arXiv:2304.09068v1 [cs.DB] 18 Apr 2023

M ETA M：目标导向的数据发现

Sainyam Galhotra、Yue Gong、Raul Castro Fernandez，芝加哥大学计算机科学系 邮箱： {yuegong, sainyam, raulcf}@uchicago.edu

摘要—**数据是机器学习和因果推断任务的核心组成部分。开放数据存储库、数据湖和数据市场等来源提供了大量数据，这为增强数据并提升这些任务的性能创造了机会。然而，数据增强技术依赖用户手动发现并筛选有用的候选增强。现有解决方案未能利用发现与增强之间的协同作用，因此未能充分利用数据。**

本文介绍了 METAM，这是一种新颖的目标导向框架，它使用候选数据集查询下游任务，形成反馈循环，从而自动引导发现与增强过程。为高效选择候选项，METAM 利用了以下属性：i）数据；ii）效用函数；以及 iii）解集大小。我们展示了 METAM 的理论保证，并通过一系列广泛的任务对其进行了实证验证。总之，我们展示了目标导向的数据发现对现代数据科学应用的潜力。

I. 引言

通过将数据集与其他数据集连接起来进行增强，可以提升因果推断和监督式机器学习等数据驱动任务的效用。开放存储库 [1]，、组织中的数据湖 [2]，，甚至数据市场 [3], [4] 中表的丰富性，带来了大量的增强候选项。在众多候选项中识别出优质的增强候选项，是一个数据发现问题。要解决这一问题，可以使用传统的数据发现系统来确定哪些表能够与输入数据连接，然后再单独确定哪些连接能够提高任务的效用。这种先发现后增强 的方法在数据发现系统返回与任务相关的候选项时能够奏效。然而，在实践中，很难保证数据发现系统能够识别出优质的增强操作，原因在于：i）相关的增强操作取决于任务；ii）分析人员可能不知道哪些属性会使增强操作具有关联性，例如，不知道哪些特征能够增强分类器的预测能力。数据发现与增强之间的脱节带来了研究机会。

我们通过一种称为**目标导向的数据发现**的新方法来把握这一机遇。在目标导向的数据发现中，数据发现并非一次性过程，而是通过执行干预来适应任务：这一过程会增强初始数据集并验证其效用。干预式查询有助于识别能够提升任务效用的增强操作，从而引导发现过程自动最大化任务效用。这实现了两个目标：第一，分析人员无需了解哪些标准能使增强操作更有效，因为这一方法是自动化的。第二，任何具有效用函数的下游任务都能从这种发现方法中受益。目标导向的数据发现意义重大；请看下面这个例子。我们利用目标导向的数据发现来预测某个地理区域的“房价”。我们的技术识别出了一些显而易见的数据集，社会科学家使用发现系统也能找到这些数据集，例如“社区居民收入”和“犯罪统计数据”。但更重要的是，它还识别出了一些与房价相关、却不那么显而易见的数据集，例如“杂货店的分布情况”和“这些地区的出租车出行次数”。事实上，许多社会学家和经济学家会利用外部数据来推断感兴趣属性之间的因果关系 [5], [6], [7]。但他们依赖领域知识和人工劳动来识别这些关系。目标导向的数据发现为自动从大型数据存储库中识别新的因果关系铺平了道路。

解决目标导向的数据发现的一种简单但计算成本极高的方法，是测量每个增强候选项的效用并选择其中最佳者。我们的论文聚焦于开发用于目标导向的数据发现的干预式查询算法，这些算法利用数据、效用函数和解的结构，自适应地优先选择待查询的候选项。

用于目标导向的数据发现的干预式查询。我们的技术贡献利用以下属性，在不损失质量的情况下优化发现的复杂度。数据属性。一个关键洞见是，相似的 增强操作在下游任务上的表现也相似。我们通过对增强项进行聚类，并从不同簇中审慎地选择增强项来利用这一洞见，从而跳过部分计算。为了对增强操作进行聚类，我们使用一组数据画像来表示每个增强操作；这些数据画像是与任务无关的数据度量，包括语义相似度、相关性和互信息等。当数据画像的某种组合与任务效用相关时，聚类可以缩小增强候选项的数量。

效用函数属性。任务的最优效用由一组增强操作决定。枚举所有子集并不可行。我们的洞见是，当效用函数具有单调性时，也就是说，新增增强操作不会降低效用，就可以通过逐一考察增强操作来高效地找到它们。我们还可以通过为用户提供的任务实现添加包装器，忽略会损害效用的增强操作，从而使任意函数都具有单调性。

解集属性。最优增强操作集是跨不同数据集的一组连接路径。大多数增强操作对于下游任务而言无关，而有用的增强操作只是候选集中的一个小子集。我们利用这一属性，通过组合测试优先选择较小的连接路径子集，而不是较大的连接路径集合

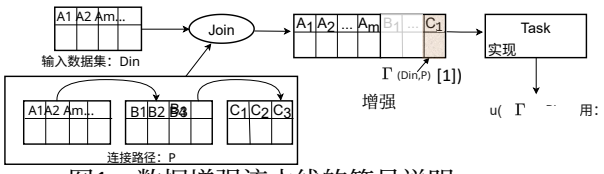


图1：数据增强流水线的符号说明。

通过组合测试优先选择较短的路径，而非较长的路径。

METAM：目标导向的数据发现 我们将上述属性结合起来，构建了一种随时算法，通过自适应地查询任务来查找增强候选项。尽管目标导向的数据发现是 NP 困难问题，但在实践中成立的合理假设下，该算法能够保证识别出近似解。我们还展示了该算法的效率保证，并将其作为端到端数据发现方法的一部分实现于 METAM 中。

评估结果。 METAM 能够在数分钟内从数百万个增强项中自动找到有用的增强项。我们在包含来自美国各城市和 Kaggle 的数据的大型代码库上评估了 METAM，并将其应用于处方分析任务，例如因果推断中的假设分析和操作指南分析、监督式机器学习中的分类和回归，以及实体链接和聚类。

大纲。 我们在第 II 节介绍符号说明并讨论问题。在第 III 节介绍干预式查询的相关见解，在第 IV 节介绍算法，并在第 V 节进行分析。我们在第 VI 节介绍评估结果，并在第 VII 节介绍相关工作。

II. 目标导向的数据发现

本节将在第 II-A 节介绍符号说明和问题陈述，并在第 II-C 节介绍预备知识。

A. 问题定义

图 1 总结了本文所使用的符号说明。令 $\mathcal{R}(A_1, \ldots\ldots, A_m)$ 表示一个包含 m 个属性的关系模式，其中 A_i 表示第 i^{th} 个属性。 D 包含一个模式 $\mathcal{R}(A_1, \ldots\ldots, A_m)$ 和一个元组列表 T，其中每个元组 $t \in T$ 都是该模式的一个具体实例。借助上述符号说明，我们现在定义含噪数据集和数据存储库。

定义 1（含噪结构化数据）。含噪结构化数据集 D 的特征是不完整的模式信息 $\mathcal{R}(A_1, \ldots\ldots, A_m)$，其中 $A_i = \phi$ 表示缺失的标头值；此外，该数据集还包含元组列表 T，这些元组可能少于 m 个值。另外，某些元组可能包含重复值。

定义 2（数据存储库）。数据存储库包含一组数据集 $\mathcal{D} = \{D_1, \ldots, D_n\}$，其中不同的数据集 D_i、D_j 可能包含重叠值。此外，这些数据集可能缺少模式并包含缺失值。

与输入数据集D_{in} 连接的不同数据集称为连接路径，定义如下。

定义 3（连接路径）。连接路径 P 定义为由含噪数据集 $P \equiv \{D_1, \ldots, D_t\}$ 构成的有序列表，其中数据集 D_j 和 D_{j+1} 对于所有 $j < t$ 都彼此连接，形成一系列连接操作。由连接这些数据集 t 所形成的数据集被视为已与原始数据集 D_{in}合并。

将连接路径 P 增强到数据集 D_{in} 记为 $\Gamma(D_{in}, P)$。使用 D_{in}增强连接路径 P可以生成多个新列，其中一些可能并非下游任务所需。为区分这些不同的增强列，我们将增强定义为对仅包含单列的连接结果进行投影。

定义 4（增强）。 增强 $\Gamma(D_{in}, P[j])$ 定义为将路径 P 与输入数据集 D_{in}连接后添加的 j^{th} 列。

请注意， $\Gamma(D_{in}, P) \equiv \{\Gamma(D_{in}, P[j]) \; \forall j \in \{1, 2, \ldots, l\}\}$，其中 l 是具体化连接路径 P后的列数。 Γ 通过具体化连接路径来添加新属性。其他类型的数据增强还包括并集（添加额外的行）、基于关系嵌入的增强、时空增强等。根据实证观察，在常见应用中，基于连接路径添加新属性的做法十分普遍。不过，我们的框架也适用于任何一般的增强集合。任务 t 是一个黑盒，由分析算法构成；该算法将数据集（D_{in} 或 D_{in}的增强版本）作为输入，并输出任务的性能度量。我们将获取任务在某个数据集上的效用这一过程称为“查询”。我们正式定义任务的效用分数如下。

定义 5（效用分数）。 给定一个作用于数据集 D 的任务 t，效用分数 $u_t(D)$ 定义为该任务在使用 D 运行时的目标（或评估指标）。

不失一般性，我们假设效用分数已归一化为 $u_t(\cdot) \in [0, 1]$，且效用分数越高表示任务质量越好。不同的增强操作集可能会提高任务效用。我们希望找到一个最小增强集，其中仅包含有助于维持高效用的增强操作。

定义 6（最小增强集）。 给定一个初始数据集 D_{in}、一个用于计算 $u_t(\cdot)$ 的机制，以及一个增强操作集 $\mathcal{P} = \{P_1[i_1], \ldots, P_j[i_j]\}$，如果从 $\Gamma(D_{in}, \mathcal{P})$ 中移除任意属性都会降低任务效用，则称该增强集是最小的。

问题 II.1（目标导向发现）。 给定一个初始数据集 D_{in}，以及一个具有机制、能够为任意表 D' 计算 $u(D')$ 的任务 t，识别能够优化该任务的增强操作这一问题，就是识别一个最小增强集 $\mathcal{T}$，使得增强数据集具有 $u(\Gamma(D_{in}, \mathcal{T})) \geq \theta$. 的效用。

B. 任务实现与效用计算

任务将数据集作为输入，并返回取决于输入数据集（至少部分取决于它）的效用分数。示例包括：1）预测分析：任务在输入训练数据集上训练机器学习模型。训练代码可以很简单，例如使用随机森林；也可以很复杂，例如使用自动机器学习解决方案。效用对应于与模型相关的指标，例如准确率、F1 分数，还可能包括公平性和鲁棒性指标。

2）处方分析涉及假设分析（假设分析和操作指南查询）、因果推断以及可解释性这几个关键组成部分
[8], [9], [10], [11]。效用对应于一个指标，用于概括所生成解释的质量、所识别因果关系的支持程度以及属性的总体因果效应。
3）数据准备与可视化：包括实体解析、实体链接、数据清洗和可视化。效用通常对应F分数和准确率。

请注意，这些只是任务实现的一些示例；特征工程、数据转换和数据分析方面的任何进展都可以纳入效用函数。

C. 预备知识与问题讨论

我们利用以往的数据发现技术 [12], [13],[14], [15]从大型数据存储库中获取一组增强操作；我们使用 Aurum [12]。尽管现有解决方案可能会生成含噪声的候选项（由于使用近似技术和存在语义歧义），但即使存在不同来源的噪声，我们的方法仍然有效。我们还利用现有的分析技术 [16], [17], [18], [19]来描述增强候选项 $\Gamma(D_{in}, \mathcal{P})$。例如，数据剖面 $\langle Corr, C_1, C_2 \rangle$ 指列 C_1 与 C_2之间的相关性。我们用术语数据剖面值 来表示该特征/属性的值。数据剖面可以指单个属性（例如属性的取值域）或多个属性（属性之间的相关性），而这些属性可能与下游效用相关。

定义 7（数据画像）。 数据画像 X 定义为数据集 D 的属性，其中 D 中的元组满足 X。
在我们的实现中，我们考虑以下数据画像。

- **相关性和互信息（MI）**： MI 常用于评估属性之间的因果依赖 [20]。相关性和 MI 也是机器学习任务中衡量属性质量的预测指标。此画像估算候选增强操作 P 与 D_{in} 的属性之间的皮尔逊相关性（MI）。

- **基于语义嵌入的距离**画像刻画所考虑数据集之间的语义相似度。此画像通过计算两个数据集嵌入之间的余弦相似度得出；这些嵌入由 BERT 等预训练模型构建 [21]。数据集嵌入通过对表中出现的词元的嵌入向量取平均值构建。

- **数据集元数据/属性**画像根据数据集来源和属性计算数据集之间的相似度。与语义嵌入不同，此画像刻画属性与数据集来源之间的句法相似性，这通常用于估算增强操作的质量 [22]
- **数据集重叠**画像用于计算增强后所识别最终数据集的基数。

扩展至其他数据画像。 Metam 可以扩展并加入新的画像，以满足新的下游任务需求（例如，异常检测 [23]，用于公平性的条件独立性 [24]），或利用分析技术的进展 [23], [25], [26], [27]。

问题讨论： 我们讨论各种“先发现、后增强”的基线方法为何无法高效解决这一问题：
- **连接所有数据**：此技术将 D_{in} 与识别出的每条连接路径进行连接。这种方法可能会引入无关属性，从而降低效用。此外，在包含数千种增强操作的大规模场景中，这种方法不可行。

- **均匀采样**：此技术从已识别的可连接数据集中均匀随机地采样连接路径，并用它们来增强 D_{in}。这种方法不能保证所选样本会提高效用分数。

- **连接路径重叠度排序**：一种常见技术是根据增强后数据集的基数对连接路径进行排序（S4 [14]和 Ver [22]使用了该方法）。此技术能够识别缺失值较少的数据集，但不能保证优化任务。

- **使用数据画像进行选择**：分析人员可以凭直觉，根据预期有用的画像对增强操作进行排序。如果画像准确且领域科学家的直觉正确，这种方法可能有效；否则效果不佳。该方法还依赖于对画像进行准确估计，而在大规模数据存储库中，这类估计难以计算，为了实现可扩展性，通常需要采用近似技术。

问题的核心在于，这些基线方法不了解任务，因此无法保证所选择的增强操作会提高效用。要解决目标导向的数据发现，需要采用干预式方法，以探究哪些增强操作会使任务的效用增加。

III. 基于干预的查询

本节首先介绍基线技术的局限性，然后讨论数据、效用函数和解的性质如何帮助我们设计高效的技术。

A. 基线干预技术

目标导向的数据发现可以通过枚举候选增强的每个子集、计算其效用，并选择达到至少为 θ 的效用的最小子集来解决。对于 n 个候选增强，此过程最多可能需要 $O(2^n)$ 次查询（对于 $n > 30$ 无法完成），因此这种方法不可行。我们考虑两种改进方法：
- **基于效用的选择：** 给定增强操作的排序（例如，基于某些数据画像），该方案按照排序顺序迭代查询，直到获得的效用高于所需的阈值 θ。每当数据剖面与任务无关时，该方案都会效率低下。而且，对于所考虑的许多有趣的数据驱动任务而言，选择与下游任务相关的数据剖面并不显而易见。

- **专家建议预测：** 为避免先验选择某个画像，可以将不同的数据画像视为专家，并使用专家选择技术来解决这一问题。例如，乘法权重更新方法（MW） [28]会估计不同专家提升效用的能力。MW 保证在事后选择表现最佳的专家。在最简单的形式中，将每个画像视为不同的专家，无法识别出最适合

排序增强的画像组合。将画像组合枚举为专家，会给决策技术（包括 MW） [28]带来已知的组合问题。

这两种技术有两个缺点：i）它们无法保证找到能够优化任务效用的最小增强集；ii）它们需要 $O(n)$ 次查询，其中 n 表示候选增强的数量。此外，在最坏情况下，没有任何算法能够在少于 $(2^n - 1)$ 次查询的情况下找到最优解（定理 2）。

B. 面向高效干预式查询 g

尽管无法设计出能够优化最坏情况的技术，但我们发现了实际场景的若干特性，这些特性有助于高效地找出最优解（或近似最优解）。

P1（最优解通常只包含少量增强操作）。大多数增强操作对任务并无帮助。最优解中的增强操作数量 k 远低于增强操作总数 $k << n$。因此，考虑规模较小的候选增强子集，很有可能找到最优解。组合测试领域的文献此前已研究过这一思路。我们利用这一洞见，优先考虑规模较小的子集，而非规模较大的子集。

如果 P1 不成立，会怎样？ k 并不比 n 小很多的情形并不现实，因为增强数千个新属性会导致空间规模急剧膨胀，并降低整体效率。即使在这种情况下，该算法仍能识别出最优解，因为它最终会考虑包含大量元素的子集。

实证验证。对第 VI 节中考虑的 10 个不同任务进行评估后发现，每种情形都有超过 5000 个候选增强（$n > 5000$）。然而，在其中 8 种情形下，最佳解包含的增强少于 5 个，其余情形少于 25 个。因此，实际上只有不到 0.5% 的候选增强有助于提高效用。

P2（相似数据集很可能对效用分数产生相似的影响）。包含相同元组集合的两个数据集对效用分数具有相同的影响。尽管重复项的存在属于极端情况，但我们观察到，开放数据集通常包含重复信息。通过考虑数据集之间的相似度，我们将这一观察扩展到一般数据集。预计增强两个具有相似画像值的不同数据集会以超过 0.5 的概率对效用分数产生相似的影响。这一性质促使我们考虑数据集属性，对相似增强进行聚类，并在分析时整体考虑所有簇内增强。我们使用第 II-C 节中介绍的数据画像。数据画像不仅可用于对增强进行聚类，还可用于为其评分（我们将这些分数称为质量分数）。

如果 P2 不成立呢？METAM 利用这一性质来确定查询增强操作的优先级。如果数据集相似度与其效用分数之间不存在关联，那么所确定的增强操作排序将与随机排序相同。在最坏情况下，这会使所需的查询次数增加到 $O(nk)$，其中 n 是候选增强的总数，而 k 是最优解中的增强操作数量。

实证验证。我们比较了相似度在 $[0.9, 1]$ 范围内的增强操作的效用差异，发现其中超过 85% 的效用差异小于 0。02。此外，随着相似度降低，效用差异会增大。因此，高度相似的数据集通常具有相似的效用。

P3（效用函数的单调性）。效用函数通常满足相对于增强操作的单调性，即向数据集中添加新列不会降低任务的效用。例如，因果推断任务通常会估计已识别属性的总因果影响，这具有单调性。贝叶斯最优分类器的准确率和 F 分数 [29] 也具有单调性（因为如果新增特征无助于预测，贝叶斯最优分类器会忽略它）。然而，由于各种原因，例如新添加属性中的缺失值或噪声，某些效用指标可能不具备单调性。仍可通过一种机制来确保单调性：如果某项增强操作降低了效用，该机制便会忽略它。我们的框架通过单调性认证组件（图 2）来保证单调性。

如果 P3 不成立会怎样？单调性认证组件会忽略那些会降低效用的增强操作。为了验证添加新的增强操作是否会降低效用而进行的额外检查，可能需要更多查询。

实证验证。我们评估了所考察场景中效用的单调性，并发现因果推断任务（假设分析和操作指南）始终具有单调性，而超过 60% 的分类任务查询也具有单调性。在其余 40% 的情况下，单调性认证组件会发起额外的 ≈ 20 查询，并忽略会降低效用的增强操作，以确保单调性。

IV. 面向目标的发现算法

本节将在第 IV-A 节概述 METAM，并在第 IV-B 节讨论其子程序。图 2 展示了 METAM 的各个组件。

A. METAM 算法概述

METAM 的输入包括初始数据集 D_{in}、一项任务 t 和一个数据集集合 $\mathcal{D}$，输出则是一个最小增强集，确保增强数据集的任务效用至少为 θ （见算法1）。该算法包含两个主要组件：i）候选生成与似然估计；ii）自适应查询策略。第一个组件负责识别候选增强，并计算数据画像向量。第二个组件在两种互补机制之间交替运行，利用属性 P1–P3 对任务进行查询。交替使用这两种流程，使 METAM 能够发挥两种技术各自的优势，并在不对效用函数作任何假设的情况下找到解决方案。

候选生成和似然估计。首先，METAM 识别由传统发现技术（第 II-C 节）为 D_{in} 提供的 $\mathcal{P}$，即候选增强（第 1 行）。系统会处理每个增强操作，以计算其

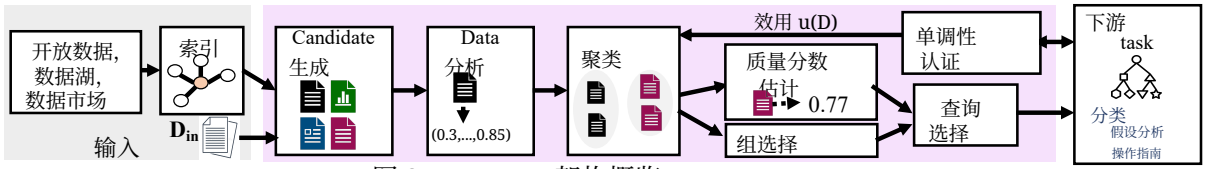


图 2：METAM 架构概览。

算法1： METAM

输入： 数据集 D_{in}、数据存储库 $\mathcal{D}$、效用阈值 θ
输出： $\mathcal{T}$：增强操作列表

1 $\mathcal{P} \leftarrow$ 生成候选项（D_{in}，$\mathcal{D}$）
2 评估配置文件（$\mathcal{P}$，D_{in}）
3 $\mathcal{T}_m^*$ $\mathcal{T}_c^* \leftarrow \phi$、$t \leftarrow 1$、$D \leftarrow D_{in}$
4 JPSCORE $\leftarrow$ 估计质量分数（$\mathcal{P}$，D_{in}）
5 $\mathcal{C} \leftarrow$ CLUSTER-PARTITION($\mathcal{P}$，D_{in}，ϵ)
6 **当** $u(\Gamma(D, \mathcal{T}^*)) < \theta$ 且 $u(\Gamma(D, \mathcal{T}_c^*)) < \theta$ **执行**
7 $\mathcal{X}$，$Q \leftarrow \phi$，$i \leftarrow 0$
8 **当** $i < \tau$ 或 最大值$_{P \in Q} u'[P] \leq u(D)$ **</style>执行**
9 $P_{max} \leftarrow$ 最大值$_{P_i \in \mathcal{P} \setminus (\mathcal{X} \cup \mathcal{T}^*)}$ JPS核心(P_i)
10 $u'[P_{max}] \leftarrow u(\Gamma(D, \{P_{max}\}))$
11 $\mathcal{X} \leftarrow \mathcal{X} \cup$簇$\{P_{max}\}$
12 更新质量分数($\mathcal{P}$，P_{max}，u')
13 $\mathcal{P} \leftarrow$ IDENTIFY-GROUP($\mathcal{C}, t$)
14 **如果** $u(\Gamma(D_{in}, \mathcal{P})) > u(\Gamma(D_{in}, \mathcal{T}_c^*))$ **则**
15 $\mathcal{T}_c^* \leftarrow \mathcal{P}$, $u'[\mathcal{T}_c^*] \leftarrow u(\Gamma(D_{in}, \mathcal{T}_c^*))$
16 $i \leftarrow i + 1$，$Q \leftarrow Q \cup \{P_{max}\}$
17 $P'_{max} \leftarrow$使$_{P \in Q \cup \{\mathcal{T}_c^*\}} u'[P]$最大化
18 **如果** $u(\Gamma(D, \{P'_{max}\})) > u(D)$ **则**
19 $\mathcal{T}^* \leftarrow \mathcal{T}^* \cup \{P'_{max}\}$
20 $D \leftarrow \Gamma(D, \{P'_{max}\})$
21 **如果**CHECK-STOP-CRITERION() **则**
22 中断
23 $\mathcal{T}^* \leftarrow$使$_{\mathcal{T} \in \{\mathcal{T}^*, \mathcal{T}_c^*\}}$最大化$_{\mathcal{T}}$ $u(\Gamma(D_{in}, \mathcal{T}))$
24 $\mathcal{T} \leftarrow$ 识别最小值($\mathcal{T}^*, \theta$)
25 返回 $\mathcal{T}$

数据画像向量（第 2 行）。首先，系统根据增强相似度对候选增强进行聚类；预计同一簇中的增强操作会以相似方式影响效用分数（性质 P2）。查询策略利用这一性质，从每个簇中选择代表元，从而减少对效用函数的查询次数。数据画像还用于计算增强操作的质量分数，以估计其对任务效用的贡献。直观地说，质量分数是一种似然估计，有助于更早查询相关的增强操作。我们将在实现细节中解释评分技术。系统利用这些簇和质量分数迭代查询任务，以评估增强操作的效用。

自适应查询策略。 在每次迭代中（第 8–16 行），算法交替采用 顺序查询 和 分组查询，以评估最有希望的查询的效用分数。顺序机制（ 以蓝色突出显示）通过依次查询任务来估计候选增强的质量分数，并选择最佳增广。该过程通过在各个簇之间分散查询来处理增强操作。具体而言，某项增强的质量分数是各个画像分数的加权平均值，其中画像的重要性被用作权重。画像的重要性通过评估较高轮廓值确实能带来较高效用的似然来估计。质量分数按非递增顺序排序，以便查询任务，同时限制最多从一个增强操作中考虑

算法 2： CLUSTER-PARTITION

输入： 增强操作 $\mathcal{P}$、 D_{in}，聚类半径 ϵ**输出：** $\mathcal{C}$：增强操作簇
1 $c_1 \leftarrow$ 随机选择（$\mathcal{P}$）
2 $S \leftarrow \{c_1\}$，$\mathcal{C} \leftarrow \{\mathcal{P}\}$
3 **当** DISTANCE($\mathcal{C}$，S) $> \epsilon$ **时执行**
4 $c \leftarrow$ 选择最远项（S）
5 $S \leftarrow S \cup \{c\}$，$\mathcal{C} \leftarrow$ 分配（$\mathcal{C}$，S）
6 返回 $\mathcal{C}$

每个簇。在对某个增强操作 P进行每次查询后，已识别的效用分数 $u(\Gamma(D, \{P\}))$ 用于更新数据画像的重要性以及其他增强操作的相应分数。在选择效用最大的增强操作之前进行查询的增强操作数量由参数 τ控制。经过 τ 次查询后，将效用增益最大的增强操作（P'_{max}）添加到解集 $\mathcal{T}^*$，并将 D 更新为 $\Gamma(D, \{P'_{max}\})$，以用于后续迭代。分组机制（红色标出）考虑大小为 t 的子集（初始化为 $t = 1$，见第 3 行）并评估其效用。该方法使用伪贪森采样 [30], [31]机制构造子集，从而优先选择有用的簇，而非相关性较低的簇。我们将在第 IV-B 节介绍此机制。当所有大小小于 t 的集合都已查询后，t 的值会增加。

METAM 会继续执行查询过程，直到增强数据集的效用不低于 θ，或者所有增强操作都已查询，且没有任何操作能够提升任务效用。

最小性检查。 在最后阶段，从$\mathcal{T}^*$ 和 $\mathcal{T}_c^*$中选出最佳解，并对已识别的增强操作进行后处理，以找出一个能够达到至少 θ效用分数的最小集合。该组件每次从解集中迭代移除一个增强操作，并评估效用分数 $u(\Gamma(D_{in}, \mathcal{T}^* \setminus \{T\}))$。如果该修改后的数据集的效用达到或超过 θ，则从解集中移除 T。

停止准则。 算法1是一种随时算法，当任务效用达到阈值 θ时停止。如果未提供 θ，算法将继续执行，直到：i) 搜索空间已被探索完毕；或 ii) 用户找到一个满意的解。

我们指定了一个 CHECK-STOP-CRITERION 子程序，用于检查用户指定的所有停止要求，例如时间约束、查询预算、解大小约束等。

B. METAM 子程序与泛化

现在，我们将深入探讨算法1中使用的 METAM 细节：如何对增强操作进行聚簇、估算质量分数，以及将查询的输出传播到其他增强操作。

CLUSTER-PARTITION。 我们使用一种基于距离的聚类技术，以确保同一簇中的增强操作，其所有画像值与其代表元之间的距离均在 ϵ 以内。换言之，算法 2 将该空间划分

为由 2ϵ 个宽度立方体组成的空间（维度为 l，其中 l为所考虑的画像数量），其中簇代表元构成中心。这一过程通常被称为基于由其画像构建的嵌入对增强操作进行的 ϵ-覆盖。参数 ϵ 在簇的数量（会影响查询复杂度）与簇的质量之间引入了权衡。算法 2 给出了聚类算法的伪代码，该算法调整贪心 k 中心聚类算法 [32]，逐步增加 k，直到所有簇的半径均小于 ϵ。增强操作 P_1 与 P_2 之间的距离计算为 $d(P_1, P_2) = \max_{i \in R} d(r_1^i, r_2^i)$，其中 R 表示画像集合。

给定候选增强 $\mathcal{P}$ 和作为输入的初始数据集 D_{in}，算法选择簇代表（也称为中心）来初始化每个簇，然后将所有增强分配给相应的聚类中心。S 表示这些中心，$\mathcal{C}$ 表示将 $\mathcal{P}$ 划分到相应簇中的划分结果。算法首先随机选择一个增强作为第一个中心，并将所有候选增强分配给该中心（第 1 行）。随后，通过选择与其中心距离最大的增强来确定后续中心（第 3–5 行）。确定新中心后，所有增强都会被重新分配到更新后的中心集合中。这个最远增强识别与重新分配过程会持续进行，直到最远增强与其中心的距离不超过 ϵ。

质量分数估计。增强的质量分数根据其提高任务效用的预期对其进行排序。该分数由两个部分组成。第一部分是基于画像的分数，即其画像值的加权平均值。该分数等价于根据数据集属性估计出的先验。流水线初始化时为所有画像分配相同的权重，并随着查询数量的增加逐步改进这些权重。形式上，画像 p 的重要性权重定义为使用 p预测增强效用时的特征重要性。第二部分是基于效用的分数，表示增强一条连接路径所带来的效用分数增益。如果增强 P之前已经被查询过，则将效用增益视为其效用分数。如果 P 尚未被查询，但来自同一簇的另一条连接路径（例如 P'）已经被查询过，则 P的效用分数计算为 $(1-d(P, P'))$ 乘以 P'的分数，其中$d(P, P')$表示增强之间的距离。

增强操作的质量分数定义为基于画像的分数与基于效用的分数之和。该分数取决于画像的重要性权重，以及同一簇中其他增强操作的效用分数。
UPDATE-QUALITY-SCORES 机制执行此更新。

识别-分组。从高层次来看，每个簇都由一个概率表示，该概率表示通过增强操作获得更高效用的可能性。在没有查询的情况下，我们无法准确估计这些概率。从某个簇中查询增强操作可以准确估计该概率，但挑战在于：是选择查询单个簇以获得准确估计，还是在多个簇之间进行多样化探索，以探索其他选项。这遵循伯努利老虎机中的传统探索-利用困境 [31], ，其中可能性可以建模为后验概率。

我们将每个簇建模为一个老虎机，其中拉动一条臂等价于采样一个增强操作。任务效用的增量就是给定簇的奖励。我们将每条臂的概率初始化为 $1/|\mathcal{C}|$，并维护获得奖励的次数，以评估后验概率。大小为 k 的子集中的每个元素都根据该概率从这些簇中随机采样。

泛化：当画像没有用时该怎么办？ 聚类算法识别出画像值相似的增强操作组，这会影响算法1的查询选择。现在，我们分析数据画像中的两种不同噪声对 METAM 的影响。第一种情况是：如果两个效用相似的增强操作没有相似的数据画像，那么这些增强操作会被置于不同的簇中，并分别进行查询以计算效用。这种情况类似于不使用聚类的方法。它不会影响 METAM 的质量，但可能会导致查询复杂度增加。第二种情况是：当不相似数据集具有相似的数据画像时，画像相似的两个增强操作可能不会产生相似的效用。为处理这类情形，我们提出一种根据画像质量进行调整的策略。METAM 不再对每个簇进行一次查询，而是从一个簇中随机选择 $\log|C|$ 个增强操作 C 进行查询，以高概率准确估计同质性，即检查采样增强操作中的大多数，其效用是否处于所查询增强操作平均效用的 $(1+\epsilon)$-近似范围内。如果不满足同质性条件，则质量分数估计会忽略效用分数组件，并在算法的后续迭代中将每个元素视为一个独立的簇。**选择画像**。开发者会纳入任何可能与下游任务相关的画像。他们可能不知道是否存在这种相关性，也不知道相关性是否强，因此会“广撒网”，希望某些画像组合能够呈现相关性。如果确实存在这种情况，
METAM 会在搜索过程中识别出来。如何平衡所纳入画像的数量和类型不在本文讨论范围内；默认情况下，我们纳入对机器学习和因果推断任务有效的画像。此外，采样技术可以降低画像计算的成本。

影响 τ。 τ 决定在选择效用增益最大的增强操作之前所考虑的簇的数量。该参数的值越大，就越能保证在选择效用增益较低的增强操作之前，先选取效用增益最大的增强操作。直观而言，与任意最小集合相比，这种方法能够识别出规模更小的最小增强操作集。默认情况下，我们选择 $\tau = |\mathcal{C}|$，以确保每个簇至少有一个增强操作被查询。选择较小的 τ，等同于依靠质量分数来选取质量分数最高的簇。在极端情况下，$\tau = 1$ 等同于按质量分数非递增的顺序查询增强操作

分数，并选择任何能够提升解集中任务效用的增强操作（无论其效用增益如何）。当无需优化解集大小时，应考虑采用 τ 的这些设置。我们在第 VI-A 节中展示了 τ 的影响。

注。 无法提升效用的增强操作可能是错误的（例如，因键不正确而导致连接错误），也可能是正确的，但对下游任务没有帮助。METAM 在 $\mathcal{P}$中能够稳健地处理错误增强项。

风险与脆弱性。 METAM 会选择能够最大化效用的属性。当效用函数未能反映应用需求时，例如，当结果将用于无法完全自动化的关键决策软件时，数据科学家可以手动核验 METAM 返回的增强操作，并标记不符合要求的增强操作。标记这些增强操作后，可能需要重新运行 METAM，以便重新优化并找出另一组最小增强集。

V. METAM 的性能保证

本节将证明，目标导向的数据发现是 NP 困难（定理 1），并且任何算法都无法在少于 2^n-1 次查询的情况下确定最优解。由于最坏情况极为人为构造，我们将分析 METAM 在实际场景中的效率。我们将证明，METAM 能够通过 $O(\log n)$ 次查询找到最优解的常数近似。我们还将讨论噪声环境，在这些环境中，前文讨论的性质（第 III-B 节）可能不成立。

NP 困难性。 解决目标导向的数据发现的一种蛮力方法是考虑增强操作的所有子集，并选择效用最大的子集。为了证明该问题是 NP 困难的，我们首先考虑该问题的判定版本，然后展示从集合覆盖问题到该问题的归约。考虑一个判定版本：如果存在一个大小为 k 的集合，使得增强该集合的效用至少提高到 θ，则输出“是”。

定理 1。目标导向的数据发现是 NP 困难的。

证明。 考虑集合覆盖问题的一个实例：给定集合 $U=\{v_1,\ldots,v_n\}$ 上元素的集合族 $\mathcal{S}=\{S_1,\ldots,S_m\}$。我们希望确定是否存在 $\mathcal{S}$ 中的 k 个集合，使得它们的并集为 U。

我们证明，集合覆盖问题是目标导向数据发现问题判定版本的一个特例。考虑一个接收增强操作集合并输出其效用的算法。对于每个集合 $S_i \in \mathcal{S}$，定义一个增强操作 P_i。令效用函数为 $u(\mathcal{P}) = |\bigcup_{P_i\in\mathcal{P}} S_i|/n$。如果存在一个算法，能够识别出一个大小为 $\leq k$ 的最小集合，使得 $u(\mathcal{P}) = 1$，那么集合覆盖问题就存在一个解。这表明集合覆盖问题可以被最优地求解，从而产生矛盾。 □

定理 2。在最坏情况下，除非执行 $O(2^n)$ 次查询，否则每个算法都只能识别出任意差的解。证明。 令 $\mathcal{T}$ 表示这样一组增强操作： $\mathcal{T}$ 的任意 2^n-1 个子集都是有效增强项。考虑一个对抗性效用函数，对于所有查询，它都会输出 $u(\Gamma(D_{in}, T)) = u(D_{in})$，直到该算法查询了 2^n-2 个 $\mathcal{T}$ 的子集。对于最后一次查询，它输出 $u(\Gamma(D_{in},\ T^*)) = \theta$。该效用函数将需要 2^n-1 次查询才能识别出最优解。 □

A. 质量保证

我们表明，即使在最坏情况下，METAM 在允许运行 n^k 次查询时也能保证找到最优解。随后，我们表明，对于第 III-B 节所讨论的实际场景，METAM 能在 $O(k)$ 次查询中识别出一个常数近似解。

最优性保证。现在我们将证明，Metam 能够识别出一组增强操作，并保证这些增强操作具有高效用。

定理 3。如果 $\exists\ \mathcal{T}^*$ 使得 $u(\Gamma(D_{in},\ \mathcal{T}^*)) \geq \theta$，则 METAM 的输出 $\mathcal{T}$ 满足 $u(\Gamma(D_{in},\ \mathcal{T})) \geq \theta$。证明。 METAM 查询策略中的组合测试组件会探索增强操作的所有可能子集，直到识别出一个有效解集。因此，可以保证 METAM 的输出至少达到 θ 的效用。由于该方法会考虑所有大小至多为 k 的子集，并逐步增加 k，因此它能在 $O(n^k)$ 次查询中识别出该解，其中 k 是最优解的大小。 □

近似保证。 我们首先证明，METAM 能够在 $O(\log n)$ 次查询内识别出最优解的常数近似。我们假设最优解包含 k 个增强操作，其中 k 为常数。为证明这一结果，我们首先考虑性质 P2 和 P3 成立的情形。随后，我们放宽这些假设，并证明即使相似数据集产生不同的效用（P2 不成立），且效用是非单调的（必须通过包装器确保 P3），METAM 仍能识别出一个近似解。

符号说明。 $\mathcal{C}=\{C_1,\ldots,C_t\}$ 表示聚类中心的集合，$\mathcal{T}^*=\{T_1^*,\ldots,T_k^*\}$ 表示最优解，$\mathcal{C}^*$ 表示聚类中心上的最优解，而 METAM 返回的解为 $\mathcal{T}=\{T_1,\ldots,T_k\}$。

定理 4。 METAM 的查询方法能够在 $O(\log\ n)$ 次查询内识别出具有 $\frac{1}{\alpha}(1-e^{-\alpha\gamma})\ -k\epsilon$ 近似的解。

证明。 首先，我们说明，代表元（聚类中心）集合上的最优解是所有增强操作上的最优解的 $(1-k\epsilon)$ 近似解 $\mathcal{P}$（使用引理 1）。其次，我们证明，METAM 识别出的解是中心上的最优解的 $\frac{1}{\alpha}(1-e^{-\alpha\gamma})$-近似解。结合这两个证明，我们即可得到所需结果。形式化地，

$$
\begin{aligned}
&u(\Gamma(D_{in},\mathcal{T}) \geq 1/\alpha(1-e^{-\alpha\gamma})u(\Gamma(D_{in},\mathcal{C}^*))\\
&\geq \frac{1}{\alpha}(1-e^{-\alpha\gamma})(1-k\epsilon)u(\Gamma(D_{in},\mathcal{T}^*)\\
&\geq \left(\frac{1}{\alpha}(1-e^{-\alpha\gamma})-k\epsilon\right)u(\Gamma(D_{in},\mathcal{T}^*) = \left(\frac{1}{\alpha}(1-e^{-\alpha\gamma})-k\epsilon\right)OPT
\end{aligned}
$$

这表明，METAM 在进行 k 轮最佳增广搜索后识别出近似解。此外，METAM 通过从每个簇中查询 $O(\log n)$ 个查询的随机样本来测试性质 P2（每个簇的组成）。此步骤会增加 $O(|\mathcal{C}|\log\ n)$ 的复杂度，用于测试簇的同质性。引理 2 表明，$|\mathcal{C}| = O(1/\epsilon^l)$，从而得到总体复杂度为 $O(\log n)$。

□

引理 1. 基于代表元 $\mathcal{C}^*$,$u(\Gamma(D_{in},\mathcal{C}^*)) \geq (1-k\epsilon)OPT.$

证明。 为证明该引理，我们利用簇的性质：中心的效用与簇中任何其他增强的效用之比在 $(1+\epsilon)$ 以内。

$$\begin{aligned} u(\Gamma(D_{in},\mathcal{T}_i^*)) &= u(\Gamma(D_{in},\mathcal{T}_{i-1}^*\cup\{P_i\})) \\ &\quad \text{Cluster property of adding a new augmentation} \\ &\le (1+\epsilon)u(\Gamma(D_{in},\mathcal{T}_{i-1}^*\cup\{C_i^*\})) \\ &\le (1+\epsilon)u(\Gamma(D_{in},(\mathcal{T}_{i-2}^*\cup\{C_i^*\})\cup P_{i-1})) \\ &\quad \text{Recursively applying the cluster property} \\ &\le (1+\epsilon)^k u(\Gamma(D_{in},\mathcal{C}_i^*)) \end{aligned}$$

因此， $u(\Gamma(D_{in},\mathcal{C}_i^*)) \ge (1+\epsilon)^{-k}OPT \approx (1-k\epsilon)OPT$

□

现在，我们界定 METAM 的 CLUSTER-PARTITION 组件所生成的簇的数量，以表明查询数量随解集大小线性增长。

引理 2。METAM 生成的簇的数量为$O(\frac{1}{\epsilon^l})$，其中 l 是所考虑的画像数量。

证明。 每条连接路径由 l 个画像表示，且每个轮廓值都在 $[0,1]$内。每当聚类半径大于 ϵ时，聚类算法就会选择一个新的中心。画像所在的 l 维空间具有单位体积。因此，该空间可以由半径为 ϵ的 $O(2^l/\epsilon^l)$ 个球体覆盖。 □

引理 3。 $u(\Gamma(D_{in},\mathcal{T})) \ge \frac{1}{\alpha}(1-e^{-\alpha\gamma})u(\Gamma(D_{in},\mathcal{T}^*))$，其中 α 和 γ 分别表示 u 的曲率和子模比率。

证明。 这里，我们分析一种特殊情况，其中 $\gamma=1$，也就是说，该函数始终是子模的，并在命题之后给出一般结果的见解。

$$\begin{aligned} &OPT = u(\Gamma(D_{in},\mathcal{T}^*)) \le u(\Gamma(D_{in},\mathcal{T}^*\cup\mathcal{T}_i)) \text{ (using monotonicity)} \\ &\le u(\Gamma(D_{in},\mathcal{T}_i)) + \sum_{j=1}^{k}\Big(u(\Gamma(D_{in},\mathcal{T}_i\cup\{T_j^*\})) - u(\Gamma(D_{in},\mathcal{T}_i))\Big) \\ &\text{Since } T_{i+1} \text{ has the maximum marginal gain} \\ &\le u(\Gamma(D_{in},\mathcal{T}_i)) + \sum_{j=1}^{k}\left(u(\Gamma(D_{in},\mathcal{T}_i\cup\{T_{i+1}\})) - u(\Gamma(D_{in},\mathcal{T}_i))\right) \\ &= u(\Gamma(D_{in},\mathcal{T}_i)) + k\left(u(\Gamma(D_{in},\mathcal{T}_i\cup\{T_{i+1}\})) - u(\Gamma(D_{in},\mathcal{T}_i))\right) \end{aligned}$$

因此， T_{i+1} 的边际增益至少为 $\frac{1}{k}(OPT - u(\Gamma(D_{in},\mathcal{T}_i)))$。令 $\delta_i = OPT - u(\Gamma(D_{in},\mathcal{T}_i))$。因此， $T_{i+1} = \delta_i - \delta_{i+1}$的边际增益。将其代入上面的方程，得到 $\delta_i - \delta_{i+1} \ge \frac{1}{k}\delta_i$。因此， $\delta_{i+1} \le (1-1/k)\delta_i$。递归应用这一过程，我们得到

$\delta_k \le (1-1/k)^k\delta_0 \le \frac{1}{e}\delta_0 = (\quad)$。因此，$(\quad)(\quad)$

$\frac{1}{e}OPT \qquad u(\Gamma(D_{in},\mathcal{T}_k) = (1-1/e)OPT$

□

这个证明扩展自子模优化的传统结果 [33]。对于一般的效用函数，可以使用 [34] 中的技巧扩展上述证明。

对 α 和 γ的依赖性。 METAM 不使用 α 和 γ 作为输入，这些概念仅用于刻画可分析的效用函数类别。该结果的近似比取决于效用函数的曲率和子模比率，这两个指标反映了效用函数满足子模性的可能性。近期研究已经针对多个现实世界函数研究了这些参数的估计问题，例如线性回归、稀疏特征选择、贝叶斯 A-最优性、行列式函数，以及带有组合约束的线性规划 [35], [34]。然而，对于一般函数，如何界定这些比率仍是一个开放问题。

实证依据。第VI节中的实验表明，大多数簇是同质的，即相似数据集具有相似的效用。因此，METAM 可以在 1000 次查询内识别出所有有用的增强操作；如果不进行聚类，则需要超过10× 次查询。

含噪簇。 当所识别的簇不具备同质性（不满足性质 P2），因此 METAM 在第一轮检查同质性后忽略所有簇（将每个增强操作视为其自身的簇）时，我们证明，经过 $O(\log n)$ 次查询，METAM 的质量分数估计能够准确地对候选增强操作进行排序（证明见引理4）。因此，METAM 能够在 $O(\log n)$ 次查询内识别有用的增强操作。此外，在最坏情况下，当由于数据中的噪声导致估计分数也不准确时，METAM 需要 $O(n)$ 次查询来选择一个增强操作；这一过程会重复 k 次，从而得到总体复杂度 $O(nk)$。

符号说明 令 $\beta^* = \langle\beta_1^*,\ldots,\beta_l^*\rangle$ 表示用于质量估计的最优轮廓分数。因此，某个增强操作的最优质量分数为 $q_i = \sum_{i=1}^{l}\beta_i^* p_i + \delta$，其中 p_i 是第 i 个轮廓值， δ_i 是均值为 0、方差为 σ^2 的高斯噪声。对于该分析，我们利用估计的轮廓分数的闭式表达式 [?], $\beta = (m\Sigma)^{-1}\sum_i p_i q_i$，其中 $\Sigma = 1/m\sum p_i p_i^T$。

引理4。 推断出的概况重要性，在 m 增强样本上的 $\beta = \langle\beta_1,\ldots,\beta_l\rangle$具有 $E[||\beta^*-\beta||_2^2] \le c\frac{l}{|\mathcal{T}|}$，其中 c 是一个常数。

证明。 我们首先估计 β^* 与估计的 β之间的差异。

$$\begin{aligned} q_i &= \langle\beta^*, p_i\rangle + \delta_i \\ \implies p_i(q_i-\delta) &= (p_i p_i^T)\beta^* \\ &\text{Summing over all augmentations} \\ \sum_{i\in\mathcal{T}} p_i(q_i-\delta_i) &= \sum_i (p_i p_i^T)\beta^* \\ (m\Sigma)^{-1}\sum_i p_i(q_i-\delta_i) &= \beta^* \\ \implies \beta^*-\beta &= (m\Sigma)^{-1}\sum_i p_i\delta_i \end{aligned}$$

由于 $||x|| = Trace(xx^T)$，因此，

$$\begin{aligned} &E[||\beta^*-\hat{\beta}||_2] = E[\,Trace\left(\left((m\Sigma)^{-1}\sum_i p_i\delta_i\right)\left((m\Sigma)^{-1}\sum_i p_i\delta_i\right)^T\right)] \\ &\text{Using } (AB)^T = B^T A^T \\ &= E\left(Trace\left((m\Sigma)^{-1}\sum_i p_i\delta_i)(\sum_i p_i\delta_i)^T(m\Sigma)^{-T}\right)\right) \\ &\text{By linearity of expectation} \\ &= Trace\left((m\Sigma)^{-1}\left(\sum_{i,j} p_i p_j^T E(\delta_i\delta_j)\right)(m\Sigma)^{-T}\right) \\ &\text{Since} E[\delta_i\delta_j] = 0, \forall i\ne j \\ &= Trace\left((m\Sigma)^{-1}\left(\sum_i p_i p_i^T E(\delta_i^2)\right)(m\Sigma)^{-T}\right) \\ &= Trace\left((m\Sigma)^{-1}\left(\sum_i p_i p_i^T\sigma^2\right)(m\Sigma)^{-T}\right) \end{aligned}$$

数据集	#表	#列	#可连接列	Size
开放数据	69K	29.5M	28.6M	119G
Kaggle	1950	91231	6.7M	18G

表 I：数据集特征

$$= Trace\left((m\Sigma)^{-1}\left(m\Sigma\sigma^2\right)(m\Sigma)^{-T}\right) = \sigma^2 Trace\left((m\Sigma)^{-T}\right)$$
$$\leq \frac{\sigma^2 Trace(\Sigma^{-1})}{m} \leq c\frac{l}{m}$$

对于某个常数 c。 □

该引理表明，只要 $\frac{1}{|T|} = O(1/\log n \equiv$ $o\ 1$) ( )，β就是 β^*的近似值。

VI. 实验评估

本节将回答以下研究问题。

- **RQ1**：METAM 能否识别出可以提升下游任务效用的增强操作？（第 VI-A 节）。
- **RQ2**：METAM 能否随着增强操作数量和数据画像数量的增加而扩展？（第VI-B节）。
- **RQ3**：METAM 对数据画像信息量和参数选择的敏感度如何？（第VI-C节）

数据集。我们从两个真实世界的数据存储库中选取了多种数据集（学校、出租车、杂货、药房、犯罪、房价等）。

- **开放数据**约 6.9 万个数据集，其中包括来自 OpenData PortalWatch 的数据集；该平台收录了 262 个开放数据门户，例如 NYC Open Data、finances.worldbank.org 等。
- **Kaggle** [36] 包含约 1950 个通过爬取不同竞赛识别出的数据集。

设置。 我们使用 Python 实现 METAM，并在一台配备 187GB RAM 的服务器上运行所有实验。连接路径索引使用 Aurum [12]的默认参数设置离线计算得出。除非另有说明，ϵ 设为 0.05， τ 设为已识别簇的数量。我们在第 VI-C 节研究这些参数的影响。我们从包含 100 条记录的随机样本中生成所有数据画像（第 II-C 节）。

A. 现代数据科学工具包

预测性（机器学习）和处方性（因果推断）分析构成了现代数据分析的核心。我们对这两种类型都进行了 METAM 评估。

- **机器学习**：我们考虑：i）两项分类任务；ii）一项旨在确保公平性并维持较高准确率的公平机器学习任务；以及 iii）一项回归任务。

- **因果推断**：我们考虑假设分析和操作分析。假设分析回答假设性推理问题，例如：“如果平均收入增加 20%，会有什么影响？”。操作分析回答诸如“我该如何实现收入超过 20%？”之类的问题。这些任务构成了处方分析的核心 [8], [9], [10], [11]。

首先，我们衡量 METAM 在这些任务上的表现，并将其与其他基线方法进行比较。其次，我们通过实现实体链接、聚类和公平机器学习，展示 METAM 能够泛化到其他任务。我们证明，METAM 无需人工干预即可提升上述所有任务的质量，并优于基线方法。

基线方法。 目前尚无研究探讨目标导向的数据发现，因此我们将先前的技术（见第 III-A 节）改编为基线方法。

- 专家建议预测（MW）：我们使用乘法权重更新算法的随机化版本，该算法根据专家的权重按比例选择专家。

- 基于重叠度排序的方法（重叠度）按照与 D_{in} 的重叠度非递增的顺序查询增强操作，这是一种先前方法中常用的机制 [14]。

- 均匀采样方法（均匀）从候选集中均匀随机地采样查询。

我们将 ARDA [37] 视为分类和回归任务的特定任务基线方法。

1) 专注于最小化解集的大小： 我们配置 METAM，以找到能够提升任务效用的最小增强操作。较小的解集更易于解释，因此非常理想。图 3 显示，METAM在全部四项任务中均优于所有基线方法。它所需的查询数量最低，同时实现了最高的效用。

分类（价格）。 我们预测纽约州、加利福尼亚州和伊利诺伊州不同地区的房价。我们通过在 Redfin 上搜索这些州中随机选取的邮政编码，下载了约 1000 条记录 [38]。该任务训练了一个随机森林分类器，用于预测房价是低还是高。在大约 270 次查询中，METAM 实现的效用从 0.69 提升至 0.84。METAM 识别出的三个最有前景的增强操作包括：i）同一邮政编码区域内是否有沃尔玛；ii）出租车出行次数；以及 iii）犯罪信息（尤其适用于芝加哥地区的预测）。其中一些已识别的增强操作，例如“附近是否有杂货店”，非常引人关注；我们发现许多研究 [5], [39] 已表明，这些因素对于预测房价非常重要。最重要的是，这些结果是在没有人工干预的情况下获得的，除了一开始将 METAM 指向数据集并为其指定任务。

分类（学校）。目标是预测学校在标准化考试中的表现 [37]，数据集约包含 1800 条记录。该任务训练一个随机森林分类器，并计算验证数据集上的 F分数，将其作为效用分数。连接路径索引从数据存储库中识别出超过 9000 个增强操作。我们抽样了 100 个候选增强，发现其中 60% 存在错误，例如使用不正确的键连接数据集。除了先验基线方法外，我们还将一种用于机器学习的先验增强技术 ARDA [37]， 适配到干预设置中（记为 iARDA），在该设置下，增强操作会按照 [37] 返回的排名降序进行查询。

METAM 的表现优于所有基线方法。在 200 次和 700 次查询中，其效用分别达到 0.65 和 0.75。相比之下，MW 和 iARDA 各自都需要超过 1200 次查询。在此数据集上，METAM 识别出增强列与目标属性之间的相关性和互信息是最具信息量的数据画像。在每次迭代中，MW 只考虑单个数据画像来选择查询，而 METAM 则结合所有画像来估计质量分数。因此，METAM 会考虑那些

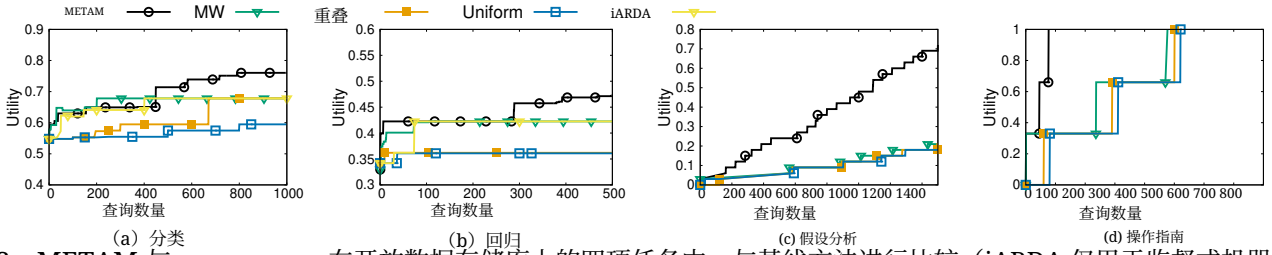


图 3：METAM 与 在开放数据存储库上的四项任务中，与基线方法进行比较（iARDA 仅用于监督式机器学习）

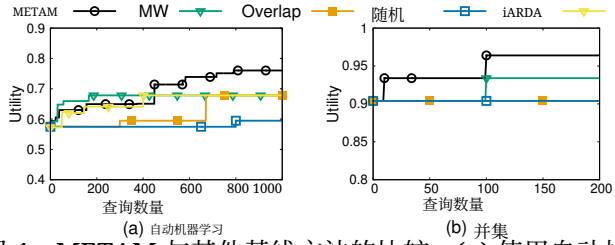


图 4：METAM 与其他基线方法的比较：(a) 使用自动机器学习进行分类；(b) 并集（添加记录）。在发现过程的早期，互信息和相关性值更高。

使用自动机器学习进行分类：我们采用学校表现数据集，并使用三个自动机器学习库（Autosklearn [40], TPOT [41], 和 PyCaret [42]）来实现分类器。图 4（a）展示了 TPOT 库的结果。在少于 1000 次查询的情况下，METAM 将训练所得分类器的效用从 0.57 提升至 0.75，而所有其他基线方法的效用均为 0.70。对于其他自动机器学习任务实现，我们也观察到了类似的结果。

添加额外记录（并集）：先前的实验考虑了添加新属性。在本实验中，我们考虑添加额外记录和属性的情形。在数据发现文献中，添加记录通常称为并集。我们针对纽约市的房屋租金预测任务进行了研究。我们使用 [15] 中的方法，为并集创建额外的增强候选项。图 4 (b) 比较了基础分类器通过添加额外行进行增强后，训练所得分类器的效用。METAM 在大约 100 次查询内将效用从 0.90 提升至 0.96，而所有其他基线方法的效用均达到 0.93。在此基础上，进一步为增强数据集添加新属性后，效用提高了 1%，达到 0.974。

借助专家优化分类器：我们在初始数据集上手动进行特征工程（通过反复试验），历时3小时，尝试优化分类器。我们成功将基础分类器的效用从0.57（通过自动机器学习实现）提升至0.62。然而，在此优化后的分类器上补充外部数据后，其效用进一步提升至超过0.75。这表明，补充外部数据可以引入可能具有很强预测性的新增属性。当初始特征集不具备极强的预测能力时，仅优化特征工程和模型学习可能无法取得类似的提升。

回归。目标是使用每日出租车行程数量等数据，预测纽约市的碰撞次数 [37]，数据集包含350条记录。该任务使用随机森林回归器，并计算平均绝对误差（平均绝对误差），将其作为效用返回 $1 - MAE$。METAM的表现优于所有基线方法。仅使用300次查询，METAM就将平均绝对误差从0。66降至0。55。其他技术需要多三倍的查询次数才能达到类似的平均绝对误差。值得注意的是，这些任务所实现的效用与先前使用这些数据集的研究所报告的数值相当 [37]，这证明了METAM能够在广泛的任务中实现具有竞争力的质量并保持良好的泛化能力。我们进一步评估了具有信息丰富的领域特定数据画像（例如特征重要性） [37] 以及无信息的通用数据画像的预测分析（分类和回归）任务，以了解METAM在第VI-C节中的灵活性。

假设分析。该任务接收一个输入数据集及一项假设更新，并输出更新查询对其他属性的因果影响。我们考虑一张包含450名学生SAT成绩的初始表格 [43]，并询问如果更新学生的“批判性阅读成绩”，哪些属性会受到因果影响。理解这些属性有助于揭示哪些因素会影响学生的阅读成绩，为实施干预措施铺平道路。该任务实现了一种因果发现算法 [44]，其质量以正确识别属性的比例（p值 ≤ 0.05）进行衡量。

图 3(c) 显示，METAM 在少于 1500 次查询内便达到 0.70 的效用，而所有其他基线方法的效用分数均低于 0.25。METAM 在约 2000 次查询时达到 1 的效用分数，而 MW 则需要超过 8000 次查询。METAM 性能提升的主要原因之一是聚类过程，该过程能够筛除预计具有与先前查询的增强操作相似效用的查询（性质 P2）。

操作分析。我们希望确定需要更新哪些属性，以最大化 450 名学生的 SAT 成绩。该任务计算因果依赖关系，并返回正确识别属性的比例。候选集包含 240 个增强操作。最优解包含三个对阅读成绩具有强因果影响的属性，METAM 在少于 100 次查询内便能识别出这一属性集合，而所有其他基线方法则需要超过 500 次查询。

2）允许解集的大小任意：我们放宽了必须找到最小增强集的要求。算法可以选择第一个能提升任务效用的增强操作，而不必选择最佳增广。无论停止准则如何，METAM 的表现始终优于基线方法。与优化解集相比，MW 所需的查询更少（ $\tau = 1$ ），但 METAM 的表现仍优于这一基线方法。在这种放宽的设置下，解集包含

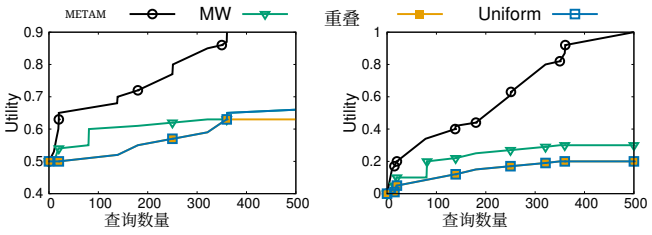


图 5：半合成评估的平均结果

数据集	METAM	MW	Overlap	Uniform
学校（C）	**0.80**	0.20	0.0	0.20
出租车（C）	**1**	0.5	0.5	0.5
犯罪（C）	**0.90**	0.20	0.1	0.1
房价（C）	**0.75**	0.25	0.0	0.25
药房	**0.95**	0.43	0.43	0.25
杂货店	**0.92**	0.37	0.10	0.17

表 II：在查询次数少于 1000 次时，METAM 与其他基线方法的效用。标记为 (C) 的数据集执行因果分析任务，其他数据集执行数据分析。

$\approx$ 9 增强操作与实验中使解集大小最小的 2 个增强操作的比较。

3）结果具有显著性：在本实验中，我们从存储库中随机选择一个数据集（例如 D），随机抽取五种不同的增强操作，并使用这些增强操作在 D 中合成一个新列，将其分别作为 i）分类任务的预测属性，以及 ii）how-to 分析任务的结果变量。我们将 D 的这一修改版本作为输入表，以测试所提出的技术。我们考虑了 100 种不同的实例化方式，并报告平均统计数据。图 5 表明，METAM 始终优于其他技术，并且能够高效地快速发现有用的增强操作。MW 优于其他基线方法，因为只要某种增强操作在任一数据画像中排名靠前，它就能够识别出该增强操作具有实用价值。

表 II 展示了 METAM 和其他基线方法在一组多样化数据集上取得的效用。这些数据集是通过筛选各数据源中访问量最高的数据集选出的。例如，犯罪数据是芝加哥开放数据中访问量最高的数据，药房信息则是宾夕法尼亚州开放数据门户中访问量最高的数据之一。在所有数据集上，METAM 都能够识别出实现最高效用分数的增强操作。在基线方法中，MW 即使使用了多得多的查询，其表现仍远逊于 METAM。

4) 泛化到其他任务：**实体链接**旨在将输入数据集 [45]中的实体链接到知识图谱（如维基数据或DBPe- dia）中的相应实体 [46]。我们考虑了一个美国疾病控制与预防中心数据集 [36]，其中包含美国各地不同城市的统计数据。每条记录都包含城市名称、州缩写，例如 阿拉斯加州的 AL。该任务是将所有城市名称链接到相应实体，并将准确率作为效用指标进行评估。该任务会在维基数据上搜索每个词元，并在存在唯一匹配项时选择相应实体。然而，像伯明翰这样的城市名称对应多个维基数据实体，因为英国和阿拉巴马州都有一座名为伯明翰的城市，而且阿拉巴马州也有一座。向该数据集中添加州名称可以为实体链接方法提供上下文，从而实现更好的链接。我们考虑了 Kaggle 数据存储库，并确定了约 185 种增强操作。METAM 在 4 个查询中识别出有用的增强操作，而 MW 需要 10 个查询，所有其他基线方法则需要 40 多个查询。

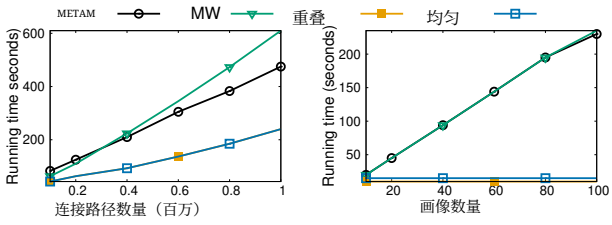
公平分类。我们考虑使用信贷数据集 [47]，根据个人的人口统计属性预测其收入，其中年龄被视为敏感属性。除初始数据集外，我们还按照先前工作 [48], [49]所采用的机制，通过转换部分特征生成新的数据集。转换后的数据集被添加到 Kaggle 存储库中，以供发现。任务实现 [49]在内部执行公平性感知特征选择，以舍弃存在偏差的属性，并使用剩余子集训练分类器。它返回训练后分类器的 F 分数作为效用分数。我们观察到，与目标高度相关的属性通常具有很低的公平性，而高度公平的属性（与敏感属性的相关性较低）并不能提高分类准确率。因此，所有基于单一概况对属性进行排序的基线方法都无法在 50 次查询内识别出有用的数据集。然而，METAM 会计算概况分数的加权平均值，从而有效地对增强操作进行排序，并因此在少于 10 次查询内识别出提升效用所需的最小集合。

聚类。本评估考察了从健康博客网站收集的不同原材料及其类别（蔬菜、水果、香料等）列表。下游任务根据产品的饱腹感评分对产品进行聚类，并将最大聚类半径的加法逆元作为效用。该数据集在存储库中识别出 8 种增强操作，其中一种会增强每种配料的最佳营养摄入（ONI）评分。该评分与真实聚类高度相关，因此有助于提升聚类质量。鉴于候选集较小，所有技术在该数据集上的表现相近，需要 ≈ 4 次查询。

B. 效率与可扩展性

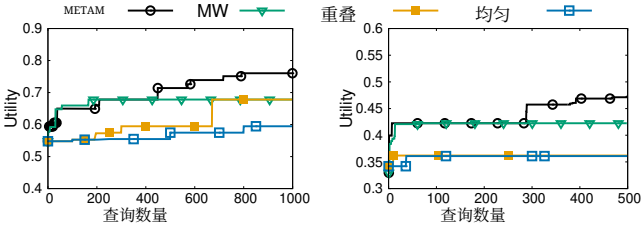
上述实验所考察的查询数量是评估运行流水线所需时间的替代指标。METAM 具有效率，能够在所有数据集上用时不到 10 分钟识别出增强操作集。大部分时间都花在识别候选增强及其数据画像上。需要注意的是，这些画像的复杂度各不相同，例如，相关性和互信息比数据集重叠更复杂。在 10 分钟中，大约有 4 分钟用于为候选增强生成数据画像。对于固定数量的查询，MW 与 METAM 所需的时间相同。其他基线方法运行速度略快，用时 7 分钟。

为评估可扩展性，我们改变候选增强操作和数据画像的数量，并测量运行 1000 个查询所需的时间。图 6 (a) 显示，所有技术的运行时间都随存储库中可连接数据集数量的增加而线性增长。由于对增强操作进行排序需要 $O(n\log\ n)$ 的时间，MW 所需的时间增长速度快于 METAM。



(a) 变化的连接路径数量　(b) 变化的画像数量

图 6：METAM 与其他基线方法的执行情况：(a) 改变可连接数据集集合；(b) 改变数据画像。



（a）分类　(b) Reg回归

图 7：使用来自 [37] 的任务特定数据画像，对比 METAM 与其他技术。

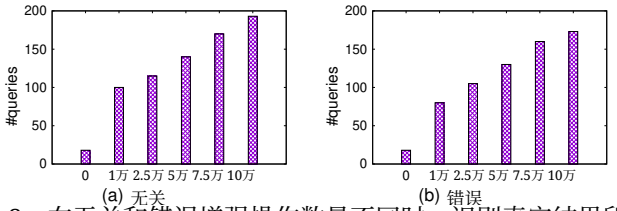
相比之下，METAM 在 $O(n)$ 的时间内对增强操作进行聚类，并利用识别出的簇进行高效计算。尽管 Overlap 和 Uniform 的速度更快，但如前所示，它们的性能远低于其他基线方法。图 6 (b) 显示，METAM 和 MW 的运行时间随画像数量线性增长，而 Overlap 和 Uniform 所需的时间没有变化，因为它们不使用任何输入画像。总之，METAM 可在 10 分钟内处理 $1M$ 增强操作，并且其运行时间随增强操作数量线性增长。

C. 数据画像信息量和参数的影响

在本节中，我们首先评估错误的候选增强（假阳性）对 METAM 整体性能的影响。随后，我们评估 (a) 添加信息丰富的画像、(b) 添加信息不足的画像，以及 (c) 移除信息丰富的画像，对 METAM 提升下游效用和整体效率能力的影响。接着，我们评估所有画像都信息不足的极端情形。最后，我们进行消融分析，以了解 METAM 各组成部分及其参数的影响。

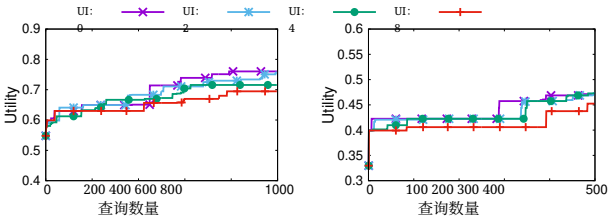
错误的增强会影响 METAM 的性能吗? 在本次实验中，我们针对分类和回归数据集各考虑了一个真实增强项。图 8 (a) 将错误增强项（候选生成算法识别出的错误连接路径）固定为 100，并改变正确但无关的增强项（无法提升任务效用）的数量；图 8 (b) 则固定无关增强项的数量，同时改变错误增强项的数量。我们观察到，METAM 在 200 次查询内识别出真实增强项，并舍弃所有错误和无关增强项。随着无关或错误增强项数量增加，所需查询次数也随之增加。

添加信息丰富的配置文件。 在这项实验中，我们使用 Arda [37] 添加了任务专用的数据画像。图 7 显示，METAM 所需的查询次数比 MW 和其他基线方法少。事实上，METAM 在不到



(a) 无关　(b) 错误

图 8：在无关和错误增强操作数量不同时，识别真实结果所需的查询数量。



(a) 分类　(b) 回归

图 9：METAM 在无信息量逐渐增加时的性能

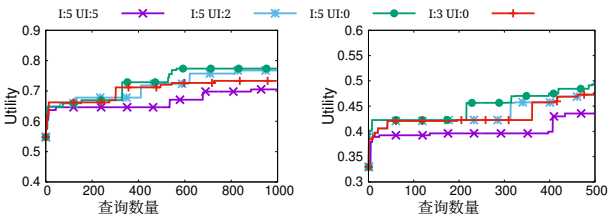
信息丰富的配置文件。信息丰富的配置文件数量固定为 5，无信息配置文件用 UI 表示。

200 次查询内便达到了 0.68 的效用，而未使用这些专用数据画像时则需要 400 多次查询。信息丰富的配置文件提升了 METAM 的性能。

添加无信息配置文件。 本实验采用初始画像集合（如第 IV 节所述），并生成额外的无信息配置文件。图 9 显示，在所有设置下，METAM 都能实现相近的效用增益；添加噪声配置文件不会影响解决方案的质量，因为 METAM 学会了忽略它们，只需付出运行少量额外查询的代价。

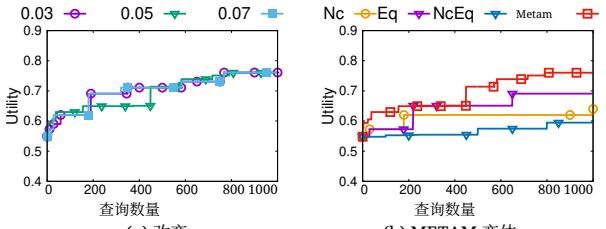
移除画像。 我们考虑包含 10 个数据画像的分类和回归任务，其中 5 个为无信息配置文件。首先移除无信息配置文件，随后移除其他信息丰富的配置文件。图 10 表明，移除无关特征（直到移除的特征数量少于 5 个）会提升任务效用随查询数量增加而逐步增长的趋势。进一步移除任何相关配置文件后，达到类似效用所需的查询数量会增加。

如果所有画像都不包含有效信息呢? 在本实验中，我们针对分类任务使用了学校数据集，其中所有画像均为随机生成。在这种情况下，METAM 估计的质量分数与随机排序相似。不过，METAM 能识别出最优的增强操作集合，但查询数量与均匀基线相近。



（a）分类　（b）回归

图 10：移除画像对 METAM 质量与查询权衡的影响。I 表示信息丰富的配置文件数量。



(a) 改变 ε (b) METAM 变体

图 11：消融分析 ϵ及 METAM 的变体。

METAM 及其变体。图 11（b）将 METAM 与三个变体进行了比较：（i）Eq：该变体相当于忽略汤普森采样过程，并以同等重要性对每个簇进行排序。（ii）Nc：将每种增强操作视为一个不同的簇，相当于忽略性质 P2。（iii）NcEq：将每种增强操作视为同等重要，并忽略聚类。METAM 的表现优于所有变体，因为 Eq 和 NcEq 会以随机顺序处理增强操作的候选集，而 Nc 则将查询浪费在冗余的增强操作上；这些操作预计无助于下游效用（METAM 会将它们聚为一类）。因此，将聚类与排序机制结合起来，能够使 METAM 优先处理重要的增强操作，并高效地忽略冗余操作。

影响 ϵ。图 11 (a) 测量了改变聚类的 ϵ 时 METAM 的质量。增大 ϵ 的值会减少数据集中的簇数量，并增大簇内增强之间的距离。因此，在这种情况下，性质 P2 可能被违反，METAM 也可能在质量分数估计时忽略簇结构。相反，选择较小的 ϵ 值意味着同一簇中的所有增强操作在所考虑的增强操作方面都非常相似，因此，簇内所有增强操作对下游任务效用的影响可能相似。图 11 显示，改变 ϵ 时，METAM 的查询数量不会发生显著变化。

VII. 相关工作

数据发现方法。数据发现系统和库 [22], [50], [12], [51], [17], [18], [16], [52], [53],[54] 会自动从数据存储库中计算连接路径。因此，给定一个查询数据集，这些系统会告知用户还有哪些数据集可以与该查询数据集连接。许多现代技术允许用户通过指定数据画像、摘要和解释等形式的要求，交互式地搜索有用的可连接数据集 [55] [56]。然而，若不了解用户的目标，就难以确定哪条连接路径对下游任务有用，正如我们在评估部分所展示的那样。

人在回路中的数据集成。人在回路中的方法 [57], [58], [59] 利用用户反馈来引导搜索过程。这些技术已应用于实体解析 [60], [61], [62], [63], 连接选择和实体对齐等问题 [64]。这些方法旨在减少寻找解决方案所需的工作量。相比之下，METAM 专注于自动化数据发现和增强，无需用户干预。

数据与机器学习（ML）。有一些为机器学习任务 [37], [65], [66], [67], [68], [69] 设计的数据增强技术，它们要么选择对预测有用的属性和数据点，要么学习关系嵌入，将数据嵌入向量空间。这些技术不是干预式的，并且会根据可用数据集的重要性对其进行排序，这些排序结果可作为 METAM 的数据画像。因此，本工作的贡献与先前根据特征或增强数据集的重要性对其进行排序的技术正交。此外，METAM 不受机器学习任务的限制。相反，它实现了目标导向的数据发现，因此能够针对任何任务引导数据发现，只要开发人员能够提供一个计算效用分数的函数即可。此外，METAM 会动态调整不同画像的权重，使其能够更稳健地应对连接路径和画像中的数据质量问题。

查询驱动任务。查询驱动方法 [70], [71]着重根据具体查询的需求，减少实体解析等任务的计算量。这些方法解决了在线环境中的效率问题，并且需要预先定义明确的查询。然而，在数据发现问题中，不可能拥有明确的查询；也就是说，分析人员通常并不清楚数据集合中的哪一部分对其任务有用。

VIII. 结论

我们提出了一种用于执行面向目标的数据发现的新方法，该方法运行干预式查询，以高效识别有用的增强操作。我们的方法利用数据属性、效用函数和解集来优化查询策略。我们分析了该方法的最优性和查询复杂度。在来自不同应用场景的多样化数据集上进行的实证分析，以及与基线方法的比较，证明了该方法在高效选择有用增强操作方面的通用性和有效性。

参考文献

[1] “Data.gov https://www.data.gov。”[在线]。可访问：https://www.data. gov [2] A. Y. Halevy、F. Korn、N. F. Noy、C. Olston、N. Polyzotis、S. Roy 和 S. E. Whang，《管理谷歌的数据湖：Goods系统概述》。*IEEE*数据工程公报，第39卷，第3期，第5–14页，2016年。[3] “Dawex, https://www.dawex.com/en/。”[4] R. C. Fernandez、P. Subramaniam 和 M. J. Franklin，《数据市场平台：交易数据资产以解决数据问题》，*PVLDB*，第13卷，第12期，第1933–1947页，2020年。[5] D. G. Pope 和 J. C. Pope，《沃尔玛进驻小镇时：房价总是低廉吗？真的总是吗？》《城市经济学杂志》，第87卷，第1–13页，2015年。[6]“经济学家发现沃尔玛门店会推高附近房价 https://abcnews.go.com/Business/economists-find-walmart-stores-increase-nearby-home-prices/story?id16560451 。”[7] “杂货店会影响房屋价值 https://www.homes.com/blog/2020/08/how-do-grocery-stores-affect-home-values/。”[8] A. Meliou 和 D. Suciu，《Tiresias：面向操作指南查询的数据库预言机》，载于*2012*年*ACM SIGMOD*数据管理国际会议论文集，2012年，第337–348页。[9] T. Software, “Tableau https://www.tableau.com/。”[在线]。可访问：https://www.tableau.com/ [10] “Microsoft Power BI https://powerbi.microsoft.com/en-us/。”[在线]。可访问：https://powerbi.microsoft.com/en-us/

[11]“Domo https://www.domo.com/。” [在线]。可访问：https://www.domo.com/[12] R. C. Fernandez、Z. Abedjan、F. Koko、G. Yuan、S. Madden 和 M. Stonebraker，《Aurum：一个数据发现系统》，载于第34届IEEE国际数据工程会议（ICDE 2018），法国巴黎，2018年4月16日至19日。IEEE计算机学会，2018年，第1001–1012 [ ] 10 1109 2018 00094 页。在线。[13]可访问：https://doi.org/ . /ICDE. . E. Zhu、D. Deng、F. Nargesian 和 R. J. Miller，《Josie：用于在数据湖中查找可连接表的重叠集合相似性搜索》，载于2019年国际数据管理会议论文集，丛书：SIGMOD ’19。美国纽约州纽约市：计算机协会，2019年，第847–864页。[14] F. Psallidas、B. Ding、K. Chakrabarti 和 S. Chaudhuri，《S4：用于查询发现的前k个电子表格式搜索》，载于2015年ACM SIGMOD数据管理国际会议论文集，澳大利亚维多利亚州墨尔本，2015年5月31日至6月4日，T. K. Sellis、S. B. Davidson 和 Z. G. Ives 编。ACM，2015年，第2001–2016页。 [ ] 10 1145 2723372 2749452 在线。可访问：[15]https://doi.org/ . / . F. Nargesian、E. Zhu、K. Q. Pu 和 R. J. Miller，《开放数据上的表联合搜索》，《VLDB基金会论文集》，第11卷，第7期，第813–825页，2018年。[16] A. Santos、A. Bessa、F. Chirigati、C. Musco 和 J. Freire，《用于近似连接相关性查询的相关性草图》，2019年国际数据管理会议论文集，2021年6月。 [ ] 10 1145 3448016 3458456 在线。可访问：[17]http://dx.doi.org/ . / . R. C. Fernandez、J. Min、D. Nava 和 S. Madden，《Lazo：一种基于基数的方法，用于联合估计Jaccard相似度和包含度》，载于第35届IEEE国际数据工程会议（ICDE） *2019,* *8 11,* 澳门，中国，*2019* 2019, 1190 1201 [ ] 4月–。IEEE，第–页。在线。可访问：https: //doi.org/ 10 1109 2019 00109 . /ICDE. . E. Zhu、F. Nargesian、K. Q. Pu 和 R. J. Miller，《LSH集合：互联网规模域搜索》，《PVLDB》，第9卷，第12期，第1185–1196页，2016年。 [在线]。可访问：http: //www.vldb.org/pvldb/vol9/p1185-zhu.pdf[19] I. Trummer，《深度神经网络能否根据列名预测数据相关性？》，CoRR，第abs/2107.04553卷，2021年。 [在线]。可访问：https://arxiv.org/abs/ 2107 04553 . K. Hlava˘ckova-Schindler、M. Palu˘s、M. Vejmelka 和 J. Bhattacharya，《基于信息论方法的时间序列分析中的因果关系检测》，《物理学报告》，第441卷，第1期，第1–46页，2007年。[21] J. Devlin、M.-W. Chang、K. Lee 和 K. Toutanova，《BERT：用于语言理解的深度双向变换器预训练》，arXiv预印本arXiv:1810.04805，2018年。[22] Y. Gong、Z. Zhu、S. Galhotra 和 R. C. Fernandez，《Ver：真实环境中的视图发现》，CoRR，第abs/2106.01543卷，2021年。 [在线]。可访问：https://arxiv.org/abs/ 2106 01543 . S. Brin、R. Motwani 和 C. Silverstein，《超越市场篮子：将关联规则推广到相关性》，载于1997年ACM SIGMOD数据管理国际会议论文集，1997年，第265–276页。[24] B. Salimi、L. Rodriguez、B. Howe 和 D. Suciu，《干预公平性：面向算法公平性的因果数据库修复》，载于2019年国际数据管理会议论文集，2019年，第793–810页。[25] C. Chen、B. Golshan、A. Y. Halevy、W.-C. Tan 和 A. Doan，《Biggorilla：用于数据准备和集成的开源生态系统》。IEEE数据工程公报，第41卷，第2期，第10–22页，2018年。[26] J. d. J. Flores Herrera、S. Nadal Francesch 和 ´O. Romero Moral，《迈向可扩展的数据发现》，载于《数据库技术进展：EDBT 2021》，第24届数据库技术扩展国际会议：塞浦路斯尼科西亚，2021年3月23日至26日：会议录。Open- Proceedings，2021年，第433–438页。[27] Z. Abedjan、L. Golab 和 F. Naumann，《关系数据剖析：综述》，《VLDB期刊》，第24卷，第4期，第557–581页，2015年。[28] S. Arora、E. Hazan 和 S. Kale，《乘法权重更新方法：一种元算法及其应用》，《计算理论》，第8卷，第1期，第121–164页，2012年。[29] R. O. Duda、P. E. Hart 和 D. G. Stork，《模式分类与场景分析》。Wiley，纽约，1973年，第3卷。[30] W. R. Thompson，《根据两组样本的证据，一个未知概率大于另一个的似然》，《Biometrika》，第25卷，第3–4期，第285–294页，1933年。[31] D. J. Russo、B. Van Roy、A. Kazerouni、I. Osband、Z. Wen 等，《汤普森采样教程》，Foundations and Trends® in Machine Learning，第 11 卷第 1 期，1–96 页，2018 年。[32] T. F. Gonzalez，《通过聚类最小化最大簇间距离》，Theoretical Computer Science，第 38 卷，293–306 页，1985年。[33] G. L. Nemhauser、L. A. Wolsey 和 M. L. Fisher，《最大化次模集合函数的近似分析——I》，Mathematical Programming，第 14 卷第 1 期，265–294 页，1978年。[34] A. A. Bian、J. M. Buhmann、A. Krause 和 S. Tschiatschek，《非次模函数贪心最大化的保证及其应用》，载于《国际机器学习会议》，PMLR，2017年，第 498–507 页。[35] A. Das 和 D. Kempe，《近似次模性及其应用：子集选择、稀疏近似与字典选择》，The Journal of Machine Learning Research，第 19 卷第 1 期，74–107 页，2018年。[36] “Kaggle，https://www.kaggle.com/。”[37] N. Chepurko、R. Marcus、E. Zgraggen、R. C. Fernandez、T. Kraska 和 D. Karger，《ARDA：面向机器学习的自动关系数据增强》，Proc. VLDB Endow.，第 13 卷第 9 期，1373–1387 页，2020年。[在线]。网址：http://www.vldb.org/pvldb/vol13/p1373-chepurko.pdf[38] “Redfin https://www.redfin.com/。”[39] P. Caporaso，《出租车活动作为纽约市住宅租金的预测指标》，博士论文，麻省理工学院，2019年。[40] “Auto-sklearn 自动机器学习 https://automl.github.io/auto-sklearn/master/。”[41] “TPOT 自动机器学习 http://epistasislab.github.io/tpot/using/。”[42] “PyCaret https://pycaret.org/。”[43] “纽约市开放数据，https://opendata.cityofnewyork.us/。”[44] “causal-learn https://causal-learn.readthedocs.io/en/latest/。”[45] S. Galhotra 和 U. Khurana，《结构化数据上的语义搜索》，载于《第 29 届 ACM 信息与知识管理国际会议论文集》，2020年，第 3381–3384 页。[46] “DBpedia，https://www.dbpedia.org/。”[47] “德国信用风险 https://www.kaggle.com/uciml/german-credit。”[48] R. S. Diaz、F. Neutatz 和 Z. Abedjan，《面向算法公平性的自动化特征工程》，PVLDB，2021年。[49] S. Galhotra、K. Shanmugam、P. Sattigeri 和 K. R. Varshney，《公平数据集成》，2020年。[50] M. Yakout、K. Ganjam、K. Chakrabarti 和 S. Chaudhuri，《InfoGather：通过与网页表格进行整体匹配实现实体增强和属性发现》，SIGMOD ’12 系列，2012年，第 97–108 页。[51] S. Castelo、R. Rampin、A. S. R. Santos、A. Bessa、F. Chirigati 和 J. Freire，《Auctus：用于数据发现与增强的数据集搜索引擎》，Proc. VLDB Endow.，第 14 卷第 12 期，2791–2794 页，2021年。 [在线]。网址：http://www.vldb.org/pvldb/vol14/p2791-castelo.pdf[52] M. Esmailoghli、J. Quian´e-Ruiz 和 Z. Abedjan，《COCOA：感知相关系数的数据增强》，载于《第 24 届数据库技术扩展国际会议论文集（EDBT 2021）》，塞浦路斯尼科西亚，2021年3月23日至26日，Y. Velegrakis、D. Zeinalipour-Yazti、P. K. Chrysanthis 和 F. Guerra 编。OpenProceedings.org，2021年，第 331–336 页。 [在线]。网址：https://doi.org/ 10 5441 002 2021 30 。//edbt. . F. Nargesian、K. Q. Pu、B. G. Bashardoost、E. Zhu 和 R. J. Miller，《数据湖组织》，arXiv 预印本 arXiv:1812.07024，2018年。[54] F. Nargesian、K. Q. Pu、E. Zhu、B. Ghadiri Bashardoost 和 R. J. Miller，《为便于导航而组织数据湖》，载于《2020年 ACM SIGMOD 数据管理国际会议论文集》，2020年，第 1939–1950 页。[55] Y. Zhang 和 Z. G. Ives，《在数据湖中查找相关表以支持交互式数据科学》，载于《2020年 ACM SIGMOD 数据管理国际会议论文集》，2020年，第 1951–1966 页。[56] A. Santos、S. Castelo、C. Felix、J. P. Ono、B. Yu、S. R. Hong、C. T. Silva、E. Bertini 和 J. Freire，《Visus：用于自动机器学习模型构建与整理的交互式系统》，载于《人在回路中的数据分析研讨会论文集》，2019年，第 1–7 页。[57] G. Li，《人在回路中的数据集成》，Proc. VLDB Endow.，第 10 卷第 12 期，2006–2017 页，2017年。 [在线]。网址：http://www.vldb.org/pvldb/vol10/p2006-li.pdf[58] Y. Zhuang、G. Li、Z. Zhong 和 J. Feng，《Hike：一种用于大规模知识库实体对齐的人机混合方法》，载于《2017年 ACM 信息与知识管理会议论文集（CIKM 2017）》，新加坡，2017年11月6日至10日。ACM，2017年，第 1917–1926 页。 [在线]。网址：https://doi.org/10.1145/3132847.3132912

[59] A. Bonifati、R. Ciucanu 和 S. Staworko，《从用户示例中学习连接查询》，*ACM*《数据库系统汇刊》，第 40 卷第 4 期，第 24 1 24 38, 2016 [ ] 10 1145 2818637 : – : . 在线 . 网址：https://doi.org/ . / [60] N. Vesdapunt、K. Bellare 和 N. Dalvi，《用于实体解析的众包算法》，*PVLDB*，第 7 卷第 12 期，1071–1082 页，2014年。[61] D. Firmani、B. Saha 和 D. Srivastava，《使用预言机进行在线实体解析》，*PVLDB*，第 9 卷第 5 期，384–395 页，2016年。[62] J. Wang、G. Li、T. Kraska、M. J. Franklin 和 J. Feng，《利用传递关系进行众包连接》，载于《*2013*年 *ACM SIGMOD* 数据管理国际会议论文集》，2013年，第 229–240 页。[63] C. Gokhale、S. Das、A. Doan、J. F. Naughton、N. Rampalli、J. Shavlik 和 X. Zhu，《Corleone：用于实体匹配的免干预众包》，载于《*2014*年 *ACM SIGMOD* 数据管理国际会议论文集》，2014年，第 601–612 页。[64] G. Li，《人在回路中的数据集成》，*PVLDB*，第 10 卷第 12 期，2006–2017 页，2017年。[65] Y. Li、X. Wang、Z. Miao 和 W. Tan，《面向机器学习驱动的数据准备与集成的数据增强》，*Proc. VLDB Endow.*，第 14 卷第 12 期，3182–3185 页，2021年。 [在线]。网址：http://www.vldb.org/pvldb/vol14/p3182-li.pdf [66] A. Kumar、J. Naughton、J. M. Patel 和 X. Zhu，《该不该连接？在特征选择之前对连接进行再三思考”，载于*2016*年国际数据管理会议论文集，2016年，第19–34页。[67] J. Liu、F. Zhu、C. Chai、Y. Luo 和 N. Tang，“深度学习的自动数据获取”，*VLDB*基金会论文集，第14卷，第12期，第2739–2742页，2021年。 [68] C. Chai、J. Liu、N. Tang、G. Li 和 Y. Luo，“面向模型充能的真实环境选择性数据获取”，*PVLDB*，第15卷，1478，2022年 7，1466 no. pp. – 。[69] Z. Zhao 和 R. Castro Fernandez，“Leva：通过关系嵌入数据增强提升机器学习性能”，载于*2022*年国际数据管理会议论文集，2022年，第1504–1517页。[70] H. Altwaijry、D. V. Kalashnikov 和 S. Mehrotra，“查询驱动的实体解析方法”，*VLDB*基金会论文集，第6卷，第14期，第1846–1857页，2013年。 [在线]。可访问：http://www.vldb.org/pvldb/vol6/p1846-altwaijry.pdf [71] Y. Altowim、D. V. Kalashnikov 和 S. Mehrotra，“关系实体解析的渐进式方法”，*VLDB*基金会论文集，第7卷，第11期，第999–1010页，2014年。 [在线]。可访问：http: // www.vldb.org/pvldb/vol7/p999-altowim.pdf