020）；而且这些概念往往彼此矛盾，因此没有任何一种可以被视为定论（Krishna 等，2022）。尽管存在这些局限性，这些概念如今仍在实践中得到广泛应用；正如我们在此提出的，它们仍可作为诊断工具，用于揭示可能值得关注的偏差。最后，和几乎所有此前关于公平性的研究一样，我们的方法要求使用预先定义了敏感特征的表格数据集；而在文本或图像数据集中，这一过程更为困难，尽管偏差问题同样值得关注（Buolamwini 和 Gebru，2018；Bolukbasi 等人，2016）。总的来说，这里开发的方法构成了算法工具包的一部分，领域专家可借此排查偏差。

B. 可复现性

实验所用超参数等具体细节见附录 I。用于这些实验的源代码已随补充材料提供。具体而言，run_separable.py 和 run_linear.py 是运行前用于指定重要性度量（附录 I.2）、数据集（附录 I.3）及其他参数的脚本。experiments/ 目录包含用于比较丰富子群体与边际子群体的脚本（见附录 K），以及第 5.5 小节中公平性比较实验所用的脚本。

C. 其他相关工作

机器学习中的公平性。机器学习公平性领域的大量研究通常关注在给定的学习场景中实现一种新的公平性概念；它可以是个体公平性概念（Dwork 等人，2012；Joseph 等人，2018年）、基于使受保护子群体间的统计率相等的概念（Hardt 等人，2016b；Pleiss 等人，2017年），或基于底层因果模型的概念（Kusner 等人，2017年）。给定一种公平性概念后，学习公平分类器的方法通常可分为“处理中”，即尝试同时学习分类器并满足公平性约束；“后处理”，即对已学习的分类器进行后处理，使其满足公平性定义（Hardt 等人，2016b）；或者，与本文的动机最为相关的“预处理”，即对数据进行预处理以消除偏差。现有关于数据集偏差的研究为我们的工作提供了宏观层面的动机。

特征重要性概念。第2节中提到的局部解释方法包括 LIME 或 SHAP 等模型无关方法（Ribeiro 等人，2016；Lundberg 和 Lee，2017）、显著性图等方法（Simonyan 等人，2013；Sundararajan 等人，2017；Baehrens 等人，2010），这类方法要求 h 在 x 中可微；此外，还有依赖分类器的模型专用方法。除了这些解释方法之外，还有试图解释整个模型行为的全局方法，因此可以在整个子群体上运行。附录 H 中描述的 LIN-FID 方法是一种全局方法，它通过在子群体上训练一个内在可解释模型（线性回归），并检查其系数来实现。其他可用于定义子群体重要性度量的内在可解释模型包括决策树（Quinlan，1986年）和广义加性模型（Liu 等人，2022年）。

公平性与可解释性。尽管目前尚无研究考察特征重要性概念在检测丰富子群体差异方面的作用，但已有少量研究探讨了公平性背景下的可解释性。近期，Grabowicz 等人（2022年）将诱导歧视形式化为分配给敏感特征的 SHAP 值的函数，并提出一种方法，用于学习受保护属性影响较低的分类器。Begley 等人（2020年）采用了类似的方法，利用 Shapley 值将模型整体的不公平性归因于其单个特征，并提出一种改善公平性的干预措施。Ingram 等人（2022年）研究了用于预测再犯的机器学习模型，并通过实证表明，模型准确率、公平性与可解释性之间存在权衡。

此外，(Lundberg，2020年) 将特征归因解释和公平性指标分解为加性分量，并观察公平性指标与输入特征之间的关系。我们的工作并不试图将公平性

指标分解为加性分量，同时也关注非加性的特征解释。此外，我们对丰富子群体的考量是对这一领域的一项创新。

D. 计算最优子群体

算法 1迭代约束优化

1: **输入：** 数据集 X^n，$|X^n| = n$、假设 h、特征约束 α_L 和 α_U、规模违反指示函数 Φ_L 和 **准确度** ν。**2：初始化：3：特征重要性向量 C** $= (F(f_j, X_i, h))_{i=1}^n$**4：梯度权重参数** $\theta_1 =$ **(0, 0) 5：学习率** $\eta = \frac{\nu}{2n^2 B}$**6：对于** $t = 1$**, 2, ... 执行 7：# 指数梯度权重 8：**$\lambda_{t,0} = B \frac{exp(\theta_t,0)}{1+exp(\theta_t,1)}$**,** $\lambda_{t,1} = B \frac{exp(\theta_t,1)}{1+exp(\theta_t,0)}$ **9：10：# 代价向量 11：**$c_t^1 = (\mathbf{C}_i - \lambda_{t,0} + \lambda_{t,1})_{i=1}^n$**12：13：# 通过 CSC 预言机获取具有最大差异的 g 14：**$g_t = CSC_{\mathcal{G}}(0, c_t^1)$ **15：16：# 计算拉格朗日函数**17 $p_{\mathcal{G}}^t = \frac{1}{t}\sum_{t'=1}^t g_{t'}$**：ˆ：((ˆ)** $B\Phi_U(\hat{p}_{\mathcal{G}}^t))$ 18 $\lambda_t' = B\Phi_L \; p_{\mathcal{G}}^t$ **19：**$L = L(\hat{p}_{\mathcal{G}}^t, \lambda_t')$ **20：21：**$\hat{p}_\lambda^t = \frac{1}{t}\sum_{t'=1}^t(\lambda_{t'},0,\lambda_{t'},1)$ **22：**$g_t' = CSC_{\mathcal{G}}(0, (\mathbf{C}_i - \hat{p}_{\lambda_0}^t + \hat{p}_{\lambda_1}^t)_{i=1}^n)$ **23：**$L = L(g_t', \hat{p}_\lambda^t)$ **24：25：**$v_t = \max\left(|L(\hat{p}_{\mathcal{G}}^t, \hat{p}_\lambda^t) - L|, |L - L(\hat{p}_{\mathcal{G}}^t, \hat{p}_\lambda^t)|\right)$**26：27：# 检查终止条件 28：如果** $v_t \le v$ **则 29：返回** $\hat{p}_{\mathcal{G}}^t, \hat{p}_\lambda^t$**30：结束 if 31：32：# 指数梯度更新 33：设置** $\theta_{t+1} = \theta_t + \eta(\alpha_L - |g_t|, |g_t| - \alpha_U)$ **34：结束 for**

感兴趣特征 f_j、可分离的特征重要性函数 F、规模 Φ_U，规模惩罚上界 B，用于 $\mathcal{G}$ 的 CSC 预言机，$CSC_{\mathcal{G}}(c^0, c^1)$，

我们首先证明，对于无约束问题，当 F 可分离时，只需调用 CSC$_{\mathcal{G}}$ 两次，即可计算出使 g_j^* 上的 FID(f_j, g, h) 在 $\mathcal{G}$ 上最大化的子群体。

引理 D.1. 若 F 可分离，且 *CSC*$_{\mathcal{G}}$ 是针对 $\mathcal{G}$ 的 *CSC* 预言机，那么对于任意特征 f_j，只需调用预言机两次即可计算出 g_j^*。

证明。 根据定义，$g_j^* =$最大值点$_{g\in\mathcal{G}}$特征重要性差异(j, g) $=$最大值点$_{g\in\mathcal{G}}|F(f_j, X^n, h) - F(f_j, g, h)| =$最大值点$_{g\in\{g^+,g^-\}}$特征重要性差异($j$, g)，其中 $g^+ =$最大值点$_{g\in\mathcal{G}}F(f_j, g, h)$ $g^- =$最小值点$_{g\in\mathcal{G}}F(f_j, g, h)$。根据可分离性的定义，我们可以写成

$$F(f_j, g(X^n), h) = \sum_{X\in g(X^n)} F'(f_j, X, h) = \sum_{i=1}^n g(X_i)F'(f_j, X_i, h)$$

然后令 $c_k^0 = 0$ 和 $c_k^1 = -F'(f_j,\ X_k,\ h)$ 对于 $k = 1,\ \ldots n$，我们看到 $g^+ = \text{CSC}_g((c_k^0,\ c_k^1))$ $g^- = \text{CSC}_g((c_k^0,\ -c_k^1))$。这就证明了该结论。 □

定理 4.1：设 F 是一个可分概念，固定分类器 h、子群类 $\mathcal{G}$和预言机 CSC$_\mathcal{G}$。然后固定一个感兴趣特征 f_j，我们将运行算法 1 两次：一次使用由 F给出的 FID，另一次使用由 $-F$给出的 FID。设 $\hat{}p_\mathcal{G}^T$是算法 1 经过 $T = O(\frac{4n^2B^2}{\nu^2})$次迭代后返回的分布，且其 $\mathbb{E}[\text{FID}(j,\ g)]$值较大。那么：

$$\begin{aligned} \texttt{FID}(j, g_j^*) - \mathbb{E}_{g\sim\hat{p}_\mathcal{G}^T}[\texttt{FID}(j,g)] \leq \nu \\ |\Phi_L(g)|, |\Phi_U(g)| \leq \frac{1+2\nu}{B} \end{aligned} \tag{3}$$

证明。 我们首先将约束优化转化为一个 min – max 目标。最小化玩家（下文称为子群体玩家 ）将在每次迭代中针对类 $\mathcal{G}$ 求解一个 CSC 问题，而最大化玩家（称为对偶玩家）则使用指数梯度算法（Kivinen & Warmuth，1997）调整两个约束上的对偶权重 λ。根据引理 D.2（Freund & Schapire，1996），我们知道，如果每个玩家都采用无遗憾策略，那么经过 T轮后找到的子群体误差至多比最优值差两位玩家的平均累计遗憾。指数梯度下降的遗憾界保证这一过程在多项式*(n)* 轮内完成。

与（Kearns 等人，2018；Agarwal 等人，2018a）一样，我们首先放宽方程 1，以便在子群体上优化整体分布，并确保约束在该分布下以期望形式成立。由此得到新的优化问题：

$$\begin{aligned} \min_{p_g\in\Delta(\mathcal{G})} \quad & \mathbb{E}_{g\sim p_g}[\sum_{i=1}^{n} g(x_i)F'(f_j, x_i, h)] \\ \text{s.t.} \quad & \mathbb{E}_{g\sim p_g}[\Phi_L(g)] \leq 0 \\ & \mathbb{E}_{g\sim p_g}[\Phi_U(g)] \leq 0 \end{aligned} \tag{4}$$

我们注意到，尽管 $|\mathcal{G}|$ 可能是无限的，但由 $\mathcal{G}$ 中的元素对 X 进行标记的不同方式数量有限；我们用 $|\mathcal{G}(X)|$ 表示这些标记方式的数量。由于方程 4 是一个含有 $|\mathcal{G}(X)|$ 个变量的有限线性规划，因此满足强对偶性，我们可以写出：

$$\begin{aligned} (p_g^*, \lambda^*) = \text{argmin}_{p_g\in\Delta(\mathcal{G})}\text{argmax}_{\lambda\in\Lambda}\mathbb{E}_{g\sim p_g}[L(g,\lambda)] = \text{argmin}_{p_g\in\Delta(\mathcal{G})}\text{argmax}_{\lambda\in\Lambda}L(p_g,\lambda) \\ \text{with} \quad L(g,\lambda) = \sum_{x\in X} g(x)F(f_j,x,h) + \lambda_L\Phi_L + \lambda_U\Phi_U, \quad L(p_g,\lambda) = \mathbb{E}_{g\sim p_g}[L(g,\lambda)] \end{aligned}$$

与（Kearns 等人，2018）一样，$\Lambda = \{\lambda \in \mathbb{R}^2 \mid \|\lambda\|_1 \leq B\}$ 的选取是为了使定义域紧致；只要 B 足够大，即 $\|\lambda^*\|_1 \leq B$，就不会改变最优参数。在实践中，这是算法 1 的一个超参数，与（Agarwal 等人，2018a；Kearns 等人，2018）中的做法类似。接着，我们沿用（Agarwal 等人，2018a；Kearns 等人，2018）的推导，通过为子群体玩家（p_g）和对偶玩家（λ）实现无遗憾策略，证明可以高效地计算（$p_g^*,\ \lambda^*$）。

形式上，由于 $\mathbb{E}_{g\sim p_g}[L(g,\ \Lambda)]$ 关于 $p_g,\ \lambda$ 是双线性的，且 $\Lambda,\ \Delta(\mathcal{G})$ 是凸且紧的，根据 Sion 极小极大定理（Kindler，2005）：

$$\min_{p_g\in\Delta(G)}\max_{\lambda\in\Lambda} L(p_g,\lambda) = \max_{\lambda\in\Lambda}\min_{p_g\in\Delta(G)} L(p_g,\lambda) = \text{OPT} \tag{5}$$

那么，根据 (Kearns 等人，2018) 中的定理4.5，我们知道，如果 $(p_g^*,\ \lambda^*)$ 是方程 5 的一个 ν-近似极小极大解，具体而言，

$$\begin{aligned} \text{if:} \quad & L(p_g^*,\lambda^*) \leq \min_{p\in\Delta(\mathcal{G})} L(p,\lambda^*) + \nu, L(p_g,\lambda) \geq \max_{\lambda\in\Lambda} L(p^*,\lambda), \\ \text{then:} \quad & F(f_j,p_g^*,h) \leq OPT + 2\nu, \quad |\Phi_L(g)|, |\Phi_U(g)| \leq \frac{1+2\nu}{B} \end{aligned} \tag{6}$$

因此，为了计算近似最优的子群体分布 p_g^*，只需计算方程 5 的一个近似极小极大解。为此，我们依赖 Freund & Schapire（1996）的经典结果，该结果指出：如果子群体玩家采取最佳响应，且对偶玩家的遗憾较低，那么随着平均遗憾趋近于零，迄今为止找到的平均策略的次优性也会趋近于零。

引理 D.2 ((Freund & Schapire，1996年))。 设 $p_1^\lambda, \dots p_T^\lambda$ 是对偶玩家在 Λ 上进行的一系列分布，设 $g^1, \dots g^T$ 分别是子群体玩家针对这些分布的最佳响应。令 $\hat{}\lambda_T = \frac{1}{T}\sum_{t=1}^T p_t^\lambda$，$\hat{}p_g = \frac{1}{T}\sum_{t=1}^T g_t$。那么，如果

$$\sum_{t=1}^{T} \mathbb{E}_{\lambda \sim p_t^\lambda}[L(g_t, \lambda)] - \min_{\lambda \in \Lambda} \sum_{t=1}^{T} [L(g_t, \lambda)] \leq \nu T,$$

$\hat{}$那么，$(\lambda_T, \hat{}p_g)$ 是该博弈的 ν-近似极小极大均衡。

为证明定理 4.1，我们需要说明：（i）可以使用 CSC$_\mathcal{G}$ 高效地实现子群体玩家的最佳响应；（ii）需要利用引理 D.2，将对偶玩家最佳响应的遗憾界转化为关于最优性的论述。建立 (*i*) 是显而易见的，因为在每一轮 t，如果 $\lambda_{t,\ 0} = \mathbb{E}_{p_t^\lambda}[\lambda_L]$，$\lambda_{t,\ 1} = \mathbb{E}_{p_t^\lambda}[\lambda_U]$，那么最佳响应问题为：

$$\text{argmin}_{p_g \in \Delta(G)} \mathbb{E}_{g \sim p_g}[\sum_{x \in X} g(x) F(f_j, x, h) + \lambda_{t,0}\Phi_L + \lambda_{t,1}\Phi_U]$$

这可以进一步简化为：

$$\text{argmin}_{g \in G} \sum_{x \in X} g(x)(F(f_j, x, h) - \lambda_L + \lambda_U) \tag{7}$$

如预期，这可以通过一次 CSC$_\mathcal{G}$ 调用来计算。为证明 (ii)，即对偶玩家的分布所采用的无遗憾算法，我们注意到，在每一轮中，对偶玩家都在 2 维空间上进行在线线性优化。Algorithm1 实现了指数梯度算法（Kivinen & Warmuth，1997），该算法具有以下保证。此保证已在（Agarwal 等人，2018a）的定理 1 中得到证明，并可直接由指数梯度的遗憾界（Kivinen & Warmuth，1997）和引理 D.2 推出：

引理 D.3 （（Agarwal 等人，2018a））。 设定 $\eta = \frac{\nu}{2n^2 B}$，算法 *1* 在至多 $O(\frac{4n^2B^2}{\nu^2})$ 次迭代内返回一个 ν-近似极小极大点 $\hat{}p_\lambda^T$。

将此结果与方程 5 结合，即可完成证明。

□

E. AVG-FID 原语的证明

在第 4 节中，我们介绍了这样一种方法：在一系列子群体规模范围内进行约束优化 FID，即可高效地优化 AVG-FID。我们在此提供该结论更完整的证明：

设 g^* 为使 AVG-FID 最大的子群体。不失一般性，$g^* = $最大值点$_{g \in \mathcal{G}} \frac{1}{n|g|}\sum g(x)F'(f_j, X, h)$（我们省略绝对值，因为还可以设置 $F' = -F$）。因此必然有，g^* 也能解决约束优化问题最大值点$_{g \in \mathcal{G}} \frac{1}{n}\sum g(x)F'(f_j, X, h)$，使得 $|g| = |g^*|$；在目标函数中，我们省略了归一化项 $\frac{1}{|g|}$，因此我们优化的是约束 FID。

现在考虑一个区间 $I = [|g^*| - \alpha, |g^*| + \alpha]$，并设我们求解 $g_I^* = $最大值点$_{g \in \mathcal{G}} \frac{1}{n}\sum g(x)F'(f_j, X, h)$，使得$g \in I$。那么由于 $g^* \in I$，我们知道 $\frac{1}{n}\sum g^* F'(f_j, X, h) \leq \frac{1}{n}\sum g_I^*(x)F'(f_j, X, h)$。这意味着：

$$\begin{aligned}\texttt{AVG-FID}(g_I^*) &\geq \frac{1}{|g_I^*|}\frac{1}{n}\sum g^*(x)F'(f_j,X,h)\\ &= \texttt{AVG-FID}(g^*) + (\frac{1}{|g_I^*|+\alpha} - \frac{1}{|g_I^*|})\texttt{FID}(g^*)\\ &= \texttt{AVG-FID}(g^*) - \frac{\alpha}{|g^*|(|g^*|+\alpha)}\cdot\texttt{FID}(g^*)\end{aligned}$$

根据上述推导，由于 $\alpha \to 0$，我们有 AVG-FID(g_I^*) $\to$ AVG-FID(g^*)。

因此，如果找到一个足够小且围绕 $|g^*|$ 的区间I，我们就可以计算出一个子群体 g，使其近似优化 AVG-FID。由于第 4 节中的离散化覆盖了单位区间，只要 n 足够大，我们就一定能找到这样的区间。

F. 成本敏感分类器，CSC$_\mathcal{G}$

定义 F.1.（代价敏感分类）对于假设类 $\mathcal{G}$，代价敏感分类（CSC）问题由一组 n 个元组 $\{(X_i,\ c_i^0,\ c_i^1)\}_{i=1}^n$ 给出，其中 c_i^0 和 c_i^1 分别是将标签 0 和 1 分配给 X_i 的成本。CSC 预言机会找到使所有点的总成本最小化的分类器 $\hat{}g \in \mathcal{G}$：

$$\hat{g} = \operatorname*{argmin}_{g\in\mathcal{G}} \sum_i \Big(g(X_i)c_i^1 + (1-g(X_i))c_i^0\Big) \tag{8}$$

算法 2CSC$_\mathcal{G}$

输入： 数据集 $X \subset \mathbb{R}^{d_{sens}} \times \mathbb{R}^{d_{safe}}$，成本 $(c^0,\ c^1) \in \mathbb{R}^n$令 X_{sens} 由每个 $(x,\ x') \in X$ 的敏感属性 x 组成。# 学习预测成本 c^0在数据集 $(X_{sens},\ c^0)$ 上训练线性回归器 r_0：$\mathbb{R}^{d_{sens}} \to \mathbb{R}$ # 学习预测成本 c^1在数据集 $(X_{sens},\ c^1)$ 上训练线性回归器 r_1：$\mathbb{R}^{d_{sens}} \to \mathbb{R}$ # 如果估计值满足条件，则预测 0 $c_0 < c_1$定义 $g((x,\ x')) := \mathbf{1}\{(r_0 - r_1)\ (x) > 0\}$返回 g

G. NP完全性

我们将在下文证明，该问题的完全一般版本（允许任意多项式时间 F）是NP完全的。首先，我们将定义该问题的一个判定变体:

$$\delta_{X,F,h,A} = \max_{g\in\mathcal{G},f_j}(|F(f_j,g,h) - F(f_j,X,h)|) \geq A$$

请注意，原问题的一个解显然也能解决该判定变体。首先，我们将证明该判定变体属于NP，随后将通过归约到最大割问题来证明它是NP困难的。

引理G.1。 该问题的判定版本属于*NP*。

证明。我们的证据将是子集 g 和特征 f_j，使得

$$(|F(f_j,g,h) - F(f_j,X,h)|) \geq A$$

鉴于这两点，在 F 可在多项式时间内完成的前提下，绝对值的计算也可在多项式时间内完成，因此该解可以在多项式时间内得到验证。 □

引理 G.2. 该问题的判定版本是 *NP*困难。

证明。 我们将定义变量，将该问题归约为 maxcut(Q, k)。给定一个图，其中 V, E 分别是 Q 的顶点集和边集（边定义为顶点对），我们将如下定义 F、 X、 G、 A 和 h：

$$\begin{aligned} X &= V \\ h &= \text{constant classifier, maps every value to 1} \\ \mathcal{G} &= \mathcal{P}(V) \text{ i.e. all possible subsets of vertices} \\ F(f_j, g, h) &= |x \in E : x[0] \in g, x[1] \in g^c| \\ &\quad \text{–i.e. } F(j, g, h) \text{ returns the number of} \\ &\quad \text{edges cut by a particular subset, ignoring} \\ &\quad \text{its first and third argument.} \\ &\quad \text{(this is trivially computable in polynomial} \\ &\quad \text{time by iterating over the set of edges).} \\ A &= k \end{aligned}$$

注意，根据定义， $F(f_j, X, h) = 0$，且 $F \geq 0$。因此， $|F(f_j, g, h) - F(f_j, X, h)| = F(f_j, g, h)$，由此可知，($|F(f_j, g, h) - F(f_j, X, h)|$) $\geq A$ 当且仅当 g 是 Q 的一个子集，且能切割至少 $A = k$ 条边。因此，能够求解特征重要性问题判定变体的算法也能求解最大割。 □

H. 线性特征重要性差异

本文所考虑的不可分离特征重要性差异（FID）概念，对应于仅使用子群体 g 中的数据训练一个本身具有可解释性的模型，并将特征 j 的影响与使用整个数据集训练时的影响进行比较。由于子群体中的所有样本点都可能相互作用，共同产生可解释模型，因此这一概念通常是不可分离的。下文将在线性回归的情形下对其形式化；这是我们在实验中研究的不可分离概念。

定义 H.1。（线性特征重要性差异）。给定一个子群体 g，设 $\theta_g = \inf_{\theta \in \mathbb{R}^d} \mathbb{E}_{(X,y)\sim\mathcal{R}}[g(X)(\theta' X - y)^2]$, 和 $\theta_{\mathcal{R}} = \inf_{\theta \in \mathbb{R}^d} \mathbb{E}_{(X,y)\sim\mathcal{R}}[(\theta' X - y)^2]$。那么，如果 e_j 是 $\mathbb{R}^d$ 中的 j^{th} 基向量，我们将线性特征重要性差异（LIN-FID）定义为

$$\texttt{LIN-FID}(j, g) = |(\theta_g - \theta_{\mathcal{R}}) \cdot e_j|$$

LIN-FID(j, g) 定义为：在子群体 g 上训练模型时，特征 j 对应的系数，与使用来自 $\mathcal{R}$ 的样本点训练模型时该系数之间的差值。使用标准加权最小二乘估计量（WLS）展开定义 H.1，可得给定特征 f_j 和子群体 $g(X)$ 的特征重要性为：

$$F_{lin}(j, g) = \left((X g(X) X^T)^{-1} (X^T g(X) Y)\right) \cdot e_j, \tag{9}$$

其中， $g(X)$ 是子群函数输出的对角矩阵。数据集 X 上的线性回归模型系数可利用普通最小二乘（OLS）的结果计算得出：$(XX^T)^{-1}(X^T Y) \cdot e_j$。

我们通过求出 $F_{lin}(j, g)$ 的最小值和最大值，并选择差值较大的那个，来计算 $\text{argmax}_{g\in G}$LIN-FID $= \text{argmax}_{g\in G}|F_{lin}(j, X^n) - F_{lin}(j, g)|$。对于第 5 节中的实验，我们使用逻辑回归作为 g 的假设类，因为它具有足够的非线性，能够捕捉数据中的复杂关系，同时又能通过其系数保持可解释性；更重要的是，此时方程 9 关于 $g(X) = \sigma(X \cdot \theta)$ $\sigma(x) = \frac{1}{1+e^{-x}}$ 的参数 θ 可微。由于方程 9 关于 θ 可微，我们可以使用随机梯度下降（SGD）或 Adam 等非凸优化器，在 θ 上最大化方程 9。

尽管这一概念因其简洁而颇具吸引力，但只有当矩阵 $Xg(X)X^T$ 满秩时，它才有意义。我们首先通过大小惩罚项 $P_{size} = \max(\alpha_L - |g(X_{train})|, 0) + \max(|g(X_{train})| -$ 对 g 的大小进行下界约束，以确保这一点

$\alpha_U, 0)$，这样我们就能像可分离方法那样施加 α 约束。我们还在 $X^T g(X)X$ 中加入一个较小的正则化项 l_2 regularization term ϵI。这会使矩阵可逆，从而避免子群体过小带来的问题。纳入这些正则化项后，方程 9 变为:

$$F_{lin}(j,g) = \lambda_s \cdot \left((X\sigma(X \cdot \theta_L^T)X^T + \epsilon I)^{-1}(X^T\sigma(X \cdot \theta_L^T)Y) \cdot e_j\right) + \lambda_c \cdot P_{size} \tag{10}$$

我们注意到，LIN-FID 与 LIME（Ribeiro 等人，2016）是相似的概念，但 LIME 估计每个点周围的局部效应，然后将其求和以得到子群体中的效应，因此它是可分离的。此外， F_{lin} 也是非凸的，如下所示:

引理 H.2。 F_{lin} （如方程 *9* 所定义）是非凸的。

证明。 我们将通过反证法证明这一点。假设 F_{lin} 是凸的，这意味着海森矩阵处处是半正定的。首先，我们将固定 $(Xg(X)X^T))^{-1}$ 为单位矩阵；通过将 g 缩放一个常数即可做到这一点，且不失一般性。该缩放不会影响 F_{lin}的凸性。

现在，我们得到了 $F_{lin} = (X^T g(X)Y) \cdot e_j$的更简单形式。随后我们可以计算海森矩阵的值:

$$\frac{\partial F^2}{\partial^2 g} \quad = \quad (X^T g''(X)Y) \cdot e_j$$

考虑如下情形： X^T是一个 2×2 矩阵，其行分别为 1，0 和 0，-1，且 Y是一个全 1 向量。如果 g 对第二列（即特征）的赋权大于第一列，那么输出的海森矩阵将是半正定的。但如果 g 对第一列的赋权大于第一列，那么它将是半负定的。由于海森矩阵并非处处半正定，因此 F_{lin} 在 g的空间上必定是非凸的。 □

这意味着，通过梯度下降收敛到的驻点可能只是局部最优。在第 5 节中，我们使用 ADAM 优化器（Kingma 和 Ba，2015 ）优化方程 10。有关实现和参数选择的更多细节，请参见附录 I。尽管只能保证局部最优，我们仍能为所有数据集找到 LIN-FID 较高的（特征，子群体）对。

I. 实验细节

I.1. 算法细节

可分情形。为了在一系列 $[\alpha_L, \alpha_U]$ 值上运行算法 1，我们需要指定对偶范数 B、学习率 η、迭代次数 T、丰富子群类 $\mathcal{G}$以及相应的预言机 CSC$_{\mathcal{G}}$。请注意，对于每个特征 f_j，算法 1 都会运行两次：一次用于最大化 FID(f_j, g, h)，另一次用于最小化该值。注意，这两种情况下我们的问题都是最小化问题，但在最大化时，只需将所有逐点特征重要性值 $F(f_j, x_i, h) \to -F(f_j, x_i, h)$取反即可。在所有实验中，我们的子群类 $\mathcal{G}$均由敏感特征上的线性阈值函数构成： $\mathcal{G} = \{\theta \in \mathbb{R}^{d_{sens}} : \theta((x, x')) = \mathbf{1}\{\theta' x > 0\}$。我们按照 (Agarwal 等人，2018a；Kearns 等人，2018) 中的方法，通过线性回归实现 CSC$_{\mathcal{G}}$ ，详见附录 F 中的算法 2。为确保对偶玩家的响应足够有力，从而施加所需的规模约束，我们通过实证发现，在所有数据集上，将超参数 $B = 10^4 \cdot \mu(f_j)$ 设为该值效果良好，其中 $\mu(f_j)$ 是特征 j 在 X上的平均绝对重要性值。我们将指数梯度下降的学习率设为 $\eta = 10^{-5}$。实证测试表明，为确保适当收敛， $\eta \cdot B$ 应为 $\mu(f_j)$量级或更小。我们发现，将误差容限超参数 $\nu = .05 \cdot \mu(f_j) \cdot n \cdot \alpha_L$ 设为该值，有助于在所有数据集和 α值上取得良好结果，并保持较短的收敛时间。对于所有数据集和方法，我们最多运行 $T = 5000$ 次迭代；根据实证观察，这足以使特征重要性差异值稳定，并确保 $\frac{1}{T}\sum_{t=1}^{T} |g_t| \in [\alpha_L, \alpha_U]$，而该方法通常在不超过 $T = 3000$ 次迭代后收敛。收敛图示例见附录 N。

不可分情况。对于不可分离方法，我们再次将数据集划分为训练集和测试集。学生数据集按 50-50 划分，而 COMPAS、银行和 Folktables 数据集按 80-20 划分为训练集和测试集。选择将学生数据集按 50-50 划分，是为了能在较小的 $g(X_{test})$ 上恰当地拟合线性回归模型。逻辑回归分类器的参数向量 θ 使用 PyTorch 随机种子 0 随机初始化，以确保结果可复现。我们使用 ADAM（Kingma & Ba，2015）优化器，学习率设为.05，作为损失函数的启发式求解器。

为了满足子群体规模约束，$\lambda_s P_{size}$ 必须显著大于 $\lambda_c F_{lin}(j,\ g)$。实证测试发现，取值为 $\lambda_s = 10^5$ 和 $\lambda_c = 10^{-1}$ 时，既能得到合适的子群体规模，也能确保平稳收敛。优化器持续运行，直至收敛到最小化的线性回归系数，同时满足规模约束。实验中，迭代次数最多为 1000 次；收敛图示例见附录 O。在分别求得最小值和最大值 $F_{lin}(j,\ g)$ 两次，并选定子群函数 g 后，我们在 X_{test} 和 $g(X_{test})$ 上拟合线性回归，以得到最终的特征重要性差异。

I.2. 特征重要性差异概念

LIME：在数据集 X^n 上训练了一个随机森林模型 h。随后，将每个数据点及分类器输出的相应概率输入 LIME 表格解释器 Python 模块。该模块返回相应的 LIME 解释值。

SHAP：这与 LIME 使用相同的方法，只是改用 SHAP 解释器 Python 模块。 .

原始梯度：在图表中标记为*GRAD*，原始梯度重要性度量是使用 OpenXAI 库中的梯度方法计算的（Agarwal 等，2022b）。该度量仅适用于可微分类器，因此本例中，h 是一个逻辑回归分类器。我们发现，在使用其他重要性概念进行测试时，随机森林或逻辑回归的选择对 h 没有实质性影响（附录 L）。由于计算时间受限，此方法仅在 COMPAS 数据集上进行了测试（以两年内再犯作为目标变量）。

线性回归：对于线性回归这一度量，选择子群体 g 使其属于逻辑回归假设类。对于给定的子群体 g（X），求得加权最小二乘法（WLS）解，其线性系数 θ_g 随后定义特征重要性值 $e_j \cdot \theta_g$。

有关这些重要性度量一致性的详细信息，请参阅附录 P。

I.3. 数据集

选择这四个数据集依据的是三项标准：(i) 它们都使用了可被视为敏感的特征，用于在偏差构成重大关切的情境下对个体进行预测；(ii) 它们都是可解释性和公平性研究中广泛使用的数据集，因此，数据集中的偏差问题应受到学界重视；(iii) 它们涵盖了不同数量的数据点、特征和敏感特征，相关汇总见表 4。对于每个数据集，我们都指定了哪些特征属于“敏感”特征。也就是说，在搜索特征重要性差异（FID）较高的子群体时，我们只考虑由通常受平等保护或隐私法保护的特征（例如种族、性别、年龄、健康数据）定义的丰富子群体。

学生：该数据集旨在利用人口统计和家庭数据，预测葡萄牙小学学生的表现。在本实验中，目标变量为学年末的数学成绩。四个数据集中，Student 的规模最小，只有 395 个数据点。Student 中的敏感特征包括性别、父母状况、居住地址（城市或农村）、每日饮酒量、每周饮酒量和健康状况。年龄通常会被视为敏感特征，但在学校的情境中，年龄主要反映年级，因此未将其列为敏感特征。分类特征居住地址、母亲职业、父亲职业和法定监护人均采用独热编码。

COMPAS：该数据集利用审前被告的个人背景和过往犯罪记录，预测其再次犯罪的风险。为提高泛化能力，我们删除了出现次数少于 10 次的刑事指控特征。我们舍弃了二元计数特征（例如，25–45 岁或 5+ 项轻罪），改用对应的连续特征。此外，分类变量“种族”采用了独热编码。这使特征总数达到 95 个。COMPAS 中的敏感特征包括年龄、性别和种族（高加索人、非裔美国人、亚裔、西班牙裔、美洲原住民和其他）。对于 COMPAS，我们对所有方法分别运行了两次：一次将二元变量“两年内再犯”作为目标变量，另一次将连续变量“十分位数评分”作为目标变量。“两年内再犯”是模型旨在预测的内容，在结果中标记为 *COMPAS R*。与此同时，“十分位数评分”是 COMPAS系统在实际应用中用于向法官提出建议的指标，在结果中标记为*COMPAS D*。

银行：此数据集考察潜在客户在营销人员联系后是否开设了银行账户。银行中的敏感特征是年龄和婚姻状况（已婚、单身或离异）。年龄特征

在银行中是一个二元变量，用于表示该个人是否年满 25 岁。

Folktables：该数据集源自美国人口普查数据。Folktables 涵盖多种任务，但我们使用了 ACSIncome 任务，该任务用于预测个人年收入是否超过 5 万美元。ACSIncome 任务旨在仿照广受欢迎的 Adult 数据集，但做了修改以解决抽样问题。本文使用的是 2018年 密歇根州的数据。为减少数据集的稀疏性，我们删除了出生地特征，并按照官方人口普查词典（Bureau，2020年）中的规定，将职业特征合并为工作类别（例如，在美国陆军、美国空军、美国海军等部门工作的人都被合并到职业=军事类别中）。Folktables 中的敏感特征包括年龄、性别、婚姻状况（已婚、丧偶、离异、分居、从未结婚/未满 15 岁）和种族（高加索人、非裔美国人、亚裔、夏威夷原住民、美洲原住民单一部族、美洲原住民泛称、其他以及 2+种族）。

表 4：数据集摘要

数据集	数据点	特征数量	敏感特征数量
学生	395	32	6
COMPAS	6172	95	8
Bank	30488	57	4
Folktables 收入	50008	52	16

J. 合成实验

除了针对真实世界数据集的实证实验外，我们还生成了两个合成数据集，并在受控环境中用它们验证我们的方法。在基线实验中，我们创建了一个结果变量y 与敏感特征相互独立的数据集，以确认我们的算法不会导致任何假发现。接着，我们修改了部分个体的结果变量分布，在子群体 g 中针对特征 f_j注入了较大的特征重要性差异。随后，我们确认算法能够发现这种特征重要性差异。本文将在此讨论这两个实验的数据集生成、实验设置和结果。

J.1. 基线情形

实验设置： 我们生成了一个规模为 $n = 4000$ 的合成数据集。数据集中的每个人都具有随机生成的敏感特征：年龄、性别和种族。性别和种族根据美国人口普查数据抽取，年龄 $\sim \mathcal{N}$（50， 7）。我们为每个个体另外生成了三个变量：一个二元变量 x_1 以及正态分布变量 $x_2 \sim \mathcal{N}$（100， 5）和 $x_3 \sim \mathcal{N}$（100， 5）。这三个附加变量均独立于敏感特征及彼此生成。随后，我们生成结果变量 $y \sim x_1 + x_2 + x_3 + \eta$， $\eta \sim \mathcal{N}$（0， 1）；请注意，对于任何敏感群体， y均由同一个模型生成，因此不应存在 AVG-FID 较大的子群体。接着，我们在该数据集上训练了一个随机森林模型，并使用 SHAP 计算特征重要性值。

结果： 我们在表5中总结了 AVG-FID 最大的 (f_j, g) 对。正如预期，算法 1 在此基线情形下未发现任何显著的 AVG-FID。AVG-FID 并非恰好为零，这是意料之中的，因为 AVG-FID 是以绝对值衡量的，这意味着随机波动导致的任何特征重要性差异都会产生非零值。

J.2. 注入情形

实验设置： 在第二个实验中，我们以基线情形中生成的相同个体为起点，为由年长的西班牙裔个体组成的子群体 g 注入了特征重要性差异（FID），该子群体约占总体的 13%。对于这个数据集，如果某个个体属于 g，那么我们生成 $y \sim x_1 + 50 * x_2 + x_3 + \eta$， $\eta \sim \mathcal{N}$(0， 1)。否则， y的生成方式与基线情形相同。与基线情形一样，我们使用随机森林模型，并将 SHAP 作为特征重要性度量。

结果： 如表5所示，发现的特征重要性差异子群中，特征重要性差异最高的三个特征是西班牙裔、年龄和 x_2。发现 x_2 是预期之中的，但在我们的合成示例中， y 对于 g 中的样本点依赖于这两个特征，而对于 X/g则完全独立，因此西班牙裔和年龄也被发现并不罕见。发现的第四大 AVG-FID

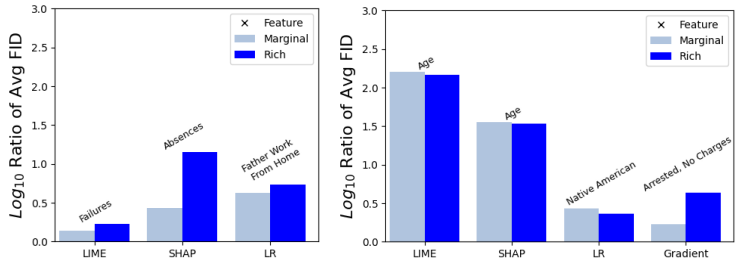
表5：两个合成数据集实验中发现的排名靠前的（f_j, g）对的汇总。 μ（F）是指定群体上的平均特征重要性值。可以看到，在基线实验中，AVG-FID 很小。在注入的情况下，算法 1 在受注入影响的三个特征上发现了非常大的 AVG-FID 子群。注入情况下排名第二的对，其 AVG-FID 与基线情况相当。

实验	特征 f_j	$\lvert g\rvert$	$\mu(F(f_j,\ X))$	$\mu(F(f_j, g))$	AVG-FID
基线	x_3	.13	$-.61$	$-.48$	**.13**
	x_2	.11	-1.16	-1.23	**.06**
	x_1	.12	$-.05$	$-.01$	**.04**
注射的	西班牙裔美国人	.15	.86	28.0	**27.13**
	Age	.17	3.57	-7.07	**10.6**
	x_2	.15	-1.56	.22	**1.78**
	非裔美国人	.15	.01	.04	**.03**

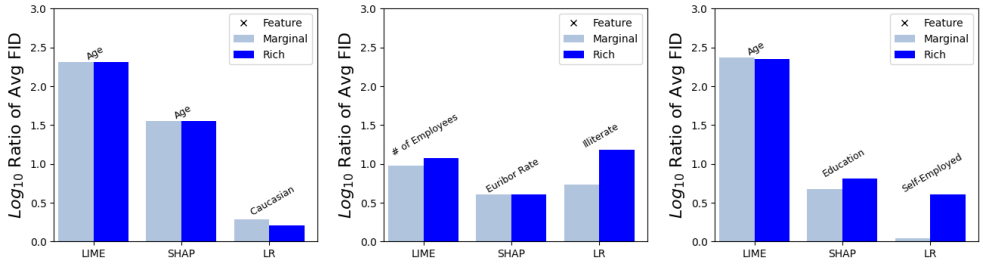
明显小于排名前三的子群体，且其幅度与基线情况相当。算法 1 针对排名最高的特征所找到的子群体，涵盖了发生差异注射的较年长西班牙裔子群体的 64%。这并非完美结果，但无需大量调优即可获得，并说明我们的方法能够在受控环境中检测出注射的差异。

K. 丰富子群与边际子群的FID值比较

本附录扩展介绍了第 5.4 节的内容。我们在此通过在边际子群体空间中寻找 AVG-FID 最大的子群体，说明采用丰富子群体的合理性。边际子群体由单个敏感特征定义，因此易于搜索。在图6中，我们将图2所示的 AVG-FID 最大的丰富子群体与具有相同特征的 AVG-FID 最大的边际子群体进行比较。约有一半情况下，将子群体类别扩展为包含由敏感属性的线性函数定义的丰富子群体，使我们能够找到 AVG-FID 更高的子群体。在其他情况下，边际子群体的 AVG-FID 与丰富子群体相近。有时，边际子群体的表现优于丰富子群体；这是因为采用丰富子群体并未比采用边际子群体带来显著优势，而且算法 1 中的经验误差容限使收敛提前停止。



(a) 学生 (b) COMPAS R



(c) COMPAS D (d) 银行 (e) Folktables

图6：将图2中的最大FID丰富子群体与同一特征上的最大FID边际子群体进行比较。结果显示为 $|log_{10}(R)|$，其中，R 对于可分离概念而言是每个数据点的平均重要性比值，对于线性系数概念而言则是系数比值。与子群体相关的特征标注在每个柱形图上方。

L. 结果的统计有效性：FID 和 $|g|$ 的泛化

在确认研究结果的有效性时，我们面临两个潜在问题：（1）在样本内找到的子群体规模在测试集上是否大致相同？（2）在训练集上找到的特征重要性差异能否推广到样本外？综合来看，（1）和（2）足以保证我们的最大 AVG-FID 值能够推广到样本外。

在图 7 中可以看到，当我们针对每个特征 f_j、g_j^* 取所发现的最大子群体，并在测试集上计算其大小 $|g_j^*|$ 时，对于可分方法和不可分方法，其结果几乎总是落在指定的 $[\alpha_L, \alpha_U]$ 范围内；在所有特征重要性概念和所有数据集上，$|g_j^*(X_{train})|$ 与 $|g_j^*(X_{test})|$ 的平均差异均小于 .005，只有 Student 例外，其差异更接近 .025，这是由于其规模较小。少数罕见子群体显著超出了期望的 α 范围，这通常是因为出现了退化情形，即相关特征的特征重要性值全部为 0。所有（数据集、概念）组合的其他图表见图 9、10。

在图 8 中，我们比较 AVG-FID(f_j，g_j^*，X_{train}) 与 AVG-FID(f_j，g_j^*，X_{test})；对于线性回归度量，则比较 LIN-FID，以观察特征重要性差异如何泛化。所有可分离概念的泛化效果都很好，在样本内和样本外测试中得到的 AVG-FID 值非常相似。不可分离的方法也能泛化，但稳健性远不如前者，并出现了异常值。这是因为较小子群体的设计矩阵病态，导致最小二乘估计量拟合不稳定。在附录 P 中，我们研究了在整个数据集上评估的特征重要性概念的稳健性。我们发现，线性回归的系数不够稳定，这表明图 8 中缺乏泛化能力，可能是因为特征重要性度量本身缺乏稳健性，而不是对 g_j^* 进行了过拟合选择。

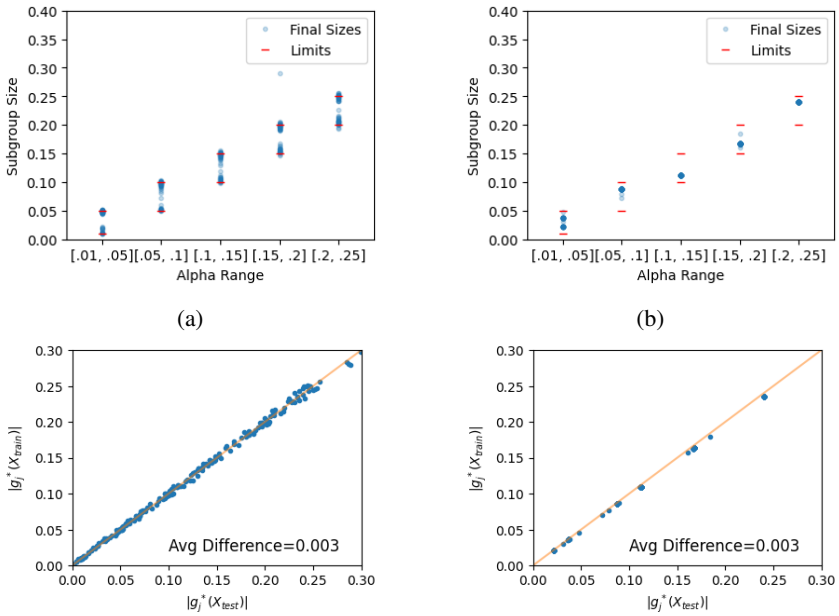


图 7: $|g|$ 在 Folktables 数据集上的泛化能力。(a) 算法 1 对所有特征和可分离概念输出的规模，以及 (b) 优化方程 10 得到的 LIN-FID 结果，表明我们的规模约束在样本内成立。(c) 绘制可分离概念对应的 $|g_j^*(X_{train})|$ 与 $|g_j^*(X_{test})|$ 的值，以及 (d) LIN-FID 的对应值，表明子群规模在样本外也具有泛化能力。

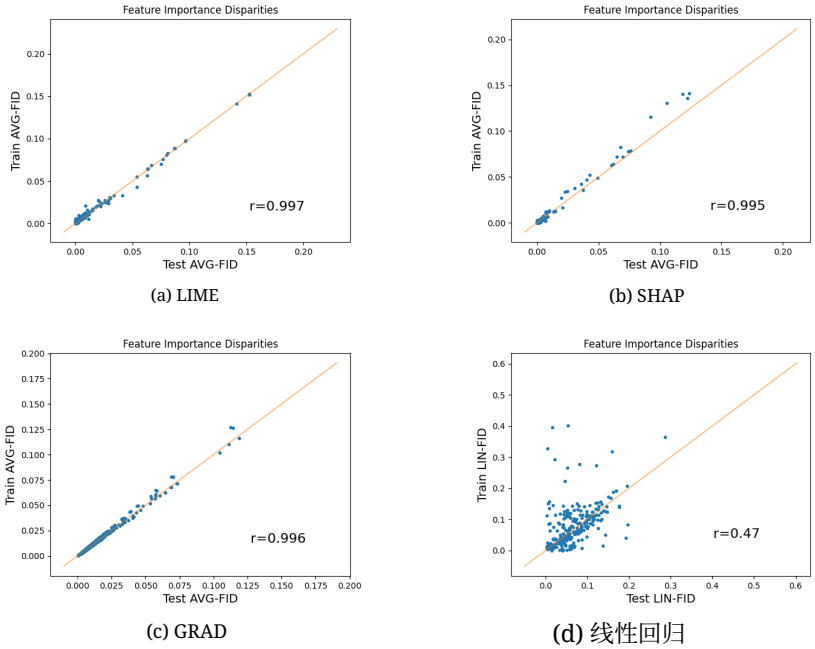


图 8: 各方法的样本外泛化。每个点代表一个特征，绘制的是在 X_{test} 与 X_{train} 上的特征重要性差异。除 (c) 使用 COMPAS R 计算外，其余均使用 Folktables 数据集计算。对角线表示完美泛化，图中还显示了皮尔逊相关系数。不可分离的方法受到加权最小二乘法不稳定性的影响。

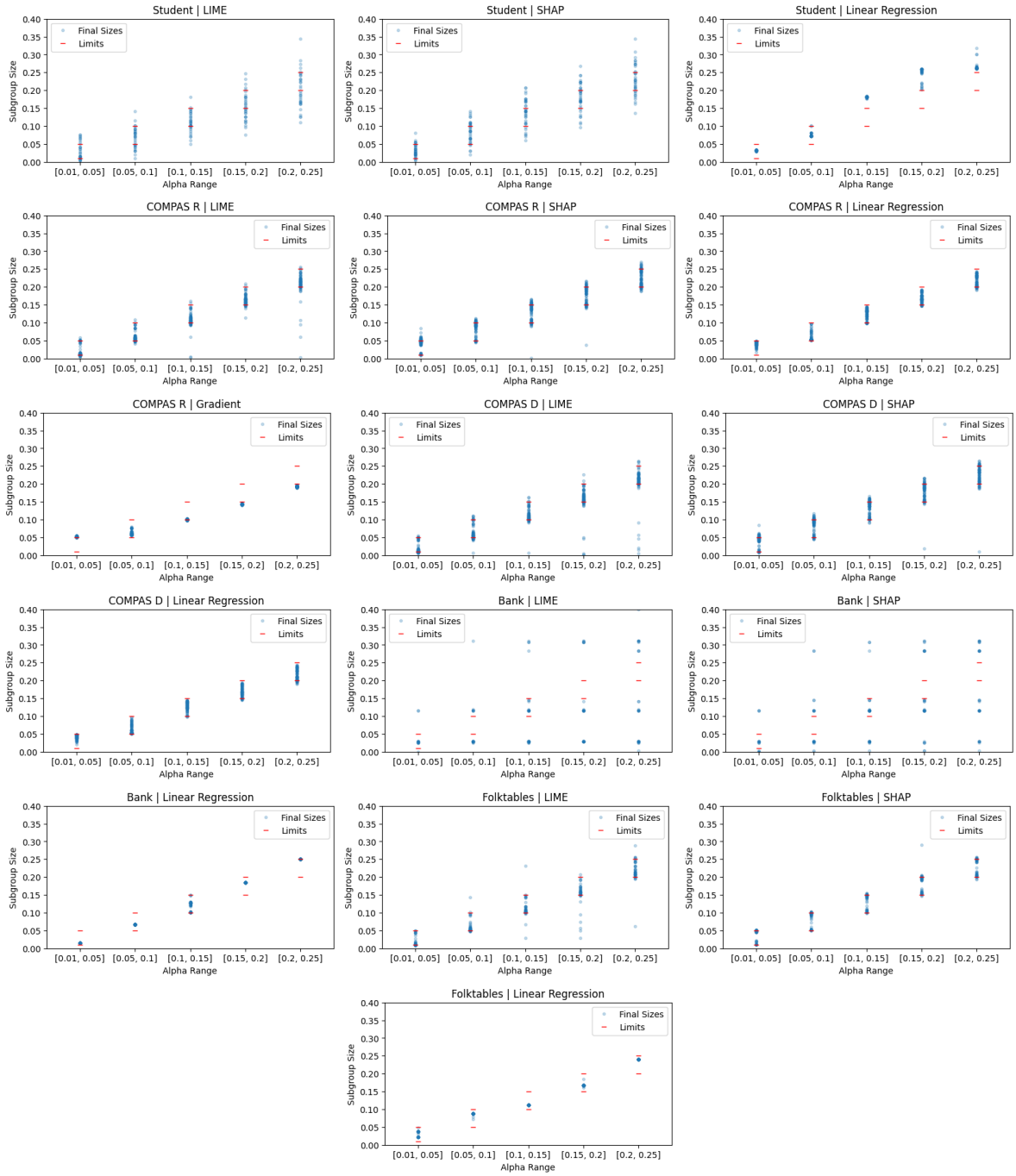
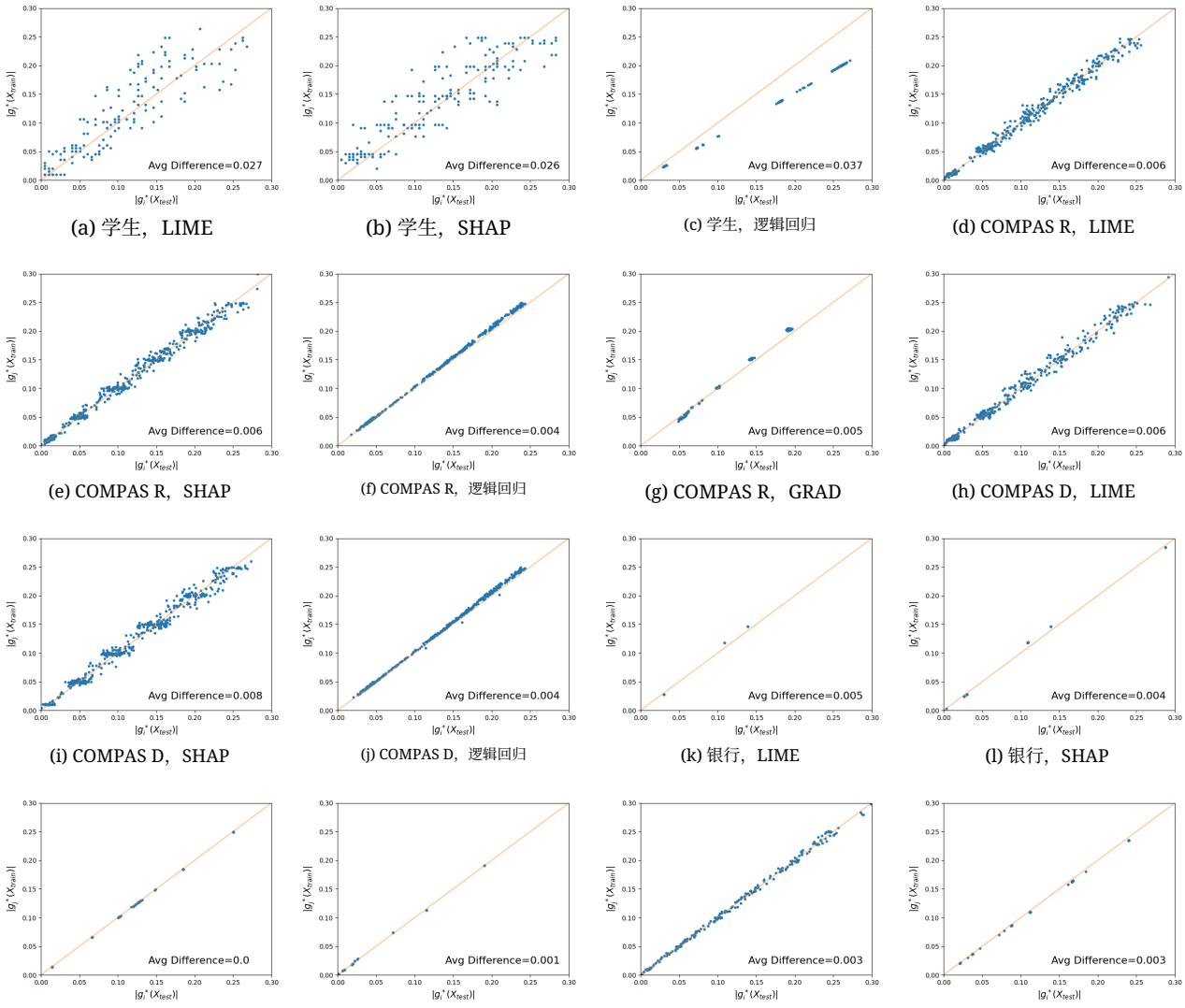


图9： $g(X_{test})$的最终子群体规模与 α 范围的比较。这些规模几乎总是落在正确的范围内。由于数据集规模较小，Student 的误差最大。一些子群体明显超出了预期范围，这主要是因为对于某个特征，许多重要性值 $F(f_j,\ X,\ h)$为零。



(a) 学生，LIME (b) 学生，SHAP (c) 学生，逻辑回归 (d) COMPAS R，LIME

(e) COMPAS R，SHAP (f) COMPAS R，逻辑回归 (g) COMPAS R，GRAD (h) COMPAS D，LIME

(i) COMPAS D，SHAP (j) COMPAS D，逻辑回归 (k) 银行，LIME (l) 银行，SHAP

(m) 银行，逻辑回归 (n) Folk，LIME (o) Folk，SHAP (p) Folk，逻辑回归

图10：比较 $|g_j^*(X_{train})|$ 和 $|g_j^*(X_{test})|$。我们可以看到，训练集和测试集中子群体的规模保持一致，表明模型具有良好的泛化能力。由于 Student数据集规模较小，只有在该数据集中平均差异才较为明显。

表 6：比较以随机森林和逻辑回归作为分类器 h 的假设类，并在 COMPAS R 数据集上使用 LIME 作为重要性度量时的结果。此处展示 AVG-FID 最高的特征、子群规模 $|g|$以及 AVG-FID。可以看出，假设类的选择 h 不会对结果产生实质性影响。我们的所有主要实验均使用了随机森林。

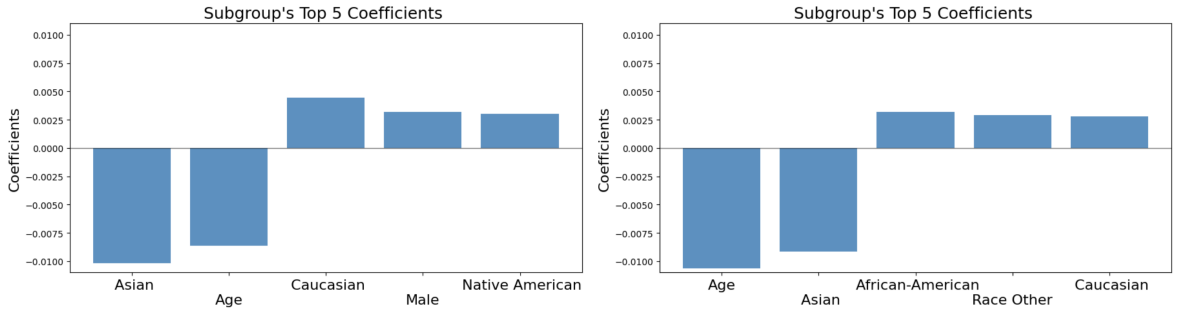
h = 随机森林			h = 逻辑回归		
特征	Size	AVG-FID	特征	Size	AVG-FID
Age	$.05-.1$	.144	Age	$.05-.1$	.21
先验计数	$.01-.05$	.089	先验计数	$.01-.05$	.092
少年其他罪行数	$.01-.05$	.055	少年其他罪行数	$.01-.05$	.055
其他特征	-	$<.025$	其他特征	-	$<.025$

h =随机森林			h =逻辑回归		
特征	Size	AVG-FID	特征	Size	AVG-FID
Age	$.01-.05$	.4	Age	$.01-.05$	.21
先验计数	$.01-.05$	.11	先验计数	$.01-.05$	.14
其他特征	-	$<.05$	其他特征	-	$<.05$

表 7：与表 6 相同，但使用 SHAP 作为重要性度量。使用 SHAP 时，在显著性下降之前，具有显著 AVG-FID 的特征较少；但在这两种情况下，h的选择都没有显著影响结果变量。

M. 假设类的选择

我们探讨的一项消融研究是分类模型的选择 h。主要实验使用了随机森林模型，我们也探索了使用逻辑回归模型。逻辑回归模型采用 sklearn 的默认超参数实现。我们发现，无论选择哪种 h，结果大致一致。在表 6 和表 7 中可以看到，不同假设类下，AVG-FID 最高的特征、其子群体规模以及 AVG-FID 值均保持一致。随后，我们进一步考察了用于定义这些子群体的特征。在图 11 中可以看到，特征“年龄”的 AVG-FID 较高的子群体都由年轻的非亚洲人构成。在所有特征重要性概念和数据集中，我们都发现了类似的一致结果。



(a) h = 随机森林， f_{j^*} = 年龄　　(b) h = 逻辑回归， f_{j^*} = 年龄

图 11：比较 h 的假设类选择。此处展示了在 COMPAS R 数据集上使用 LIME 作为特征重要性度量时，发现的 AVG-FID 最高的子群体所对应的定义性系数。对于年龄这一特征，无论选择哪种 h，年轻且非亚裔都是 g^* 最具定义性的两个系数。

我们承认，这项消融研究并非最全面的研究。不过，鉴于这些初步结果显示，假设类的选择之间没有显著差异，我们决定在主要实验中继续使用随机森林作为假设类来呈现所有结果。这无疑是未来研究可以进一步探索的方向。

N. 算法 1 优化收敛

以下是更多图表，展示算法 1 收敛过程的示例。我们每隔 10 次迭代记录一次数据，分别在图 12 中绘制拉格朗日值（用于计算误差 $v_t = max(|L(\hat{p}_{\mathcal{G}}^t, \hat{p}_{\lambda}^t) - L|, |L - L(\hat{p}_{\mathcal{G}}^t, \hat{p}_{\lambda}^t)|)$）、子群规模和 AVG-FID 值。我们可以看到，AVG-FID 值不断上升，除非子群规模超出 α 范围；拉格朗日误差则逐渐收敛至设定的误差界 v，随后算法终止。

虽然定理 1 指出收敛时间可能呈二次增长，但在实践中，我们发现计算时间并不是一个重要问题。收敛时间会根据数据集略有变化，但大多数情况下，给定特征的收敛只需几次迭代，计算耗时几秒。对于需要数千次迭代的特征，在较大的数据集上计算可能需要约 30 分钟。

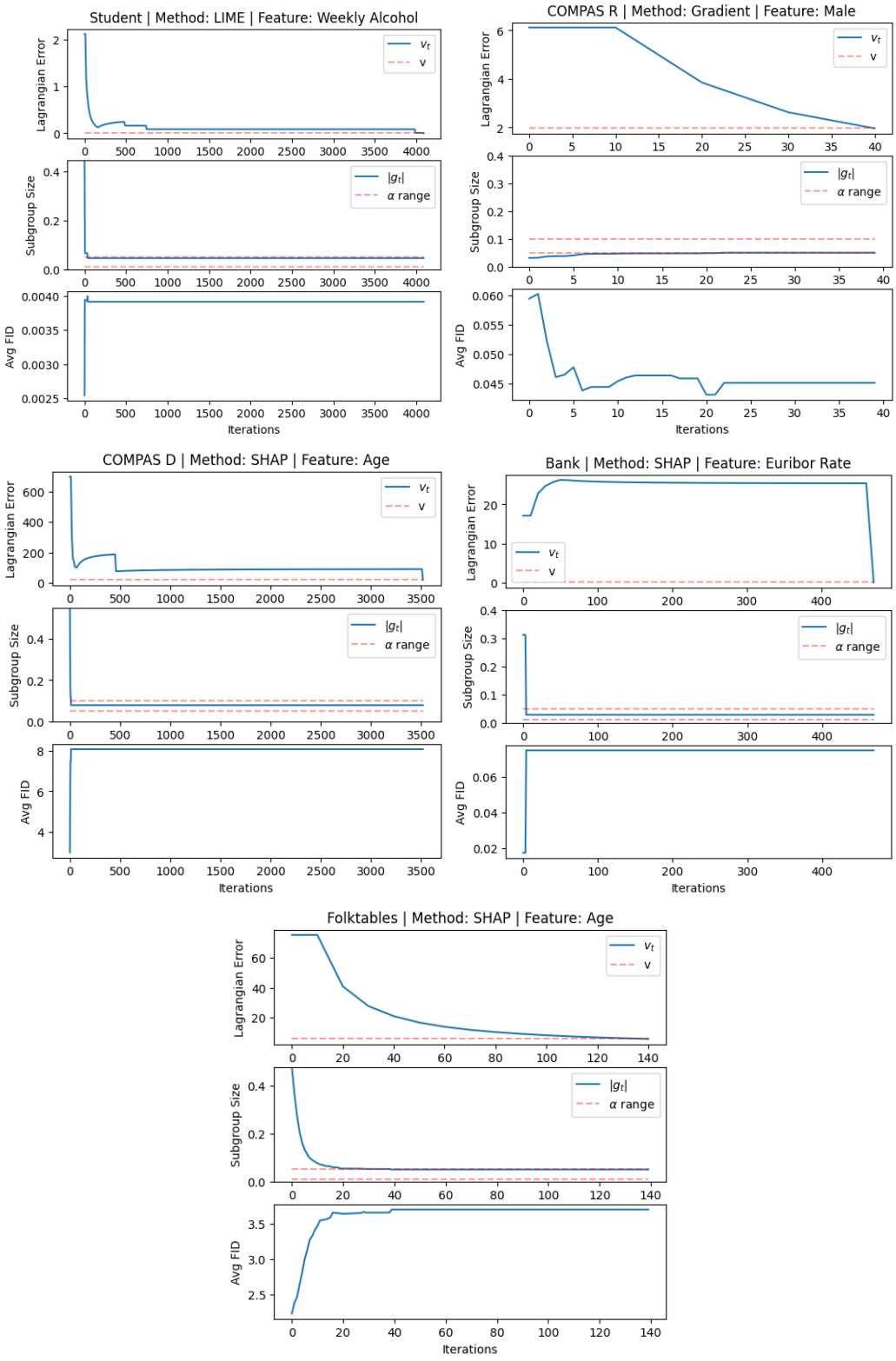


图 12：详细展示算法 1 收敛情况的图表。顶部图表展示误差收敛情况，即我们的解与最小值/最大值参与者解之间拉格朗日值的最大差异。其他两幅图表展示该解的子群规模和 AVG-FID。收敛几乎总是在少于 5000 次迭代内发生，缓解了人们对理论运行时间的担忧。

O. 非可分优化收敛

以下是展示不可分离方法中收敛情况的其他图表。使用能够奖励最小化线性回归系数（或最大化该系数），并使规模处于α约束范围内的损失函数时，我们通常会在几百次迭代后达到收敛。在图13中，我们可以看到，在相应的上方图表中，子群规模收敛至指定的 α 范围并保持在该范围内。同时，在下方图表中，我们看到 LIN-FID 尝试进行最大化，但随着适当规模的确定而发生振荡。

使用这种方法几乎总是在 < 1000 次迭代后收敛。在最大的数据集上，对所有特征运行该方法的计算时间约为 2 小时。优化过程使用 NVIDIA Tesla V100 上的 GPU计算 完成。

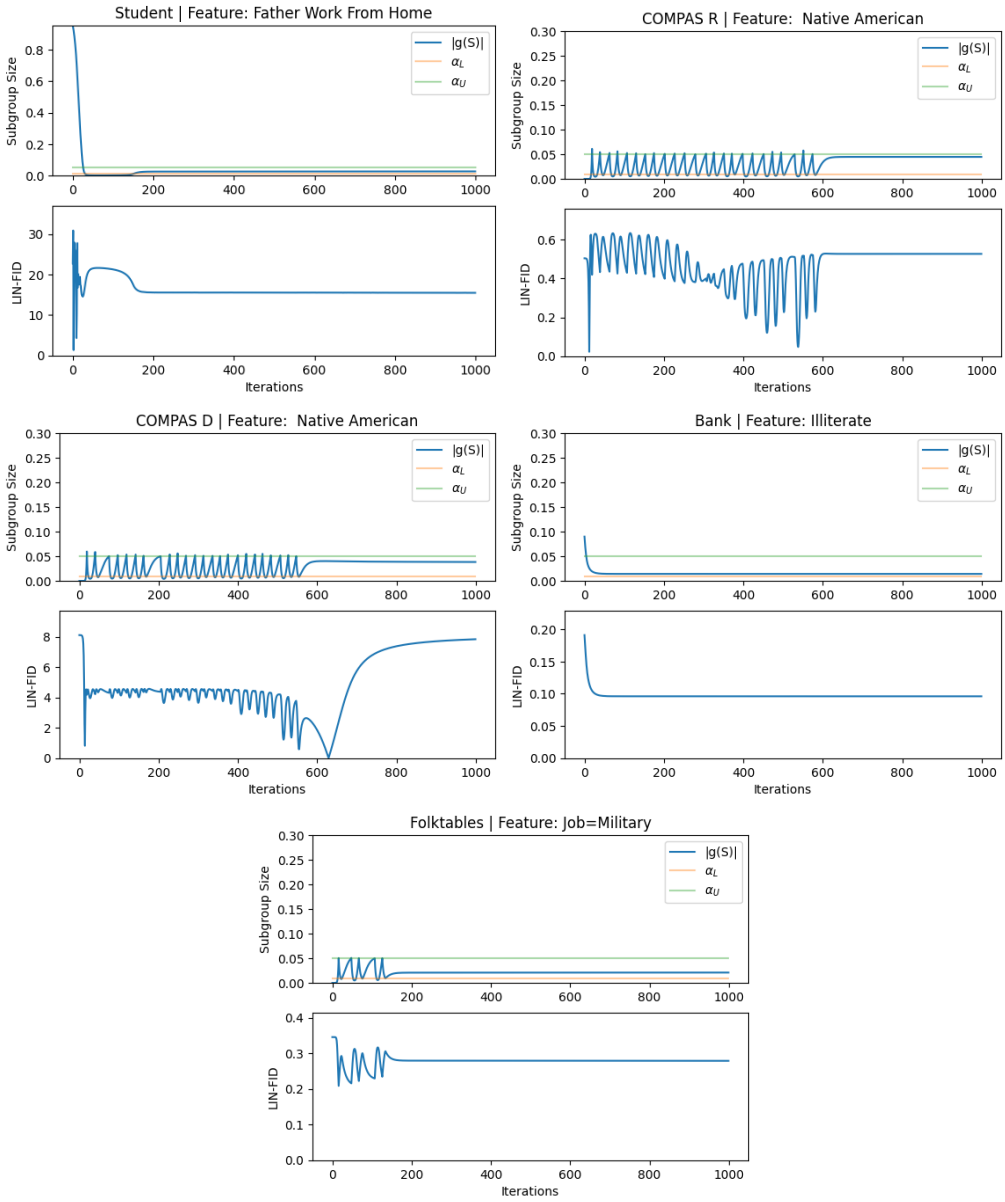


图13： g 在 Adam 优化器训练迭代过程中的子群规模和线性回归系数图。对于每个数据集，均展示了 LIN-FID 最高的特征。

P. 重要性概念一致性

为了了解不同重要性度量方法的一致性，我们绘制了 $F(f_j,\ X_{test},\ h)$ 与 $F(f_j,\ X_{train},\ h)$ 的值，其中每个点代表 COMPAS 数据集的一个特征 f_j。这些点越贴近对角线，该方法给出的重要性值就越一致。如图 14 所示，LIME 和 GRAD 的一致性极高。线性回归的一致性较差，这是因为在病态设计矩阵上拟合最小二乘估计量时存在不稳定性。SHAP 的特征重要性归因也不一致，不过，如图 8 所示，AVG-FID 仍然具有良好的泛化能力。这可能意味着，尽管 SHAP 在不同数据集之间不一致，但相对于自身而言是一致的。也就是说，如果 $F(j,\ X_{train}) > F(j,\ X_{test})$，那么 $F(j,\ g(X_{train})) > F(f,\ X_{test})$，这意味着 AVG-FID 值将保持不变。

正如其他研究所指出的，这些不一致似乎是某些可解释性方法固有的（Krishna 等，2022；Dai 等，2022；Agarwal 等，2022a；Alvarez-Melis 与 Jaakkola，2018；Bansal 等，2020）。探索这些泛化属性将是本研究未来一个令人振奋的方向。

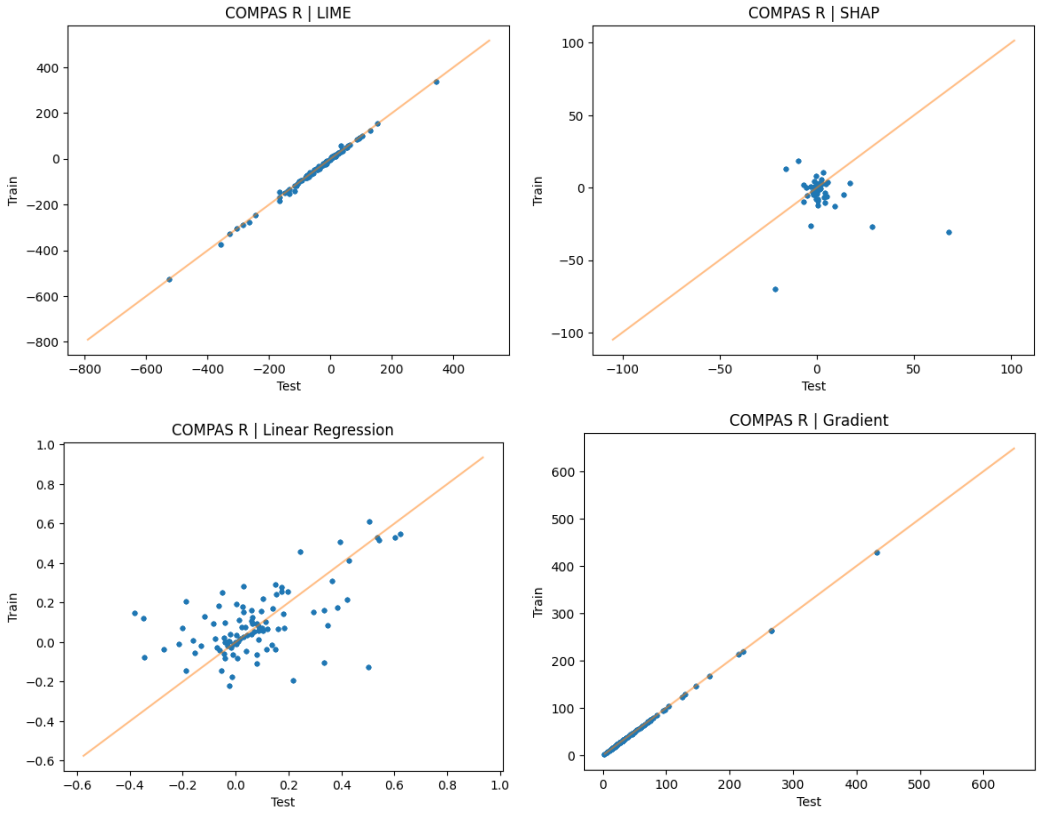


图 14：重要性概念的一致性。每个点代表一个特征，x 值为 $F(j,\ X_{test})$，y 值为$F(j,\ X_{train})$。点越接近对角线，该概念就越一致。

Q. 伦理审查

本研究过程中没有出现实质性的伦理问题。所使用的数据集均为公开数据、已去标识化，并且常用于公平机器学习研究。本研究没有任何部分旨在重新识别数据，或将其用于测试我们的方法以外的任何用途。