排名中偏倚代表性群体的检测

Jinyang Li密歇根大学
jinyli@umich.edu

Yuval Moskovitch内盖夫本-古里安大学yuvalmos@bgu.ac.il

H. V. Jagadish密歇根大学
jag@umich.edu

arXiv:2301.00719v2 [cs.LG] 7 Jul 2023

摘要——**许多关键领域中用于决策的现实工具都基于排序结果。随着人们对算法公平性的认识不断提高，近期研究提出了用于衡量排名公平性的指标。许多定义会考量在任何合理的 k 下，排名前 k 的排序项中不同“受保护群体”的代表性。给定受保护群体后，确认算法公平性是一项简单的任务。然而，群体的定义可能事先并不为人所知。**

本文研究如何检测排名前 k 的排序项中具有偏倚代表性的群体，从而无需预先定义受保护群体。此类群体的可能数量呈指数级增长，使问题难以解决。我们针对两种不同的公平性度量提出了高效搜索算法：全局代表性界限和比例代表性。随后，我们提出一种方法，利用 Shapley 值的概念解释群体代表性中的偏差。最后，我们通过实验研究展示了该方法的可扩展性，并证明了所提出算法的实用性。

索引词—**代表性、排名、公平性、偏差**

I. 引言

排名是广泛应用领域中常用的一种操作，例如，在网络搜索引擎中呈现结果 [8], 建立信用评分 [7], 学校录取 [29]以及招聘 [17]。尽管这些工具便捷且实用，但它们可能存在偏差。因此，它们可能以不理想的方式影响决策，甚至影响人的生活 [2], [6], [11]。这一问题引起了研究界的广泛关注，近期的一系列研究致力于衡量和缓解排名中的偏差与不公平性 [3], [10], [17], [22], [34], [36], [39]。

算法公平性的概念已在广泛的一类模型中得到深入研究 [25], [30]。公平性度量通常针对数据中的特定“受保护群体”，该群体依据某些敏感属性（例如性别、种族、年龄或这些属性的组合）的取值来界定，通常也会考虑社会歧视的历史。分析系统针对给定群体的公平性度量是一项简单的任务。然而，“非标准”的受保护群体并不总能预先指定，在考察系统性能时，这类群体可能会被忽略。

例如，一种用于给学生评定成绩的模型（取代因 COVID-19 疫情而取消的考试）被证明对成绩优异的学生存在偏差，这些学生来自贫困学区[1]。例如，即使低收入家庭的学生以西班牙语为母语，模型仍预测他们无法通过西班牙语考试。在这种情况下，模型歧视了来自贫困学区的西班牙裔学生。偏差的一个主要来源是，模型使用各学校的历史考试成绩来预测学生的表现。然而，将学校（以学校编号标识）作为界定受保护群体的依据并不直观，因此可能不会有人考虑这样做。此外，即使我们将西班牙裔学生这一组视为受保护群体，也可能发现不了任何公平性问题，因为这一学生子群体仅占所有西班牙裔学生的一小部分。

本文研究如何检测受到排序算法不公平对待的群体。换言之，我们希望让数据揭示（潜在的）不公平性，而不要求人工建模者事先确定受保护属性。遵循排名公平性相关文献中提出的公平性定义（例如，参见 [10], [30], [36]），我们将任意合理范围内的 k 对应的前k 个排序项中的群体代表性视为一种公平性度量。

近期研究探讨了如何在无需事先指定受保护属性的情况下，自动检测数据中“有问题”或存在偏差的子群体 [9], [12], [21], [23], [28]。然而，这些研究仅考虑了分类模型。在 [27]中， [28] 的作者将其框架扩展至排名问题。与我们的工作不同，我们基于文献中的排名公平性度量，而他们使用差异度来衡量数据子组之间的性能差异。这一区别导致各方法返回的结果集有所不同（详情参见第六节-D）。

接下来，我们概述本文的主要贡献。

问题形式化： 我们正式定义以下问题：对于给定的排序算法 R、数据集 D 以及可能的 k 范围，检测在排名前-k 个排序项中具有偏倚代表性的组。组通过对一组属性进行值赋值来定义，我们将其称为 模式（见第 II 节）。为了提供简洁且有意义的结果，我们对返回组的大小设置阈值，并仅报告结果集中不被任何其他组包含的组（称为最一般模式，见第 III 节）。我们

[1]https://www.nytimes.com/2020/09/08/opinion/international-baccalaureate-algorithm-grades.html

从 [10], 的基本定义开始，该定义使用上界和下界来限制数据中不同组在排名前-k 个排序项中的元组数量。我们的目标是报告这样的组：对于给定的可能 k 范围，其代表性（即元组数量）在排名前-k 个排序项中不处于给定界限之内。我们将此问题称为全局界限代表性偏差 问题。随后，我们考察采用比例代表性的主流公平性度量类别（例如，见 [36]）。直观而言，每个组在排名前-k 个排序项中的代表性应与其在数据中的规模成比例。基于这一概念，我们定义比例代表性偏差 问题。我们证明，不存在能够解决这两个问题中任一个的多项式算法。

检测具有偏倚代表性的群体：我们提出了用于解决以下问题的算法：在排名前-k 的排序项中，找出所有具有显著规模的群体（就其在数据中的大小以及被其他群体包含的程度而言），且这些群体具有偏倚代表性。我们首先提出一种简单的基线方案，该方案利用了 [5] 中提出的模式图概念。我们展示了如何以自上而下的方式遍历该图，以查找具有偏倚代表性的群体。随后，对于给定范围内的每个 k，重复执行该搜索。鉴于该问题的复杂性，我们重点对搜索进行优化。我们的优化方案基于以下事实：排名前-k 的元组集合与排名前-($k+1$)的元组集合仅相差一个元组。因此，后续 k 值对应的搜索空间通常非常相似。优化方案利用这一观察结果，避免搜索树中的部分节点。

结果分析：给定一个在排名前-k 的排序项中具有偏倚代表性的组，分析人员可能希望了解偏差的原因。为此，我们提出了一种利用 Shapley 值概念的方法，用于识别会显著影响所检测群体排名的属性。为了分析所检测群体与排名前-k 的排序元组之间的差异，我们将这些属性的值分布进行可视化。Shapley 值已被用于为回归模型和分类模型提供类似的解释。我们的创新之处在于，为排名问题开发了一种相应的方法。

实验研究：我们通过全面的实验研究补充算法开发。我们评估算法的性能和特性，即可扩展性和参数设置效应。我们使用三个真实世界数据集，考察属性数量、群体大小阈值以及 k 的范围所产生的影响。结果表明，尽管这些问题在理论上具有复杂性，我们的方案仍适用于实践；与基线方案相比，优化算法也很有用。随后，我们通过实验展示了该方法在结果分析中的实用性。最后，我们通过一项案例研究，将我们的算法所得结果与 [27] 提出的方法所得结果进行比较。

论文结构：本文其余部分安排如下。第二节介绍问题定义所需的预备知识。接着，我们在第三节正式定义检测数据中代表性存在偏差的群体这一问题，并证明其难度。第四节介绍我们的解决方案。第五节介绍一种用于分析和解释算法结果的方法。第六节介绍实验评估，第七节概述相关工作，并在第八节作出总结。

II. 预备知识

接下来，我们介绍用模式表示数据组所需的背景知识，以及排名公平性的概念。我们将以下示例作为贯穿全文的示例，以阐释本文提出的观点。

示例 *2.1*：学生表现数据集 [13] 包含来自葡萄牙阿连特茹地区两所葡萄牙中学——Gabriel Pereira（GP）和 Mousinho da Silveira（MS）——的信息。数据收集于 2005—2006学年，包含 1044 名学生的数学和葡萄牙语考试成绩，以及人口统计、社会和学校相关信息。图1展示了数据中的一个样本，其中包含以下属性：性别、学校、地址（城市或乡村）以及失败次数（过去课程不及格的次数）。成绩属性表示学生的成绩，评分范围为 $0-20$。设有一个优秀学生项目委员会，希望根据学生的学业表现选出奖学金获得者。为此，他们使用排序算法 R 按成绩对学生进行排名。成绩相近时，失败次数较少的学生排名更高。获奖学者名单将公开发布，并且应体现多样性与包容性。

A. 数据组

我们假设数据以单个关系数据库的形式表示，并且用于群体定义的关系属性值为类别型。若要在群体定义中纳入取自连续域的属性值，我们会将其分桶为若干范围，使其转化为类别型；这是实践中呈现聚合结果时非常常见的做法。我们使用模式这一概念，即对一组属性进行赋值，来定义数据 [5]中的群体。

定义 *2.2*（模式）：设 D 是一个具有类别型属性 $\mathcal{A}=\{A_1,\ldots,A_n\}$ 的数据库，并设 $Dom(A_i)$ 为 $i\in[1..n]$中 A_i 的活动域。D 上的一个模式 p 是满足以下条件的集合：$\{A_{i_1},\ldots,A_{i_k}\}\subseteq\mathcal{A}$且对每个 $A_{i_j}\in p$，$a_j\in Dom(A_{i_j})$。我们用 $Attr(p)$表示 p中的属性集合。

我们称元组 $t\in D$ 满足模式 p，即 $t.A_i=a_i$对于每个 $A_i\in Attr(p)$。模式 p 的大小 $s_D(p)$是 D 中满足 p的元组数量。给定排序算法 R，我们使用 $R^k(D)$表示D中的前k 个排名项。最后，我们使用 $s_{R^k(D)}(p)$表示 p 在 $R^k(D)$ 中的大小。

例*2.3*：考虑图1所示的数据集。$p=$\{School = GP\}是一个模式的示例。元组3、4、7、8、12、13、15和16满足 p，因此 $s_D(p)=8$。对于 R 排序算法（其结果如Rank列所示），我们有$s_{R^5(D)}(p)=1$，因为排名前5的排序项中只有一个元组满足 p。

#	性别	学校	地址	失败次数	成绩	Rank
1	F	MS	R	1	11	8
2	M	MS	R	1	15	3
3	M	GP	U	1	8	10
4	M	GP	U	2	4	16
5	M	MS	R	0	19	2
6	F	MS	U	1	4	15
7	F	GP	R	1	7	11
8	M	GP	R	1	6	13
9	F	MS	R	0	14	4
10	F	MS	R	2	7	12
11	M	MS	R	2	13	6
12	F	GP	U	0	20	1
13	F	GP	U	2	12	7
14	M	MS	U	1	13	5
15	F	GP	U	1	5	14
16	M	GP	U	0	9	9

图1：学生数据。Rank列展示了根据成绩和以往的不及格次数得出的排名。排名前5的学生已突出显示

B. 公平性度量

排名公平性问题已在一系列研究中得到探讨（综述请参见 [30] ）。一项基础定义见 [10], ，它针对排名前k的排序项中不同群体的代表性设定了约束。该定义针对排名前k的位置中具有属性 l （即属于受保护群体）的项数，设定了上界 U_{kl} 和下界 L_{kl} 。根据 [10], 的定义，如果排名前k 中来自受保护群体的所选项数量处于给定界限之内，则排序算法是公平的。

示例 *2.4*：再次考虑图1中给出的数据集，以及结果显示在 Rank 列中的排序器。考虑每所学校的学生人数下界为 2 的情况 $k = 5$ （即，$L_{5,\ school=MS} = L_{5,\ school=GP} =2$）。在这种情况下，在排名前 5 的学生中，只有一名来自 GP 学校，因此该排序器不满足约束。

排名中另一类重要的公平性度量，关注不同群体在排名前 k 个排序项中的比例代表性（例如，参见 [36]）。直观地说，这些定义可以看作是 [10] 定义的变体：对于每个群体 g 和每个 k，来自 g 的项在排名前 k 个排序项中的出现次数界限，是根据该群体在数据集中的大小来定义的。

示例 *2.5*：接续示例 2.4，每所学校（MS 和 GP）的学生总数均为 8。数据集总大小为 16，因此，每所学校在排名前 5 个排序项中应大致占有 $5 \cdot \frac{8}{16} \approxeq 2$ 的比例代表性。

III. 问题定义

我们的目标是针对给定的排序算法 R、数据集 D 以及一系列 k，检测前-k 个排序项中存在偏倚代表性的群体。我们借鉴文献中针对排序算法提出的公平性度量来定义问题。具体而言，我们提出了两个问题定义，采用了具有代表性的公平性度量；两者都会考察在不同的 k 值下，前-k 个排序项中不同群体的代表性。直观来说，考虑一系列 k 值可确保排名在任意位置上都公平。例如，若前 10 个项目中有 5 名学生的住址在城市，另有 5 名学生的住址在农村，但城市住址的学生占据了排名第 1 至第 5 位，那么如果我们只关心选出前 10 名学生，输出看起来或许“公平”；但如果前 10 名中的名次也很重要（例如，排名会影响奖励金额），那么显然这个排名是不公平的。

第一个定义直接采用了 [10] 的定义。[10] 的公平性定义通过上界和下界限制前-k 中不同群体（即模式）的数量。根据此定义，如果对于某个 k，前-k 个元组中某个模式 l 的大小超过上界 U_{kl} 或低于下界 L_{kl}，则结果存在偏差。我们取消了预先定义 l 的要求，并使用 U_k 和 L_k 分别表示给定排序算法的前-k 个排序元组中每种模式的上界和下界。

如果某个群体在排序算法 R 的输出中具有偏倚代表性，我们称该群体具有偏倚代表性；具体而言，由 R 排序的前-k 个排序项中，该群体的大小对于给定范围内任意可能的 k，都不在给定界限之内。直观而言，我们希望避免报告对群体描述得“过于具体”，并为用户提供一组简洁的属性，以刻画具有偏倚代表性的有意义群体（就其大小而言）。为此，我们提出最一般模式的概念。给定群体在前-k 个排序项中的代表性界限，我们称模式 p 为具有偏倚代表性的最一般模式，条件是使用 p 表示一个代表性不足的群体（依据给定界限），且 $R^k(D)$ 中 p' 的计数 $\forall p' \subsetneq p$ 落在给定界限之内。现在，我们可以正式定义这个问题。

问题 *3.1*（全局界限代表性偏差）：给定数据库 D、排序算法 R、大小阈值 τ_s[2]、范围 $[k_{min}, k_{max}]$，以及每个 $k_{min} \leqslant k \leqslant k_{max}$ 的下界 L_k 和上界 U_k，对于每个 $k_{min} \leqslant k \leqslant k_{max}$，找出所有大小为 $\geqslant \tau_s$ 且满足 $s_{R^k(D)}(p) < L_k$ 或 $s_{R^k(D)}(p) > U_k$ 的最一般模式 p。

请注意，排序算法被视为黑箱，因此该问题是模型无关的。沿着比例代表性研究的思路，我们通过细化上述定义来考虑另一个问题定义，以纳入数据集中各群体的大小。直观地说，前 k 个排序项中来自每个群体的项数，应当与该群体在数据中的代表性成比例。

问题 *3.2*（比例代表性偏差）：给定数据库 D、排序算法 R、大小阈值 τ_s 和范围 $[k_{min}, k_{max}]$，对于每个 $k_{min} \leqslant k \leqslant k_{max}$，找出所有大小为 的最一般模式 p $\geqslant \tau_s$ $s_{R^k(D)}(p) <$，使得 $\alpha \cdot s_D(p) \frac{k}{|D|}$ 或 $s_{R^k(D)}(p) > \beta \cdot s_D(p) \frac{k}{|D|}$。 $\alpha < \beta \in \mathbb{R}$

比例代表性为用户提供了一种直观的界限定义。然而，全局代表性的定义允许用户主动控制不同群体在前-k 个排名项中的代表性界限，

[2]我们使用绝对值作为大小阈值。等价地，也可以将其定义为数据集大小的一定比例。

	A_1	A_2	$\cdots$	A_{n-1}	A_n	Rank
t_1	1	0	$\cdots$	0	0	1
t_2	0	1	$\cdots$	0	0	2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
t_n	0	0	$\cdots$	0	1	n
t_{n+1}	0	0	$\cdots$	0	0	$n+1$

图 2：定理 3.3 证明中的数据集 D

即使这些群体在整体数据中的代表性较低或较高。例如，考虑一种用于男性占主导地位领域（如科技公司）求职者的排序算法。如果公司希望提高所聘用女性的代表性，但女性申请人数较少，而且只有排名前-k 的求职者会获邀面试，那么采用比例代表性时，前-k名中的女性人数也会较少，因此获邀面试的女性人数同样会较少。 [10] 的基本定义允许用户针对不同的 k值，为数据中受保护群体的代表性设定界限。根据此定义，我们提出了全局界限代表性问题。该问题假设事先不知道受保护群体的身份，并旨在识别所有代表性存在偏差的群体。请注意，我们的目标是仅报告最一般模式（群体），以简洁地描述这些群体。

虽然具有偏倚代表性的最一般模式通常远少于所有具有偏倚代表性的模式，但在最坏情况下，其数量可能呈指数级增长。即使我们只考虑下界，这一点仍然成立（例如，如果 $U_k = |D|$）。

定理 3.3：给定一个数据集 D 和一个排序算法 R，不存在多项式时间算法能够保证枚举出 R 在 D上的结果中所有具有偏倚代表性的最一般模式。

证明 *3.4*：我们通过构造来证明该定理。考虑一个数据集 D，包含 n（假设 $n \geq 2$）个二元属性$\{A_1 \ldots, A_n\}$和 $n+1$ 个元组 $t_1, \ldots, t_n, t_{n+1}$，如图 2 所示。也就是说，$\forall i \in [1, n]$，$t_i[A_i] = 1$，且 $\forall j \neq i$，$t_i[A_j] = 0$。t_{n+1} 的所有属性均为零。令 R 为一个排序算法，使得 D 中的前k 个元组是图 2 中的元组$t_1, \ldots, t_k$。对于某个 $n \geq 2$，令 $k_{min} = k_{max} = n$对应问题 3.1（全局代表性界限），令 $\alpha = \frac{n+3}{n+4}$对应问题 3.2（比例代表性）。

考虑一个模式 p，它有 $m \leq n$ 个属性 A_i，其赋值为 $A_i = 0$。令 I 为这些属性的索引集合。在 t_1、$\cdots$、t_n中，p的大小为 $n - m$，因为 $\forall i \in I$、$t_i[i] = 1$、t_i 不满足 p，而其余所有元组都满足 p。令 $m = \frac{n}{2}$，因此在排名前 k的排序项中，p 的大小为 $s_{R^k(D)}(p)$，这是全局代表性界限下的情况。对于比例代表性，我们有 $s_D(p) = \frac{n}{2} + 1$，因为 t_{n+1}也满足 p。因此我们得到 $s_{R^k(D)}(p) = \frac{n}{2} < \alpha \cdot s_D(p) \cdot \frac{k}{|D|} = \frac{n+3}{n+4} \cdot \frac{n}{2} + 1 \cdot \frac{n}{n+1} = \frac{n}{2} \cdot \frac{n+3}{n+4} \cdot \frac{n+2}{n+1}$。为证明其是要报告的最一般模式，我们检查 p 的父模式 p'，它有 $m - 1$ 个属性其赋值为 0。对于全局代表性界限，我们有 $s_{R^k(D)}(p') = n - m + 1 = \frac{n}{2} + 1 = L_k$。对于比例代表性，我们有 $s_{R^k(D)}(p') = \frac{n}{2} + 1 > \alpha \cdot s_D(p') \cdot \frac{k}{|D|} = \frac{n+3}{n+4} \cdot \frac{n}{2} + 2 \cdot \frac{n}{n+1} = \frac{n+2}{2} \cdot \frac{n+3}{n+1} \cdot \frac{n}{n+2}$。因此，每个有 $\frac{n}{2}$ 个属性被赋值为 0 的模式都应属于结果集。这类模式的数量为 $\binom{n}{n/2} > \sqrt{2^n}$，呈指数级增长。因此，任何枚举这些模式的算法都具有指数级复杂度。

上界：最一般模式这一概念的提出，源于其所提供信息的效用。例如，在全局代表性界限的情形下，如果前-k 名中的女性人数低于下界，那么黑人女性人数显然也低于该界限。与下界对应的最一般模式不同，在上界的情形下，最具体模式提供的信息更丰富。例如，如果黑人女性人数超过上界，那么黑人总人数和女性总人数也会超过上界。一种合理的问题定义可以考虑最具体的实质性模式。类似于最一般模式的定义，如果模式 p 的规模高于给定阈值 τ_s，且对于满足 $p \subsetneq p'$的每个模式 p'，p' 的规模都低于阈值 τ_s，那么模式 p 就是最具体的实质性模式。因此，目标是找出不满足下界的最一般模式，以及超过上界的最具体实质性模式。为便于阐述，在本文余下部分中，我们将重点讨论仅考虑下界时问题 3.1 和 3.2 的解法。我们指出，我们的解法可以调整，以支持此类问题定义（以及其他定义，例如上界对应的最一般模式和下界对应的最具体模式）。

虽然定理 3.3 表明，具有偏倚代表性的最一般群体数量可能呈指数级增长，但我们的实验评估显示，实际上这类群体的数量要少得多。在 97.58% 的情况下，报告的群体数量少于 100。值得注意的是，呈现大量结果可能会让用户不知所措。用户友好界面可以按 k 值组织输出，并根据群体在数据中的总体规模或其代表性偏差（所需表示与实际表示之间的差异）对群体进行排名。

IV. 检测具有偏倚代表性的群体

接下来，我们介绍用于检测问题 3.1 和 3.2 中定义的具有偏倚代表性的群体的算法。我们首先提出一种简单方案，可根据这两种问题定义检测群体。随后，我们介绍两种优化算法，分别用于优化针对这两个问题的搜索。

A. 迭代式自顶向下搜索（基线方案）

第一个简单解决方案利用 [5] 中介绍的算法遍历可能模式的集合，从最一般的模式开始，计算数据中排名前 k 的排序项里每个组的代表性（针对给定范围内的每个 k）。这是通过使用 模式

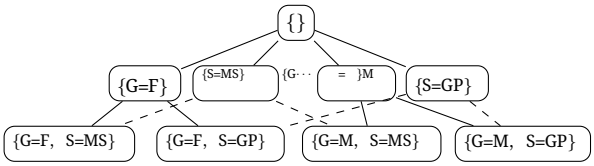


图 3：贯穿全文的示例的部分模式图。搜索树的边以实线标出。

图 [5] 的概念实现的。简而言之，图中的节点是所有可能模式的集合；如果 $p \subset p'$ 和 p' 可以通过向 p 添加一个属性值对得到，则模式 p 和 p' 之间存在一条边。在这种情况下，我们称 p（p'）为 p'（p）的父节点（子节点）。如 [5]，所示，模式图可以采用自顶向下的方式遍历，同时每个模式最多访问一次。也就是说，遍历模式图的一棵生成树，我们将其称为搜索树，并正式定义如下。

定义 *4.1*：设 D 为一个具有类别属性 $\mathcal{A} = \{A_1, \ldots, A_n\}$的数据集。我们假设属性有序，对于给定的属性集 $S \subseteq \mathcal{A}$，用 $idx(S)$ 表示 S中所有属性的最大索引值。设 p 为 D的模式图中的一个节点。搜索树 T 中 p 的子节点为 p'，满足：p' 是模式图中 p 的子节点，且 $idx(Attr(p') \setminus Attr(p)) > idx(Attr(p))$。也就是说，$p'$ 由 p 添加一个属性值对得到，且新添加属性的索引大于 p中属性的最大索引值。

示例 *4.2*：贯穿全文的示例所用数据集对应的模式图中的一部分如图 3 所示，其中 G 和 S 分别是 Gender 和 School 的简写。在模式图中，模式 {G=F, S=GP} 是模式 {G=F} 和 {S=GP} 的子节点；然而，在搜索树中，它只是 {G=F} 的子节点。这是因为 $idx(\{G\}) = 1 < 2 = idx(\{G, S\} \setminus \{G\})$，而 $idx(\{S\}) = 2 > 1 = idx(\{G, S\} \setminus \{S\})$。

算法 1： 自顶向下搜索

输入： 一个数据集 D、一个排序算法 R、一个大小阈值 τ_s, k 以及下界 L_k
输出： $Res = \{p_1, \ldots, p_n\}$ 其中 $\forall p_i \in Res\ s_D(p_i) \geq \tau_s$ 以及 p_i 是一个具有 $s_{R^k(D)}(p) < L_k$的最一般模式
/* 对于比例界，α 作为输入给出 以及 $L_k = \alpha \cdot s_D(p)\ \frac{k}{|D|}$ */

1 $Res \leftarrow \emptyset$
2 $\mathcal{S} \leftarrow$ {生成子项({})}
3 **当** $\mathcal{S}$ 不为空 **时**
4 $p = \mathcal{S}.pop()$
5 **如果** 模式大小 $(p, D) > \tau_s$ **则**
6 $top\text{-}k_c \leftarrow$ 模式大小 $(p, R^k(D))$
/* 对于比例界，请使用 $top\text{-}k_c < \alpha \cdot s_D(p)\ \frac{k}{|D|}$ */
7 **如果** $top\text{-}k\ c < L_k$ **then**
8 更新 (Res, p)
9 **else**
10 $\mathcal{S}.push$(生成子项(p))
11 **返回** Res

算法 1 用于检测数据中规模足够大（即高于给定阈值 τ_s）且代表性较低（低于 L_k）的模式，这些模式出现在给定数据集 D、排序算法 R 和给定 k 所对应的 top-k 排名项中。它自顶向下遍历模式图的搜索树，并使用队列 $\mathcal{S}$，同时维护已识别模式的集合 $s_{R^k}(D) < L_k$。首先，它将结果集 Res 初始化为 $\emptyset$（第 1 行），并将 $\mathcal{S}$ 设为最一般（空）模式的子节点集合（第 2 行）。当队列 $\mathcal{S}$ 非空时（第 3–10 行），算法取出队列 p 中的第一个模式（第 4 行），并计算其在 D 中的大小。若该大小大于 τ_s（第 5 行），则计算 p 在 $R^k(D)$、$s_{R^k}(D)$ 中的大小（第 6 行）。若 $s_{R^k}(D)$ 低于下界 L_k（第 7 行），则使用 update 过程更新 Res（第 8 行）；该过程会检查模式图中 p 的任一祖先是否已在 Res 中（这是可行的，因为算法遍历的是搜索树，而非模式图）。否则（第 9 行），$s_{R^k}(D)$ 超过下界 L_k，于是使用 generateChildren 过程将 p 的子节点加入队列（第 10 行）；该过程按照定义 4.1 所述生成节点的子节点。最后，返回 Res（第 11 行）。

ITERTD 算法（基线）：给定一个数据集 D 和一个排名算法 R、一个大小阈值 τ_s、一个范围 $[k_{min}, k_{max}]$，以及针对每个 $k_{min} \leqslant k \leqslant k_{max}$的下界 L_k，一种基于全局代表性界限定义检测代表性存在偏差的群体的简单方法是使用算法 1，对每个$k_{min} \leq k \leq k_{max}$执行自顶向下搜索。然后，在每次迭代中，报告在排名前k 的排序项中代表性较低的模式。类似地，通过一些细微修改（以算法 1 中的注释形式展示），算法 1 也可用于比例代表性的情形。目标是报告大小足够、但在前k 个元组 $R^k(D)$ 中代表性不足的模式 p，其中 $R^k(D)$ 中的代表性应与 D 中的代表性成比例。在这种情况下，下界 L_k 不作为输入给出，而是根据每个模式在数据中的大小和一个值α计算其界限。请注意，模式的大小在第 5 行计算；给定α 后，我们可以将第 7 行中的条件替换为条件$top\text{-}k\ c < \alpha \cdot s_D(p)\ \frac{k}{|D|}$。

接下来，我们将针对这些问题提出更高效的算法。

B. 全局代表性界限

关键观察是，当 k和 $k+1$[3]的下界保持不变时，针对 k 和 $k+1$的 R 排名结果中具有偏倚代表性的模式，其搜索空间通常非常相似。这是因为前 k 个元组与前 $(k+1)$ 个元组仅相差一个元组。也就是说，增大k 只会对搜索空间进行局部修改。设 T_k为用于查找前 k 个元组中受到不公平对待的模式集合而生成的搜索树，$R(D)[k+1]$为使用 R对 D 进行排名所得结果中的第 $(k+1)$ 个元素。我们可以界定 T_k 中因增大 k而大小受到影响的节点数量。

[3]我们假设 $L_k \leqslant L_{k+1} \forall k\ \ k_{min} \leqslant k < k_{max}$。这是一个合理的假设，因为随着 k 增加，top-k中的元组数量也会增加，因此界限相应增加也合乎逻辑。

命题 *4.3*： $R(D)[k+1]$ 在 T_k中最多可满足 $\frac{|T_k|}{2}$ 个模式（节点），其中 $|T_k|$ 表示 T_k中的节点数量。

证明 4.4：（概要）该证明的基本思路是：每当搜索过程中生成一个模式 p 时，也至少会生成一个与 p 具有相同属性集合（$Attr(p) = Attr(p')$）、但在单个属性的赋值上与之不同的模式 p'。这一结论成立的前提是每个属性至少有 2 个取值。例如，模式 {Gender = M, School = MS} 和 {Gender = M, School = GP} 都是模式 {Gender =M} 的子节点，并且在调用过程 generateChilder({Gender =M}) 时生成。令 A_i 为满足 $A_i \in Attr(p)$ 且 $\{A_i = a_i\} \subseteq p$、同时 $\{A_i = a_i'\} \subseteq p'$ 的属性。$R(D)[k+1]$至多能满足其中一个模式，因为 $R(D)[k+1]$ 中 A_i的取值要么是 a_i，要么是 a_i'（如果$|Dom(A_i)| > 2$，也可能是其他取值）。由于搜索过程中生成的每个节点都满足这一性质，$R(D)[k+1]$至多能满足 $\frac{|(T_k)|}{2}$ 个 T_k 中的模式（节点）。

在这种情况下，从搜索 k 的端点开始搜索 $k+1$，可以显著缩小搜索空间（从而缩短运行时间，见第 VI-B 节）。

算法 2: GLOBALBOUNDS。检测具有偏倚代表性的群体
基于全局界限

输入： 数据集 D、排序算法 R、大小阈值 τ_s、一个范围 $[k_{min}, k_{max}]$，以及每个对象的下界 L_k $k_{min} \leqslant k \leqslant k_{max}$
输出： Res，满足对于每个 $k_{min} \leqslant k \leqslant k_{max}$ $Res[k] = \{p_1, \ldots, p_n\}$ 其中 $\forall p_i \in Res[k]$ $s_D(p_i) \geq \tau_s$ 并且 p_i 是具有 $s_{R^k(D)}(p) < L_k$ 的最一般模式

```
1  Res ← ∅
2  Res[k_min], DRes ← TopDownSearch(D, R, τ_s, k_min, L_{k_min})
3  对于 k = k_min + 1 到 k_max 执行
4      如果 L_{k-1} < L_k 则
5          Res[k], DRes ← TopDownSearch(D, R^k(D), τ_s, L_k)
6      else
7          Res[k] ← Res[k − 1]
8          foreach b ∈ {p ∈ Res[k − 1] | R(D)[k] 满足 p} ∪ DRes 执行
9              Res[k], DRes ← searchFromNode(b, Res[k], DRes)
10 返回 Res
```

GLOBALBOUNDS 算法：我们针对在 top-k 中检测具有全局代表性偏差的群体这一问题提出了优化算法 GLOBALBOUNDS，如算法 2 所示。GLOBALBOUNDS 首先初始化结果集映射 Res（第 1 行）。随后，对于 $k = k_{min}$ 的情况执行自顶向下搜索（第 2 行）。这通过过程 TopDownSearch 完成，该过程与算法 1 中所示的算法类似，仅有一处小改动。它会维护一个集合 $DRes$，其中包含搜索过程中遇到的模式 p：这些模式在 top-k 中的大小低于下界，因此不属于结果集，因为结果集中已经包含了它们的祖先 p。TopDownSearch 会返回搜索的结果集 Res，以及集合 $DRes$。当 k 增大（且 L_k 保持不变）时，算法将利用此集合在模式图中启动局部搜索。

接下来，算法对每个 k 从$k = k_{min} + 1$ 到 $k = k_{max}$ 执行搜索（第 3–9 行）。对于每个 k，如果界限增大，则使用 TopDownSearch 执行新的自顶向下搜索。否则，算法只考虑来自 $DRes$ 的模式，以及满足新插入元组 $R(D)[k]$ 的来自 $Res[k-1]$ 的模式（第 8 行）。这是因为新元组只会影响这些模式在顶端k 中的大小（根据命题 4.3，至多影响树的一半）。对于每个这样的模式，算法都会应用过程 searchFromNode（第 9 行），以在图中相关部分恢复搜索。此搜索会更新 $Res[k]$ 和$DRes$。最后，返回 Res（第 10 行）。

命题 *4.5*： GLOBALBOUNDS 返回给定范围内每个 k 在顶端k 中使用全局界限的、具有偏差表示的所有最一般模式 p 的集合。

证明通过对 k 进行归纳法得出，基本情形为 $k =k_{min}$。由于篇幅限制，省略具体细节。

例 4.6：再次考虑贯穿全文的示例中的 D 和 R。假设给定大小阈值 τ_s =4, $k_{min} = 4, k_{max} = 5$ 以及下界 $L_4 = L_5 = 2$。对 $k = 4$ 完成自顶向下搜索后，结果集 $Res[4]$ 中包含（除其他项外）模式 {Address =U} 和 {Failures = 1}，后者在前 4 个元组中只出现一次（即低于下界）。$DRes$ 中包含的模式包括 {Gender = F, Address = U}、{Gender =M, Address = U}、{Gender =F, Failures = 1} 和 {Address=R, Failures = 1}。这些模式是在自顶向下搜索过程中生成的，并且在 $Res[4]$ 中有祖先（{Address =U} 和 {Failures = 1}）。接下来，算法开始计算 $k = 5$ 的偏倚代表性模式。前 5 个元组中的新元组是元组 14。它匹配 $Res[4]$ 中的模式 {Address =U} 和 {Failures = 1}。因此，算法从这些节点开始搜索。它们在前 5 个元组中的大小超过了下界。在此次搜索中，这两个模式从结果集中移除，并添加模式 {Address =U, Failures = 1}。从集合 $DRes$ 中，将模式 {Gender = F, Address = U}、{Gender = M, Address = U}、{Gender = F, Failures = 1} 和 {Address = R, Failures = 1} 添加到结果集 $Res[5]$ 中，因为它们在前 5 个元组中的大小仍低于阈值 L_5，且各自的祖先已从结果集中移除。

C. 比例表示

接下来，我们考虑检测具有偏差的比例代表性群体这一问题，如问题 3.2 所示。输入包括数据集 D、排序算法 R、范围$[k_{min}, k_{max}]$、大小阈值 τ_s 和 $\alpha \in \mathbb{R}$。目标是报告在 D 中具有足够大小、但在前 k 个元组 $R^k(D)$ 中的代表性不足的模式 p，其中 $R^k(D)$ 中的代表性应与 p 在 D 中的代表性成比例。

首先请注意，对于第四节 IV-B 节中所述的全局代表性界限情形，所提出的优化解并不

适用于此情形。回想一下，GLOBALBOUNDS（算法 2）旨在通过避免搜索连续迭代之间未发生变化的模式图区域来缩小搜索空间。当界限保持不变（$L_k = L_{k+1}$）时，排名中不满足（$k+1$）元组的模式不会受到影响，因此可以从搜索中排除。比例代表性则并非如此，因为每个模式的界限也取决于 k。

回顾一下，目标是找到模式 p 使得$s_{R^k(D)}(p) < \alpha \cdot s_D(p) \cdot \frac{k}{|D|}$，并注意到在计算过程中，$\alpha$和 $s_D(p)$不会发生变化。因此，不等式是否成立取决于 k的值。给定 R, D, α、一个模式 p以及一个值 k，使得 $s_{R^k(D)}(p) \geqslant \alpha \cdot s_D(p) \cdot \frac{k}{|D|}$，我们用 $\mathcal{k}$ 表示固定 $s_{R^k(D)}(p)$的值时使不等式不成立的 k 的最小值。也就是说，使 $s_{R^k(D)}(p) < \alpha \cdot s_D(p) \cdot \frac{k}{|D|}$成立的最小值。 —

例4.7：设 $\alpha = 0.9$。对于贯穿全文的示例中的 D 和 R，p ={性别 = F} 满足 $k = 4$ 的这个不等式，因为 $s_{R^k(D)}(p) = 2 > 1.8 = 0.9 \cdot 8 \cdot \frac{4}{16}$。在这种情况下，$\tilde{k} = 5$，因为 $0.9 \cdot 8 \cdot \frac{5}{16} = 2.25$。

直观而言，如果将 k 增加至 k，但满足 p 的元组数量保持不变，那么 p 在排序结果中的代表性就存在偏差。如果结果集中没有 p 的祖先，那么就应将 p 添加到结果集中。为此，我们优化后的算法 PROPBOUNDS 会计算搜索树中每个模式对应的 k 值。它维护一个集合 $\mathcal{K}$，其中包含可能需要添加到结果集中的模式：如果这些模式不满足排名中的 k 元素，就应将其添加到结果集中。$\mathcal{K}$ 包含搜索树某一分支中的模式，其 k 值单调递减。当计算到达 $k = k$ 时，如果$s_{R^k(D)}(p) = s_{R^{k-1}(D)}(p)$（即，$p$不满足 $R(D)[k]$），则应从 $\mathcal{K}$ 中提取 $\mathcal{K}$ 中的一个模式 p 及其对应的 k 值，并将其添加到结果集中。

算法 3： PROPBOUNDS。检测具有偏差比例代表性的群体

输入： 一个数据集 D、一个排序算法 R、一个大小阈值 τ_s、一个范围 $[k_{min}, k_{max}]$ 和 $\alpha \in \mathbb{R}$

输出： Res，使得对于每个 $k_{min} \leqslant k \leqslant k_{max}$ $Res[k] = \{p_1, \ldots, p_n\}$ 其中 $\forall p_i \in Res[k]$ $s_D(p_i) \geq \tau_s$ 和 p_i 是一个最一般模式，具有 $s_{R^k(D)}(p) < \alpha \cdot s_D(p) \frac{k}{|D|}$

```
1 Res ← ∅
2 Res[k_min], K, DRes ← TopDownSearch(D, R, τ_s, k_min, α)
3 对于 k = k_min + 1 到 k_max 执行
4     Res[k] ← Res[k − 1]
5     Res[k], K, DRes ← selectiveTD (D, R, τ_s, k, α, R(D)[k], K, DRes)
6     对于 p ∈ DRes ⋃_{K[p']=k} {p'} 满足条件 R(D)[k] 不满足 p do
7         更新 (Res, p)
8 返回 Res
```

PROPBOUNDS 算法：PROPBOUNDS（算法 3）的运行过程如下。与 GLOBALBOUNDS 类似，它首先初始化结果集映射 Res（第 1 行），并对 k_{min} 执行自顶向下搜索（第 2 行），如过程 TopDownSearch 所示（其中需要针对比例界进行相应修改）。此外，除了集合 Res、$DRes$之外，它还维护集合 $\mathcal{K}$。然后，算法遍历 k 的取值，从 $k_{min} + 1$ 到 k_{max}（第 3–7 行）。每次迭代中，它首先用上一次迭代的结果 $Res[k-1]$ 初始化 $Res[k]$（第 4 行）。接着，算法使用过程 selectiveTD 从根节点开始进行（部分）搜索（第 5 行）。该搜索会忽略树中不受 $R(D)[k]$影响的区域。过程的结果用于更新结果集 $\mathcal{K}$ 和 $DRes$。随后，算法遍历 $DRes$ 中的模式（搜索过程中访问到且其祖先位于结果集中的模式），以及 $\mathcal{K}$ 中所有具有 $k = k$ 且不受$R(D)[k]$影响的模式，以确定结果集的变化（第 6–7 行）。最后，返回 Res（第 8 行）。

命题 *4.8*：PROPBOUNDS 返回给定范围内每个 k 的 top-k 中具有偏差比例代表性的所有最一般模式 p。

与命题 4.5 类似，证明采用归纳法，基本情形为 $k = k_{min}$。由于篇幅所限，具体细节略去。

例 4.9：再次考虑贯穿全文的示例中的 D 和 R。假设给定大小阈值 τ_s =5、$k_{min} = 4$、$k_{max} = 5$和 $\alpha = 0.9$。在对 $k = 4$ 进行自顶向下搜索结束时，结果集 $Res[4]$ 由以下模式组成：{学校 = GP}、{地址 = U} 和 {失败次数 =1}。对于其中的每一个，$s_D(p) = 8$，因此 $s_{R^k}(D)(p)$ 的上界为 $\alpha \cdot s_D(p) \cdot \frac{k}{|D|} = 0.9 \cdot 8 \cdot \frac{4}{16} = 1.8$，但 $s_{R^k}(D)(p)$ 仅为 1。集合 $\mathcal{K}$ 由 {性别 =M} 和 {性别 = F} 组成，它们的 k 均为 5；此外还包括 {学校 = MS} 和 {地址 R}，其 $k = 7$。请注意，模式 {学校 = MS、地址=R} 是在第一次自顶向下搜索中生成的，但由于其 k 值为 9，高于其在搜索树中的父节点 {学校 = MS}，因此未被添加到 $\mathcal{K}$ 中。

当 k 增加到 5 时，算法只会重新检查受 $R(D)[5]$（元组 14）影响的模式：{性别 = M}、{学校 = MS}、{地址 =U}和 {失败次数 = 1}。模式 {地址 = U} 和 {失败次数 = 1} 仍保留在 $k = 5$ 的结果集中，尽管它们在前 5 名中的大小更大，因为 $k = 5$ 的界限也随之增大。最后，根据来自 $\mathcal{K}$ 的信息，将模式 {性别 = F} 添加到 $Res[5]$中（它存储在带有 $k = 5$的 $\mathcal{K}$ 中）。

V. 结果分析

掌握算法的结果后，分析人员可能希望了解检测到的群体在代表性方面产生偏差的原因。我们提出一种利用 Shapley 值概念来提供此类解释的方法 [33]。Shapley 值是博弈论中的一个概念，用于解释不同属性对模型在给定输入下的输出所产生的影响。近年来，Shapley 值在可解释性领域的应用日益广泛，

并用于解释机器学习模型 [24], [35]。对于回归模型（或具有概率输出的分类器） M，Shapley 值用于评估给定输入 t 时，每个属性对 M 输出的贡献。这是通过使用属性值的所有可能子集，计算每个属性值的加权边际贡献来完成的。

直观而言，偏差的一种解释可能是影响给定群体中元组排名的属性值。为此，我们提出一种由两部分组成的解释方法：第一部分识别对给定群体中元组排名影响最大的属性（使用 Shapley 值），然后比较这些属性在 top-k 和具有偏倚代表性的群体中的值分布。为了采用 Shapley 值，我们需要应对两项挑战。第一，我们需要调整问题设定：给定的是排序算法 R （作为黑箱），而不是回归模型。第二，Shapley 值用于解释单个元组的属性值所作的贡献，而我们感兴趣的是解释元组群体（在 top-k中）代表性不足的问题。

为应对第一个挑战，我们计算一个回归模型 M_R ，用于模拟 R 的过程，并通过计算 M_R(t)的 Shapley 值，近似给定元组t 的属性值对 t排名的影响。为此，我们定义 $D_R = \{(t,\ R(D)[t])\ \mid t \in D\}$，其中$R(D)[t]$ 表示 t 在 $R(D)$中的排名，并用它来训练回归模型 M_R。接着，给定一个模式 p ，使得 p是我们的某个算法针对给定 k检测偏倚代表性群体时返回的结果；为了对此结果作出解释，我们计算每个满足 p的元组 t的 Shapley 值（$s_1^t, \ldots, s_m^t$），也就是对检测到的群体中的每个元组都进行计算。随后，我们将结果汇总为模式 p的单个 Shapley 值向量（$s_1, \ldots, s_m$），使得

$$s_i = \frac{\sum_{t \text{ s.t. } t \text{ satisfies } p} s_i^t}{s_D(p)}$$

为了展示模式 p 与前k 个模式之间的差异，我们将满足模式p 的元组中具有较大 Shapley 值的属性的值分布，与这些属性在前k个元组中的分布进行可视化比较。在第六节 C 部分，我们展示了当排序模型作为黑箱提供时，使用我们的方法能够揭示用于排名的属性（因此会影响前k 个元组中群体代表性的属性）。此外，我们还展示了，对于我们的算法检测到的群体，排名中最重要属性的值分布与前k 个元组中的值分布不同，这表明所识别属性的取值能够解释这些结果。

VI. 实验

我们使用三个真实世界的数据集对所提出的解决方案进行实验检验。我们先介绍实验设置，然后展示一项定量实验研究，旨在评估我们算法的可扩展性。具体而言，我们考察算法在每种公平性定义下的性能，并分析其与属性数量、模式大小阈值、以及 k 的范围之间的关系。随后，我们展示所提出的方法，并对第五节中呈现的结果进行分析。最后，我们通过案例研究，将我们的算法与 [27] 的算法进行比较，展示两者结果的差异，并将其与 [27], 中介绍的框架进行对照。

A. 实验设置

a) 数据集：我们使用了三个规模（元组数和属性数）各不相同的真实数据集，具体如下。

- **COMPAS 数据集**[4] 由 ProPublica 收集并发布，是其调查犯罪风险评估软件中种族偏见的一部分。该数据集包含 6,889 名个体的人口统计信息、 COMPAS 软件生成的再犯评分以及犯罪信息。我们最多使用 16 个属性，并剔除了姓名、ID、日期等属性。

- **学生表现数据集（学生数据集）**[5] 展示了如例 2.1 所述的两所葡萄牙学校中接受中等教育的学生的表现。在实验中，我们考虑了包含数学考试相关信息的数据片段（ 395 个元组和 33 个属性）。

- **德国信用数据集**[6] 包含 1,000 名贷款申请人的财务和人口统计信息，共有 20 个属性。该数据集最初用于分类任务，在这一任务中，每份申请都会被判定为信用风险良好或不良。

比较的算法：我们根据运行时间评估所提出算法的性能。

- **IterTD（基线）**。用于检测具有偏倚代表性的群体的简单解决方案，通过迭代应用第四节 A 部分所示的自顶向下搜索。

- **GLOBALBOUNDS （算法 2）**。该算法基于第四节 B 部分所述的全局界限，检测具有偏倚代表性的群体。

- **PROPBOUNDS （算法 3）**。该算法基于第四节 C 部分所述的比例代表性界限，检测具有偏倚代表性的群体。

参数设置：由于空间限制，除非另有说明，否则我们报告以下默认参数集合下的结果： $\tau_s = 50$、 $k_{min} = 10$、 $k_{max} = 49$；对于全局界限， $10 \le k < 20$、 $20 \le k < 30$、$30 \le k < 40$ 和 $40 \le k < 50$ 的下界分别为 10、20、30 和 40；对于比例代表性， $\alpha = 0.8$ 。报告的结果反映了算法在预期和典型使用场景下的性能。根据排名公平性的目标（确保任意位置的公平性），随着 k 的增加，我们逐步提高前k 个排序项中群体代表性的界限。由于我们的目标是向用户报告检测到的群体，因此我们设置参数，使得大多数情况下报告的群体数量介于 1 到 100 之间。属性数量设置为基线方案能够处理的最大数量，而连续属性（例如，

[4]https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis [5] https://archive.ics.uci.edu/ml/datasets/student+performance [6] https://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data)

年龄）根据其域和值等分为 $3 - 4$ 个分桶。我们注意到，分桶的选择会影响模式图的大小，也可能影响可能的群体定义及其在前k个排序项中的代表性，进而也可能影响运行时间。然而，在本研究中，我们假设用于群体定义的属性值是类别型的（即分桶已给定）。我们使用不同的参数设置对算法进行了实验评估，并观察到相似的趋势。

排序算法：学生数据集根据属性 G3 的值进行排序，该属性反映学生的数学期末成绩。对于 COMPAS 数据集，我们采用了与 [4] 中类似的排序方法：将属性值 c days from compas、juv other count、days b screening arrest、start、end、age 和 priors count 归一化后作为评分属性。数值按 $(val - min)/(max - min)$ 进行归一化。除年龄外，数值越高，得分越高。元组按得分降序排列。对于德国信用数据集，我们采用了 [36] 中基于信用 worthiness 的排序方式。

所有实验均在一台搭载 2.8 GHz 四核 Intel Core i7 CPU 和 8GB 内存的 macOS 设备上进行。算法使用 Python 3.7 实现。

B. 实验结果

GLOBALBOUNDS 和 PROPBOUNDS 的运行速度都远快于基线，尤其是随着属性数量增加、基线的开销呈指数级增长时更是如此。详情如下。

属性数量：第一组实验旨在研究属性数量对运行时间的影响。为此，我们将数据集中的属性数量从 3 调整到 $|\mathcal{A}|$，其中 $\mathcal{A}$ 是数据集中所有属性的集合。属性数量（及其基数）决定了可能模式的数量，进而决定了搜索空间的大小。因此，随着属性数量增加，我们预计运行时间会急剧增长。结果（采用 10 分钟超时设置）如图 4–5 所示。事实上，在所有情况下，我们都观察到运行时间迅速增加，同时 GLOBALBOUNDS 和 PROPBOUNDS 的表现优于 ITERTD。

大小阈值：在下一组实验中，我们评估了大小阈值 τ_s 对运行时间的影响。为此，我们将大小阈值从 10 调整到 100，其余参数采用默认值。结果如图 6 和图 7 所示。我们观察到算法的运行时间有所缩短。这是因为随着阈值增大，满足大小阈值的模式数量会减少，从而搜索空间也随之缩小。在所有情况下，GLOBALBOUNDS 和 PROPBOUNDS 的表现都优于 ITERTD。

范围 k：我们考察了算法所考虑的 k 范围对可扩展性的影响。我们将 k_{min} 设为 10，并将 k_{max} 从 50（50）逐步增加至 1000（350），从而将 k 的范围从 40（40）扩大至 990（340），分别针对 COMPAS、Student 和德国信用数据集观察其对运行时间的影响。由于数据集大小不同（COMPAS 为 6889，Student 为 395，德国信用为 1000），我们为 k 设置了不同的最大范围。结果如图8和图9所示。在所有情况下，优化算法的表现均优于 ITERTD，这说明搜索空间得到了有效缩减。

请注意，与 ITERTD 相比，GLOBALBOUNDS 和 PROPBOUNDS 通过利用 k 的迭代搜索结果来计算 $k+1$ 的结果集，从而优化搜索空间。因此，随着 k 的范围增大，我们预计优化算法相较于基线方案的性能提升会更加显著。图8和图9展示了这一趋势。为进一步证明该方法的实用性，我们比较了各算法在搜索过程中检查的模式数量。在使用全局界限检测偏倚代表性群体时，COMPAS 数据集、学生数据集和信用卡数据集中的搜索量最多分别减少了 39.35%、56.87% 和 29.27%；而在检测比例代表性时，分别减少了 39.60%、20.49% 和 56.83%。

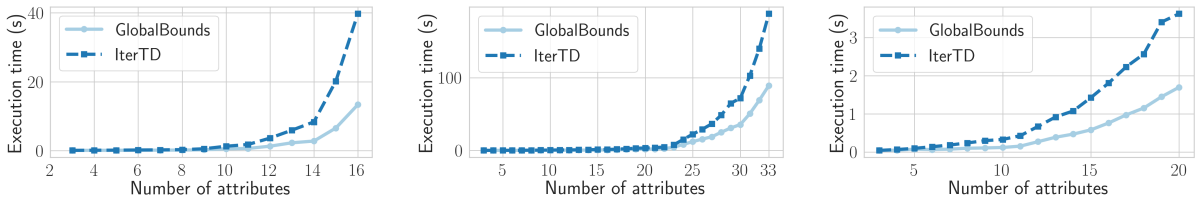
C. 使用 *Shapley* 值进行结果分析

接下来，我们将展示第 V 节提出的基于 Shapley 值的结果分析方法的实用性。本实验有两个目标。首先，我们将说明，当排序算法作为黑箱给出时，我们提出的基于 Shapley 值的方法能够评估属性对排名的影响，并揭示排序实际使用的属性信息。然后，我们通过比较 top-k 中各值的分布与检测到的群体中的值分布，并着重分析二者的差异，展示这些属性的值分布如何用于解释代表性偏差。

我们使用每个数据集的排名数据训练了一个回归模型，并考察了算法检测出的群体的 Shapley 值。我们展示了学生数据集中模式（群体） $p_1 = \{$母亲的教育程度 $=$ 初等教育（四年级）$\}$ 、COMPAS 中 $p_2 = \{$年龄 $=$ 小于 $35\}$，以及德国信用数据集中 $p_3 = \{$现有账户状态$=$（$0 \leqslant \cdots < 200$）DM[7]$\}$ 的结果；这些模式由 GLOBALBOUNDS 算法针对$k = 49$ 和 $L_k = 40$检测得出。对于算法检测出的其他群体及其他参数，我们也观察到了类似结果。图 10a、10b 和 10c 展示了各群体对应的聚合 Shapley 值，具体说明见第五节。我们展示了每个群体中数值最大的六个属性的 Shapley 值，因为其余属性的值显著更低（分别低于 p_1、 p_2 和 p_3 对应的最大聚合 Shapley 值的 5.79%、3.31% 和 9.54%）。

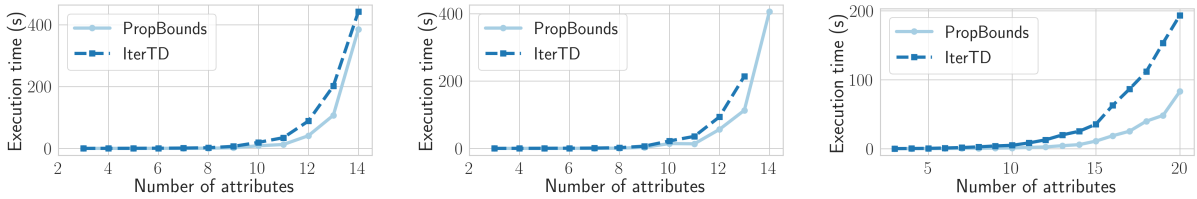
对于母亲受教育程度为初等教育的学生组，我们的算法检测出该组在 Student 数据集中具有偏倚代表性，

[7]公司的银行对账单上的借记通知单（DM）是指银行从公司银行账户中扣除款项。换句话说，银行借记通知单会减少银行账户余额，效果类似于从该银行账户开出的支票。



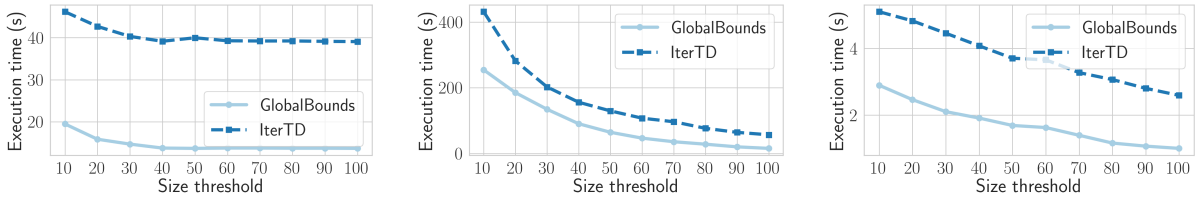
(a) COMPAS 数据集　(b) 学生数据集　(c) 德国信用

图 4：具有全局界限的排名中，运行时间随属性数量的变化



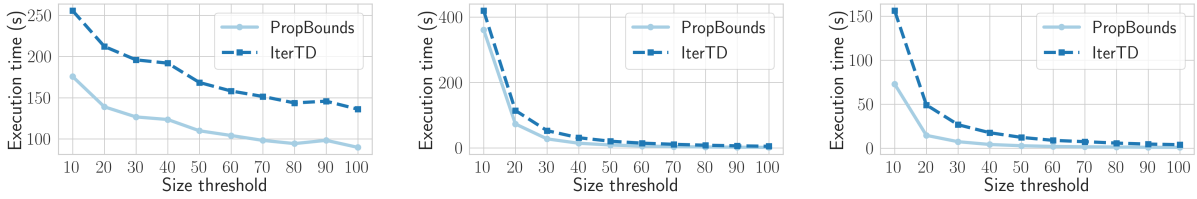
(a) COMPAS 数据集　(b) 学生数据集　(c) 德国信用

图 5：具有比例代表性的排名中，运行时间随属性数量的变化



(a) COMPAS 数据集　(b) 学生数据集　(c) 德国信用

图 6：采用全局界限的排名中，运行时间随大小阈值 τ_s 的变化



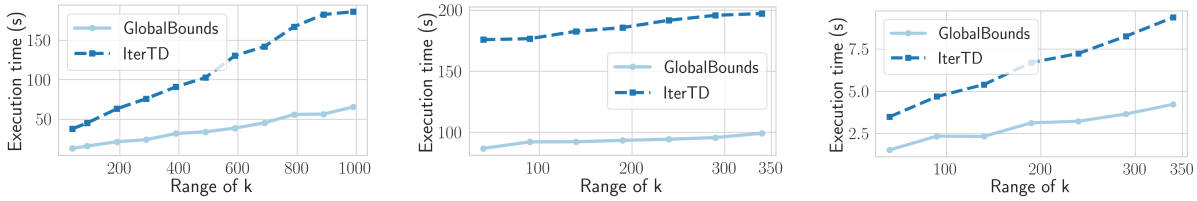
(a) COMPAS 数据集　(b) 学生数据集　(c) 德国信用

图7：大小阈值 τ_s 的运行时间——具有比例代表性的排名

最终成绩在排名上的聚合 Shapley 值最大（图10a）。这一结果符合实际情况：排序算法确实使用最终成绩进行排名（事实上，它是唯一使用的属性）。除最终成绩外，第一阶段和第二阶段成绩也具有显著的聚合 Shapley 值（尽管远低于最终成绩的聚合 Shapley 值）。这是因为这些属性与最终成绩高度相关 [13]。我们还注意到，结果中包含母亲的教育程度这一属性。这可能表明母亲的教育程度与最终成绩存在某种相关性。不过，我们也注意到，其他属性（例如父亲的教育程度）的 Shapley 值聚合后并未呈现明显模式：一些值产生正面影响，另一些则产生负面影响，而且组内不同元组的值各不相同。相比之下，组内所有元组的母亲的教育程度属性值都相同（均为初等教育）（因为它用于定义该群体）。因此，群体中不同元组的该属性的 Shapley 值相近。与其他属性相比，这也可能会提高聚合值。我们还在算法检测到的其他群体中观察到了这一现象：用于定义检测到的群体的属性，其值也略高。

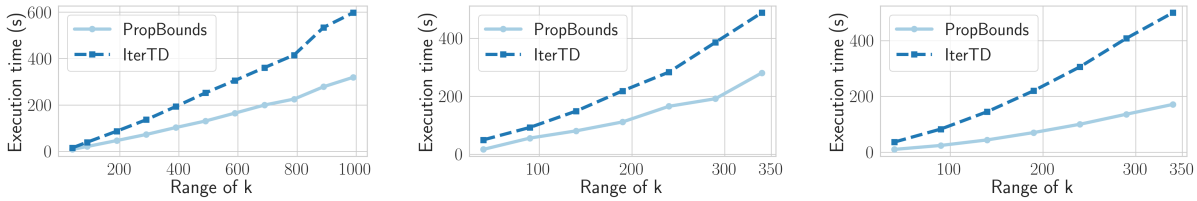
对于 COMPAS 数据集，元组根据一项综合评分进行排名，该评分基于七个属性：距 COMPAS 的天数、其他未成年定罪记录的数量、距筛查逮捕的天数、开始日期、结束日期、年龄以及先前犯罪的次数。在图10b中，展示了低于35岁人群的聚合 Shapley 值，其中上述七个属性中的六个是 Shapley 值最大的六个属性。在这种情况下，结束日期和先前犯罪的次数被确定为影响检测到的群体的最重要因素。

对于德国信用数据集，元组按照 [36]，中的排名进行排序；然而，实际的排序算法



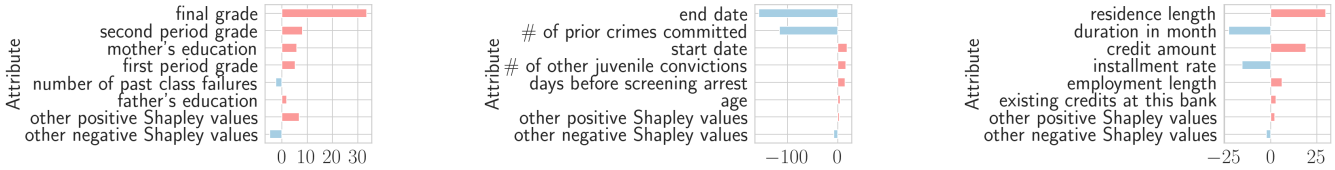
(a) COMPAS 数据集　　(b) 学生数据集　　(c) 德国信用

图 8：在全局界限下， k-排名的运行时间随范围变化的情况



(a) COMPAS 数据集　　(b) 学生数据集　　（c）德国信用

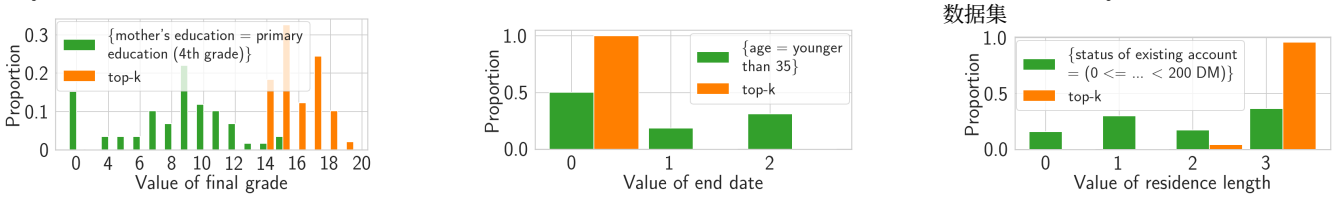
图 9：运行时间与 k -比例代表性排名的范围的关系



（a）组的聚合 Shapley 值 p_1 ={母亲的教育程度 = 小学教育-程度}在 Student 数据集中

（b）组的聚合 Shapley 值 p_2 ={年龄 = 年轻于 35} 在 COMPAS 数据集

(c) 组的聚合 Shapley 值 p_3 ={现有账户状态 =(0 ⩽ ··· < 200DM)} 德国信用数据集



(d) 学生数据集中最终成绩属性的值分布

(e) COMPAS 数据集中结束日期属-性中的值分布

(f) GermanCredit 数据集中居住时长的值分布，

图10：使用 Shapley 值进行结果分析

未知。也就是说，我们没有真实值，无法验证检测到的属性是否确实被排序算法用于解释排名前 k 中检测到的组在代表性方面存在偏差。图10c显示，居住时长、月数、信贷金额和分期付款率这几个属性的 Shapley 值最大。所有这些属性都是用于判断一个人信用worthiness的合理特征。

Shapley值表示不同属性对群体排名的影响。为分析检测到的群体与排名前k 个元组在这些属性上的差异，我们在图10d、10e和10f中展示了 Shapley 值最大的属性的值分布。由于排名前k 的元组数量与检测到的群体数量不同， y轴表示具有x轴所示值（该属性的可能值集合）的元组比例，而非元组数量。

对于全部三个数据集，我们观察到，具有最大 Shapley值的属性的值分布在前k 个元组与检测到的具有偏倚代表性的组中的元组之间存在巨大差异。对于学生数据集（图10d），前 k 个元组的最终成绩全部落在 $15-20$范围内，而检测到的组中的大多数元组的最终成绩低于15。在 COMPAS 数据集（图10e）中，所有前k个元组的结束日期值均为0，而检测到的组中只有一半元组具有相同的值，且其中将近30%的元组的值为2。在图10f所示的居住时长属性的值分布中，也观察到了类似结果。

D. 与现有解决方案的比较

在 [27] 中，研究了如何识别数据中表现与整体数据集不同的子群体。与我们的问题定义不同，我们的问题定义依赖于排名公平性度量，以定义在排名前-k 的排序项中具有偏倚代表性的群体； [27] 的工作则使用

差异度这一概念来衡量数据子组之间的性能差异。数据 $t \in D$ 中的每个数据项都关联一个结果 $o(t)$，其中结果函数是根据排序算法对 t 的排名定义的。群体 $o(G)$ 的结果，是其中每个数据项 $t \in G$ 的结果的平均值。数据 D 中子群体 G 的差异度，是 G 与整体数据之间的差异，即 $o(G) - o(D)$。给定子组大小阈值 s，[27] 的解会计算大小大于 s 的所有 子群体的差异度。

为更好地展示各定义及其对应算法所识别出的群体之间的差异，我们使用 Student 数据集进行了实验。我们采用默认大小阈值 $\tau_s = 50$ （ [27] 中的支持度为 0.13，即数据的 13%）。由于 [27] 不考虑 k的范围，我们固定了 $k_{min} = k_{max} = 10$ （即比较 $k = 10$时的结果）。为便于比较，我们仅使用数据中的前 4 个属性：学校、性别、年龄和地址。我们采用结果函数 $o(t)$，为 t 位于前 k 的元组赋值 1，其余元组赋值 0（如 [27]所示）。最后，对于我们的算法，排名采用全局界限时的默认下界为 10，比例代表性的默认参数为 $\alpha = 0.8$ 。

PROPBOUNDS 输出了 2 个模式：{性别=F} 和{地址=R}，GLOBALBOUNDS 也返回了这两个模式。此外，GLOBALBOUNDS 还返回了模式{学校=GP}、 {性别=M} 和 {地址=U}。这些模式在排名前 10 的排序项中分别少于 10 个实例（分别为 9、7 和 9 个）；但考虑到它们的总体规模（分别为 349、208 和 307），它们在排名前 k 中的代表性足够，因此 PROPBOUNDS 不会返回这些模式。 [27]的算法返回了 28 个群体，其中包括我们的解决方案检测出的群体。由于报告的群体数量可能极其庞大， [27] 的算法会根据群体的差异度对其进行排名。差异度最高的 5 个模式都包含属性 $3 - 5$ ，其赋值为性别=M；也就是说，它们是 GLOBALBOUNDS 返回的模式 {性别=M} （在模式图中）的后代。模式{性别=M} 的差异度排名为第 17。

我们的算法与 [27] 的解决方案主要区别在于如何定义目标组。两种解决方案处理的问题相似，但我们的方法倾向于采用简洁组（最一般模式），而 [27] 的解决方案旨在识别在整体数据中代表性充足且差异度较高（衡量“不公平性”的指标）的所有 群体。因此， [27] 的输出通常更大，并包含彼此相互包含的子群体。最后， [27] 的工作只考虑单个 k ，而我们考虑一系列 k，这与文献中的公平性定义相符，使得该解决方案对排名前 k 中的任意位置都公平。

VII. 相关工作

排名算法中的公平性概念已在一系列研究中得到探讨，并提出了不同的公平性定义 [10]–[20], [30], [34], [36], [38]。这些定义通常关注前 k 个位置，因为它们往往最为重要的位置。本文考察其中两种定义：一种基础定义 [10], ，通过限制数据中不同群体的代表性来衡量公平性；另一种经过改进的定义则考虑群体代表性的比例关系。按照算法公平性领域的惯例，这些定义针对的是一些已知的受保护群体。我们利用这些定义来界定检测具有偏倚代表性的群体这一问题，从而不必预先定义受保护群体。生成公平排序结果的问题已在 [4], [38] 中得到研究。这些研究考察了排名公平性的各种定义，而这些定义都依赖于受保护群体这一概念。该研究方向与本文定义的问题相互独立；当受保护群体事先未知时，我们提出的方法可用于识别这类群体。

近期研究探讨了如何在分类 [9], [12], [27], [28] 的背景下，自动检测数据中“有问题”或存在偏差的子群体，而无需事先指定受保护属性。在 [28], 中，作者提出了差异度的概念，用来衡量分类器在不同数据子组上的行为差异。其目标是报告规模超过给定阈值且差异度较高的子群体。在 [27] 中，他们将框架扩展到排名领域，并将平均结果值作为衡量组结果的指标；该值根据各组中实例的排名来定义。与 [27], 不同，我们的问题定义依据群体在排名前 k 个排序项中的代表性，将其作为排名的公平性度量。我们在第 VI-D 节展示了各定义所识别群体的差异。交互式系统 MithraCoverage [21] 研究了交叉群体中的人口偏差。该系统引入覆盖度的概念，用以识别数据集中代表性不足的交叉子群体。与此不同，我们的工作只报告在数据中代表性充足、但在排序算法输出中代表性不足的模式。

已有一系列研究探讨了如何使用 Shapley 值为机器学习模型提供解释（例如，参见 [24]–[35]）。在这些研究中，针对分类或回归模型的单个输入实例计算 Shapley 值。随后，将某个特征的 Shapley 值解释为该特征对给定输入的输出所作的贡献。与此不同，我们旨在解释排序算法输出中一组元组的代表性。在 [28] 中，作者提出了一种利用 Shapley 值衡量各项对群体差异度贡献的方法。然而，该方法仅考虑用于定义群体的属性所作的贡献。相比之下，我们的方法将所有属性都视为可能的解释。这需要在 Shapley 值的计算中增加一个聚合步骤。如 VI-C 所示，解释通常隐藏在用于排名的属性值中。此外，我们对 Shapley 值所作的调整，以解释排序算法，是一项新的工作。

我们的基线方案采用了 IV-A 节中介绍的自顶向下搜索，其基础是 [5]中提出的算法

（针对一个较简单的问题），该算法又借鉴了 Apriori算法 [1], 中用于以最佳优先方式枚举集合的集合枚举树 [32], 中用于发现函数依赖（FD）的方法 [19], [26] 以及用于发现频繁项集和进行关联规则挖掘的方法 [1], [37]。与 [5],类似，我们的工作与其他工作的关键区别在于解中遍历的图结构：模式图（在我们的工作中）与其他工作中的幂集格。条件函数依赖（CFD） [15] 通过考虑用于描述仅在数据的子群体上成立的依赖关系的模式，扩展了函数依赖的概念。与基线方案采用的自顶向下搜索类似，用于发现 CFD 的算法 [14], [16], [18], [31] 也采用了模式和模式格的概念。然而，最终目标不同（发现 CFD 与识别具有偏倚代表性的群体），因此基线方案中的剪枝技术也有所不同。随后，我们针对所定义的每个问题提出了两种新颖的优化算法。如第四节所述，这些算法缩小了搜索空间，并且如实验评估所示，其性能显著优于基线方案。

VIII. 结论

本文研究了如何检测排序算法结果中具有偏倚代表性的群体。我们基于文献中先前定义的公平性度量，考察任意合理 k范围内，前k 个排序项中受保护群体的代表性。我们的问题定义消除了预先定义受保护群体的需要。我们考虑了该问题的两种变体。第一种基于前k 个排序项中不同群体代表性的全局界限；第二种则根据每个群体在数据中的总体代表性，限制其在前k个排序项中的代表性。

我们从理论上分析了该问题的复杂度，表明在最坏情况下，群体数量可能随数据集属性数量呈指数级增长。我们提出了一种可处理这两种定义的基线算法，以及两种旨在提升各公平性度量下性能的优化算法。此外，我们还提出了一种解释算法输出的方法。未来研究有许多值得探索的方向，包括扩展该框架以支持其他公平性度量，以及进一步研究阈值的自动建议。

致谢

我们谨此感谢匿名审稿人提出的评论和建议。本研究部分得到了美国国家科学基金会（NSF）项目资助，项目编号为1741022、1934565和2106176。

参考文献

[1] Rakesh Agrawal 和 Ramakrishnan Srikant。《大型数据库中挖掘关联规则的快速算法》。载于 VLDB 1994，第487–499页。Morgan Kaufmann，1994年。[2] Kristen M. Altenburger、Rajlakshmi De、Kaylyn Frazier、Nikolai Avte- niev 和 Jim Hamilton。《职业自我推销中是否存在性别差异？对近期工商管理硕士毕业生 LinkedIn个人资料的实证案例研究》。载于《第十一届国际网络与社交媒体会议》、ICWSM 2017，蒙特利尔，魁北克省，加拿大，2017年5月15日至18日，第460–463页。AAAI出版社，2017年。[3] Abolfazl Asudeh、H. V. Jagadish、Julia Stoyanovich 和 Gautam Das。《设计公平排序方案》。载于《2019年国际数据管理会议论文集》、SIGMOD 2019会议，阿姆斯特丹，荷兰，2019年6月30日至7月5日，第1259–1276页。ACM，2019年。[4] Abolfazl Asudeh、HV Jagadish、Julia Stoyanovich 和 Gautam Das。《设计公平排序方案》。载于《2019年国际数据管理会议论文集》，第1259–1276页，2019年。[5] Abolfazl Asudeh、Zhongjun Jin 和 H. V. Jagadish。《评估并补救给定数据集的覆盖度》。载于 ICDE，2019年。[6] Solon Barocas 和 Andrew D Selbst。《大数据的差别影响》。《加利福尼亚州法律评论》，104:671，2016年。[7] Tobias Berg、Valentin Burg、Ana Gombovi´c 和 Manju Puri。《金融科技公司的崛起：使用数字足迹进行信用评分》。《金融研究评论》，33(7):2845–2897，2020年。[8] Sergey Brin 和 Lawrence Page。《大型超文本网络搜索引擎的剖析》。《计算机网络》，30(1-7):107–117，1998年。[9] ´Angel Alexander Cabrera、Will Epperson、Fred Hohman、Minsuk Kahng、Jamie Morgenstern 和 Duen Horng Chau。《Fairvis：用于发现机器学习中交叉性偏见的可视分析》。载于 VAST，2019年。[10] L. Elisa Celis、Damian Straszak 和 Nisheeth K. Vishnoi。《带公平性约束的排名》。载于《第45届国际自动机、语言与编程研讨会》、ICALP 2018，2018年7月9日至13日，布拉格，捷克共和国，第107卷 LIPIcs，第28:1–28:15页。Schloss Dagstuhl - Leibniz-Zentrum f¨ur Informatik，2018年。[11] Le Chen、Ruijun Ma、Anik´o Hann´ak 和 Christo Wilson。《研究性别对简历搜索引擎中排名的影响》。载于《2018年计算机系统中的人为因素会议论文集》、CHI 2018，蒙特利尔，魁北克省，加拿大，2018年4月21日至26日，第651页。ACM，2018年。[12] Yeounoh Chung、Tim Kraska、Neoklis Polyzotis、Ki Hyun Tae 和 Steven Euijong Whang。《用于模型验证的自动化数据切片：一种大数据—人工智能集成方法》。《IEEE知识与数据工程汇刊》，32(12):2284–2296，2020年。[13] Paulo Cortez 和 Alice Maria Gonc, alves Silva。《使用数据挖掘预测中学生学业表现》。2008年。[14] Thierno Diallo、N¨oel Novelli 和 Jean-Marc Petit。《发现（频繁的）常量条件函数依赖》。《国际数据挖掘、建模与管理期刊》，4 223，2012 3 205( ): – 。[15] Wenfei Fan、Floris Geerts、Xibei Jia 和 Anastasios Kementsietsidis。《用于捕获数据不一致性的条件函数依赖》。《ACM数据库系统汇刊》（TODS），33(2):1–48，2008年。[16] Wenfei Fan、Floris Geerts、Jianzhong Li 和 Ming Xiong。《发现条件函数依赖》。《IEEE知识与数据工程汇刊》，23(5):683–698，2010年。[17] Sahin Ce m Geyik、Stuart Ambler 和 Krishnaram Kenthapadi。《搜索与推荐系统中的公平性感知排名及其在 LinkedIn人才搜索中的应用》。载于《第25届 ACM SIGKDD知识发现与数据挖掘国际会议论文集》、KDD 2019，安克雷奇，阿拉斯加州，美国，2019年8月4日至8日，第2221–2231页。ACM，2019年。[18] Lukasz Golab、Howard Karloff、Flip Korn、Divesh Srivastava 和 Bei Yu。《为条件函数依赖生成近似最优的表》。《VLDB学报》，1 390，2008 1 376( ): – 。[19] Yk¨a Huhtala、Juha K¨arkk¨ai nen、Pasi Porkka 和 Hannu Toivonen。《TANE：发现函数依赖和近似依赖的高效算法》。《计算机学报》，42(2):100–111，1999年。[20] Kalervo J¨arvelin 和 Jaana Kek¨al¨ainen。《基于累积增益的信息检索技术评估》。《ACM信息系统汇刊》（TOIS），20(4):422–446，2002年。[21] Zhongjun Jin、Mengjing Xu、Chenkai Sun、Abolfazl Asudeh 和 HV Ja- gadish。Mithracoverage：一个用于研究交叉公平性中群体偏差的系统。载于*2020*年 *ACM SIGMOD* 数据管理国际会议论文集，第2721–2724 页，2020 年。[22] Caitlin Kuhlman、MaryAnn Van Valkenburg 和 Elke A. Rundensteiner。FARE：使用成对误差度量诊断公平排序。载于万维网大会，*WWW 2019*，旧金山，加利福尼亚州，美国，*2019*年*5*月*13*日至*17*日，第 2936–2942 页。ACM，2019年。[23] Jinyang Li、Yuval Moskovitch 和 H. V. Jagadish。DENOUNCER：检测分类器中的不公平性。*VLDB Endowment* 会议录，14(12):2719–2722，2021年。[24] Scott M. Lundberg 和 Su-In Lee。一种解释模型预测的统一方法。载于神经信息处理系统进展第*30*卷：*2017*年神经信息处理系统大会，*2017*年*12*月*4*日至*9*日，长滩，加利福尼亚州，美国，第 4765–4774 页，2017年。[25] Shira Mitchell、Eric Potash、Solon Barocas、Alexander D’Amour 和 Kristian Lum。算法公平性：选择、假设与定义。《统计学及其应用年度评论》，8:141–163，2021年。[26] Thorsten Papenbrock、Jens Ehrlich、Jannik Marten、Tommy Neubert、Jan-Peer Rudolph、Martin Sch¨ onberg、Jakob Zwiener 和 Felix Nau- mann。函数依赖发现：对七种算法的实验评估。*VLDB Endowment* 会议录，8(10):1082–1093，2015年。[27] Eliana Pastor、Luca de Alfaro 和 Elena Baralis。识别排名与分类中的偏差子群。*arXiv* 预印本 *arXiv:2108 07450*，2021年。。[28] Eliana Pastor、Luca de Alfaro 和 Elena Baralis。寻找问题：通过模式偏离分析分类器行为。载于*2021*年数据管理国际会议论文集，第1400–1412 页，2021年。[29] Christopher Peskun、Allan Detsky 和 Maureen Shandling。医学院招生标准在预测四年后的住院医师排名方面的有效性。《医学教育》，41(1):57–64，2007年。[30] Evaggelia Pitoura、Kostas Stefanidis 和 Georgia Koutrika。排名与推荐中的公平性：综述。*VLDB J.*，31 458，2022 3 431( ): – 。[31] Joeri Rammelaere 和 Floris Geerts。重新审视条件函数依赖发现：从“fd”中拆分出“c”。载于欧洲机器学习与数据库知识发现联合会议，第 552–568 页。施普林格，2019年。[32] Ron Rymon。通过系统化集合枚举进行搜索。载于*KR*。Morgan Kaufmann，1992年。[33] L Shapley。7. n人博弈的一个值。《博弈论的贡献 II》（1953）307-317。载于《博弈论经典著作》，第 69–79 页。普林斯顿大学出版社，2020年。[34] Ashudeep Singh 和 Thorsten Joachims。排名中的曝光公平性。载于 Yike Guo 和 Faisal Farooq（编），第*24*届 *ACM SIGKDD*知识发现与数据挖掘国际会议论文集，*KDD 2018*，伦敦，英国，*2018*年*8*月*19*日至*23*日，第 2219–2228 页。ACM，2018年。[35] Erik Strumbelj 和 I gor Kononenko。利用特征贡献解释预测模型与个体预测。知识与信息系统，41(3):647–665，2014年。[36] Ke Yang 和 Julia Stoyanovich。衡量排序输出的公平性。载于第*29*届科学与统计数据库管理国际会议论文集，芝加哥，美国，*2017*年*6*月*27*日至*29*日，第 22:1–22:6 页。ACM，2017年。[37] Mohammed Javeed Zaki。可扩展的关联挖掘算法。《*IEEE*知识与数据工程汇刊》，12 390，2000 3 372( ): – 。[38] Meike Zehlike、Francesco Bonchi、Carlos Castillo、Sara Hajian、Mo- hamed Megahed 和 Ricardo Baeza-Yates。Fa* ir：一种公平的 top-k 排名算法。载于*CIKM 2017*，第 1569–1578 页，2017年。[39] Meike Zehlike、Francesco Bonchi、Carlos Castillo、Sara Hajian、Mo- hamed Megahed 和 Ricardo Baeza-Yates。Fa*ir：一种公平的 top-k 排名算法。载于*CIKM 2017*，新加坡，*2017*年*11*月*6*日至*10*日，第 1569–1578 页。ACM，2017年。