arXiv:2301.02053v1 [cs.DS] 5 Jan 2023

带公平性约束的最大-最小多样化：精确算法与近似算法

Yanhao Wang* Michael Mathioudakis† Jia Li‡ Francesco Fabbri§

摘要

多样性最大化旨在从大型数据集中选出多样且具有代表性的项目子集。这是一项基础优化任务，应用于数据摘要、特征选择、网络搜索、推荐系统等领域。然而，当数据项根据性别或种族等敏感属性归属不同分组时，如果不加约束，此任务的算法解决方案可能会导致某些分组代表不足或过度。因此，我们致力于解决 *max-min diversification with fairness constraints*问题，旨在选出 k 个项目，使任意一对所选项目之间的最小距离最大化，同时确保从每个组中选出的项目数量处于预先设定的上下界之间。在这项工作中，我们提出了一种基于整数线性规划的精确算法，适用于小型数据集；此外，还提出了一种 $\frac{1-\varepsilon}{5}$ -近似算法，适用于任意 $\varepsilon \in (0, 1)$，可扩展至大型数据集。在真实世界数据集上进行的大量实验表明，我们提出的算法优于现有算法。

关键词： 最大-最小多样化，算法公平性

1 引言

近年来，算法日益广泛地应用于日常生活的许多领域，以数据驱动自动化决策。这引发了人们的担忧：算法可能会对特定人口群体作出不公平且带有歧视性的决策，尤其是在投票、招聘、银行业、教育和刑事司法等敏感的社会计算领域 [12, 25]。为缓解这些担忧，大量研究致力于将公平性纳入自动化决策任务中的算法，包括分类 [14], 聚类 [10], 排序 [24, 32], 匹配 [28], 和数据摘要 [8, 20]。

本文研究多样性最大化问题，并探讨其公平性感知变体。该问题旨在从给定数据集中选取一个多样化的项目子集，常见于数据摘要 [8, 23], 网络搜索 [2], 推荐 [21], 特征选择 [31], 等领域 [34]。关于多样性最大化问题的现有文献主要关注两个目标，即最大-最小多样化 （MMD），其目标是最大化任意一对所选项目之间的最小距离；以及最大-总和多样化 （MSD），其目标是最大化所选项目之间成对距离的总和。如图 1 所示，MMD 倾向于均匀覆盖数据范围，而 MSD 倾向于挑选“离群点”，并可能在解中包含高度相似的项目。由于 MMD 所刻画的多样性更能体现数据摘要、特征选择以及许多其他任务的解所追求的属性，本文将仅考虑 MMD。具体而言，给定度量空间中的一个集合 V，其中包含 n 个项目，以及一个正整数$k \leq n$，MMD 要求从中选出大小为-k 的子集 S，使 V中的最小成对距离最大化。

具体而言，我们研究公平最大-最小多样化（FMMD）问题，这是 MMD 的一种变体，其目标不仅是最大化上述定义的多样性度量，还要确保满足下文所述的群体公平性约束。设 V中的所有项目根据性别或种族等敏感属性划分为 C个互不相交的组 V_1，.....， V_C。为确保公平代表性，从每个组 V_c中选取的项目数量（其中 $c \in [1, \ldots, C]$）须限制在作为输入指定的上下界之间。这一定义的群体公平性约束涵盖并推广了多种现有的群体公平性概念，包括比例代表制 [7, 17],、平等代表制 [19, 20], 和统计均等 [14, 33],，并已广泛用于多种优化问题，例如前 k 项排序 [9], 、子模最大化 [17], 和多胜者投票 [7]。

*华东师范大学。yhwang@dase.ecnu.edu.cn
†赫尔辛基大学。michael.mathioudakis@helsinki.fi
‡华东师范大学。jiali@stu.ecnu.edu.cn§Spotify。
francescof@spotify.com



(a) MMD (b) MSD

图 1：最大-最小多样化 （MMD）与最大-总和多样化（MSD）的示意图。解中的项目 以蓝 色标记。

1.1 相关工作 Erkut [15]证明了度量空间中的 MMD 问题是 NP-困难的。Ravi 等人 [26]提出了用于 MMD 的 $\frac{1}{2}$-近似贪心算法 [16] ，并证明除非 P=NP，否则不存在能够达到更优近似因子的多项式算法。近年来，针对各种场景下的 MMD，人们提出了许多不同的算法。Indyk 等人 [18] 基于核心集的概念，提出了用于 MMD 的 $\frac{1}{3}$-近似分布式算法。Drosou 和 Pitoura [13] 设计了一种用于动态数据中 MMD 的 $\frac{b-1}{2b^2}$-近似基于覆盖树的算法，其中 b 是覆盖树的基数。Ceccarello 等人 [6] 提出了适用于有界倍增维数度量空间中 MMD 的 $(\frac{1}{2}-\varepsilon)$-近似 MapReduce 和流式算法。Borassi等人 [5] 提出了一种用于 MMD 的滑动窗口算法。然而，上述算法均不适用于 FMMD，因为它们无法保证满足公平性约束。

Moumoulidou 等人 [23] 首次提出了公平 MMD 变体的近似算法。Addanki 等人 [1] 改进了文献 [23] 中算法的近似比。Wang 等人 [30] 为公平 MMD 变体提出了两种流式算法。然而，这些算法适用于精确规模群体公平性约束，而这只是我们所研究的有界规模群体公平性约束的一种特殊情况。此外，实证结果表明，这些算法得到的解质量低于我们的方法。

除 MMD 外，许多其他优化问题也有类似的群体公平性感知变体，例如行列式点过程 [8]、 k-中心 [11, 19, 20], 、top-k 排序 [9], 、子模最大化 [17, 29],以及多胜者投票 [7]。然而，由于这些问题的目标函数与 MMD 不同，针对其公平变体提出的算法无法直接用于 FMMD。

1.2 我们的结果 </style}本文的主要结果是提出了两种用于公平最大-最小多样化（FMMD）问题的新算法。该问题从一个数据集 V 中选取一个大小为 k 的子集 S ，在满足群体公平性约束的同时最大化多样性值。

尽管该问题是 NP-困难的，我们首先提出 FMMD-E，这是一种适用于在小型数据集上求解 FMMD 的精确算法。该算法利用了 MMD 问题与最大独立集（MIS）问题之间的联系。它将 FMMD 表述为在无向图上寻找满足群体公平性约束的顶点独立集问题。随后，通过整数线性规划（ILP）求解化简后的问题，即可在 $O(n^k \log n)$ 时间内得到 FMMD 的最优解。

由于 FMMD-E 无法扩展至大型数据集，我们提出了 FMMD-S，这是一种可扩展性更强的 FMMD 近似算法。具体而言，对于任意 $\varepsilon \in (0, 1)$，FMMD-S 可在 $O\big(Ckn + C^k \log \frac{1}{\varepsilon}\big)$ 时间内为 FMMD 提供 $\frac{1-\varepsilon}{5}$-近似解。在假设 $k = o(\log n)$ 且 $C = O(1)$ 的情况下，FMMD-S 的时间复杂度可降至 $O\big(n(k + \log \frac{1}{\varepsilon})\big)$。FMMD-S 的基本思路是：首先运行贪心算法 [16, 26]，将计算限制在原始数据集的一个小得多的子集（即 核心集）上；然后，使用与 FMMD-E 类似的基于整数线性规划的方法，从该子集中获得 FMMD 的近似解。

最后，我们比较了我们的算法与 [1, 23, 30]中最先进的算法在真实世界数据集上求解 FMMD问题的性能。结果表明：*i)* FMMD-E 能在合理的时间内为小型数据集（例如 $n = 1,000$）提供精确解；*ii)* FMMD-S 在可扩展至包含数百万项的大型数据集的同时，能在相近的时间内返回质量高于现有近似算法的解。

2 预备知识

本节首先正式定义 FMMD问题，即最大-最小多样化（MMD）的公平性感知变体，然后给出其困难性结果。

最大-最小多样化（MMD）。设 V 为由 n 个项目组成的集合，且 $d: V \times V \to \mathbb{R}_{\geq 0}$ 是一个距离度量，用于刻画 V中项目之间的差异性。根据定义，对于任意 u、v、$w \in V$， $d(\cdot,\cdot)$ 满足以下性质：i) $d(u,v) = 0 \Leftrightarrow u = v$（不可分辨者同一性）；ii) $d(u,v) = d(v,u)$（对称性）；iii) $d(u,v) + d(v,w) \geq d(u,w)$（三角不等式）。对于MMD，子集 $S \subseteq V$ 的多样性值 $div(S)$ 定义为其中不同项目之间所有成对距离的最小值——即，$div(S) = \min_{u,v \in S: u \neq v} d(u,v)$。给定集合 V、距离函数 $d(\cdot,\cdot)$ 和正整数 $k \leq n$，MMD问题要求从 V 中选取一个大小为 k 的子集 S，使 $div(S)$ 最大化。

公平最大-最小多样化（FMMD）。设集合 V按某个敏感类别属性（如性别或种族）划分为 C个互不相交的组 $V_1, \ldots, V_C$。对于 FMMD（MMD 的公平性感知变体），群体公平性约束限制从每个组 V_c 中为 $c \in [C]$ 选取的项目数量，使其处于 l_c 至 h_c 的范围内（包括两端）。同时，还要求所选项目的总数为 k。形式上，FMMD 的所有可行解构成的集合 $\mathcal{F}$ 为

$$\mathcal{F} = \{S \subseteq V : |S| = k \wedge l_c \leq |S \cap V_c| \leq h_c, \forall c \in [C]\}$$

并且，为了排除显然为空的集合 $\mathcal{F}$，我们进一步假设 $l_c \leq h_c \leq |V_c|$ and$\sum_{c=1}^{C} l_c \leq k \leq \sum_{c=1}^{C} h_c$。FMMD问题要求从 V中选取一个子集 S，使得 S满足群体公平性约束（即 $S \in \mathcal{F}$），并使 $div(S)$ 最大化；形式上，即 $S^* = \text{argmax}_{S \in \mathcal{F}} div(S)$，其中 S^*和 $\text{OPT} = div(S^*)$ 分别表示 FMMD 的最优解及其多样性值。

FMMD 的困难性。通过从 CLIQUE 问题进行归约，已证明 MMD 在一般度量空间中是 NP-困难的，且除非 P=NP [15, 26]，否则无法以优于 $1/2$ 的因子进行近似。尽管如此，贪心算法仍能在 $O(kn)$ 时间内提供最优的 $1/2$-近似解 [16]。虽然贪心算法无法直接用于 FMMD，因为它可能给出不满足 $\mathcal{F}$ 的解（即“不公平”的解），但在第 3.2 节中，它将作为 FMMD-S 算法的数据约简子程序。由于当 $C = 1$ 且 $l_1 \leq k \leq h_1$ 时，MMD 是 FMMD 的一个特例，因此 MMD 的困难性结果可推广到 FMMD，如下所示：

THEOREM 2.1. *FMMD 是 NP-困难的，且对于任何参数 $\varepsilon > 0$，除非 P=NP，否则无法以 $\frac{1}{2} + \varepsilon$ 的因子进行近似。*

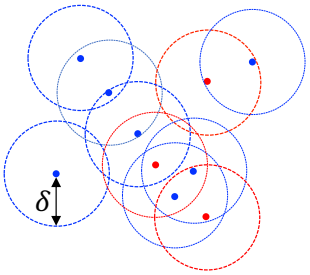
3 算法

本节介绍我们为 FMMD 提出的算法。首先，我们提出 FMMD-E，这是一种精确算法，运行时间为 $O(n^k \log n)$（第 3.1 节）。其次，我们提出 FMMD-S，这是一种 $\frac{1-\varepsilon}{5}$-近似算法，对于任意误差参数 $\varepsilon \in (0,1)$，其运行时间为 $O\left(Ckn + C^k \log \frac{1}{\varepsilon}\right)$ time（第 3.2 节）。

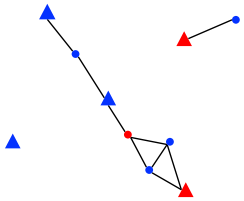
3.1 一种基于整数线性规划的精确算法 为了构建 FMMD 的精确算法，我们采用了与 [3] 将 MMD 归约为最大独立集（MIS）类似的思路。给定可行解集合 $\mathcal{F}$ 和一个正实数 δ，FMMD 的判定版本询问是否存在集合 $S \subseteq V$，使得 $S \in \mathcal{F}$ 且 $div(S) \geq \delta$。给定一个 FMMD 判定问题的实例，我们按如下方式构建无向图 $G = (V, E)$：G 中的顶点集合与 V 相同，并且当且仅当 $d(u,v) < \delta$时，两个顶点 $u, v \in V$之间存在一条边。我们提醒，若顶点集合 S 中任意两个顶点都不相邻，则称其为一个独立集 。此外，我们定义公平独立集 （FIS）问题，以判定图 G上是否存在独立顶点集 $S \in \mathcal{F}$。根据上述定义，下面的引理指出 FMMD 与 FIS 等价。

LEMMA 3.1. *FMMD 等价于 FIS。*

证明。先证明一个方向：假设 FMMD 的答案为‘是’——也就是说，存在 $S \in \mathcal{F}$ 属于 V，且$div(S) \geq \delta$。那么，对于任意 u, $v \in S$，都有 $d(u,v) \geq \delta$。因此，根据构造，不存在边$(u,v) \in E$，并且 S是 G的独立集。因此，FIS 的答案也为‘是’。反过来，假设 FIS 的答案为‘是’——也就是说， $S \in \mathcal{F}$ 是一个独立集。根据定义，任意两个顶点之间都不存在边



(a) FMMD



(b) FIS

图2：将公平最大-最小多样化（FMMD）归约为公平独立集（FIS）的示例：数据集中有 $n = 10$ 个点，分属蓝色和红色的 $C = 2$ 个组。对于 $c = 1,2$，包含 $k = 5$、$l_c = 2$ 和 $h_c = 3$ 的 FMMD 实例被归约为一个 FIS 实例，其中公平独立集由三角形表示。

在 G中，这根据构造意味着，对于任意 $u, v \in S$，都有 $d(u, v) \geq \delta$，因此，对于给定的 $S \in \mathcal{F}$，有 $div(S) \geq \delta$。所以，FMMD 的答案也为‘是’。由此我们证明，FMMD 的答案为‘是’当且仅当 FIS 的答案为‘是’，证毕。□

此外，根据 FMMD 的定义，我们可以得到两个容易验证的观察结果。

FACT 1. (MONOTONICITY) 如果存在一个满足 $div(S) \geq \delta$ 的集合 $S \in \mathcal{F}$，那么对于任意 $\delta' \leq \delta$ 都存在这样的集合；如果不存在任何满足 $div(S) \geq \delta$ 的集合 $S \in \mathcal{F}$，那么对于任意 $\delta' \geq \delta$ 都不存在这样的集合。

FACT 2. (DISCONTINUITY) *FMMD* 的最优多样性值 OPT 始终等于某一对项目 u、$v \in V$ 之间的距离 $d(u, v)$。

F综上所述，以下定理断言 FMMD 可归约为 FI S.

THEOREM 3.1. 通过求解 $O(\log n)$ 个 *FIS* 实例，可以得到 *FMMD* 的精确解。

证明。 考虑以下算法。首先，计算并排序 V 中所有项目对之间的距离。然后，对成对距离的排序数组进行二分搜索，找到最大的 d^*，使其对应的 FIS 实例的答案为‘是’。由于成对距离的数量为 $O(n^2)$，二分搜索可在 $O(\log n)$ 步内找到 d^*。根据引理 3.1，有 $div(S^*) \geq d^*$。观察1保证，由于 d^* 的最大性，不存在任何满足 $div(S) > d^*$ 的 $S \in \mathcal{F}$。观察2确保 d^* 恰好等于最优值。因此，上述过程可以求出 FMMD 的精确解。□

FMMD的整数线性规划公式。 根据定理3.1，要获得 FMMD 的精确算法，剩下的工作是为 FIS 设计精确算法。我们注意到，不带公平性约束的 FIS 等价于最大独立集 (MIS)问题。因此，我们对 MIS 的基于边的整数线性规划（ILP）公式进行调整，加入公平性约束来定义一个 FIS 实例，如公式 3.1–3.5 所示。

$$\max \quad z = \sum_{i=1}^{n} x_i \tag{3.1}$$

$$\text{s.t.} \quad x_i + x_j \leq 1, \forall (v_i, v_j) \in E \tag{3.2}$$

$$\sum_{i=1}^{n} x_i \leq k \tag{3.3}$$

$$l_c \leq \sum_{v_i \in V_c} x_i \leq h_c, \forall c \in [C] \tag{3.4}$$

$$x_i \in \{0, 1\}, \forall i \in [n] \tag{3.5}$$

算法 1FMMD-E

输入：具有 $n = |V|$ **的数据集** $V = \bigcup_{c=1}^{C} V_c$**；** $c \in [C]$ **的上下界** l_c**、** $h_c \in \mathbb{Z}^+$**；大小约束** $k \in \mathbb{Z}^+$**。输出：一个集合** $S^* \subseteq V$**，使得** $S^* \in \mathcal{F}$**。1：计算** V **中所有项目对之间的距离，并按升序排列为** $D[1, \ldots, {}^{n}(_{n-1})_{2}]$ **2：令** $L \leftarrow 1$**、** $H \leftarrow {}^{n}(_{n-1})$ **2、** $cur \leftarrow \frac{L+H}{2}$ **、** $S^* \leftarrow \emptyset$ **3：当** $H > L$ **时执行**

4 $G = \overline{V\ E}$ $E = \{\ u\ v\ \in V \times V \mid d\ u\ v\ < D[cur]\}$**：构建一个无向图 (,),**

5 **其中 (,) (,): 求解公式 . – . 中 ILP 的解 x。6：基于 x 找到一个集合** $S = \{v_i \in V :\ x_i = 1$**、** $i \in [n]\}$ **7：如果** $|S| = k$**，则 8：如果** $div(S) > div(S^*)$ **或** $S^* = \emptyset$**，则** $S^* \leftarrow S$**9：令** $L \leftarrow cur + 1$ **和** $cur \leftarrow \frac{L+H}{2}$ **10：否则 11：令** $H \leftarrow cur - 1$ **和** $cur \leftarrow \frac{L+H}{2}$ **12：返回** S^*

其中， x_i 是一个二元变量，用于指示 $v_i \in V$ 是否包含在解中；公式 3.1中的目标函数和公式 3.2中的第一个约束与 MIS 的基于边的整数线性规划形式相同；公式 3.3中的第二个约束将解的大小限制为至多 k；公式 3.4中的第三个约束则规定了从每个组 V_c中选取的项目数量的上下界。通过对公式 3.1–3.5中的整数线性规划求得最优解，我们将找到一个公平独立集 $S = \{v_i \in V :\ x_i = 1,\ i \in [n]\}$，其 G 为 $z = k$；或者确认不存在这样的集合（若 $z < k$）。

算法描述与复杂度。结合定理3.1的构造性证明和FMMD的整数线性规划公式，我们得到FMMD的精确算法FMMD-E，如算法1所示。首先，该算法用 O（n^2）步计算 V中所有 ^{n}（$_{n-1}$）对不同项目之间的距离，并用 O（$n^2 \log\ n$）步按升序将其排序到数组 $D[1, \ldots, {}^{n}(_{n-1})_{2}]$中。然后，通过在 D上进行二分搜索，用O（log n）步求得最优值。对于最优值的每个猜测 $D[cur]$，算法用 O（n^2）步构建无向图 G，并用 $\binom{n}{k} = O(n^k)$ 步通过求解公式 3.1–3.5中的整数线性规划找到一个集合 S。如果 $|S| = k$，则 $S \in \mathcal{F}$ 且 div（S） $\geq D[cur]$。此时，搜索空间缩小至上半部分，以检查是否存在更优解。否则，或者若$|S| < k$，则最优值 $< D[cur]$，并将搜索空间缩小至下半部分。最后，二分搜索结束时，该算法已找到FMMD的精确解 S^*。FMMD-E的时间复杂度为O（$n^k \log n$）。此外，由于 $|D| = |E| = O$（n^2），其空间复杂度为 O（n^2）。

3.2 一种更具可扩展性的近似算法 FMMD-E 的主要缺点是由于复杂度呈指数级，无法处理大型数据集。标准优化库（如 CPLEX[1] 和 Gurobi[2],）只能在合理时间内求解变量数量达数千个的整数线性规划，并得到最优解。解决这一挑战的自然方法是构造一个“核心集”，即从原始数据集中选取一个小型子集，并在该子集上运行精确算法以寻找近似解。形式化地说，子集 $V' \subseteq V$ 称为 FMMD 的一个 α-核心集（$0 \leq \alpha \leq 1$），当且仅当 OPT$[V'] \geq \alpha \cdot$ OPT，其中 OPT$[V']$ 是 FMMD 在 V'上的最优多样性值。

接下来需要回答*i*）如何构造这样的核心集，以及*ii*）能够获得怎样的近似因子。对于*i*），我们的思路受到流式处理和分布式环境下 MMD 的可组合核心集 [18, 31] 概念的启发。其基本思路是先对数据集进行划分，并在每个分区上运行 [16] 的贪心算法，以获得一个部分解；然后从所有部分解的并集中计算最终解。在 FMMD 中，数据集自然可以划分为 C 个组。因此，我们先从每个组中找出一个解，再将所有分组专用解的并集作为我们的核心集，最后使用 FMMD-E 从核心集中求得一个解。由于核心集规模较小，这样做是可行的。我们将所得算法称为 FMMD-S。对于*ii*），我们证明，对于任意 $\varepsilon \in (0,\ 1)$，所得到的解具有 $\frac{1-\varepsilon}{5}$ 的近似因子。

算法描述。 FMMD-S 如算法 2 所述。首先，在不考虑公平性约束的情况下调用贪心算法，对V进行计算，得到初始解 U （第 1–3 行）。注意，贪心算法对于 MMD 是$_{1\ 2}$-近似的，并且 FMMD 的任何可行解也必须是 MMD 的可行解。因此，最优 –

[1] www.ibm.com/products/ilog-cplex-optimization-studio
[2] www.gurobi.com/products/gurobi-optimizer/

算法 2FMMD-S

输入：数据集 $V = \bigcup_{c=1}^{C} V_c$**，包含** $n = |V|$**；** $c \in [C]$**的上下界** l_c**、** $h_c \in \mathbb{Z}^+$**；大小约束** $k \in \mathbb{Z}^+$**；误差参数** $\varepsilon \in$ **（0, 1）**。**输出：集合** $S \subseteq V$**，使得** $S \in \mathcal{F}$。1 u_1 V $U = \{u_1\}$**：从中任意选择一个项并设置：对于，，执行**3 $u_i \leftarrow$ 2 $i \leftarrow 2 \ldots k$ $v \in V$ $u \in U$ d u v $U \leftarrow U \cup \{u_i\}$**：** **arg max min**（，）**以及：设置，（）5：重复6：对于** $c \leftarrow 1$**、.....、** C**执行 7：当** $|U_c| < k$ **且存在某个项** $v \in V_c$ **使得** div（$U_c \cup \{v\}$）$\geq d'$ **时执行 8：** $u'_c \leftarrow$ **arg max**$_{v \in V_c}$ **min**$_{u \in U_c} d$（u，v）**9：** $U_c \leftarrow U_c \cup \{u'_c\}$ 10 $G = V' E$ $V' = \bigcup_c U_c$ $E = \{ u v \in V' \times V' \mid d u v < \frac{d'}{2} \}$**：构建无向图** 11 3 1 3 5 **（，），其中 且（，）（，）：根据公式**。–。**计算整数线性规划的解 x**。**12：找到集合** $S = \{v_i \in V$**：** $x_i = 1$**、** $i \in [n]\}$**，基于 x 13：如果** $|S| < k$ **则 14：** $S \leftarrow \emptyset$ **和** $d' \leftarrow$ $(1 - \varepsilon)$ $\cdot d'$**15：直到** $S \neq \emptyset$**16：返回** S

FMMD 的多样性最优值 OPT 以 $2 \cdot div(U)$ 为界。随后，该算法按组将 U 划分为 U_1、.....、U_C，并猜测 OPT 为其上界 $d' = 2 \cdot div(U)$（第 4 行）。对于每个 $c \in [C]$，算法运行贪心算法，从 V_c 中添加新项目至 U_c，直到 $|U_c| = k$ 或者不存在任何 $v \in V_c$ 能使 $div(U_c \cup \{v\}) \geq d'$（第 5–9 行）。此时，每个 U_c 都是一个部分分组专用解，而部分解的并集 $V' = \bigcup_c U_c$ 就是核心集。接下来，采用与 FMMD-E 类似的过程，在 V' 上构建一个多样性阈值为 $\frac{d'}{2}$ 的图 G，并在 G 上求解公式 3.1–3.5 对应的整数线性规划，以获得解 S（第 11–12 行）。最后，如果 $|S| = k$，我们就得到一个解 $S \in \mathcal{F}$，其 $div(S)$ 为 $\geq \frac{d'}{2}$，并将 S 作为最终解返回；否则，将 d' 按因子 $1 - \varepsilon$ 缩小，其中 $\varepsilon \in (0, 1)$ 是误差参数，然后对更小的 d' 再次执行上述过程，直到找到可行解 S（第 13–16 行）。

理论分析。接下来，我们给出 FMMD-S 的复杂度和近似因子。

THEOREM 3.2. FMMD -S 是一个用于 *FMMD* 的 $\frac{1-\varepsilon}{5}$ -近似算法，运行时间为 $O\big(Ckn + C^k \log \frac{1}{\varepsilon}\big)$。

证明。如果存在任意集合 $S' \subseteq V'$ 使得 $S' \in \mathcal{F}$ 且 $div(S') \geq \frac{d'}{2}$，那么 FMMD-S 会从公式 3.1–3.5 中整数线性规划的精确解中识别出这样的 S'（第 12 行）。此外，由于贪心算法（第 1–3 行）是 1 2-近似的，[26], 的初始值至少为 $2 \cdot \frac{\text{OPT}}{2}$ = 最优值 $> \frac{2}{5} \cdot$ 最优值。因此，为了证明近似因子，只需证明存在某个 $S' \subseteq V'$ 使得 $S' \in \mathcal{F}$ 且 $div(S') \geq \frac{d'}{2}$，当 $d' \in [\frac{2(1-\varepsilon)}{5} \cdot$ 最优值 $\frac{2}{5} \cdot$ 最优值]时。

为此，我们接下来从 V' 构造这样一个集合 S'。令 S^* 为 V 上 FMMD 的最优解，并令 $S^*_c = S^* \cap V_c$ 为其来自组 c 的子集。首先，我们初始化 $S' = \emptyset$。然后，我们针对不同的组考虑两种情况。我们先以任意顺序考虑情况 #1 的组，再以任意顺序考虑情况 #2 的组，并从每个组 c 中选择 $|S^*_c|$ 个项目加入 S'。

情况 #1（$|U_c| < k$）：令 f：$V_c \to U_c$ 将每个项目 $v \in V_c$ 映射到其在 U_c 中的最近邻 f（v）。注意，第 7 行中的条件确保对于任意 $v \in V_c$ 和 $u \in U_c$，d（v，u）$< d'$。对于每个项目 s_c，$_i \in S^*_c$，我们将项目 f（s_c，$_i$）加入 S'。现在我们证明所加入的项目彼此不同。事实上，如果对于 $i \neq j$，f（s_c，$_i$）$\equiv f\, s_c$（$_j$），则根据三角不等式可得

$$d\ s_{c\,i}\ s_{c\,j} \leq d\ s_{c\,i}\ f\ s_{c\,i} + d\ s_{c\,j}\ f\ s_{c\,j} = d\ s_{c\,i}\ f\ s_{c\,i} + d\ s_{c\,j}\ f\ s_{c\,i} < (,,,)\,(,,(,))\,(,,$$

$2 \cdot d' \leq \frac{4}{5}$ $<$ $d\ s_{c\,i}\ s_{c\,j} \geq$ $s_{c\,i}\ s_{c\,j} \in S^*$ (,))(,,(,))(,,(,)) OPT OPT；然而，同时我们有 (,,,) OPT，因为 ,,,, 这就导致矛盾。此外，由于我们为每个 s_c、$_i \in S^*_c$ 在 U_c 中确定了一个不同的项目，因此我们有 $|U_c| \geq |S^*_c|$。处理完情况 #1 中的所有组后，对于任意 f（s^*_i）fs^*_j，我们有 d（f（s^*_i）f（s^*_j））$\geq d$（s^*_i，s^*_j）$- d$（s^*_i，f（s^*_i））$- d$（s^*_j，f（s^*_j））$>$OPT $- 2d' \geq \frac{\text{OPT}}{5}$，从而 $div(S') > \frac{\text{OPT}}{5}$。

情况 #2（$|U_c| = k$）：令 g：$U_c \to S'$ 将每个项 $u \in U_c$ 映射到当前 S'实例中的最近邻 $g(u)$。我们从 U_c 中移除所有满足 $d(u, g(u)) < \frac{d'}{2}$ 的 $u \in U_c$。由于第 7 行的条件保证对任意 $i \neq j$都有$d(u_{c,i}, u_{c,j}) \geq d'$，因此对于 S' 中的每个项，至多有一个项被移除——否则，三角

表1：我们实验中使用的数据集统计信息

数据集	组	C	n	dim	距离度量
Adult	Sex	2	48,842	6	l_2-距离
	Race	5			
	S+R	10			
CelebA	Sex	2	202,599	25,088	l_1-距离
	Age	2			
	S+A	4			
人口普查	Sex	2	2,426,116	25	l_1-距离
	Age	7			
	S+A	14			
Twitter	Sex	3	18,836	1,024	角距离

不等式将给出 $d(u_{c,i}, u_{c,j}) < d'$，从而导致矛盾。因此， U_c中至少还剩下 $k - |S'|$ 个项目。此外， $|S_c^*| \leq k - |S'|$ 因为在整个构造过程中， S' 始终包含与 S^* 相同数量的、来自每个所考虑分组的项目。我们从剩余项目中选取 $|S_c^*|$ 个，并将其加入 S'。由于我们先前从 U_c中移除项目时已确保，对于每个加入的 $u_{c,i}$，都有 $d(u_{c,i}, g(u_{c,i})) \geq \frac{d'}{2}$，因此执行此操作后，我们仍有 $\overline{div}(S') \geq \frac{d'}{2} \geq \frac{1-\varepsilon}{5}$ ·最优值。最后，在情况 #2 中处理完所有分组后，我们得到一个集合S'，其中来自每个组 $c \in [C]$ 的项目数量与 S^*相同，这意味着 $S' \in \mathcal{F}$，且$div(S') \geq \frac{1-\varepsilon}{5}$ ·最优值。因此，我们得出结论：FMMD-S 是 FMMD 的一个 $\frac{1-\varepsilon}{5}$ -近似算法。

由于计算 U 以及为每个 $c \in [C]$ 计算 U_c 需要 $O(nk)$ 的时间，因此计算 V' 的总时间为 $O(Ckn)$ 和 $|V'| \leq Ck$。接着，由于需要考虑的可能解最多有 $\binom{Ck}{k} = O(C^k)$ 个，求解 FMMD-S 的公式 3.1–3.5 中的整数线性规划需要 $O(C^k)$ 的时间。此外， d' 的迭代次数为 $O(\log \frac{1}{\varepsilon})$，因为 d' 的第一个值与最后一个值之比为 $O(1)$。因此，FMMD-S 的时间复杂度为 $O\big(Ckn + C^k \log \frac{1}{\varepsilon}\big)$。当 $k = o(\log n)$ 且 $C = O(1)$ 时，其时间复杂度可降至 $O\big(n(k + \log \frac{1}{\varepsilon})\big)$。此外，其空间复杂度为 $O(n + C^2k^2)$，因为 G 中的边数为 $O(C^2k^2)$。

4 实验评估

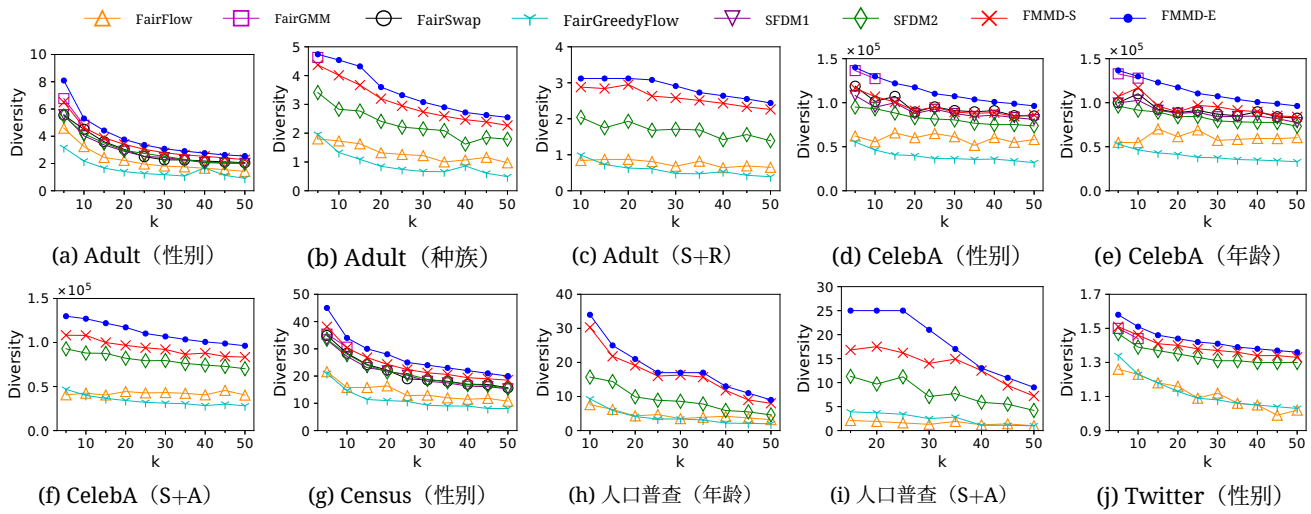
4.1 实验设置 本节开展广泛实验，以评估我们提出的算法（即 FMMD-E 和 FMMD-S）的性能。我们将其与最先进的 FMMD 算法进行比较，包括 FairSwap、FairFlow、FairGMM、 [23], FairGreedyFlow [1], 以及 SFDM1 和 SFDM2 [30]。我们使用 Python 3 实现了上述所有算法，并使用 NetworkX 库构建和操作图，使用 Gurobi 优化器求解整数线性规划。所有实验均在一台配备 Intel Core i5-9500 3.0GHz 处理器和 32GB 内存、运行 Ubuntu 20.04.3 LTS 的台式机上完成。每个算法均使用单线程运行。所有数据和代码均已公开发布于 https://osf.io/te34m/。

我们使用了表1中列出的四个公开真实世界数据集，其中 dim 表示特征向量的维度。每个数据集的详细信息和预处理流程见附录A。我们实验中的公平性约束根据比例代表制 [7, 17]定义：对于每个组$c \in [C]$，我们设置 $l_c = \max(1, (1-\alpha)k \cdot \frac{|V_c|}{n})$，并设置 $h_c = (1+\alpha)k \cdot \frac{|V_c|}{n}$，其中在 FMMD-E 和 FMMD-S 中使用 $\alpha = 0.2$，而在其他所有算法中使用$k_c = \lceil k \cdot \frac{|V_c|}{n} \rceil$ 或 $\lfloor k \cdot \frac{|V_c|}{n} \rfloor$。所有算法在每次实验中均执行十次。我们报告每种算法所得解的平均运行时间和平均多样性值。我们使用“N/A”表示某算法在一天内未找到解，或在$C > 2$ 时无法运行（即 FairSwap 和 SFDM1）。在初步实验中（见附录B），我们发现当 ε 降低到 0.05 以下时，FMMD-S 的解的质量几乎没有提升，因此在其余所有实验中，我们将 FMMD-S 的 $\varepsilon = 0.05$ 固定为该值。

4.2 实验结果 表2展示了不同算法在“小型”数据集上实现的多样性，即从每个完整数据集中均匀随机抽取1,000个项目得到的数据集。图3–4展示了不同算法在 k取不同值时，在小型数据集上的性能。请注意，FairSwap和 SFDM1仅适用于 $C = 2$ 的情况，而当 $C > 3$ 或 $k > 10$ 时，FairGMM无法在一天内完成，因为它必须枚举 $\binom{kC}{k}$ 集合以计算解。当这些算法无法提供有效解时，后续表格和图中将不再列出它们。

表2：对于解的大小为$k = 10$的小型数据集，不同算法返回的解的多样性值。报告不满足公平性约束时的最优值 OPT*，以展示“公平代价”。

数据集	组	FairSwap	FairFlow	FairGMM	FairGreedyFlow	SFDM1	SFDM2	FMMD-E	FMMD-S	OPT*
Adult	Sex	4.51	3.24	4.81	2.18	4.03	4.18	**5.30**	4.64	5.30
	Race	N/A	1.73	N/A	1.33	N/A	2.83	**4.54**	4.01	
	S+R	N/A	0.84	N/A	0.99	N/A	2.04	**3.12**	2.88	
CelebA	Sex	101457.3	55540.7	127354.6	46266.9	94873.3	93216.6	**129818.2**	106959.3	129871.5
	Age	110098.1	54649.6	127871.2	46312.3	102762.9	91578.2	**129871.5**	116701.6	
	S+A	N/A	42412.7	N/A	39967.2	N/A	88026.7	**127974.6**	108055.3	
人口普查	Sex	28.4	15.8	29.8	14.7	27.2	28.0	**34.0**	30.3	35.0
	Age	N/A	7.6	N/A	9.3	N/A	15.7	**34.0**	30.3	
Twitter	Sex	N/A	1.23	1.44	1.23	N/A	1.39	**1.51**	1.46	1.51



(a) Adult（性别） (b) Adult（种族） (c) Adult（S+R） (d) CelebA（性别） (e) CelebA（年龄）

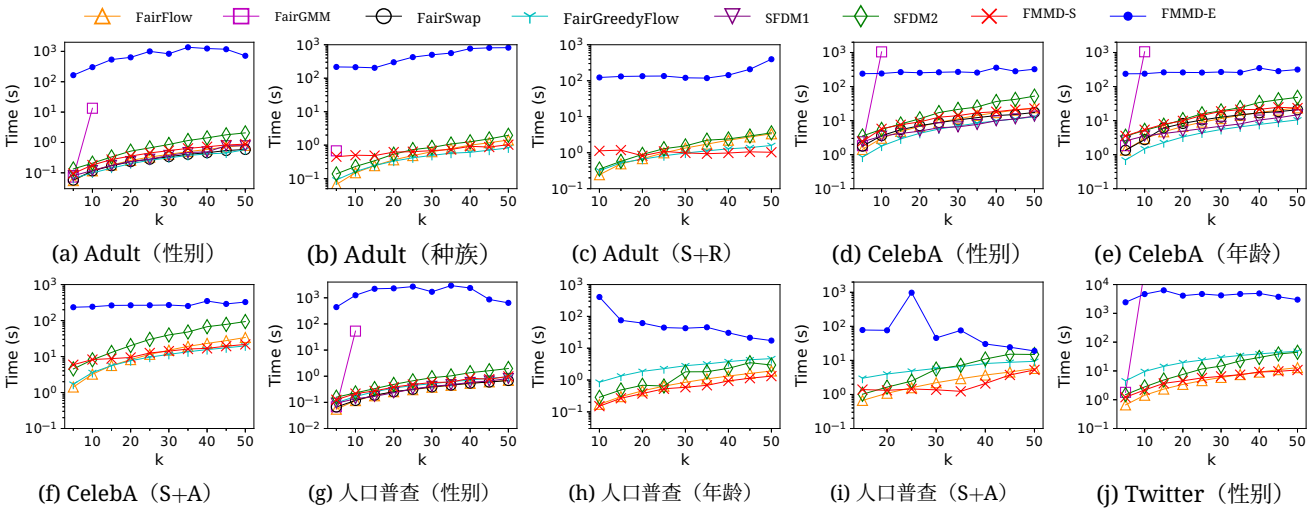
(f) CelebA（S+A） (g) Census（性别） (h) 人口普查（年龄） (i) 人口普查（S+A） (j) Twitter（性别）

图3：小型数据集上不同算法的解在解的大小 k 变化时的多样性值。

总体而言，当 $n = 1,000$ 时，FMMD-E 总能在时间限制（即 24 小时）内为 FMMD 求得最优解。除 $Adult$ （ $C \geq 5$）外，所有数据集上的公平代价（即公平性约束导致的多样性下降）都微乎其微。与所有近似算法（FairGMM 除外）相比，FMMD-S 的解质量高得多（多样性值最高高出 $3.9\times$）。尽管 FairGMM 有时能得到略优于 FMMD-S 的解，但其运行速度慢两个数量级以上。此外，所有算法的多样性值都会随着 k 而下降，因为 div 是单调非增函数。FMMD-E 的运行时间与 k 无关；FairGMM 的运行时间随 k 呈指数增长；FMMD-S 和其他算法的运行时间则随 k 线性增长，这些结论均可由它们的时间复杂度推出。

表3展示了在所有完整数据集上，不同算法针对解的大小为 $k = 50$ 时的多样性值和运行时间。图5–6展示了在完整数据集（包含所有项目）上，将解的大小 k 从 10 调整至 100 时不同算法的性能。总体而言，随着 n 和 dim的增加，所有算法的运行时间都会显著增长。FairGMM 和 FMMD-E 无法在一天内完成，因此未列入表3。在大多数情况下，FMMD-S 得出的解优于所有基线方法（多样性值最高提升 $7.2\times$ ）。唯一的例外是，当 $C = 2$时，FMMD-S 在 $CelebA$ 上的解的质量略低于 FairSwap 和 SFDM1。这是因为$CelebA$ 的维度极高（$d = 25{,}088$），导致不同点对之间的距离不易区分。在这种情况下，FMMD-S 中的阈值法不如 FairSwap 和 SFDM1 中的局部搜索方法。此外，FMMD-S 的时间效率低于基线方法，因为求解整数线性规划通常耗时较长。尽管如此，在人口普查数据集上——这是最大的一个数据集，包含两百多万个项目——FMMD-S 仍能在 3 小时内完成计算。

为了评估不同算法的可扩展性，我们在合成数据集上改变组数 C 和点数 n 。具体而言，每个数据集由十个二维高斯各向同性簇组成



(a) Adult（性别） (b) Adult（种族） (c) Adult（S+R） (d) CelebA（性别） (e) CelebA（年龄）

(f) CelebA（S+A） (g) 人口普查（性别） (h) 人口普查（年龄） (i) 人口普查（S+A） (j) Twitter（性别）

图4：小型数据集上不同算法在解的大小 k 变化时的运行时间。

表3：完整数据集上不同算法针对解的大小 $k = 50$的结果。由于 FairGMM 和 FMMD-E 无法在时间限制内（即 24 小时）给出任何解，因此未列出这两种算法。

数据集组		FairSwap 多样性	FairSwap 时间（秒）	FairFlow 多样性	FairFlow 时间（秒）	FairGreedyFlow 多样性	FairGreedyFlow 时间（秒）	SFDM1 多样性	SFDM1 时间（秒）	SFDM2 多样性	SFDM2 时间（秒）	FMMD-S 多样性	FMMD-S 时间（秒）	FMMD-E 多样性	FMMD-E 时间（秒）
Adult	Sex	2.55	28.45	2.10	26.47	1.46	99.61	2.86	8.17	3.22	13.71	**3.56**	46.38		
	Race	N/A		1.43	28.47	0.78	91.24	N/A		2.86	15.81	**3.56**	48.10		
	S+R	N/A		0.99	32.64	0.50	148.12	N/A		2.55	23.35	**3.61**	45.51		
CelebA	Sex	125865.8	3093.9	101303.9	2624.1	57696.8	3558.0	**128450.2**	1473.4	117342.5	1354.4	123639.0	4536.4		
	Age	112387.6	4141.2	74679.4	2598.8	55470.0	2260.8	**126446.9**	1024.0	121056.9	1222.3	113798.5	4626.6		
	S+A	N/A		29278.4	2589.5	34066.4	3871.3	N/A		118598.7	1141.7	**129866.3**	4752.5		
人口普查	Sex	22.8	1533.3	16.8	1250.4	N/A		22.3	328.7	24.4	459.14	**28.6**	3268.8		
	Age	N/A		5.2	1461.5	N/A		N/A		12.5	662.76	**16.0**	2728.2		
	S+A	N/A		3.4	1723.5	N/A		N/A		11.1	170.98	**15.0**	9224.8		
Twitter	Sex	N/A		1.08	120.85	1.04	783.7	N/A		1.33	85.56	**1.38**	208.52		

簇中心在 $[-10, 10]^2$ 中随机选取，且具有单位协方差矩阵。每个点都以均匀随机的方式分配到 C 个组中的一个。使用欧几里得距离作为距离度量。固定 $C = 2$ 或 10 时，我们得到六个数据集，点数分别为 $n = 10^2$，10_3、.....、10_7；固定 $n = 1$，000 时，则得到十个数据集，组数分别为 $C = 2$，4、.....、20。

图 7 展示了在合成数据集上，通过改变 n 和 C，针对解的大小 $k = 20$ 比较不同算法的性能。在解的质量方面，当 C 增大时，FMMD-E 和 FMMD-S 的多样性值保持稳定，而所有其他算法的多样性值则显著下降。在效率方面，FMMD-E 的运行时间几乎不受 C影响。其他算法在 C 较大时运行得更慢。然而，当 $C \geq 8$时，FMMD-S 的运行速度快于其他任何算法，并且随着 C的增加，其时间效率优势变得更加显著。最后，所有算法的多样性值和运行时间都随着数据集大小 n的增加而增长。由于具有指数时间复杂度，FMMD-E 无法扩展到大型数据集。总的来说，在不同的 C 或 n下，就解的质量而言，FMMD-S 优于所有其他近似算法；当 C 或n 较大时，其优势更加明显。这些结果证实了 FMMD-S 针对组大小 C 和数据集大小 n的可扩展性。

5 结论

本文研究了带公平性约束的最大最小多样化问题（FMMD）。我们针对小型数据集提出了一种基于整数线性规划的精确算法。进一步地，我们基于该精确算法和核心集的概念，为海量数据集设计了一种可扩展的 $\frac{1-\varepsilon}{5}$ - 近似算法，其中 $\varepsilon \in (0,1)$。在四个真实世界数据集上进行的大量实验结果证实了所提算法的有效性、效率和可扩展性。

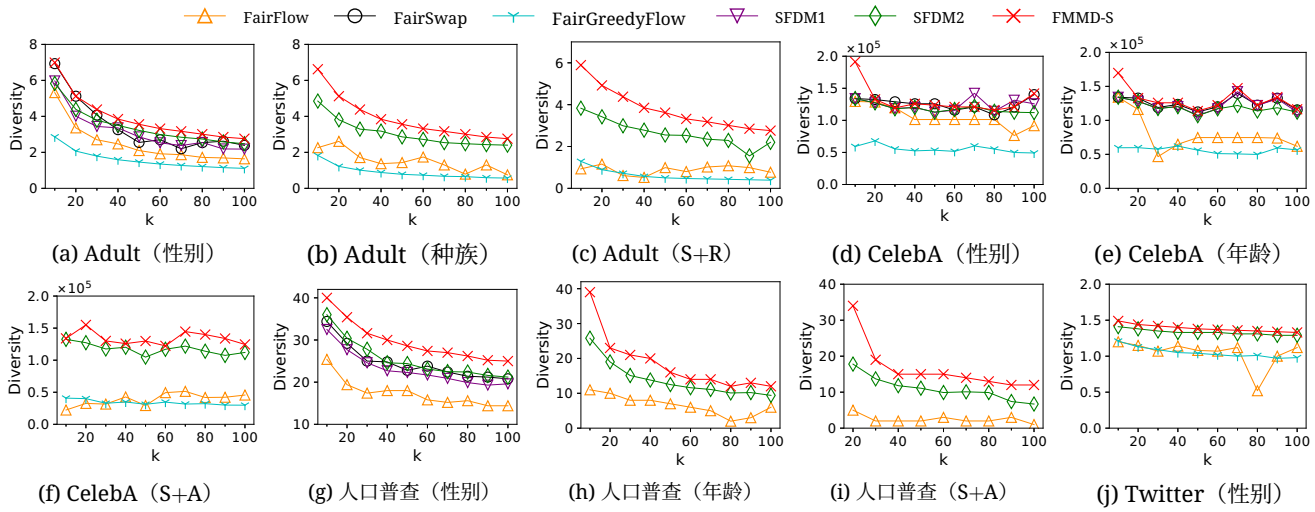


图5：完整数据集上不同算法在解大小 k 变化时的解的多样性值。

尽管我们的工作在多样性最大化公平变体算法的理论和实验方面均有所推进，但仍留下许多有待未来探索的开放问题。一个自然的问题是，FMMD是否存在多项式时间 $O(1)$-近似算法，因为我们的FMMD-S算法的近似因子为 $\frac{1-\varepsilon}{5}$，但仅在 $C = O(1)$ 和 $k = o(\log n)$ 时才以多项式时间运行；而目前已知的最佳多项式时间算法在 [1] 中只能达到 $\frac{1}{C+1}$ 的近似因子。此外，研究其他多样性度量的公平性感知变体（例如 [4, 18] 中的那些）也很有意义。

参考文献

[1] R. Addanki、A. McGregor、A. Meliou 和 Z. Moumoulidou。《公平最大-最小多样化的改进近似与可扩展性》。载于 *ICDT*，第 7:1–7:21 页，2022年。[2] R. Agrawal、S. Gollapudi、A. Halverson 和 S. Ieong。《搜索结果多样化》。载于 *WSDM*，第 5–14 页，2009年。[3] T. Akagi、T. Araki、T. Horiyama、S. Nakano、Y. Okamoto、Y. Otachi、T. Saitoh、R. Uehara、T. Uno 和 K. Wasa。《最大最小离散问题的精确算法》。载于 *FAW*，第 263–272 页，2018年。[4] A. Bhaskara、M. Ghadiri、V. S. Mirrokni 和 O. Svensson。《在度量空间中寻找多样化元素的线性松弛》。载于 *NIPS*，第 4098–4106 页，2016年。[5] M. Borassi、A. Epasto、S. Lattanzi、S. Vassilvitskii 和 M. Zadimoghaddam。《用于最大化次可加目标和多样性目标的更优滑动窗口算法》。载于 *PODS*，第 254–268 页，2019年。[6] M. Ceccarello、A. Pietracaprina、G. Pucci 和 E. Upfal。《有界倍增维数度量空间中的多样性最大化的 MapReduce 和流式算法》。*Proc. VLDB Endow.*，10:5（2017年），第 469–480 页。[7] L. E. Celis、L. Huang 和 N. K. Vishnoi。《具有公平性约束的多胜者投票》。载于 *IJCAI*，第 144–151 页，2018年。[8] L. E. Celis、V. Keswani、D. Straszak、A. Deshpande、T. Kathuria 和 N. K. Vishnoi。《公平且多样的基于DPP的数据摘要》。载于 *ICML*，第 715–724 页，2018年。[9] L. E. Celis、D. Straszak 和 N. K. Vishnoi。《具有公平性约束的排序》。载于 *ICALP*，第 28:1–28:15 页，2018年。[10] F. Chierichetti、R. Kumar、S. Lattanzi 和 S. Vassilvitskii。《通过公平小簇实现公平聚类》。载于

NIPS 5029 5037, , 第 – 页。

2017

[11] A. Chiplunkar、S. Kale 和 S. N. Ramamoorthy。《如何在海量数据模型中求解公平k中心问题》。载于 *ICML*，第 1877–1886 页，2020年。[12] A. Chouldechova 和 A. Roth。《机器学习中的公平性前沿概览》。*Commun. ACM*，63:5（2020年），第 82–89 页。[13] M. Drosou 和 E. Pitoura。《动态数据上的多样化集合选择》。*IEEE Trans. Knowl. Data Eng.*，26:5（2014年），第 1102–1116 页。[14] C. Dwork、M. Hardt、T. Pitassi、O. Reingold 和 R. S. Zemel。《通过认知实现公平》。载于

ITCS 214 226, , 第 – 页。

2012

[15] E. Erkut。《离散p-离散化问题》。*Eur. J. Oper. Res.*，46:1（1990年），第 48–60 页。

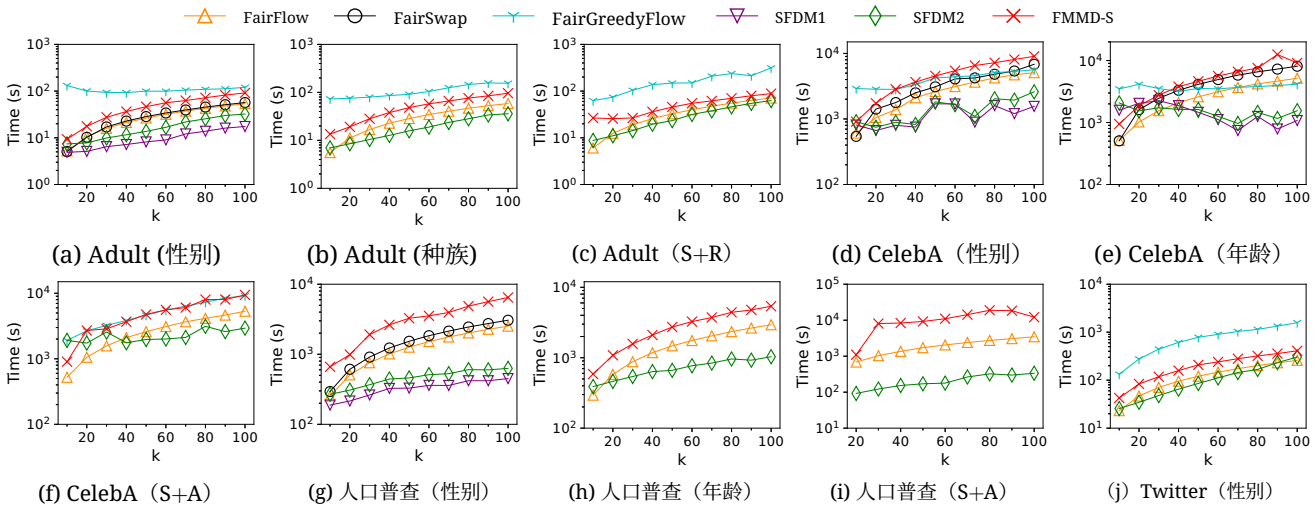


图 6：完整数据集上不同算法在解的大小变化时的运行时间 k 。

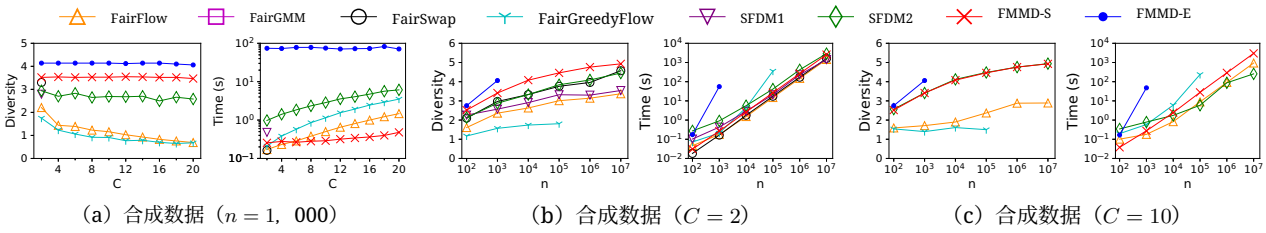


图 7：在解的大小为$k = 20$时，不同组数 C 和数据集大小 n 下不同算法的结果。由于 FMMD-E 在规模为 $n > 1,000$ 的数据集上无法在时间限制内提供任何解，FairGMM 对于 $k = 20$ 无法在时间限制内提供任何解，并且 FairSwap 和 SFDM1 无法处理 $C > 2$，因此这些情况下未绘制它们。

[16] T. F. Gonzalez。用于最小化最大簇间距离的聚类。 *Theor. Comput. Sci.*，38（1985），第 293–306 页。[17] M. E. Halabi、S. Mitrovi´c、A. Norouzi-Fard、J. Tardos 和 J. M. Tarnawski。流式次模最大化中的公平性：算法与困难性。载于 *NeurIPS*，第 13609–13622 页，2020年。[18] P. Indyk、S. Mahabadi、M. Mahdian 和 V. S. Mirrokni。用于多样性与覆盖最大化的可组合核集。载于 *PODS*，第 100–108 页，2014 年。[19] M. Jones、H. Nguyen 和 T. D. Nguyen。通过最大匹配实现公平 k 中心。载于 *ICML*，第 4940–4949 页，2020年。[20] M. Kleindessner、P. Awasthi 和 J. Morgenstern。用于数据摘要的公平 k 中心聚类。载于 *ICML*，第 3448–3457 页，2019年。[21] M. Kunaver 和 T. Pozrl。推荐系统中的多样性——综述。 *Knowl.-Based Syst.*，123（2017），第 154–162 页。[22] Z. Liu、P. Luo、X. Wang 和 X. Tang。野外场景中的深度学习面部属性。载于 *ICCV*，第 3730–3738 页，2015 年。[23] Z. Moumoulidou、A. McGregor 和 A. Meliou。公平性约束下的多样化数据选择。载于 *ICDT*，第 13:1–13:25 页，2021 年。[24] H. Narasimhan、A. Cotter、M. R. Gupta 和 S. Wang。用于排序与回归的成对公平性。载于 *AAAI*，第 5248–5255 页，2020年。[25] A. Olteanu、C. Castillo、F. Diaz 和 E. Kiciman。社交数据：偏差、方法论陷阱与伦理边界。*Front. Big Data*，2:13（2019）。[26] S. S. Ravi、D. J. Rosenkrantz 和 G. K. Tayi。离散化问题的启发式算法与特殊情形算法。

Oper.，：

Res. 42 2 1994 299 310

（），第 – 页。[27] N. Reimers 和 I. Gurevych。Sentence-BERT：使用孪生 BERT 网络的句子嵌入。载于 *EMNLP/IJCNLP*，第 3980–3990 页，2019 年。[28] G. S. Sankar、A. Louis、M. Nasre 和 P. Nimbhorkar。具有群体公平性约束的匹配：在线与离线

算法。载于 $IJCAI$，第 377–383 页，2021 年。[29] Y. Wang、F. Fabbri 和 M. Mathioudakis。数据流中的公平且具代表性的子集选择。载于 WWW，第 1340–1350 页，2021 年。[30] Y. Wang、F. Fabbri 和 M. Mathioudakis。具有公平性约束的多样性最大化流式算法。载于 $ICDE$，第 41–53 页，2022年。[31] S. A. Zadeh、M. Ghadiri、V. S. Mirrokni 和 M. Zadimoghaddam。通过分布式多样性最大化实现可扩展特征选择。载于 $AAAI$，第 2876–2883 页，2017年。[32] M. Zehlike、F. Bonchi、C. Castillo、S. Hajian、M. Megahed 和 R. Baeza-Yates。Fa*ir：一种公平的 top-k 排名算法。载于 $CIKM$，第 1569–1578 页，2017 年。[33] R. S. Zemel 、 Y. Wu 、 K. Swersky 、 T. Pitassi 和 C. Dwork。学习公平表征。载于

$ICML$ 325 333, ，第 2013 – 页。[34] M. Zhang 、 H. Li 、 S. Pan 、 X. Chang 和 S. W. Su。通过多样性最大化克服单次训练神经架构搜索中的多模型遗忘。载于 $CVPR$，第 7806–7815 页，2020年。

数据准备

我们使用的四个真实世界数据集及其预处理流程的详细信息如下。

- **Adult** 取自 UCI机器学习存储库[3]。它收录了来自1994年美国人口普查数据库的48,842条记录。我们选择六个数值属性作为特征，并将每个属性归一化，使其具有零均值和单位标准差。l_2-距离（欧几里得距离）用作距离度量。根据两个人口统计属性生成分组：性别 和种族。单独使用这两个属性及将其组合使用，分别得到2个（性别）、5个（种族）和10个（性别 +种族）组。
- **CelebA** 由 Liu 等人 [22] 在网站[4]上提供。它包含202,599张人脸图像。我们使用 Keras[5]中的预训练 VGG16模型，为每张图像提取一个25,088维（$512 \times 7 \times 7$）特征向量。特征向量之间的 l_1-距离（曼哈顿距离）用作距离度量。我们根据人工标注的类别标签“性别” {‘女性’、‘男性’}生成2个组，根据另一个人工标注的类别标签“年龄” {‘年轻’、‘非年轻’}生成2个组，并根据这两个标签共同生成4个组。
- **人口普查**也取自 UCI机器学习存储库[6]。它包含1990年美国人口普查数据中的2,426,116条记录。我们选取25个（经过归一化的）数值属性作为特征，并使用 l_1-距离（曼哈顿距离）作为距离度量。我们分别根据两个或两个兼具的人口统计属性性别、年龄生成2个、7个和14个分组。
- **Twitter** 取自 Kaggle[7]。它是一个包含用户资料的18,836条推文集合。我们使用句子转换器模型[8] [27]（“bert-large- nli-stsb-mean-tokens”）将每条推文转换为1,024维特征向量。我们使用角距离作为距离度量。我们根据用户资料中的属性“性别” {‘女性’、‘男性’、‘非人类’} 生成3个分组。

每个数据集上各分组的比例见表4。

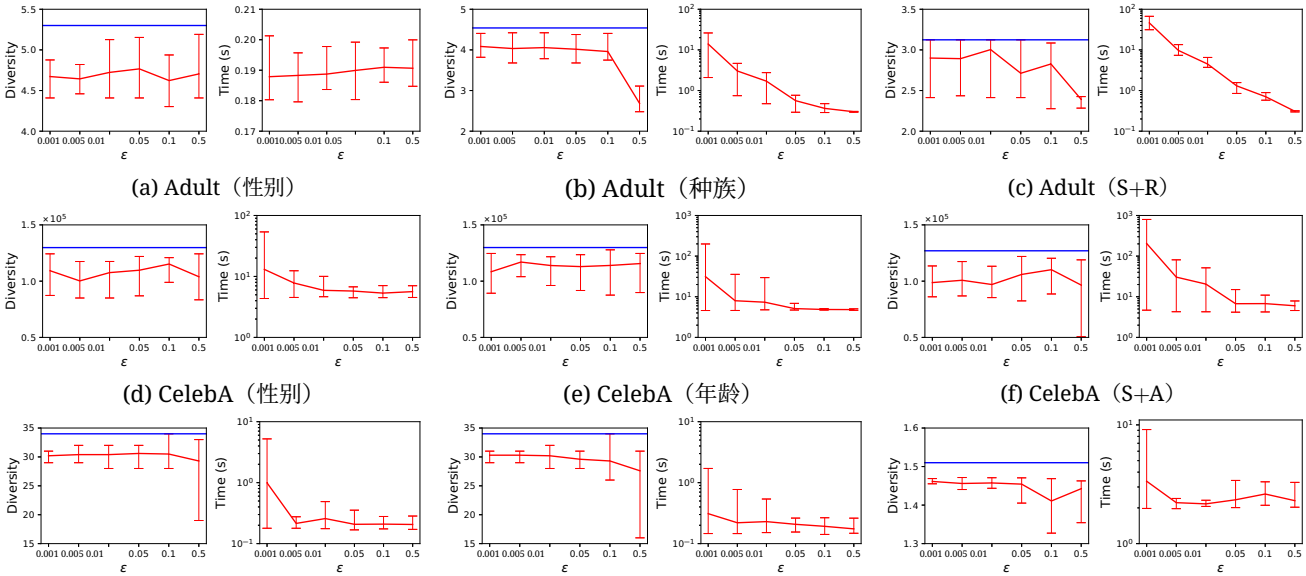
B FMMD-S的参数调优

图 8 展示了在小型数据集（$n = 1{,}000$）上，当解的大小为$k = 10$ 时，将参数 ε 从 0.001 调整到 0.5 对 FMMD-S 性能的影响。就解的质量而言，当 ε 取值较小时，FMMD-S 解的多样性值基本保持不变。然而，随着 ε增大，多样性值会下降，且稳定性变差。尤其是当 $\varepsilon = 0.5$时，FMMD-S 在大多数情况下都无法提供稳定且高质量的解。就效率而言，当 ε较小时，运行时间会显著增加，且稳定性变差，尤其是在 $\varepsilon \leq 0.01$时。这些结果与我们对 FMMD-S 的理论分析相符。以上结果表明，在大多数情况下， ε 最合适的取值范围为 $[0.01, 0.1]$ 。因此，我们决定在本文的所有其他实验中均将 $\varepsilon = 0.05$ 设为该值。

[3] archive.ics.uci.edu/ml/datasets/adult
[4] mmlab.ie.cuhk.edu.hk/projects/CelebA.html
[5] https://keras.io/api/applications/vgg/
[6] archive.ics.uci.edu/ml/datasets/US+Census+Data+(1990)
[7] www.kaggle.com/crowdflower/twitter-user-gender-classification
[8] huggingface.co/models?library=sentence-transformers

表4：各数据集组的比例。

数据集	组	各组的比例
Adult	Sex	‘女性’：33.2% ，‘男性’：66.8%
	Race	‘美洲印第安人-爱斯基摩人’：1.0% ，‘亚裔太平洋岛民’：3.1% ，‘黑人’：9.6% ，‘其他’：0.8% ，‘白人’：85.5%
	性别+种族	‘AIE+F’：0.4% ，‘API+F’：1.1% ，‘B+F’：4.7% ，‘O+F’：0.3% ，‘W+F’：26.7% ，‘AIE+M’：0.6%，‘API+M’：2.0%，‘B+M’：4.9%，‘O+M’：0.5%，‘W+M’：58.8%
CelebA	Sex	‘女性’：58.3% ，‘男性’：41.7%
	Age	‘非年轻’：22.6%，‘年轻’：77.4%
	性别+年龄	‘NY+F’：7.3%，‘Y+F’：51.0%，‘NY+M’：15.3%，‘Y+M’：26.4%
人口普查	Sex	‘女性’：51.6% ，‘男性’：48.4%
	Age	‘A1 (12-)’：18.2%，‘A2 (13-19)’：10.0%，‘A3 (20-29)’：15.3%，‘A4 (30-39)’：16.7%，‘A5 (40-49)’：12.9%，‘A6 (50-64)’：13.6%，‘A7 (65+)’：13.3%
	性别+年龄	‘A1+F’：8.9%，‘A2+F’：4.9%，‘A3+F’：7.7%，‘A4+F’：8.5%，‘A5+F’：6.6% ，‘A6+F’：7.2%，‘A7+F’：7.9%，‘A1+M’：9.3%，‘A2+M’：5.1%，‘A3+M’：7.6%，‘A4+M’：8.2%，‘A5+M’：6.3%，‘A6+M’：6.5%，‘A7+M’：5.4%
Twitter	Sex	‘女性’：35.6% ，‘男性’：32.9% ，‘非人类’：31.5%



(a) Adult（性别） (b) Adult（种族） (c) Adult（S+R）

(d) CelebA（性别） (e) CelebA（年龄） (f) CelebA（S+A）

(g) 人口普查（性别） (h) 人口普查（年龄） (i) Twitter（性别）

图 8：FMMD-S 在解的大小为 $k = 10$ 时，参数 ε 取不同值的结果。每个图中的误差棒分别表示 10 次运行中多样性值和运行时间的最大值与最小值。在所有展示多样性值的图中，蓝色水平线表示 FMMD-E 提供的最优解的多样性值，用于衡量 FMMD-S 的近似解与最优解之间的差距。