基于样本量分析得出的分类偏差度量

Ioannis Ivrissimtzis[a]、Shauna Concannon[a]、Matthew Houliston[b]、Graham Roberts[b]

[a]杜伦大学，英国[b]
*Servelegal*有限公司，英国

arXiv:2601.03453v1 [stat.ME] 6 Jan 2026

摘要

我们提出一种简单直观的原则，用于度量算法分类偏差：分类器在不同人口群体中的错误率差异，其显著性与统计检验检测出这些差异所需的样本量成反比。也就是说，如果需要较大的样本量才能在统计上确认存在偏差行为，那么该算法的偏差较小，反之亦然。在一个简单的情境中，我们假设存在两个不同的人口群体，并分别以 ε_1和 ε_2表示这两个群体错误率的非参数估计。我们采用一个广为人知的卡方检验样本量近似公式，并验证所提度量的一些基本理想性质。接下来，我们将所提度量与另外两种常用统计量进行比较：错误率之差 $\varepsilon_2 - \varepsilon_1$ 和错误率之比 $\varepsilon_2/\varepsilon_1$ 。我们证明，所提度量本质上有所不同，因为它对算法偏差的排序可能不同于其他度量；并讨论它相较于另外两种度量的一些优势。最后，我们简要讨论所提度量的一些理想性质如何源自该方法的基本特征，而非我们采用的样本量近似公式，因此预期这些性质在包含两个以上人口群体的更复杂情境中也成立。

1. 引言

我们研究在划分为若干类别（称为人口群体）的数据集上进行测试的分类器的行为。在典型应用中，测试数据集中的元素对应于人，而类别则对应于种族、性别或年龄等特征。

我们假设已获得分类器在各人口群体上的错误率的非参数点估计。通常，这些估计是通过在测试集上运行算法，并统计各人口群体中分类失败的比率而得到的经验估计。我们的目标是提出并研究一种度量，用于量化分类器的偏差，也就是衡量算法在不同人口群体上的性能差异是否显著。

这一核心思想直观、简单且具有普适性，可表述为以下原则：

> 算法的偏差与某个统计检验在给定检验效能下，以特定置信度检测出该偏差所需的样本量N成反比。

例如，若某个数据集包含两个人口群体D_1和D_2，，算法在这两个群体上的错误率分别为 ε_1 和 ε_2，要回答 ε_1 与 ε_2 之间的差异有多显著，我们需要考察从该数据集中抽取多少样本才能通过统计检验检测出这一差异。所需样本越多，分类器的偏差就越小。

请注意，在此我们并不评估算法在各个人口群体上的错误率点估计值的统计显著性。我们将这些估计值视为已知，并希望据此得出对算法偏差的评估。尽管这或许看起来是一种局限，因为这只是转移问题，而非解决问题，但我们仍认为，这是建立连贯偏差度量的一项基本属性。它将文献中常常混为一谈的两个问题区分开来：我们测量什么，以及测量的过程。事实上，如果我们将偏差定义为错误率 ε_1 和 ε_2 不同的概率度量，那么随着我们对算法进行更多测试，测得的偏差很可能会增大。然而，测得偏差的这种增加并不意味着算法的行为发生了变化，而是意味着我们对其行为的不确定性降低了。

1.1. 论文结构

第2节讨论背景和相关工作。第3节将所提出的一般原则应用于由两个人口群体组成的数据集中的分类器这一具体实例。我们采用卡方检验来对性能差异进行统计检测，并使用一种成熟的近似闭式公式估算样本量。我们表明，样本量与检验效能和置信度之间仅相差一个乘法常数，并确立了偏差度量所应具备的其他一些理想性质。

在第4节中，我们将拟议偏差度量的表现与两种广泛使用的度量指标进行比较：误差比之差 $\varepsilon_2 - \varepsilon_1$ 和误差比之比 $\varepsilon_2/\varepsilon_2$ 。首先，我们研究当错误率之差或比率保持不变时，拟议指标的表现。随后，我们通过人脸活体检测这一机器学习分类问题的实际数据来说明该方法。

在第5节中，我们进一步论证我们的方法：更一般地说，对于任意数量的人口群体，任何具备某些基本理想属性的统计检验，都会产生具有某些相应理想属性的偏差度量。例如，当所有人口群体的错误率趋近于零时，偏差也会趋近于零。

在第6节中，我们将讨论一些可作为未来研究方向的替代方案，以替代本文采用的标准样本量分析方法，并在第7节简要总结。

1.2. 贡献与局限

本文的主要贡献是提出了一种基于样本量分析的新型算法偏差度量。尽管下文将讨论一项局限，但拟议指标具备 [1] 提出并在[2]中进一步阐述的良好度量的特征。它能够连贯地捕捉所关注的性能方面，并清晰地区分算法偏差与其测量中的不确定性。它可以用单个数字简单地报告，直观易懂，而且至少在涉及两个人口群体的简单情形下，计算上可行。

作为次要贡献，在第4.2节中，我们通过将拟议指标与错误率比值进行比较，着重指出了算法偏差的一个方面；我们认为，文献中在很大程度上忽视了这一点。随着算法不断演进，并且我们逐步迈向对误分类零容忍的目标，度量在零错误率附近的表现变得十分重要，值得深入研究。具体而言，在错误率空间中，当我们从不同方向趋近总体错误率为零这一点时，偏差度量应当相对稳定。

拟议偏差度量的主要局限在于它依赖两个连续变量——统计检验的检验效能和显著性——以及一个分类变量，即统计检验的类型。尽管文献中普遍认为这些连续变量存在一些标准取值，例如检验效能为 0.80、置信度为 0.95，但为了采用更透明的方法，我们不固定这些取值，而是仅将拟议指标用于对算法的偏差进行排名，而非作为绝对度量。选择卡方检验以外的统计检验会产生何种影响，是一个更复杂的问题，超出了本文的研究范围。

第二个局限是，我们没有提出一个明确的偏差函数，将偏差表示为与样本量 N 成反比且可相互比较的量。尽管有一些显而易见的候选形式，例如 $1/N$ 或 $-\log N$，但由于缺乏明确的理由来选择其中一种，为了采用更透明的方法，我们倾向于直接使用 N 作为公平性度量，认为它与偏差成反比且可相互比较。不过，我们注意到，N 不仅直观，而且具有明确的实际意义：它告诉我们，在偏差达到统计可检测的程度之前，一个算法可以使用多少次。

2. 背景

对算法偏差的研究已发展成为一个本质上跨学科的领域，吸引了来自不同领域的研究人员，例如计算机科学、数学、社会科学以及法律领域。

[3] 是一篇颇具影响力的论文，塑造了人们从法律角度看待这一问题的方式。近期有多篇综述和范围综述聚焦于该问题的其他方面。 [4] 简要回顾了算法偏差度量，并以信用评分作为目标应用。总体而言， [5] 梳理了人们在应对算法偏差时需要斟酌的大量拟议方法。基于 PRISMA 指南，[6], 在一项全面的技术比较研究中采用了多种偏差度量和数据集。

在 [7], 中，重点在于所采用的公平性定义与其所服务的既定政策目标之间的一致性，并主张采用更贴合具体情境、而非公理化的算法偏差定义。 [8]提出了一个实用框架，以帮助选择适合具体情境的偏差度量。与之方向相反，[9], 研究了那些往往隐含在拟议偏差定义之下的假设、原则和信念。

我们注意到，尽管人们投入了大量研究精力来理解各种偏差度量在伦理、法律、政策和监管方面的影响，但大多数拟议指标都是针对研究人员的需求量身定制的。他们旨在通过本质上属于白盒分析的方法，获取有助于采取纠正干预措施的洞见，无论是在训练阶段还是推理阶段。在 [10], 中，零失效测试被提出作为一种黑盒方法，用于算法性能评估，可能更适合政策制定者和监管机构。

2.1. 衡量偏差

与我们的方法类似，已有若干拟议的偏差度量旨在评估算法对不同人群的公平性。在开创性研究中， [11], 提出了均等赔率和均等机会，作为衡量算法准确率或真正率的公平性指标。 [12] 提出了极小极大帕累托公平性，通过最坏情形定义群体公平性。该方法随后在 [13]中得到了进一步研究和评估。 [14] 提出将准确率差异除以经验标准差，作为偏差度量。 [15] 提出了一个框架，用于在具有连续结果的回归问题中使用分类偏差度量指标。

另一种研究思路是提出若干度量指标，旨在确立个体层面的公平性。 [16], 率先提出了针对个体的任务特定相似性度量这一概念。个体公平性领域的一项重要研究方向以反事实公平性为基础；这一概念由 [17], 提出，并通过结构因果模型 [18]运用了一种因果分析形式。 [19], 提出了一种后处理预测器，以实现均等赔率意义上的反事实公平性。 [20], 提出并验证了一种算法，可在反事实公平性与算法性能之间实现可控权衡。

2.1.1. 交叉偏差

近期研究越来越关注交叉偏差问题，其中我们可能面对的不仅是两个以上

由单一属性形成的群体，还包括由多个属性形成的相互重叠的群体。 [21]首次展示了这一问题的重要性。 [22], 提出了一种基于80%规则的交叉公平性度量。 [23]采用极大极小方法定义最坏情况群体公平性，这次是在交叉情境下。 [24] 扩展了 [19], 的工作，在交叉情境下引入反事实均等赔率。 [25] 提出了交叉敏感性，这是一种基于理想分布之间瓦瑟斯坦距离的度量，并将其用于系统分析偏差的交叉作用。

在这里，我们确实研究了交叉偏差问题，尽管原则上，所提出的样本量分析方法可以直接应用于该问题。

2.2. 分析并纠正算法偏差

鉴于已部署系统中存在算法偏差所带来的影响，许多研究不再局限于测量这一基础问题，而是着重识别其成因与各种表现形式，并提出相应的修正方法。

[26], 使用变分自编码器学习数据集中的潜在分布，并调整训练过程，提高数据中代表不足区域的权重。他们在验证实验中采用的偏差度量，是算法在不同人口群体上的精确率方差。 [27] 构建了一个规模大且种族分布均衡的人脸图像数据集，并报告称，在该数据集上训练的模型在不同种族和性别群体中的表现有所提升且更加一致。 [28] 研究了难度显著更高的语言偏差问题，对多个预训练大语言模型在各种度量指标上进行了比较，并提出一种更实用的方法：在下游任务中评估语言偏差。 [29]研究了在 ImageNet 上预训练的大型模型中的偏差，这些模型通常用作下游计算机视觉任务的骨干网络。最后， [30]提出了一种通过调整数据增强过程来减少计算机视觉模型偏差的方法。

3. 二元分类器的偏差度量

设C 为一个二元分类器，且$D = \{D_1 \cup D_2\}$ 是由两个互不相交的人口群体的并集构成的测试数据集，D_1 和D_2。我们的目标是定义并研究C 在这两个人口群体上的偏差度量，以落实第 1 节中讨论的一般原则。

请注意，在我们的设定中，算法C 不会将D 中的样本分类为人口群体D_1 或人口群体D_2中的元素。相反，我们在D中的元素上执行一项未具体说明的二元分类任务，该任务在D_1 和D_2, 上的错误率分别为 ε_1 和 ε_2 。为简单起见，我们还可以假设D 仅由某一类的样本组成，例如正样本；此时， ε_1 和 ε_2 分别是C 在D_1 和D_2, 上的假阴性率。不过，这一简化假设并非本研究成立的必要条件。

此处将偏差定义为分类器在人口群体D_1 和D_2, 中的表现差异，而零假设针对无偏的C 是

$$H_0 : \varepsilon_1 = \varepsilon_2$$

样本量分析将针对 χ平方检验进行；在此设定下，这是最简单、最自然的选择。

样本量的计算有多种实用方法，每种方法都基于不同的近似假设，例如，将 ε_1 和 ε_2 的离散二项分布近似为连续正态分布。在 [31], 中，给出了一个实践中常用的样本量公式 N：

$$N_{\alpha,\beta}(\varepsilon_1, \varepsilon_2) = \frac{1}{2}\left(\frac{z_\beta + z_\alpha}{\arcsin(\sqrt{\varepsilon_1}) - \arcsin(\sqrt{\varepsilon_2})}\right)^2 \qquad (1)$$

其中， α 为显著性水平， $1-\beta$ 为检验功效， z_x 表示正态分布在 x处的临界值。公式 1 中比例 ε_1 和 ε_2 的反正弦变换是一种标准统计方法；当任一比例 ε_1、 ε_2 显著偏离 0.5 时，这种方法是合理的。请注意，在单侧检验中，应使用 $\alpha/2$，而不是 α 。此外，文献中常用准确率 $1-\varepsilon_1$、$1-\varepsilon_2$ 代替错误率 ε_1、 ε_2，但两种表述只是在类别标签互换后才有所不同。

命题3.1指出，所提出的二分类偏差度量相对于 α和 β 的取值是不变的，最多相差一个乘法常数。

命题3.1。对于任意 $0 \le \varepsilon_1 \le \varepsilon_2 \le 1$，有

$$\frac{N_{\alpha_1,\beta_1}(\varepsilon_1, \varepsilon_2)}{N_{\alpha_2,\beta_2}(\varepsilon_1, \varepsilon_2)} = c$$

其中c 是仅取决于 α_1、 α_2、 β_1和β_2 的常数。

证明。由公式 1可得

$$\frac{N_{\alpha_1,\beta_1}(\varepsilon_1, \varepsilon_2)}{N_{\alpha_2,\beta_2}(\varepsilon_1, \varepsilon_2)} = \left(\frac{z_{\alpha_1} + z_{\beta_1}}{z_{\alpha_2} + z_{\beta_2}}\right)^2.$$

□

注3.1。命题3.1使我们能够在固定 α 和 β的取值时，研究由公式 1估计的样本量$N_{\alpha,\ \beta}(\varepsilon_1, \varepsilon_2)$，并知晓这些参数的任何其他取值只会使样本量估计值按一个常数缩放。实际上，这使我们能够边缘化变量 α 和 β；这些变量与偏差概念之间不存在明显的自然联系，从而得到仅取决于 ε_1 和 ε_2 的结果。特别地，如果对于 α 和 β 的一组给定取值，分类器 C_1 在D 上的偏差被认为小于分类器C_2, ，那么对于 α和 β 的任何其他取值，该结论同样成立。

注释 3.2。 我们注意到，尺度不变性是许多广泛使用的度量指标所不具备的属性，这些度量指标包括基于欧几里得距离的度量，例如均方误差。通常会通过数据归一化来解决这一问题，而归一化的类型取决于应用场景。同样，所提出的偏差度量对于 α 和 β 的取值也不具有不变性。这意味着，对于偏差度量的以下三种可能应用：

(i) 根据分类器C在两个人口群体D_1和D_2上的错误率，量化该分类器的偏差。(ii) 根据两个分类器C_1和C_2在D_1和D_2上的错误率，量化它们的偏差差异。(iii) 根据两个分类器C_1和C_2在D_1和D_2上的偏差对它们进行排序。

对于 (i) 和 (ii)，我们得到部分结果，即结果只确定到一个乘法常数；对于 (iii)，我们则得到完整结果。

在本文余下部分中，我们将显著性水平固定为默认值 $\alpha = .95$，并将检验功效固定为 $1-\beta = .90$。尽管对我们的目的而言，这两个值的选取是任意的，但它们是两个统计量的常用取值。图 1 展示了公式 1 中样本量函数 $N_{\alpha,\beta}(\varepsilon_1,\varepsilon_2)$ 的自然对数色彩图，所用参数为默认值 $\alpha = .95$ 和 $\beta = .90$。我们采用了对数变换，以压缩取值范围并改善可视化效果。

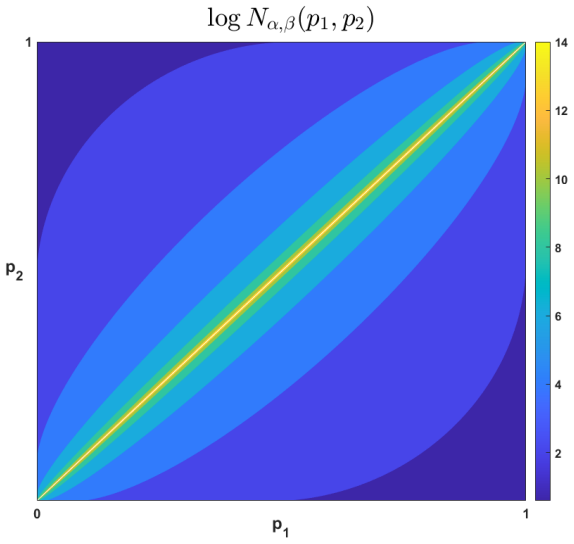

图 1：log $N_{\alpha,\beta}(\varepsilon_1,\varepsilon_2)$ 各取值的色彩图。

从图 1 可以看出，$N_{\alpha,\beta}(\varepsilon_1,\varepsilon_2)$ 的最大值出现在对角线附近 $\varepsilon_1 = \varepsilon_2$。事实上，命题 3.2 指出，当 $\varepsilon_2 \to \varepsilon_1$ 时，N 的值趋于无穷大，因此其倒数 $1/N$ 趋于 0。这是任何偏差度量都应具备的理想性质，因为如果某个分类器在两个人口群体中的错误率相同，那么它们之间就不存在偏差。

命题 3.2。对于任意 $0 < \varepsilon_1 < 1$，我们有

$$\lim_{\varepsilon \to \varepsilon_1} N_{\alpha_1,\beta_1}(\varepsilon,\varepsilon_1) = \infty.$$

证明。由于平方根和反正弦函数在 (0,1) 上连续，因此我们有

$$\lim_{\varepsilon \to \varepsilon_1} (\arcsin(\sqrt{\varepsilon}) - \arcsin(\sqrt{\varepsilon_1})) = 0$$

从而得到，

$$\lim_{\varepsilon \to \varepsilon_1} \frac{1}{2}\left(\frac{z_\beta + z_\alpha}{\arcsin(\sqrt{\varepsilon}) - \arcsin(\sqrt{\varepsilon_1})}\right)^2 = \infty$$

□

命题 3.3 指出，如果保持算法在更准确的人口群体中的误差不变，而另一个人口群体中的误差增加，那么偏差也会增加。这是任何合理的偏差度量都应具备的另一项理想性质。

命题 3.3。对于每个 $0 \le \varepsilon_1 \le \varepsilon_2 \le \varepsilon_3 \le 1$

$$N_{\alpha,\beta}(\varepsilon_1,\varepsilon_3) \le N_{\alpha,\beta}(\varepsilon_1,\varepsilon_2) \tag{2}$$

证明。根据 $N_{\alpha,\beta}$ 在公式 1 中的表达式，只需证明

$$\arcsin(\sqrt{\varepsilon_1}) - \arcsin(\sqrt{\varepsilon_2}) \le \arcsin(\sqrt{\varepsilon_1}) - \arcsin(\sqrt{\varepsilon_3})$$

or,

$$\arcsin(\sqrt{\varepsilon_3}) \le \arcsin(\sqrt{\varepsilon_2})$$

由于 $\arcsin(x)$ 因而 $\arcsin(\sqrt{x})$ 在 $[0,1]$ 中单调递增，因此这一点成立。 □

4. 与比率和差值的比较

在本节中，我们将所提出的偏差度量与另外两种自然的偏差度量进行比较。第一种是两个群体之间错误率的差值

$$D_{err}(\varepsilon_1,\varepsilon_2) = \varepsilon_2 - \varepsilon_1 \tag{3}$$

第二种是错误率之比

$$R_{err}(\varepsilon_1,\varepsilon_2) = \frac{\varepsilon_2}{\varepsilon_1} \tag{4}$$

为简化说明，我们不考虑偏差的方向，并假设 $\varepsilon_1 \le \varepsilon_2$，即分类器在群体 D_1 上的错误率较低。

4.1. 排名比较

首先，我们说明，这三种偏差度量方法本质上是不同的，因为它们可能会对分类算法产生不同的偏差排名。我们使用了 [32] 中报告的数据。其中，共有五个分类器用于人脸识别问题测试，这些分类器分别对应损失函数、模型和训练集的不同组合；测试数据集包含四个种族（亚洲人、黑人、印度人、白人）和两种性别（女性、男性）。报告了每个人口群体上的真正率，以及对于每个种族而言女性与男性之间真正率的差值。

表1展示了文献 [32] 报告的各算法在每个群体上的真正率；该论文将真正率差值用作偏差度量，并计算了样本量以供比较。两种偏差度量得出的不同排名以不同颜色标出。

亚洲人	AF	AM	D_{err}	R_{err}	N	$D_{err}\downarrow$	$R_{err}\downarrow$	$N\uparrow$
算法1	65.00	79.78	14.78	1.73	154	**5**	**1**	**2**
算法2	81.56	94.44	12.89	3.32	101	**4**	**5**	**5**
算法3	81.56	93.11	11.56	2.68	135	**2**	**2**	**3**
算法4	81.22	93.00	11.78	2.68	131	**3**	**3**	**4**
算法5	90.00	96.78	6.78	3.11	214	**1**	**4**	**1**

黑人	BF	BM	D_{err}	R_{err}	N	$D_{err}\downarrow$	$R_{err}\downarrow$	$N\uparrow$
算法1	85.56	86.78	1.22	1.09	13708	**3**	**2**	**3**
算法2	91.00	94.22	3.22	1.56	1118	**5**	**5**	**5**
算法3	91.56	94.11	2.56	1.43	1739	**4**	**4**	**4**
算法4	93.44	93.78	0.33	1.05	88612	**2**	**1**	**1**
算法5	98.00	97.67	0.33	1.17	33266	**1**	**3**	**2**

印度人	IF	IM	D_{err}	R_{err}	N	$D_{err}\downarrow$	$R_{err}\downarrow$	$N\uparrow$
算法1	86.78	90.78	4.00	1.43	1058	**5**	**4**	**5**
算法2	96.00	96.11	0.11	1.03	536364	**1**	**1**	**1**
算法3	94.56	95.56	1.00	1.23	8023	**2**	**2**	**2**
算法4	94.89	93.67	1.22	1.24	6189	**3**	**3**	**3**
算法5	98.56	97.22	1.33	1.93	1920	**4**	**5**	**4**

白人	WF	WM	D_{err}	R_{err}	N	$D_{err}\downarrow$	$R_{err}\downarrow$	$N\uparrow$
算法1	76.67	86.22	9.56	1.69	279	**5**	**1**	**3**
算法2	89.44	96.56	7.11	3.07	205	**4**	**3**	**4**
算法3	90.11	97.11	7.00	3.42	193	**3**	**4**	**5**
算法4	92.67	95.78	3.11	1.74	946	**2**	**2**	**1**
算法5	95.78	98.78	3.00	3.46	462	**1**	**5**	**2**

表1：五种算法在四个种族（亚洲人、黑人、印度人和白人）及两种性别（第2列为女性，第3列为男性）上的真正率。我们报告D_{err}、R_{err}和N（第4—6列），并根据这三个偏差指标（第7—9列）对算法的偏差进行排名。

4.2. 在误差比率或差值恒定时的行为

为了更好地了解拟议指标与误差差值或比率之间的差异，我们研究前者在后者保持不变时的表现。首先，我们有：

命题 4.1。令 $0 \le \varepsilon_1 \le \varepsilon_2 \le 1/2$，且$d \in (0, 1/2 - \varepsilon_2)$。我们有

$$N_{\alpha,\beta}(\varepsilon_1 + d, \varepsilon_2 + d) \le N_{\alpha,\beta}(\varepsilon_1, \varepsilon_2).$$

证明。根据公式 1 中$N_{\alpha,\ \beta}$的公式，只需证明

$$\arcsin(\sqrt{\varepsilon_1}) - \arcsin(\sqrt{\varepsilon_2}) \le \arcsin(\sqrt{\varepsilon_1 + d}) - \arcsin(\sqrt{\varepsilon_2 + d})$$

or,

$$\arcsin(\sqrt{\varepsilon_1}) - \arcsin(\sqrt{\varepsilon_1 + d}) \le \arcsin(\sqrt{\varepsilon_2}) - \arcsin(\sqrt{\varepsilon_2 + d})$$

或者，该

$$f(x) = \arcsin(\sqrt{x}) - \arcsin(\sqrt{x + d})$$

是关于x的递增函数，其中 $0 \le x \le 1/2$，且$d \in (0, 1/2 - x)$。为此，只需证明f关于 x 的导数为正，即

$$0 < f'(x) = \frac{1}{2\sqrt{1-x}\sqrt{x}} - \frac{1}{2\sqrt{-x-d+1}\sqrt{x+d}}$$

得到，

$$\frac{1}{2\sqrt{-x-d+1}\sqrt{x+d}} < \frac{1}{2\sqrt{1-x}\sqrt{x}}$$

or,

$$2\sqrt{1-x}\sqrt{x} < 2\sqrt{-x-d+1}\sqrt{x+d}$$

or,

$$(1-x)x < (1-x-d)(x+d)$$

这一结论成立，因为等式两边两个因子的和均为 1，而左边两个因子之差的绝对值为 $1-2x$，大于右边的 $1-2x-2d$。□

命题 4.1 突显了拟议指标与D_{err}之间的重要区别。当我们保持D_{err} 所度量的偏差不变时，$N_{\alpha,\beta}$ 所度量的偏差会随着算法准确率的提高而增加。表 2 展示了$c = 0.10$ 的这一性质。

	ε_1	ε_2	D_{err}	N	$D_{err}\downarrow$	$N\uparrow$
算法1	0.20	0.30	0.10	319	=	**1**
算法2	0.10	0.20	0.10	213	=	**2**
算法3	0.00	0.10	0.10	42	=	**3**

表 2：命题 4.1 的示例。差值 $D_{err} = \varepsilon_2 - \varepsilon_1$ 恒为 0.10。N 的值随着 ε_1、ε_2 的减小而减小。

命题 4.2 与命题 4.1 等价。这一次，错误比率保持不变；与命题 4.2 不同的是，随着算法变得更加准确，拟议指标所度量的偏差会降低。

命题 4.2. 令 $0 \le \varepsilon_1 \le \varepsilon_2 \le 1/2$，且 $r \in (1, 1/(2\varepsilon_2))$。我们有

$$N_{\alpha,\beta}(\varepsilon_1, \varepsilon_2) \le N_{\alpha,\beta}(r\varepsilon_1, r\varepsilon_2).$$

证明。 根据公式 1 中$N_{\alpha,\beta}$ 的公式，只需证明

$$\arcsin(\sqrt{r\varepsilon_1}) - \arcsin(\sqrt{r\varepsilon_2}) \le \arcsin(\sqrt{\varepsilon_1}) - \arcsin(\sqrt{\varepsilon_2})$$

or,

$$\arcsin(\sqrt{\varepsilon_2}) - \arcsin(\sqrt{r\varepsilon_2}) \le \arcsin(\sqrt{\varepsilon_1}) - \arcsin(\sqrt{r\varepsilon_1})$$

或者说，

$$f(x) = \arcsin(\sqrt{x}) - \arcsin(\sqrt{rx})$$

是x 的递减函数，其中 $0 \le x \le 1/2$ 且$r \in (1, 1/(2x))$。只需证明f 关于x 的导数为负

$$f'(x) = \frac{1}{2\sqrt{1-x}\sqrt{x}} - \frac{r}{2\sqrt{rx}\sqrt{1-rx}} < 0$$

or,

$$\frac{1}{2\sqrt{1-x}\sqrt{x}} < \frac{r}{2\sqrt{rx}\sqrt{1-rx}}$$

或者，经过化简后

$$\sqrt{1-rx} < \sqrt{r}\sqrt{1-x}$$

or,

$$\sqrt{1-rx} < \sqrt{r-rx}$$

这对 $1 < r < 1/2x$成立。 □

表3展示了命题 4.2 针对$r = 2$所描述的性质。

	ε_1	ε_2	R_{err}	N	$R_{err}\downarrow$	$N\uparrow$
算法1	0.20	0.40	2.00	88	=	**3**
算法2	0.10	0.20	2.00	213	=	**2**
算法3	0.05	0.10	2.00	463	=	**1**

表3：命题 4.2 的示例。错误$R_{err} = \varepsilon_2/\varepsilon_1$ 的比率恒为 2。随着 ε_1、ε_2 减小，N 的值会增大。

4.3. 误差值递减时的偏差度量行为

随着算法不断改进，分类问题中最先进水平的演进对应着一系列准确度越来越高的算法；理想情况下，这些算法在两个人口群体上的错误率都趋近于零：

$$(\varepsilon_1, \varepsilon_2) \to (0, 0).$$

为了深入了解这三个度量在 (0,0) 附近的行为，图 2 展示了它们作为 ε_1、ε_2 函数的等高线。我们注意到，错误比率在 (0,0) 附近存在奇点，这意味着对于精度极高的算法而言，它是一种不稳定的偏差度量。事实上，其值取决于趋近 (0,0) 的方向；而且，沿每个这样的方向，其值都是恒定的。

命题 4.3 表明，样本量度量在 (0,0) 处不存在奇点。也就是说，与误差差异类似且不同于误差比，拟议指标在 (0,0) 处确实存在极限。

命题 4.3。令 $0 \le \varepsilon_1 \le \varepsilon_2$。我们有

$$\lim_{\varepsilon_1, \varepsilon_2 \to 0^+} N_{\alpha,\beta}(\varepsilon_1, \varepsilon_2) = \infty$$

证明。由于

$$\lim_{x \to 0} \arcsin(x) = 0$$

我们有

$$\lim_{\varepsilon_1, \varepsilon_2 \to 0^+} (\arcsin(\varepsilon_1) - (\arcsin(\varepsilon_2)) = 0$$

给出

$$\lim_{\varepsilon_1, \varepsilon_2 \to 0^+} N_{\alpha,\beta}(\varepsilon_1, \varepsilon_2) = \infty$$

□

这是一个直观上预期的结果。事实上，任何合理的统计检验都需要一定数量的负样本，才能区分两个人口群体之间的性能差异。因此，随着两个人口群体上的错误率降低，我们需要更多的样本，才能以足够高的概率获得所需的负样本。

下一节中，我们将讨论可通过此类论证推导出的所提偏差度量的性质。除了这种替代方法具有普适性、不受所用统计检验的影响之外，我们还注意到，该方法并不依赖样本量的近似公式（例如公式 1），因此，在研究极值性质时，不存在与近似精度相关的有效性问题。

5. 独立于统计检验选择的性质

尽管对于何为恰当的统计检验尚无普遍认可的定义，但人们通常认为，所有合理的检验都应具备一些理想属性。最低限度的假设是，检验结果应通过对输入数据进行统计分析得出，而非任意确定。

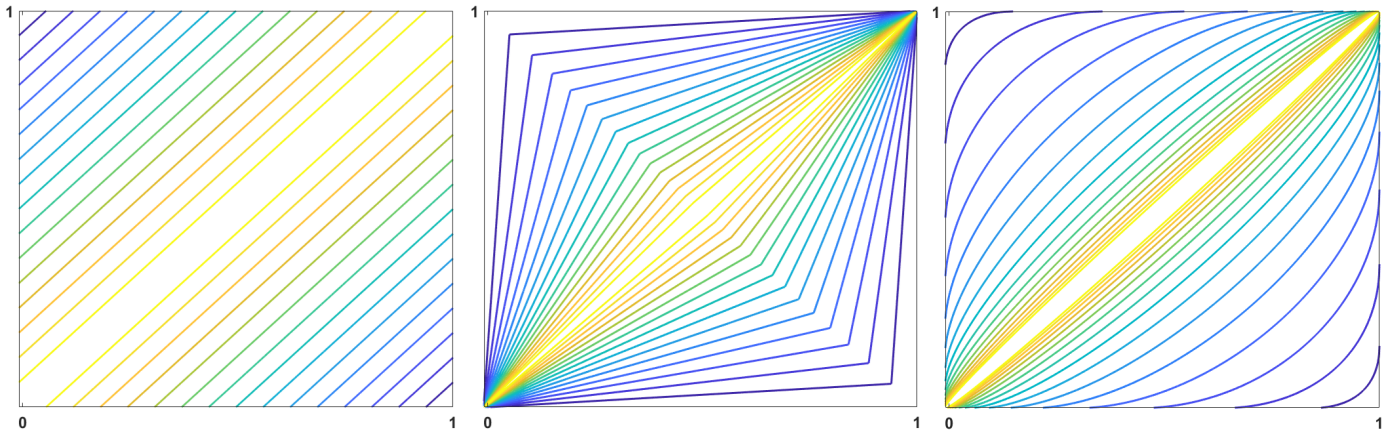

图 2：从左到右：方程中函数 $D_{err}(\varepsilon_1, \varepsilon_2) = \varepsilon_2 - \varepsilon_1$、$R_{err}(\varepsilon_1, \varepsilon_2) = \varepsilon_2/\varepsilon_1$ 和 $N_{\alpha,\beta}(\varepsilon_1, \varepsilon_2)$ 的等高线 . 1.

其中一项属性是：如果大小为N的样本中极有可能不存在误分类，那么检验就不应检测出任何偏差。我们将在人口群体多于两个的情形下，将这一属性形式化，从而推广命题 4.3。

考虑一个由 k 个人口群体组成的测试集：$D_1, D_2, \ldots, D_k$，并设分类算法在这些群体中的错误率分别为 $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_k$。根据第 1 节所述的一般原则，算法的偏差与统计检验用于判断不同人口群体之间的错误率是否存在差异所需的样本量成反比。因此，原假设为

$$H_0 : \varepsilon_1 = \varepsilon_2 = \cdots = \varepsilon_k \tag{5}$$

由此可得以下结论。

命题 5.1。 设 $N_{\alpha,\beta}(\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_k)$ 为对H_0, 进行适当统计检验所需的样本量，置信度为 α，检验效能为 $1-\beta$。则有

$$\lim_{\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_k \to 0^+} N_{\alpha,\beta}(\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_k) = \infty \tag{6}$$

证明。 设 $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_k < 1/n^2,$，且 $n \in \mathbb{N}$。大小为n 的样本不包含误分类的概率至少为

$$(\frac{n^2 - 1}{n^2})^n \to 1 \quad \text{as} \quad n \to \infty$$

因此，样本大小为n 的样本很可能不包含误分类，如 $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_k \to 0^+$ 所示。根据适当的统计检验应具备的理想性质——当样本很可能不包含误分类时，不应检测出偏差——可得公式 6。 □

命题 5.1 是命题 4.3 的与统计检验无关的推广。尽管对适当统计集合的全部理想属性进行正式定义超出了本文的范围，但我们注意到，命题 3.2、3.3 也对应于人们所期望的任何合理统计检验都应具备的属性。

例如，命题 3.2 可以被视为这样一种更一般的直观属性的实例：当错误率趋于相等时，统计检验需要更大的样本量才能将它们区分开来。也就是说，对于公式 5 中的原假设，我们应有

$$\lim_{\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{k-1} \to \varepsilon_k} N_{\alpha,\beta}(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_k) = \infty$$

类似地，命题 3.3 可视为一项更一般性质的实例：当最大错误率上升，而所有其他错误保持不变时，偏差也会增加。也就是说，如果

$$0 \le \varepsilon_1 \le \varepsilon_2 \le \cdots \le \varepsilon_{k-1} \le \varepsilon_k \le \varepsilon_{k+1} \le 1/2$$

我们应有

$$N_{\alpha,\beta}(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{k-1}, \varepsilon_{k+1}) \le N_{\alpha,\beta}(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{k-1}, \varepsilon_k).$$

6. 讨论

这里我们讨论前文未涉及的拟议指标的一些方面，以及一些潜在的替代方法；这些方法的详细阐述超出了本文的范围。

样本量计算，无论是解析计算（例如公式 1），还是通过蒙特卡罗模拟，都是基于对误差分布的假设，例如用于 χ平方检验的χ平方分布。因此，可以直接参照这类分布来制定偏差度量。不过，样本量方法借助了大量积累的理论成果和实践经验，这些成果和经验有助于确定哪些分布最合适，以及如何高效地进行相关计算。尽管如此，在白盒方法中，直接处理误差分布仍然是一种可行方案；在计算样本量时，可以纳入特定问题领域中有关误差分布的实验依据。

采用贝叶斯框架也是一种测量偏差的等效方法。不过，当错误率及其

分布不是固定不变的，而是在算法验证过程中随着样本的处理而更新时，贝叶斯方法似乎更为自然。我们在此重申，我们的目标不是提出一种计算错误率分布的方法，而是根据不同人口群体的错误率点估计来确立一种偏差度量。

拟议方法与信息论之间的联系也值得探讨。例如，正如命题 4.3、5.1 的讨论所指出的，在准确率很高的算法中，偏差检测所需的样本量更大，这可能是因为人口群体上的随机变量熵较低。也就是说，如果算法的总体准确率很高，那么在所有人口群体中，每个分类结果所提供的有关该算法偏差的信息都很少。

全文始终默认来自两个人口群体的样本量相同。其他方法也可能值得探讨，例如让样本量与人口群体规模成正比，但这也会带来复杂性；我们认为，最好结合实际应用的具体情境来处理这些问题。

6.1. 人类感知问题

随着我们越来越依赖算法来做出影响日益深远的决策，借用一句著名的法律格言 [33]，：算法不仅要公平，还必须让人们认为它公平。因此，偏差感知也是一个应当研究的方面；在此，我们假设存在韦伯定律。

图 3 展示了韦伯定律在人类数量感和数量判断中的一个实例 [34]。尽管在图 3 所示的视觉情境中，韦伯定律被认为已得到充分证实，但也有研究指出，相关研究的实证发现取决于任务与刺激的具体情况 [35]。因此，未来研究的一个有趣方向似乎是设计实验，在与人类和已部署的机器学习分类器交互相似的情境、任务和刺激中，验证韦伯定律。

此类关于人类感知的实证研究还需要解决另一个复杂问题：人们可能会过度关注实验中熵较高的结果，也就是误分类。然而在实践中，这或许情有可原，因为这类结果往往伴随着更高的代价。以图 3（上）为例，这意味着注意力会集中在黑点上，而白点基本上会被视为背景。

尽管如此，作为初步观察，我们注意到，命题 4.1 中所述的N 的行为似乎与人类对偏差的感知相当吻合，至少比差值 $\varepsilon_2 - \varepsilon_1$ 更吻合。也就是说，正如图 3 所示，我们预计，当误分类数量增加而其绝对数量之差保持不变时，人口群体之间算法性能的差异会更容易被察觉。

最后，在与传播相关的问题上，本文的见解可用于指导如何以通俗语言传达研究发现。例如，在医学研究中，分类通常对应于诊断，而

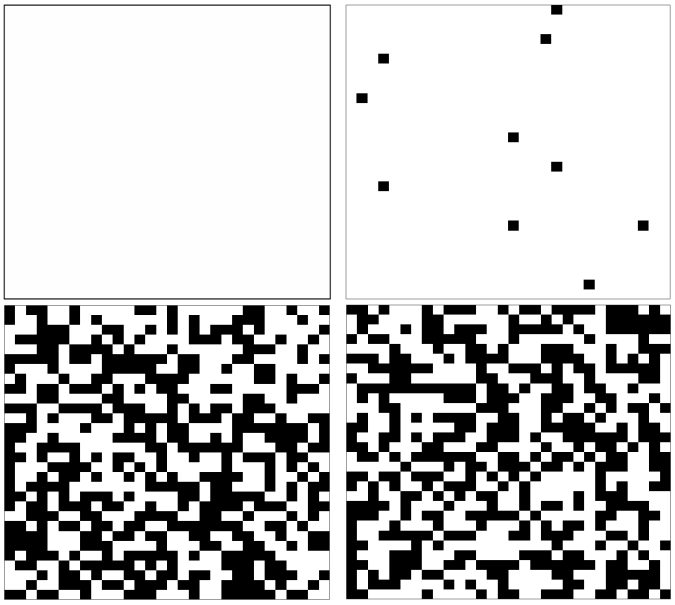

图 3：在每幅图中，30×30 网格中的黑点代表误分类样本，白点代表正确分类样本。**上方：** $\varepsilon = 0.00$（左）与 $\varepsilon = 0.01$（右）。**下方：** $\varepsilon = 0.49$（左）与 $\varepsilon = 0.50$（右）。

面向普通人的摘要有时只报告两组阳性诊断的比率$r = \varepsilon_2/\varepsilon_1$，而不报告任何 $\varepsilon_1, \varepsilon_2$，尽管这些信息有助于评估研究发现的影响。有趣的是，尽管研究规模N 常常会出现在面向普通人的摘要中，但解读这一信息所需的内容可能仍然缺失。尤其是，较大的N 可能对应于设计中较高的检验效能和置信度，即较大的 α和 $1-\beta$，也可能对应于较小的 ε_1，ε_2。

7. 结论

我们提出了一种算法偏差度量，在本文中，算法偏差指算法在不同人口群体上的性能差异。我们的方法基于一个简单而直观的原则：当错误率差异很大时，只需进行小规模的统计检验即可确认这一点；相反，如果错误率十分接近，则需要更大规模的检验。我们研究了两个人口群体这一最简单的情形，利用 χ 卡方检验样本量的闭式公式，确立了拟议方法的若干理想性质。随后，我们将其与两种常用度量指标进行比较，即错误率之差和错误率之比，并证明该方法得出的偏差排名存在实质性差异。最后，我们讨论了拟议指标的一些理想性质如何自然地源于样本量分析，以及为何应将这些性质视为与检验无关。

尽管正如本文所示，偏差排序结果与 χ-平方检验参数 α、β 的取值无关，但可以预期，采用不同的统计检验并对误差分布作出不同假设，会得到不同的排序。未来，我们计划探索这一研究方向，以进一步深入理解算法偏差这一概念。

参考文献

[1] D. J. Hand, 《分类器性能测量：ROC曲线下面积的一种一致性替代方案》, 《机器学习》77 (1) (2009) 103–123。[2] C. Anagnostopoulos、D. J. Hand、N. M. Adams, 《分类性能测量：hmeasure软件包》, 技术报告, 伦敦帝国学院 (2017)。[3] S. Barocas、A. D. Selbst, 《大数据的差异影响》, 《加州法律评论》104 (2016) 671。[4] S. Das、R. Stanton、N. Wallace, 《算法公平性》, 《金融经济学年度评论》15 (1) (2023) 565–593。[5] R. Schwartz、A. Vassilev、K. Greene、L. Perine、A. Burt、P. Hall, 《迈向人工智能偏差识别与管理标准》, 技术报告 SP 1270, 美国国家标准与技术研究院 (2022)。[6] T. P. Pagano、R. B. Loureiro、F. V. N. Lisboa、R. M. Peixoto、G. A. S. Guimar˜aes、G. O. R. Cruz、M. M. Araujo、L. L. Santos、M. A. S. Cruz、E. L. S. Oliveira、E. G. S. Nascimento, 《机器学习模型中的偏差与不公平：关于数据集、工具、公平性度量指标以及识别与缓解方法的系统综述》, 《大数据与认知计算》7 (1) (2023) 15。[7] S. Corbett-Davies、J. D. Gaebler、H. Nilforoshan、R. Shroff、S. Goel, 《公平性的度量与误测》, 《机器学习研究杂志》24 (312) (2023) 1–117。[8] C. J. Barr、O. Erd´elyi、P. D. Docherty、R. C. Grace, 《公平性综述：在机器学习中选择适合具体情境的公平性度量指标实用指南》, arXiv:2411.06624 (2024)。[9] S. A. Friedler、C. Scheidegger、S. Venkatasubramanian, 《公平性（不）可能：不同的价值体系需要不同的公平决策机制》, 《美国计算机学会通讯》64 (4) (2021) 136–143。[10] I. Ivrissimtzis、M. Houliston、S. Concannon、G. Roberts, 《二元分类器的零失败测试》, arXiv:2407.03979 (2024)。[11] M. Hardt、E. Price、N. Srebro, 《监督学习中的机会平等》, 《神经信息处理系统进展》（NeurIPS）29 (2016)。[12] N. Martinez、M. Bertran、G. Sapiro, 《极小极大帕累托公平性：多目标视角》, 载于：国际机器学习会议，PMLR，2020，第6755–6764页。[13] E. Diana、W. Gill、M. Kearns、K. Kenthapadi、A. Roth, 《极小极大群体公平性：算法与实验》, 载于：AAAI/ACM人工智能、伦理与社会会议，2021，第66–76页。[14] D. DeAlcala、I. Serna、A. Morales、J. Fierrez、J. Ortega-Garcia, 《人工智能模型中的偏差测量：一种引入 N-Sigma 的统计方法》, 载于：第47届计算机、软件与应用会议（COMPSAC），IEEE，2023，第1167–1172页。[15] J. Su´árez Ferreira、M. Slavkovik、J. Casillas, 《回归问题中公平性测量的一般流程》, 《国际数据科学与分析期刊》 (2025) 1–20。[16] C. Dwork、M. Hardt、T. Pitassi、O. Reingold、R. Zemel, 《通过认知实现公平》, 载于：第3届理论计算机科学创新会议论文集，2012，第214–226页。[17] M. J. Kusner、J. Loftus、C. Russell、R. Silva, 《反事实公平性》, 《神经信息处理系统进展》（NeurIPS）30 (2017)。[18] J. Pearl, 《统计学中的因果推断：概述》, 《统计学调查》3 (2009) 96–146。[19] A. Mishler、E. H. Kennedy、A. Chouldechova, 《风险评估工具中的公平性：通过后处理实现反事实均等赔率》, 载于：2021年ACM公平、问责与透明度会议论文集，2021，第386–400页。[20] Z. Zhou、T. Liu、R. Bai、J. Gao、M. Kocaoglu、D. I. Inouye, 《结合事实预测与反事实预测实现反事实公平性》, 《神经信息处理系统进展》（NeurIPS）37 (2024) 47876–47907。[21] J. Buolamwini、T. Gebru, 《性别阴影：商业性别分类中的交叉性准确率差异》, 载于：公平、问责与透明度会议，PMLR，2018，第77–91页。[22] J. R. Foulds、R. Islam、K. N. Keya、S. Pan, 《公平性的交叉性定义》, 载于：第36届国际数据工程会议（ICDE），IEEE，2020，第1918–1921页。[23] A. Ghosh、L. Genuit、M. Reagan, 《通过最坏情况比较刻画交叉性群体公平性》, 载于：第2届多样性研讨会，人工智能……

ficial Intelligence (AIDBEI), 《机器学习研究论文集》第142卷，2021年，第22–34页。[24] S. Wastvedt, J. D. Huling, J. Wolfson，风险预测中反事实公平性的交叉性框架, 《生物统计学》25 (3) (2024年) 702–717。[25] P. Shukla, A. Chinchure, E. Diana, A. Tolbert, K. Hosanagar, V. N. Balasubramanian, L. Sigal, M. Turk，顾此失彼？应对文生图模型中的交叉性偏见，arXiv预印本 arXiv:2505.17280 (2025年)。[26] A. Amini, A. P. Soleimany, W. Schwarting, S. N. Bhatia, D. Rus，通过学习得到的潜在结构揭示并缓解算法偏差，载于：AAAI/ACM人工智能、伦理与社会会议，2019年，第289–295页。[27] K. K¨arkk¨ainen, J. Joo，FairFace：用于偏见测量与缓解、在种族、性别和年龄方面均衡的面部属性数据集，载于：IEEE/CVF冬季计算机视觉应用大会论文集，2021年，第1548–1558页。[28] P. Delobelle, E. Tokpo, T. Calders, B. Berendt，以有偏的标尺衡量公平性：预训练语言模型偏差指标的比较研究，载于：北美计算语言学协会分会会议：人类语言技术，2022年，第1693–1706页。[29] V. Krug, S. Stober，预训练图像识别模型中的交叉偏差，arXiv:2506.03664 (2025年)。[30] T. Kumar, A. Mileo, M. Bendechache，基于显著性的指标与face-keeporiginalaugment：一种提升公平性与多样性的新方法，《多媒体系统》31 (2) (2025年) 1–14。[31] T. P. Ryan, 《样本量确定与检验效能》, 约翰·威立父子公司，2013年。[32] H. Wu, K. W. Bowyer, “均衡”的人脸识别数据集应该均衡什么？, arXiv:2304.09818 (2023年)。[33] R v Sussex Justices, ex parte McCarthy, [1924] 1 KB 256 (1924年)。[34] S. Dehaene，韦伯–费希纳定律的神经基础：对数心理数轴，《认知科学趋势》7 (4) (2003年) 145–147。[35] A. Testolin, J. L. McClelland，数量估计是否确实遵循韦伯定律？对两个案例研究的重新审视，《心理学公报与评论》28 (1) (2021年) 158–168。