

面向成本收益协同优化的 高质量数据准备方法

崔文萱¹, 丁小欧¹, 程海峰¹, 钱泽凯¹, 李勇², 王宏志¹, 王晨³

1. 哈尔滨工业大学计算学部, 黑龙江 哈尔滨 150001;

2. 哈尔滨工业大学学校办公室, 黑龙江 哈尔滨 150001;

3. 清华大学大数据系统软件国家工程实验室, 北京 100084

摘要

针对实际机器学习 (machine learning, ML) 模型训练场景, 数据准备环节资源利用率低、成本不可控及预算约束下优化不足等问题, 构建了一种面向成本收益协同优化的数据准备整体框架 CBAPrep (cost-benefit-aware data preparation)。采用数学规划方法, 将数据预处理、清洗、转换等操作统一纳入成本建模与决策框架, 并设计了面向数据管道的成本收益分析规划方法。实验结果表明, 在有限预算条件下, CBAPrep 可实现更优的成本效益权衡, 提升数据准备过程的整体性能与性价比。研究表明, 将成本收益分析和优化的思想纳入数据准备优化策略, 有助于实现资源的高效配置, 并为复杂数据处理场景中的成本控制与性能优化提供参考。

关键词

数据治理; 数据准备; 资源调度; 成本收益

中图分类号: TP311.13

文献标志码: A

doi: 10.11959/j.issn.2096-0271.BDR26172

A cost-benefit collaborative optimization method for high-quality data preparation

Cui Wenxuan¹, Ding Xiaou¹, Cheng Haifeng¹, Qian Zekai¹, Li Yong², Wang Hongzhi¹, Wang Chen³

1. Faculty of Computing, Harbin Institute of Technology, Harbin 150001, China

2. General Office of Harbin Institute of Technology, Harbin 150001, China

3. National Engineering Research Center for Big Data Software, Tsinghua University, Beijing 100084, China

Abstract

In practical machine learning (ML) model training scenarios, data preparation is still constrained by low resource utilization efficiency, uncontrollable costs, and insufficient optimization under budget constraints. To address these problems, a holistic data preparation framework named CBAPrep (cost-benefit-aware data preparation) was proposed, oriented toward cost-benefit collaborative optimization. Mathematical programming was adopted to incorporate data preprocessing, cleaning, transformation, and other operations into a unified cost modeling and decision-making framework, and a cost-benefit analysis and planning method for the overall pipeline was designed. Experimental results showed that CBAPrep achieved a better trade-off between cost and performance under limited budget conditions, thereby improving the overall performance and cost-effectiveness of the data preparation process. The results indicate that incorporating cost-benefit analysis and optimization into data preparation strategies facilitates efficient resource

allocation and provides useful support for cost control and performance optimization in complex data processing scenarios.

Key words
data governance, data preparation, resource scheduling, cost-benefit

0 引言

数据治理是保障数据质量与安全、促进数据高效流通与价值释放的基础性工作^[1-2]。随着数据要素流通和数据社会化进程的不断深化，数据治理已逐渐从组织内部的数据质量管理扩展到跨组织、跨领域的数据流通安全治理与价值释放全过程^[3]。在此背景下，数据准备在机器学习(machine learning, ML)等数据驱动应用中的重要性愈发凸显，已成为保障训练数据质量和下游模型性能的重要环节。

真实业务数据普遍存在格式不统一、字段缺失、异常值和噪声等多种数据质量缺陷，因此需要针对各类问题展开数据清洗与准备工作。随着ML技术的加速落地应用，自动化数据准备工具逐渐替代部分人工操作，大幅提升了数据处理效率。然而，在实际模型训练场景下，数据准备仍需耗费较高成本：一方面，许多环节仍依赖人工处理，带来显著的时间与人力成本^[4]；另一方面，不同自动化算法在处理效果、资源消耗和适用场景上存在差异，算法的选择、组合与决策过程本身也会产生额外的开销。因此，如何有效控制数据准备成本仍是实际ML应用场景中不容忽视的问题。

数据准备涵盖基于规则、统计、机器学习及深度学习等多类实现方法，不同方法在处理效果与适用场景上存在差异，因此需要对其进行合理选型与编排。现有研究已提出多种自动化管道优化框架：

Learn2Clean^[5]采用强化学习方法，对标准化、特征选择及多类数据清洗操作进行序列决策；HAIPipe^[6]结合人工经验与人工智能(artificial intelligence, AI)实现协同编排；DiffPrep^[7]将离散的管道选择转化为可微梯度优化问题；CtxPipe^[8]进一步利用上下文信息提升管道决策质量。除管道级优化方法外，数据清洗领域也提出了BoostClean^[9]、HoloClean^[10]、Uniclean^[11]和ActiveClean^[12]等自动化方法，分别从错误修复、概率推理、算法集成和主动清洗等角度提升数据质量。上述两类方法的对比见表1。

在追求下游模型性能的同时，还需要考虑数据准备过程的成本消耗。在低预算场景下，不能盲目选择效果最优的数据准备管道，还应综合考量各类方法的性价比。现有管道优化方法虽然通过策略设计能够提升数据准备效果或效率，但对成本收益关系的系统分析仍较不足。少数方法虽然兼顾成本，但其重点也多集中于降低管道搜索和构建阶段的算法开销，而非针对清洗算法自身的执行成本及资源投入进行优化。现有研究主要存在以下两方面局限。

(1) 聚焦单一任务，缺乏投入成本与模型收益之间的整体性关联分析。例如，BoostClean、HoloClean主要关注特定清洗算法的效果提升，对成本的考虑多局限于算法自身的运行开销，较少分析该算法对完整数据准备流程及最终下游模型性能的影响。因此，有必要对各数据准备环节的资源投入与最终收益进行统一评估，为具体算法的选择和组合提供依据。

表1 数据准备优化策略对比

数据准备优化策略		模型性能考量	成本考量	优化依据	优化目标	功能
数据管道组合优化	Learn2Clean	★★★★	★	数据质量指标	性能提升	序列决策
	HAIPipe	★★★★	★★	人类与AI协作	性能提升	管道编排
	Diffprep	★★★★	★	可微梯度优化	性能提升	流程优化
	Ctxpipe	★★★★	★	上下文感知	性能提升	流程优化
数据清洗方法	BoostClean	★★	★★	错误检测	算法鲁棒	错误修复
	HoloClean	★★	★	概率推理	数据质量	统计清洗
	Uniclean	★★★★	★★	算法集成	性能提升	按需清洗
	ActiveClean	★★	★★	主动学习	性能提升	主动清洗

(2) 缺乏成本收益视角下的数据准备方法调配和资源分配策略研究。目前，VChecker^[13]、CrowdGame^[14]已对成本收益问题进行了探索；CBAClean 综合考虑清洗效果与执行成本，推荐了具有成本效益的数据清洗方案^[15]。然而，这些方法主要面向特定任务或单一清洗方案，尚未对完整数据准备管道的算子组合、执行程度和资源分配开展系统化建模。现有管道研究通常更关注理想条件下的性能提升，忽视了实际预算约束下的成本与收益平衡问题。

针对上述局限，本文将成本收益优化引入 ML 的数据准备过程，在保证训练数据质量和下游模型性能的同时，对数据准备成本进行统一管控与分配，提出了一种面向用户需求的自动化数据准备优化框架。该框架充分考虑实际业务场景中的资源限制，通过联合优化算子组合和资源投入，寻找低成本、高收益的数据准备方案。

对数据准备全流程开展有效的成本收益优化仍存在诸多挑战，主要技术难点如下。

- 传统成本指标与实际预算量化难以对齐。现有研究多用数据量、特征维度等技术指标衡量数据准备开销，尽管这类指

标能够反映资源消耗，但难以直接与实际预算需求挂钩。若想将实际成本预算约束纳入优化条件，需要将以技术指标为主的成本测算转换为现实量纲下的成本测算，而这会大幅增加成本评估的复杂程度。

- 成本控制与模型性能目标难以兼顾。以往的数据准备管道研究一般不考虑成本开销，倾向于充分执行清洗方法以最大化模型性能，但也容易带来过高的数据准备成本。反之，若一味压缩成本，也可能造成数据质量和下游模型性能的下滑。因此，如何在满足用户需求的前提下实现成本和收益的合理兼顾，是本文需要解决的关键难点。

- 复杂组合空间优化算法的复杂度难以控制。若数据准备包含 K 个环节，每个环节包含 M 种方法和 P 种参数配置，算法的搜索空间高达 $O((M \cdot P)^K \cdot K!)$ 。若直接进行全量搜索，会导致严重的组合爆炸问题。若在传统管道搜索基础上引入成本的规划分配，将会进一步增加优化问题的求解难度。因此，设计一种高效的剪枝机制和近似算法，在控制计算开销的同时获得较优的数据准备方案，是本文面临的一个重大技术挑战。

针对上述技术瓶颈，本文提出了一套系

统性的解决方案，主要贡献和创新点如下。

(1) 提出了一个面向成本收益协同优化的数据准备整体框架 CBAprep (cost-benefit-aware data preparation)。该框架将成本建模、收益预测、组合搜索与资源分配统一纳入优化流程：采用统计回归方法评估数据准备成本，并利用梯度提升树 (gradient boosting decision tree, GBDT) 预测下游模型收益，从而在预算约束下联合优化数据准备效果与资源消耗，弥补了现有方法侧重性能提升、缺乏成本收益分析的不足。

(2) 提出数据准备算子组合的资源分配算法 CARO (cost-benefit-aware resource optimization)。与以往研究追求理想状态下数据准备方法的充分执行不同，本文聚焦解决数据准备方法的成本收益优化问题。CARO 通过引入任务执行程度变量，将方法部分执行纳入优化空间，并根据预算自适应约束分配资源。结合蒙特卡洛采样与梯度引导优化，将资源分配问题由高复杂度搜索转化为关于任务数量线性增长的近似求解过程，算法的时间复杂度从 $O(n^2)$ 降至 $O(n)$ ，实现了对复杂资源分配问题的高效求解。

(3) 设计了方法组合与资源分配的分层联动搜索方法。区别于传统仅针对方法组合开展搜索优化的研究，本文同时将方法组合决策与资源分配两类变量纳入优化体系，在候选方法组合筛选的基础上进一步进行资源投入优化，实现组合选择与资源分配的协同优化。该机制结合规则约束减少无效搜索空间，通过粗粒度组合筛选与细粒度资源规划协同提升系统整体优化效率，且显著提升了搜索效率。

(4) 实验表明，在预算受限的场景下，CBAprep 具备更优的性能与鲁棒性。基于 5 个真实数据集开展测试，相较于 4 种先进

数据准备基线，CBAprep 在有限预算约束下获得了更优性能收益。在低预算场景中，CBAprep 的准确率提升超过 6%。

1 相关工作

近年来，ML 模型训练数据准备过程中的成本问题受到了一定关注，现有工作主要从自动化清洗和管道策略优化两个方面控制数据准备成本。其中，自动化清洗主要适用于缺失值填充、异常值检测、数据一致性检查等具体数据处理环节，利用概率模型、深度学习等技术自动完成特定的清洗目标。研究者不断改进和优化算法，以提升数据清洗的效果和效率。管道策略优化方面的研究，旨在通过近似计算和先验筛选快速获得最佳管道编排，实现源数据到下游模型“端到端”的自动化流程。

1.1 数据清洗自动化算法研究

目前，有关数据清洗自动化的研究主要集中在对训练数据的不一致性、异常值、缺失值、重复项等问题的自动化处理。传统数据清洗方法包括基于完整性约束的清洗、基于规则的清洗、基于统计的清洗、基于深度学习的清洗等^[2,16]。例如，均值填补、回归插补等统计技术基于数据分布特征对缺失值进行处理^[17-20]。近年来，相关研究逐渐转向先进的 ML 技术。HoloClean 利用概率图模型、自动化规则和监督学习检测并修复数据错误，以减少人工干预^[10]；ZeroER 利用对抗生成网络实现零样本重复记录识别^[21]；DemandClean 从数据真实性与多样性的权衡出发，构建多目标学习框架^[22]。

此外,针对不同领域的清洗需求,研究者还提出了独特的自动化清洗方法。针对ML模型训练场景,ActiveClean面向凸优化模型,将数据清洗嵌入梯度下降过程,并优先处理高损失样本^[12]。在自动化算法评估方面,BoostClean采用集成学习方法从预定义方法库中筛选最优清洗策略^[9]。类似地,RLClean进一步利用深度强化学习实现清洗策略的自动化决策与优化^[23]。此外,针对高维、不平衡等复杂数据场景,已有研究通过特征重要性评估和搜索策略,降低冗余特征带来的建模与清洗成本^[24]。

1.2 自动化ML数据准备管道优化策略

随着各类自动化数据处理工具的涌现,为实现ML数据准备阶段的高效数据管道智能编排,国内外学者围绕自动化ML数据准备管道优化策略展开了一系列深入研究^[25]。现有自动化ML数据准备管道优化方法主要分为两类。

第一类方法将数据管道搜索建模为复杂组合优化问题,采用ML、强化学习、元特征和代理模型等技术降低搜索开销^[26]。例如,Learn2Clean框架通过挖掘各类数据处理方法之间的约束关系,以缩减搜索空间,并设计了基于Q学习的强化学习管道搜索机制^[5]。

第二类方法将传统离散管道搜索转化为连续优化问题,借助可微架构搜索技术大幅降低优化开销。例如,HAIPipe结合人工先验经验与AI智能编排进行管道优化^[6];DiffPrep将离散的管道搜索空间转化为连续可微空间,利用梯度下降算法完成高效寻优,整个过程仅需执行一次ML模型训练^[7];Ctxpipe捕获原始数据和转换数据的深层语义信息,引入门控上下文插

件调控模型训练与推理过程,精准筛选适配下游任务的高质量特征管道^[8]。

上述自动化数据准备管道优化方法替代了部分人工操作,降低了数据准备成本。然而,现有方法仍存在明显局限:面对复杂参数配置和高维数据场景,管道优化与执行仍可能带来较大资源开销,且现有研究主要从算法设计角度控制时空成本,缺乏对算子执行成本与下游模型收益的统一量化。为此,本文进一步研究数据准备阶段计算成本与模型收益的平衡,并据此优化方法组合与资源分配。

2 问题概述

基于上述研究成果,针对各类数据清洗方法的成本,以及这些方法组合和资源分配的不同策略对下游ML模型质量指标的影响,本文提出一种面向成本收益协同优化的数据准备整体框架CBAPrep。以用户预算有限的情况为例,CBAPrep能够在预算范围内为用户提供最高的模型性能收益。该问题以数学形式化定义如下。

给定下游模型 M_T 、原始数据集 D 、用户预算 B 、所有可能的数据准备方法组合 $\phi \in \Phi$,以及数据准备策略的资源分配方案 $\alpha \in A$ 。其中, (ϕ, α) 表示资源分配方案 α 下的数据准备流程 ϕ 。本文用成本函数 $C(\phi, \alpha)$ 来表征成本,用质量函数 $Q(\phi, \alpha)$ 评估方案 (ϕ, α) 对下游模型 M_T 性能指标的影响。本研究致力于在搜索空间 Φ 中找到一个特定的方案 ϕ^* ,并计算该方案下的资源分配方案 α^* ,使其满足:

$$\begin{aligned} \phi^*, \alpha^* = \operatorname{argmax}_{\phi \in \Phi, \alpha \in A} Q(\phi, \alpha) \\ \text{s.t. } C(\phi, \alpha) \leq B \end{aligned} \quad (1)$$

更直观地讲, 本文主要解决以下问题: 给定原始数据集 D 、所有可能的数据准备策略组成的搜索空间 Φ 、模型性能评估指标 Q 、用户提供的预算 B 。从 Φ 中寻找成本小于预算 B 的一个 ϕ , 从而得到使性能指标 Q 最大化的训练数据集 $D'=\phi(D)$ 。

CBAprep 的核心目标是, 在成本收益约束范围内寻求优化目标最大化的准备方法组合和资源分配方式。根据各类用户需求, 将优化问题建模为数学规划问题, 通过求解规划问题实现整个数据准备流程的优化。

CBAprep 整体架构如图 1 所示。CBAprep 以分类任务作为下游任务, 考虑预处理、数据清洗、特征工程等操作, 收集用户的具体需求指标, 在数据准备方法组合空间中进行决策, 并对具体组合进行资源优化分配, 从而提供满足用户需求的最优数据准备方案。

3 CBAprep 的核心方法

3.1 基于成本收益需求的数学规划模型

首先, 在 CBAprep 框架中将用户需求归纳和分类为数学规划问题。用户需求分类示意图如图 2 所示。

考虑成本收益, 用户对数据准备过程的需求可分为以下 3 类。

(1) 综合考虑成本和收益, 选择在预算范围内能够提供最佳性能提升的方案。设总预算为 B , 方法 j 的执行成本为 c_j , 二进制变量 x_j 用来表示是否选择方法 j , 其目标是最大化收益成本比, 同时满足总成本 $\sum_j c_j x_j \leq B$ 。

(2) 针对特定的精度要求, 选择能够达到该精度的最低成本方案。设目标精度

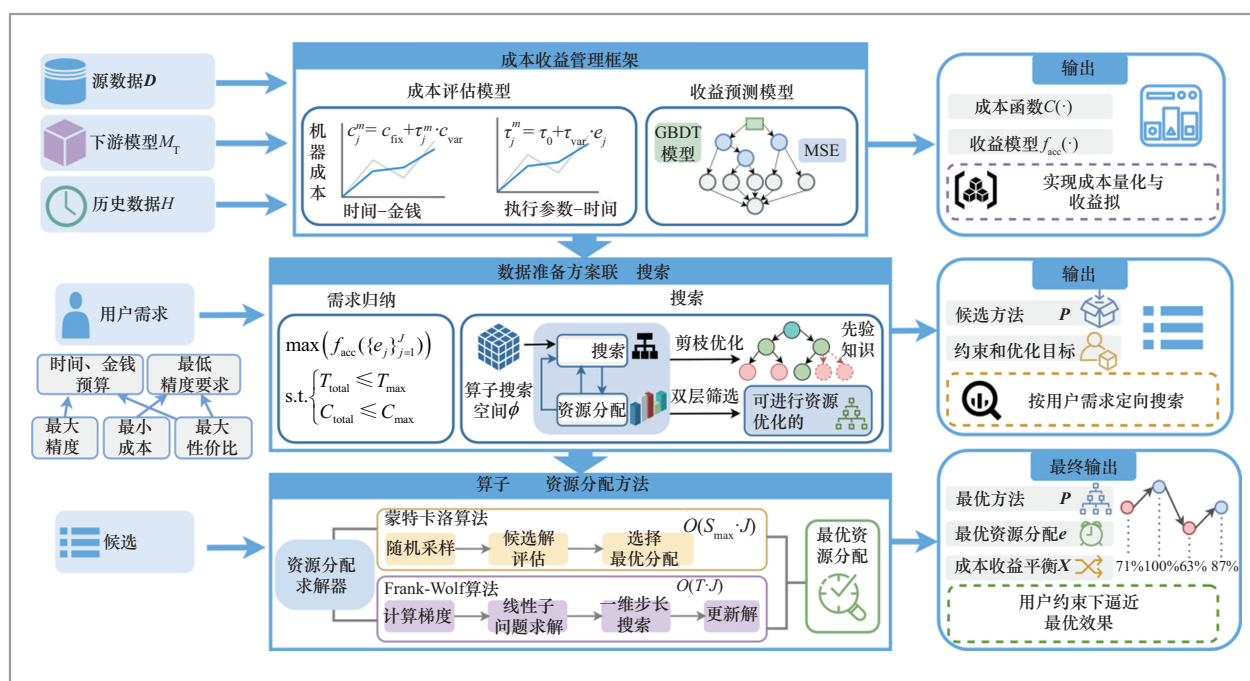


图1 CBAprep 整体架构

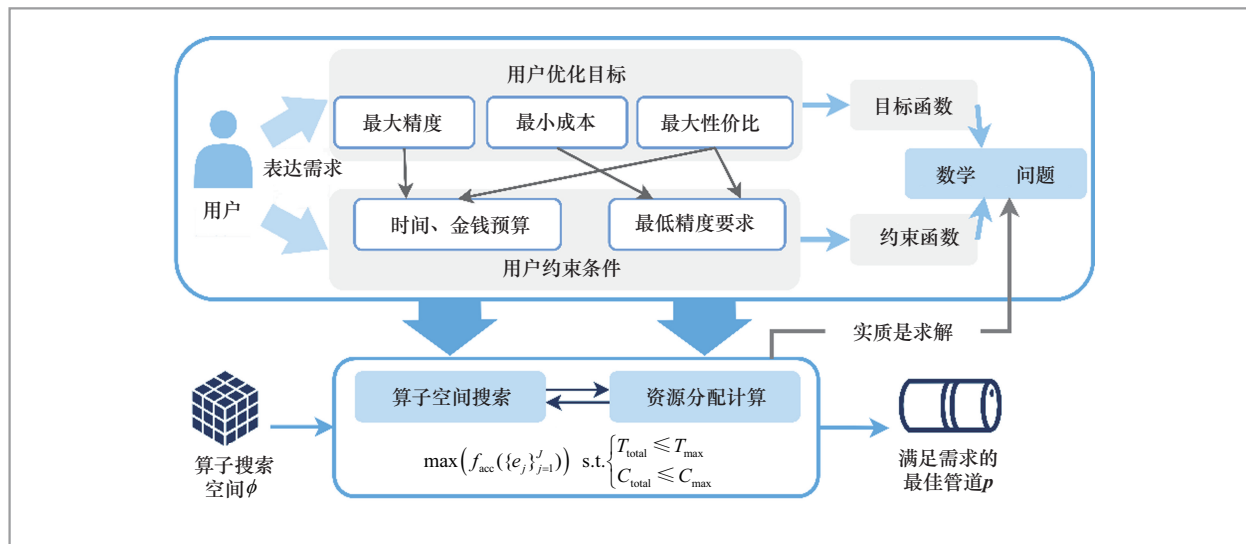


图2 用户需求分类示意图

为 δ ，二进制变量 $\mathbf{x}$ 用来表示方法选择向量， $\mathbf{e}$ 表示方法的执行程度向量， f_{acc} 是方法选择和资源分配到下游模型准确度的映射，该方案的目标是最小化总成本 $\sum_j c_j x_j$ ，同时满足精度约束 $f_{acc}(\mathbf{x}, \mathbf{e}) \geq \delta$ 。

(3) 在预算有限的情况下，优先选择性价比高的方法。该方案的目标是最小化总成本 $\sum_j c_j x_j \leq B$ ，同时最大化总收益 $\sum_j r_j x_j$ 。其中， r_j 表示方法 j 的收益。

以上3类需求对应的优化目标为最大精度、最小成本、最大性价比。在优化目标和约束条件下，将用户需求表示为多变量的数学规划问题。下面以基于预算约束的数据准备为例进行说明。

为全面反映数据准备阶段各类任务的成本，对于数据处理任务 j ，3种不同需求的优化目标不同：成本优先模式的优化目标为最小化总成本 C_{total} ，即 $\min(C_{total})$ ；精度优先模式的优化目标为最大化模型准确度 $f_{acc}(j)$ ，即 $\max(f_{acc}(j))$ ；平衡模式的优化目标为最大化收益成本比，即 $\max\left(\frac{f_{acc}(j)}{C_{total}}\right)$ 。其中， $C_{total} = \sum_{j \in J} (c_j)$ ， J 表示当前数据准备方案中采取数据处理任务的集合， $j \in J$ 表示其中一个具体的数据处理任务。

数据准备阶段用户需求模式总结见表2。

成本收益分析的核心是在一定约束条

表2 数据准备阶段用户需求模式总结

用户需求模式	优化目标	约束条件	表达式
成本优先模式	最小化总成本	达到最低精度要求	$\min(C_{total})$
精度优先模式	最大化模型准确度	时间/金钱成本 $\leq B$	$\max(f_{acc}(j))$
平衡模式	最大化收益成本比	满足资源与精度的双向边界	$\max\left(\frac{f_{acc}(j)}{C_{total}}\right)$

件下得到实现优化目标的解。这些约束条件通过数据准备框架中的用户引导机制获得，使最终优化结果充分符合用户要求，包括 $C_{total} \leq C_{max}$ 、 $f_{acc}(j) \geq \delta_{min}$ ，其中 δ_{min} 表示用户对下游模型设定的最低精度要求。

深入分析用户需求并将其抽象为特定的需求范式，进而将其转化为具体的数学规划问题。后续算法的核心就是在组合决策与资源分配的框架下，实现对该规划模型及其衍生问题的有效求解。

3.2 数据准备成本收益管理框架

本文从时间开销的视角开展数据准备成本的分析和计算，并以下游模型性能的提升为本文所提方法的收益依据。在此基础上，构建统一的成本收益管理机制，对各类数据准备任务进行成本管控与全局收益评估。

3.2.1 成本计算方法:MachineCost

根据不同数据准备方法的机器处理逻辑，将其成本分为以下3类，用于确定各类数据准备方法的成本调整空间和计算基准。数据准备方法成本计算类别见表3。

(1) 规则映射处理类。该方法可在局部数据上拟合规则，但必须将规则应用于全量数据，否则可能造成量纲失衡或语义错位。其资源投入可通过控制拟合数据

规模、候选特征规模或搜索空间大小进行调节，因此，部分执行主要表现为“部分数据拟合、全量规则应用”。

(2) 样本独立处理类。包括缺失值处理、离群值检测和重复数据去除等。该类操作以样本或记录为基本对象，不同样本间相对独立，通过控制处理样本比例调节成本，且处理时间通常与数据量近似线性相关。

(3) 全局结构处理类。包括数据集划分、数据转换等。该类操作会改变数据集整体结构，难以按样本比例拆分，其成本在特定数据集和参数配置下通常较固定，因此本文将其作为固定成本项，不参与连续执行程度优化。

基于以上数据准备方法成本计算的分类总结，本文进一步对执行程度变量的适用边界展开形式化说明。对于第 j 个数据准备算子，执行程度 $e_j \in [0, 1]$ 表示该算子在可调资源维度上的投入比例。对于规则映射处理类和样本独立处理类算子， e_j 可对应参与拟合的数据比例、处理样本比例、候选特征比例或搜索迭代比例。对于全局结构处理类算子，本文设置 $e_j = 1$ ，将其作为固定成本项纳入预算约束。

在此基础上，支持执行程度调节算子的机器成本可近似表示为执行程度的函数：

$$C_j(e_j) = a_j e_j + b_j \tag{2}$$

其中， a_j 和 b_j 表示通过历史运行数据拟合得到的成本参数， e_j 表示第 j 个算子的执行程度。

对于不支持执行程度调节的全局结构类算子，其成本记为固定值 C_j 。因此，给定数据准备方案 P 和执行程度向量 e ，总成本可表示为：

$$C(P, e) = \sum_{j \in \mathcal{J}_e} C_j(e_j) + \sum_{j \in \mathcal{J}_f} C_j \tag{3}$$

表3 数据准备方法成本计算类别

类别	处理步骤
规则映射处理类	特征编码、特征选择、特征提取、数据规范化
样本独立处理类	重复数据去除、缺失值处理、离群值检测
全局结构处理类	数据转换、数据集划分

其中, $\mathcal{J}_a$ 表示支持执行程度调节的算子集合, $\mathcal{J}_f$ 表示支持固定执行的算子集合。

MachineCost 采用基于上述算子执行逻辑分类的成本建模方式, 实现对算子成本的灵活刻画。

3.2.2 数据准备收益评估模型

本文以下游模型性能衡量数据准备收益, 并根据任务类型选择不同指标: 分类任务采用准确率 (accuracy) 和 F1 分数 (F1-score), 回归任务采用平均绝对误差 (mean absolute error, MAE)、均方根误差 (root mean square error, RMSE) 或决定系数 (coefficient of determination, R^2), 深度学习及大模型任务采用验证集损失、Benchmark 得分或人工偏好评价。因此, CBAprep 的收益模型并不依赖特定下游模型, 而是将数据准备策略到下游性能的映射视为一个可学习的预测问题。

以分类任务为例, 本文以下游分类模型的准确率为衡量数据准备收益的指标。因此, 针对分类任务训练收益评估模型,

模型输入为方法组合的 0~1 决策变量及浮点型执行程度 $\phi_i = [z_i, e_i]$, 输出为最终的收益预测准确率 $F_{acc}^{(T)}(\phi)$, 这一过程面临成本和收益关系非线性、输入数据维度过高的问题。考虑到成本收益关系具有非线性且输入维度较高, 本文采用 GBDT 模型进行收益预测。GBDT 模型采用树结构刻画复杂非线性关系, 并可通过限制树深度和叶节点数控制模型复杂度。基于 GBDT 的下游模型准确率预测模型流程如图 3 所示。其中, 训练数据集 $D = \{(\phi_i, y_i)\}_{i=1}^n; y_i \in [0, 1]$ 表示对应下游模型的实际准确率; $\mathcal{L}$ 表示损失函数, 用于衡量预测精度与实际收益之间的偏差; T 表示迭代次数、 η 表示学习率、 d_{max} 表示最大树深, 用于控制模型复杂度, 防止维度爆炸; X_{train} 和 y_{train} 分别表示处理后的训练数据及标签; $f_0(x)$ 表示初始预测模型; $r_i^{(t)}$ 表示第 t 轮第 i 个样本的伪残差; $R_j^{(t)}$ 和 $\gamma_j^{(t)}$ 分别表示第 t 棵树的第 j 个叶节点区域及其权重; $I(\cdot)$ 表示指示函数; $f_t(x)$ 和 $F(x)$ 分别表示第 t 个弱学习器和最终 GBDT 模型。

基于 GBDT 的下游模型准确率预测模

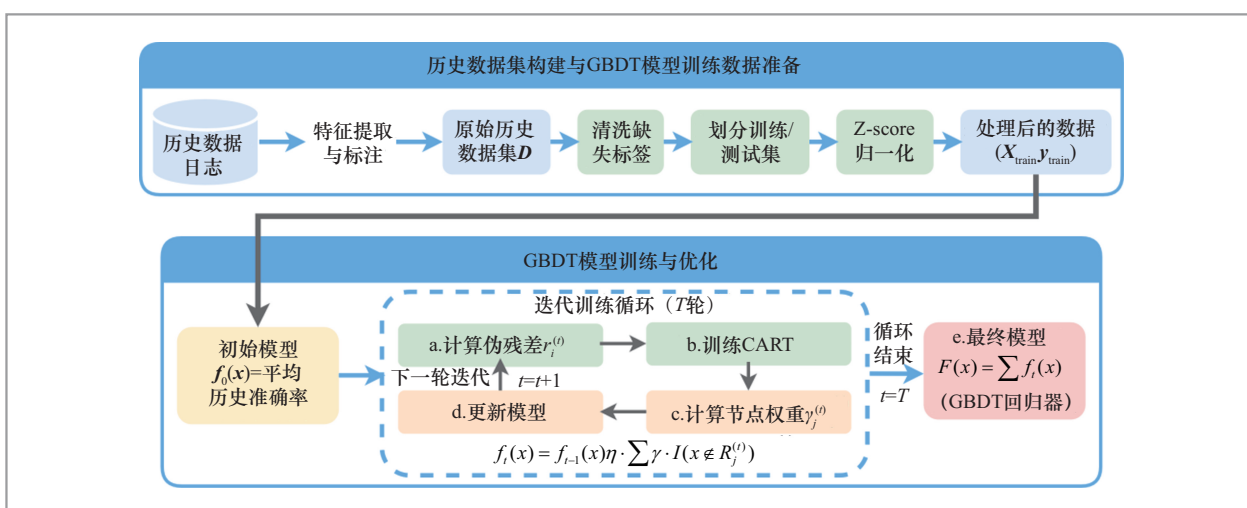


图3 基于GBDT的下游模型准确率预测模型流程

型算法如算法1所示。

算法1 基于GBDT的下游模型准确率预测模型算法

输入：训练数据集 D 、损失函数 $\mathcal{L}$ 、迭代次数 T 、学习率 η 、最大树深 $d_{\max}$

输出：最终的收益预测准确率 $F_{\text{acc}}^{(T)}(\phi)$

计算历史样本准确率的初始平均水平作为预测起点：

$$F_{\text{acc}}^{(0)}(\phi) \leftarrow \arg \min_{\gamma} \sum_{i=1}^n L(y_i, \gamma)$$

展开 T 轮迭代，每轮构建一棵新的决策树

for t from 1 to T do

计算准确率残差：

$$r_{\text{acc},i}^{(t)} \leftarrow - \left[\frac{\partial L(y_i, F_{\text{acc}}(\phi_i))}{\partial F_{\text{acc}}(\phi_i)} \right]_{F_{\text{acc}} = F_{\text{acc}}^{(t-1)}},$$

$$\forall i \in \{1, \dots, \mu\} h^{(t)} \leftarrow$$

构建深度为 $d_{\max}$ 的GART树拟合样本 $\{(\phi_i, r_{\text{acc},i}^{(t)})\}_{i=1}^n$

计算使损失函数最小化的权重：

$$\gamma_{\text{acc},j}^{(t)} \leftarrow \arg \min_{\gamma} \sum_{\phi_i \in R_j^{(t)}} L(y_i, F_{\text{acc}}^{(t-1)}(\phi_i) + \gamma),$$

$\forall j \in \{1, \dots, J^{(t)}\}$

将新生成的树以学习率 η 累加到原有模型中：

$$F_{\text{acc}}^{(t)}(\phi) \leftarrow F_{\text{acc}}^{(t-1)}(\phi) + \eta \cdot \sum_{j=1}^{J^{(t)}} \gamma_{\text{acc},j}^{(t)} I(\phi \in R_j^{(t)})$$

end for

return $F_{\text{acc}}^{(T)}(\phi)$

在GBDT训练阶段，上文时间复杂度为 $O(T \cdot (d \cdot M \cdot h + 2^h + M))$ ，由样本量 M 、迭代次数 T 、树深度 h 和特征维数 d 共同决定。其中，特征标准化复杂度为 $O(d \cdot M)$ ，单棵树构建中分裂点查找占 $O(d \cdot M \cdot h)$ ，树生长因深度限制仅需 $O(2^h)$ ，每轮残差计算的时间复杂度为 $O(M)$ 。

本文首先利用线性回归模型量化数据准备方法的执行成本，再基于历史运行记

录构建GBDT模型，用于预测下游任务收益，从而实现对资源消耗与性能增益的协同评估。作为CBAprep的关键组件，MachineCost与GBDT收益模型的有效性将在第4.4节通过成本估计和收益预测实验进行验证。

3.3 数据准备算子组合的资源分配算法:CARO

数据准备算子组合的资源分配优化实际上是将给定方案的代价降至预算范围内，并在这个过程中尽量保留最大收益。对于支持渐进式资源投入的数据准备算子，执行程度 e_j 表示其资源投入比例；对于不支持部分执行的全局结构类算子，其执行程度固定为1，并作为固定成本项计入预算。

给定数据准备方案 P 、预算 B 、成本函数 $C(P, \mathbf{e})$ 和收益预测函数 $F(P, \mathbf{e})$ ，数据准备算子组合的资源分配问题可表示为：

$$\max_{\mathbf{e}} F(P, \mathbf{e}) \quad (4)$$

$$\text{s.t. } C(P, \mathbf{e}) \leq B, 0 \leq e_j \leq 1, j \in \mathcal{J}_a, e_j = 1, j \in \mathcal{J}_f$$

其中， $\mathbf{e} = (e_1, e_2, \dots, e_J)$ 。

在该资源分配问题中，只需求解以上固定方法组合的预算资源规划问题，就能实现预算范围内高代价方法组合的成本优化。求解上述问题同样面临高维连续搜索困难、高价值资源投入方向难以动态识别的双重挑战。

(1) 基于蒙特卡洛采样的资源优化方法

蒙特卡洛方法在可行域内随机采样执行程度向量，将复杂优化问题转化为有限候选解评估问题，适用于高维资源分配，且不需要目标函数的显式梯度，对非凸、非线性收益函数具有较强适应性。本文在约束可行域内随机采样执行程度向量

$e^{(1)}, e^{(2)}, \dots, e^{(S_{\max})}$ 。其中, $S_{\max}$ 为最大采样次数。各采样方案的预算可行性及收益 $R^{(s)}$ 可表示为:

$$R^{(s)} = f_{\text{acc}}(e^{(s)}) \quad (5)$$

选择收益最大的可行解:

$$e^* = \arg \max_s R^{(s)} \quad (6)$$

若每次采样需计算 J 个任务, 则整体时间复杂度为 $O(S_{\max} \cdot J)$ 。若任务数呈线性增长, 可调节 $S_{\max}$ 以平衡搜索精度与计算开销。本文采用蒙特卡洛近似求解, 并以 Frank-Wolfe 方法为收益函数可微时的替代路径。

(2) 基于 Frank-Wolfe 算法的资源优化方法

对资源优化问题进行求解, 本质上是求解约束问题。用 e 表征任务数量占比、投入时间占比等比例关系, 因此, $T_{\text{total}} \leq T_{\max}$ 和 $C_{\text{total}} \leq C_{\max}$ 对 e_j 来说是线性约束, 可以用 Frank-Wolfe 算法求解预算资源规划问题。数据准备方案资源消耗优化算法如算法 2 所示。

算法 2: 数据准备方案资源消耗优化算法

输入: 方法组合 $\phi = \{x_1, x_2, \dots, x_J\}$, 准确度函数 $f_{\text{acc}}(e)$, $e \in [0, 1]^J$, 最大迭代次数 K , 精度参数 ϵ

输出: 近似解 e^*

```
//设置初始可行点  $e \leftarrow e^{(0)}$ 
for  $k$  from 0 to  $K-1$ 
    //计算梯度, 确定可行方向:
     $g^{(k)} \leftarrow \nabla f_{\text{acc}}(e^{(k)})$ 
    //解决线性优化子问题, 得到可行域顶点  $s^{(k)}$ :  $s^{(k)} \leftarrow \arg \max_{s^{(0)} \in [0, 1]^J} \langle g^{(k)}, s \rangle$ 
    //沿着下降方向作一维步长搜索:
     $d^{(k)} \leftarrow \langle g^{(k)}, e^{(k)} - s^{(k)} \rangle$ 
```

```
if  $d^{(k)} \leq \epsilon$  //满足收敛条件, 返回结构
return  $e^{(k)}$ 
```

```
//不满足收敛条件, 继续迭代:  $\gamma_k =$ 
 $\arg \max_{\gamma \in [0, 1]} f_{\text{acc}}(e^{(k)} + \gamma(s^{(k)} - e^{(k)}))$ 
 $e^{(k+1)} \leftarrow e^{(k)} + \gamma(s^{(k)} - e^{(k)})$ 
end for
return  $e^*$ 
```

传统搜索方法在 $[0, 1]^J$ 空间维度内, 搜索复杂度随任务数 J 呈指数级增长, Frank-Wolfe 算法将非线性优化问题分解为序列线性问题, 每次迭代仅需在可行域顶点间搜索, O/ϵ 次迭代后能够达到最优解, 解决了搜索效率问题。

面向数据准备成本收益协同优化任务, 本文构建了实现资源约束的数据准备收益最大化方法, 引入用户需求分析机制, 通过成本量化与基于 GBDT 的收益拟合, 配合双层搜索策略, 实现了数据准备算子组合与资源投入的联合优化。

4 实验设置与结果分析

4.1 数据集

针对一系列分类任务展开实验, 所选数据集及其对应的下游任务场景具体如下。

(1) Google (Google Play Store Apps): 来源于 Kaggle 的真实数据集, 包含 Google Play 商店应用信息, 存在字段缺失、噪声和类别不平衡等问题。实验任务为预测应用评分是否高于 4.0, 可用于商店推荐场景。

(2) House (House Prices): 来源于 Kaggle 的真实数据集, 包含面积、位置、建造年份和房间数量等房屋属性, 存在缺失值、异常值、特征冗余和类别不平衡等问题。实验任务是预测房价是否高于中位数, 可用于房地产估价或贷款风险评估。

(3) Shuttle：来源于加州大学欧文分校机器学习数据集库（University of California, Irvine Machine Learning Repository, UCI）的数据集，包含航天飞机飞行过程中的温度、压力和振动等传感器读数，存在噪声、字段缺失和类别不平衡等问题。实验任务是判断航天飞机状态是否异常，可用于健康监测和故障预警。

(4) Chess：来源于UCI的数据集，包含国际象棋对局中的位置和移动信息，存在字段缺失、噪声和类别不平衡等问题。实验任务是预测对局是否获胜，可用于国际象棋AI训练和策略分析。

(5) Uscensus：来源于UCI的数据集，包含美国人口普查中的年龄、教育程度和职业等信息，存在字段缺失、噪声和类别不平衡等问题。实验任务是预测个人年收入是否高于5万美元，可用于市场分析或政策优化。

上述数据集的属性数量和数据量见表4。

表4 实验数据集的属性数量和数据量

数据集	属性数量/个	数据量/条
Google	13	10 800
Chess	7	4 500
House	81	1 460
Shuttle	10	58 000
Uscensus	15	3 200

4.2 对比模型和评价指标

实验选取分类任务作为下游任务进行初步实验，选用逻辑回归（logistic regression, LR）、线性判别分析（linear discriminant analysis, LDA）、朴素贝叶斯（naive Bayes, NB），以及分类与回归树（classification and regression tree, CART）作为下游分类模型，以验证不同方法对不同模型结构的适用性。数据按照7：3的比例划分为训练集和测试集，测试集在管道搜索过程中保持不可见。考虑到各基线支持的模型存在差异，统一采用LR完成管道方法对比。

实验指标方面，采用模型性能、执行时间、实际预算比、成本收益效率4项核心指标衡量数据准备方案的收益、成本、实际预算比和成本收益效率。实验指标的含义与来源见表5。其中，模型性能采用准确率（accuracy）和F1分数两类具体指标进行衡量。

参与对比的基线方法有以下4种。

- Learn2Clean（简称L2C）：采用经典强化学习算法Q-Learning，在预定义动作空间中寻找最优数据清洗路径。
- HAIPipe-AI（简称HAI-AI）：应用HAIPipe提出的基于深度强化学习的AI数据准备管道生成算法，结合深度强化学习与启发式搜索生成数据准备管道。
- DiffPrep：将预处理过程微分化，

表5 实验指标含义与来源

指标	含义	来源
performance	模型性能	预设的模型性能衡量指标
prep_time	执行时间	数据准备过程的整体时间
cost/budget	实际预算比	实际数据准备成本与给定预算的比值
gain/cost	成本收益效率	模型性能提升与实际数据准备成本的比值

通过梯度下降联合优化预处理方案与模型参数。

- Ctxpipe：借助语义嵌入获取数据上下文特征，并通过深度强化学习进行算子决策。

进一步引入预算约束随机搜索（random-budget search, RBS）、贪心背包（greedy knapsack, GK）方法和均匀预算分配（uniform-budget allocation, UBA）3种预算感知基线。

- RBS：在预算约束下随机采样算子组合及执行程度，并选择验证集性能最优的可行方案。

- GK：按单位成本收益比排序，并采用背包式贪心策略选择算子。

- UBA：在算子组合确定后，将预算均匀分配给各可调算子。

4.3 实验设置

为展现 CBAPrep 框架在数据准备管道成本收益规划方面的优势，将 CBAPrep 框架与其他数据准备管道生成方法 L2C、DiffPrep、Ctxpipe、HAI-AI 进行对比。

设定整体数据准备管道运行时间预算 B_{exec} ，分别取 300 ms、500 ms、700 ms、1 000 ms，收集在 B_{exec} 约束下，各类数据管道生成方法在下游模型验证集上的性能指标，包括准确率和 F1 分数。

对于 CBAPrep 框架，直接输入 B_{exec} 作为约束进行管道生成即可。由于对比基线方法未设置运行时间预算约束规划模块，因此，在其生成最优管道情况下，如果生成管道运行时间超出时间预算，则对管道内容进行裁剪，以满足时间预算要求；如果生成管道的运行时间最终无法满足 B_{exec} ，则判定该生成方法无法交付可用结果（表格中用“—”表示）。

为避免非预算感知基线通过管道裁剪带来的实验偏差，进一步将 CBAPrep 框架与 RBS、GK 和 UBA 3 种预算感知基线进行比较。所有方法均在相同 B_{exec} 约束下搜索数据准备方案，并比较其平均下游模型性能、实际预算比和单位成本收益。

4.4 实验结果与分析

4.4.1 管道搜索结果在有限预算下的表现对比

在 5 个数据集下，CBAPrep 框架与 L2C、DiffPrep、Ctxpipe、HAI-AI 数据准备管道生成方法在不同预算约束下的效果对比见表 6。

根据表 6，在多数运行时间预算受限场景下，CBAPrep 均能生成满足预算要求的可用管道，并取得较优的准确率。当运行时间预算降至 300 ms 时，多数缺乏预算感知机制的基线方法无法生成有效管道，而 CBAPrep 在所有预算设置下均能稳定输出可用方案，表现出较强的资源受限适应性。以 Shuttle 数据集为例，CBAPrep 在 300 ms 运行时间预算下的准确率达到 0.966，已接近多数基线方法在 1 000 ms 运行时间预算下的性能，说明其能够优先组合高性价比算子，且在有限预算内获得较高收益。随着运行时间预算由 300 ms 增至 1 000 ms，CBAPrep 的性能平稳提升，表明其可根据资源变化调整算子组合和分配方案。相比之下，部分基线方法因缺乏成本收益规划，在运行时间预算不足时需要裁剪管道，导致性能明显下降。例如，L2C 在 Shuttle 数据集上于 700 ms 和 1 000 ms 时的准确率为 0.966，而在 300 ms 和 500 ms 时降至 0.900。总体来看，CBAPrep 能够随预算变化提供更精细、稳定的数据准备方案。

表6 不同预算约束下各数据管道生成方法的效果对比

数据集	B_{acc}/ms	L2C		HAI-AI		DiffPrep		Ctxpipe		CBAprep	
		准确率	F1分数	准确率	F1分数	准确率	F1分数	准确率	F1分数	准确率	F1分数
Google	300	—	—	—	—	0.545	0.356	0.525	0.474	0.545	0.478
	500	0.600	0.591	0.550	0.358	0.549	0.547	0.595	0.581	0.556	0.539
	700	0.600	0.591	0.550	0.358	0.549	0.547	0.595	0.581	0.586	0.554
	1 000	0.633	0.633	0.621	0.613	0.593	0.574	0.595	0.581	0.628	0.614
Jungle	300	—	—	—	—	—	—	0.674	0.641	0.662	0.645
	500	0.604	0.600	0.681	0.652	0.670	0.484	0.674	0.641	0.681	0.674
	700	0.673	0.671	0.683	0.651	0.670	0.484	0.674	0.641	0.687	0.521
	1 000	0.680	0.656	0.760	0.733	0.680	0.552	0.674	0.641	0.741	0.728
House	300	—	—	—	—	—	—	0.818	0.813	0.898	0.898
	500	0.877	0.865	0.901	0.900	0.895	0.878	0.818	0.813	0.903	0.900
	700	0.915	0.912	0.913	0.911	0.915	0.910	0.818	0.813	0.915	0.899
	1 000	0.928	0.928	0.920	0.919	0.928	/0.928	0.818	0.813	0.928	0.918
Shuttle	300	0.900	0.892	0.922	0.910	—	—	0.934	0.927	0.966	0.405
	500	0.900	0.892	0.930	/0.923	0.954	0.891	0.974	0.972	0.970	0.632
	700	0.966	0.965	0.929	0.921	0.979	0.907	0.974	0.972	0.981	0.606
	1 000	0.966	0.965	0.951	0.947	0.990	0.479	1.000	1.000	0.982	0.567
Uscensus	300	—	—	—	—	0.816	0.695	0.785	0.756	0.858	0.731
	500	0.796	0.759	0.800	0.682	0.816	0.695	0.813	0.797	0.871	0.690
	700	0.796	0.759	0.802	0.686	0.849	0.780	0.813	0.797	0.875	0.700
	1 000	0.854	0.849	0.802	0.686	0.849	0.780	0.813	0.797	0.874	0.723

4.4.2 预算感知基线对比实验

在相同预算约束和相同候选规模下，进一步对比 CBAprep 与预算感知基线 RBS、GK 和 UBA 在搜索数据准备方案上的性能，其中 RBS 从满足预算约束的随机候选中根据验证集真实性能选择最优方案。以平均准确率、平均实际预算比和平均成本收益效率为性能评估指标。其中，平均准确率和平均实际预算比越高越好；平均成本收益效率应不超过 1，在满足预算约束时越接近 1，表示预算利

用越充分；若均大于 1，则越接近 1 表示超预算程度越小。在不同数据集下，不同方案搜索数据准备方案的性能对比见表 7。

实验结果表明，CBAprep 在多数数据集上取得了更高的平均准确率，且在预算利用和单位成本收益方面表现较为稳定。与 RBS 相比，CBAprep 利用成本估计和收益预测结果减少无效搜索，能够更稳定地找到高收益候选方案。与 GK 相比，CBAprep 不仅考虑单个算子的单位成本收益，还进一步考虑算子组合之

表 7 不同方案搜索数据准备方案的性能对比

数据集	方法	平均准确率	平均实际预算比	平均成本收益效率
Google	CBAprep	0.578 8	0.733 6	0.078 6
	RBS	0.588 9	1.101 0	0.100 7
	GK	0.554 6	1.150 9	0.155 9
	UBA	0.576 1	1.139 0	0.145 3
Jungle	CBAprep	0.692 8	0.404 0	0.073 4
	RBS	0.683 3	0.752 3	0.028 9
	GK	0.679 1	1.205 1	-0.152 8
	UBA	0.678 8	1.469 9	-0.123 1
House	CBAprep	0.911 0	1.152 7	0.000 1
	RBS	0.869 4	1.439 2	-0.139 1
	GK	0.867 1	1.205 1	-0.152 8
	UBA	0.873 8	1.469 9	-0.123 1
Shuttle	CBAprep	0.974 8	0.490 6	0.228 3
	RBS	0.967 8	0.845 2	0.187 4
	GK	0.833 6	0.532 5	-0.660 6
	UBA	0.678 8	0.725 0	0.000 3
USCensus	CBAprep	0.869 5	0.825 1	0.133 5
	RBS	0.863 3	1.150 9	0.155 9
	GK	0.848 7	1.101 0	0.100 7
	UBA	0.866 1	1.139 0	0.145 3

间的相互影响和执行程度分配，因此能够避免局部贪心选择带来的次优结果。与UBA相比，CBAPrep能够根据数据进行自适应资源分配，从而取得更好的成本收益效率。

4.4.3 消融实验

本文设置了CBAPrep框架下的消融实验，用于验证系统关键组件在成本收益规划方面的作用，并探究各类超参数设置对系统整体性能的影响。

- (1) 关键组件作用分析
- 为验证算子优化选择与自适应资源分

表8 各模块的组件设置

策略	算子采用	算子优化选择	自适应资源分配
CBAPrep	✓	✓	✓
R_DIS	✓	✓	×
R_ALL	✓	×	×
NONE	×	×	×

表9 各模块的消融实验对比

下游模型	实验方法	执行时间	准确率	F1分数
NB	CBAPrep	0.702	0.717	0.702
	R_DIS	0.649	0.627	0.608
	R_ALL	0.575	0.158	0.135
	NONE	0	0.175	0.148
LDA	CBAPrep	0.665	0.678	0.667
	R_DIS	0.547	0.654	0.638
	R_ALL	0.571	0.285	0.257
	NONE	0	0.079	0.061
CART	CBAPrep	0.988	0.559	0.539
	R_DIS	0.722	0.537	0.512
	R_ALL	0.351	0.225	0.198
	NONE	0	0.253	0.224

配两个关键组件的作用，分别设置CBAPrep、R_DIS、R_ALL和NONE4种策略进行对比。其中，CBAPrep：在算子优化选择基础上进行自适应资源分配。R_DIS：仅进行算子组合的优化选择。R_ALL：完全随机产生的数据准备策略。NONE：不进行数据准备。各模块的组件设置见表8。

在满足预算要求的前提下，基于Google数据集，设置Bexec为1 000 ms，分别对比CBAPrep与R_DIS、R_ALL、NONE在下游模型NB、LDA和CART中的运行时间、运行开销和分类模型准确率。各模块的消融实验对比见表9。

整体来看，CBAPrep在NB、LDA和CART模型上均取得最高的准确率与F1分数，其在NB模型中的准确率达0.717，说明CBAPrep能够稳定提升下游模型性能。相比之下，NONE与R_ALL表现较差，R_ALL在NB模型中的准确率仅为0.158，表明随机选择算子及资源分配容易造成模型性能不稳定。R_DIS的性能表现明显优于R_ALL，验证了算子筛选模块能够帮助模型获得更合适的算子组合，但仍低于CBAPrep，说明自适应资源分配对模型性能有进一步提升作用。

从成本收益表现看，在LDA模型下，R_DIS的执行时间比R_ALL少0.024 s，同时将准确率由0.285提升至0.654，说明算子筛选能够识别高性价比方案。CBAPrep仅比R_DIS增加0.118 s，却将准确率进一步提升至0.678，表明自适应资源分配模块能够通过合理配置有限资源获得更优的数据准备效果。

- (2) 超参数设置影响

资源分配计算部分的采样次数会影响特定管道的成本与性能，而最大采样次数 $S_{\max}$ 是蒙特卡洛采样的基本参数，会对系

统整体的管道优化结果产生重要影响。为分析 $S_{\max}$ 对搜索性能的影响,开展了相关消融实验,实验固定数据集为 Google,下游模型为 LDA, Bexec 为 500 ms, $S_{\max}$ 分别取 100、800、1 000、5 000、10 000 和 50 000。每组实验独立运行 100 次,并报告下游模型平均准确率及波动范围。 $S_{\max}$ 对下游模型预测准确率的影响如图 4 所示。

图 4 表明,随着 $S_{\max}$ 增加,预测准确率整体呈上升趋势并逐渐趋于稳定。当 $S_{\max}=100$ 时,预测准确率约为 0.561,且误差波动较大,说明采样不足会导致资源分配估计不稳定。随着 $S_{\max}$ 增至 800 和 1 000,预测准确率明显提高,误差范围缩小,表明增加采样次数能够提高资源分配搜索结果的稳定性。当 $S_{\max}$ 达到 5 000 以上时,预测准确率上升幅度逐渐减小,继续增加采样仅带来有限收益和额外开销。因此,本文将 $S_{\max}=10\,000$ 设为默认值,以平衡搜索效果与计算成本。

4.4.4 成本估计与收益预测模型的有效性验证

CBAprep 框架的优化决策依赖于 MachineCost 成本估计模型和 GBDT 收益评估模型的预测结果,因此,本文基于下游模型 LR 进一步对两个模型进行独立验证。

(1) MachineCost 的有效性验证

针对成本估计模型 MachineCost,采样不同数据准备算子、不同执行程度和不同数据集规模下的运行配置,记录真实执行时间作为真实成本,并与 MachineCost 预测结果进行比较。评价指标采用 MAE、RMSE、 R^2 和平均绝对百分比误差(mean absolute percentage error, MAPE)。其中,MAE 和 RMSE 用于衡量预测成本与真实成本之间的绝对误差,

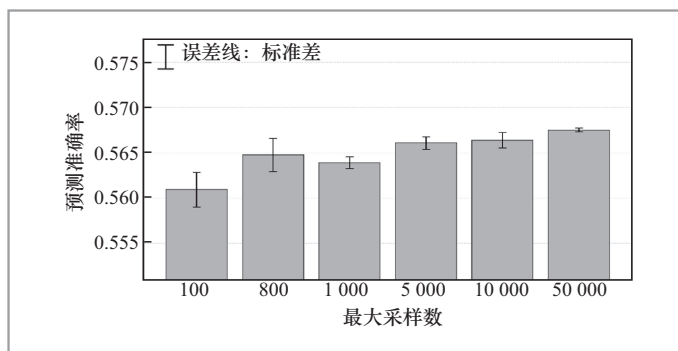


图4 $S_{\max}$ 对下游模型预测准确率的影响

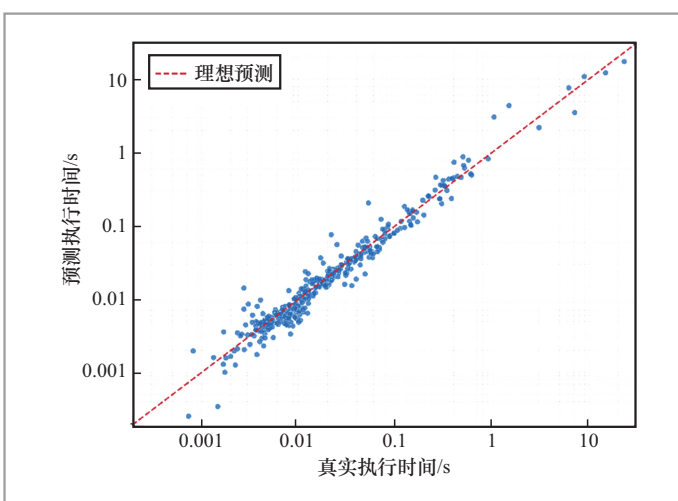


图5 MachineCost的预测执行时间与真实执行时间对比

MAPE 用于衡量相对误差, R^2 用于衡量模型对真实成本变化的拟合程度。MachineCost 的预测执行时间与真实执行时间对比如图 5 所示。

MachineCost 模型的 MAE 为 0.0685, RMSE 为 0.4395, R^2 达到 0.9216, 说明该模型能够较好地拟合真实执行成本变化趋势, 并有效解释不同数据准备方案之间的成本差异。MAPE 为 25.1122%, 其相对误差处于可接受范围内, 说明 MachineCost 的成本估计具有可靠性。

(2) GBDT 的有效性验证

针对收益评估模型 GBDT 的有效性验

证, 本文构造不同的数据准备管道配置, 以获得对应下游模型的真实性能, 并将其与GBDT预测收益进行比较。评价指标采用MAE、RMSE、 R^2 和Spearman相关系数。其中, Spearman相关系数用于衡量预测收益排序与真实收益排序之间的一致性。GBDT模型的有效性验证结果如图6所示。

根据实验数据, GBDT模型的MAE为0.0045, RMSE为0.0077, R^2 为0.8352, 说明该模型能够较准确地拟合数据准备策略与下游模型性能之间的非线性关系。进一步, GBDT模型的Spearman相关系数达0.9071, 表明GBDT模型的预测收益排序与真实收益排序具有较强一致性, 能够有效识别高收益候选管道。同时, 收益预测模型的有效性也说明CBAPrep具有一定的迁移能力, 在更换下游模型时, 只需替换任务指标并重新训练或微调收益预测模型即可。

4.5 CBAPrep框架在其他模型类型场景下的可移植性分析

上述实验主要以分类任务为代表验证

了CBAPrep框架的效果。值得注意的是, CBAPrep框架的核心思想并不依赖于特定分类模型, 而是将数据准备策略抽象为成本建模、收益预测和预算约束优化问题。因此, 在可移植性方面, CBAPrep框架可以通过改变收益指标适配不同的模型场景。对于回归任务, 可将收益指标由准确率替换为MAE、RMSE或 R^2 ; 对于深度学习任务, 可将训练时间、显存消耗和数据增强成本纳入成本模型, 并以验证集性能或损失为收益指标。

在大模型训练与微调场景中, 数据准备过程通常包含数据去重、低质量样本过滤、指令数据构造、安全过滤、人类偏好标注和样本筛选等环节^[27-28]。这些环节均存在明显的成本收益权衡关系, CBAPrep框架可将Token消耗、图形处理器计算成本、人工标注及筛选成本纳入成本模型, 并以Benchmark得分、验证集损失、奖励模型评分或人工偏好胜率衡量收益。需要指出的是, 实际大模型场景下数据规模更大、收益评估周期更长、成本构成更为复杂, 因此, 构建适用于大模型数据准备场景的高精度、低开销的成本收益预测模型, 将是后续研究的关键方向之一。

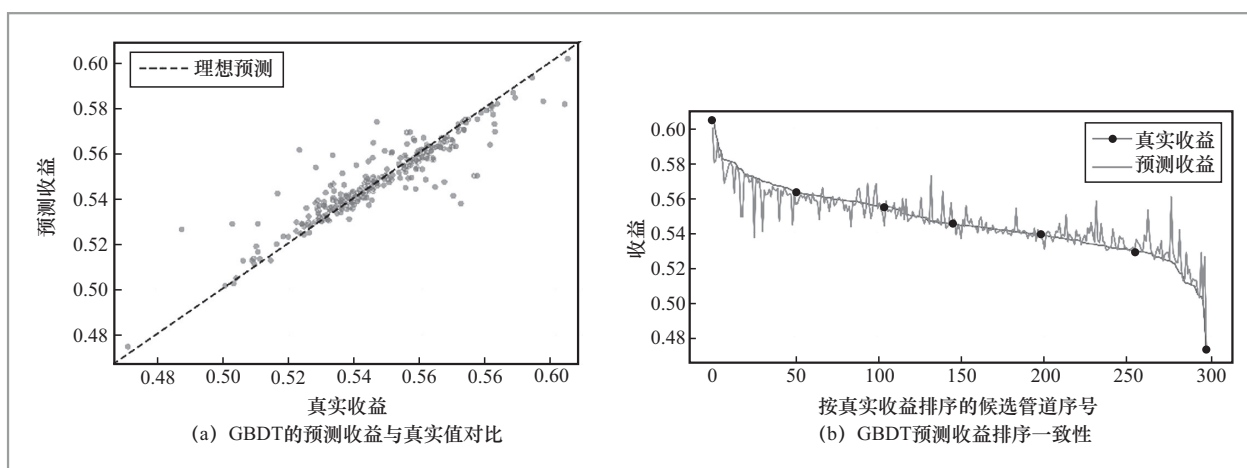


图6 GBDT模型的有效性验证结果

5 结束语

针对数据准备的成本不可控与资源利用效率不足问题,本文提出了一种面向成本收益优化的数据准备搜索框架CBAPrep。在方法层面,通过构建数据准备过程的成本收益管理框架,实现了从技术代价指标到现实预算量纲的映射,为后续优化提供了量化基础;在优化层面,提出了数据准备方案的资源分配优化方法,以线性复杂度实现近似最优的成本分配;在搜索层面,设计了方法组合与资源分配的分层联动机制,兼顾组合选择与资源优化,有效压缩组合空间,提升优化效率。在对整体数据准备管道运行时间预算约束的基础上,CBAPrep在多数数据集上的表现优于其他对比基准数据准备方法,展现了成本收益分析规划研究在数据准备领域的重要作用,为实现成本收益平衡的高质量数据准备提供了可行路径。

参考文献:

- [1] 刘艺,曹建军,翁年凤,等.数据治理及其发展研究综述[C]//第十二届中国指挥控制大会论文集(上册).北京:兵器工业出版社,2024: 170-176.
- Liu Y, Cao J J, Weng N F, et al. Research review on data governance and its development[C]//Proceedings of the 12th China Command and Control Conference(Volume I). Beijing: Military Science Press, 2024: 170-176.
- [2] 郝爽,李国良,冯建华,等.结构化数据清洗技术综述[J].清华大学学报(自然科学版), 2018, 58(12): 1037-1050.
- Hao S, Li G L, Feng J H, et al. Survey of structured data cleaning methods[J]. Journal of Tsinghua University (Science and Technology), 2018, 58(12): 1037-1050.
- [3] 黄科满,许多,杜小勇.数据社会化视角下的数据流通安全治理:五位一体框架[J].大数据, 2024, 10(6): 5-15.
- Huang K M, Xu D, Du X Y. Data circulation security governance from the perspective of data socialization: a five-sphere framework[J]. Big Data Research, 2024, 10(6): 5-15.
- [4] 范举,陈跃国,杜小勇.人在回路的数据准备技术研究进展[J].大数据, 2019, 5(6): 1-16.
- Fan J, Chen Y G, Du X Y. Progress on human-in-the-loop data preparation[J]. Big Data Research, 2019, 5(6): 1-16.
- [5] Berti-Equille L. Learn2Clean: optimizing the sequence of tasks for web data preparation[C]//Proceedings of the World Wide Web Conference. New York: ACM, 2019: 2580-2586.
- [6] Chen S B, Tang N, Fan J, et al. HAIPipe: combining human-generated and machine-generated pipelines for data preparation[J]. Proceedings of the ACM on Management of Data, 2023, 1(1): 1-26.
- [7] Li P, Chen Z Y, Chu X, et al. DiffPrep: differentiable data preprocessing pipeline search for learning over tabular data[J]. Proceedings of the ACM on Management of Data, 2023, 1(2): 1-26.
- [8] Gao H T, Cai S F, Dinh T T A, et al. CtxPipe: context-aware data preparation pipeline construction for machine learning[J]. Proceedings of the ACM on Management of Data, 2024, 2(6): 1-27.
- [9] Krishnan S, Franklin M J, Goldberg K, et al. BoostClean: automated error detection and repair for machine learning

- [PP]. arXiv (2017-11-03) [2026-04-30]. arXiv: 1711.01299.
- [10] Rekatsinas T, Chu X, Ilyas I F, et al. HoloClean: holistic data repairs with probabilistic inference[J]. Proceedings of the VLDB Endowment, 2017, 10(11): 1190-1201.
- [11] Ding X O, Qian Z K, Wang H Z, et al. UniClean: a multi-signal fusion pipeline for optimizing data cleaning workflow [C]//Proceedings of the 2025 IEEE 41st International Conference on Data Engineering (ICDE). Piscataway: IEEE Press, 2025: 4600-4603.
- [12] Krishnan S, Wang J N, Wu E, et al. ActiveClean: interactive data cleaning for statistical modeling[J]. Proceedings of the VLDB Endowment, 2016, 9(12): 948-959.
- [13] Zhang H J, Chai C L, Doan A H, et al. Manually detecting errors for data cleaning using adaptive crowdsourcing strategies[C]//Proceedings of the 23rd International Conference on Extending Database Technology (EDBT 2020). Konstanz: OpenProceedings.org, 2020: 311-322.
- [14] Yang J R, Fan J, Wei Z W, et al. A game-based framework for crowd-sourced data labeling[J]. The VLDB Journal, 2020, 29(6): 1311-1336.
- [15] Ding X O, Su H B, Qian Z K, et al. CBAClean: a comprehensive system for recommending data cleaning solutions through cost-benefit analysis in data quality management[C]//Proceedings of the 2025 IEEE 41st International Conference on Data Engineering (ICDE). Piscataway: IEEE Press, 2025: 4636-4639.
- [16] Rahm E, Do H H. Data cleaning: problems and current approaches[J]. IEEE Data Engineering Bulletin, 2000, 23(4): 3-13.
- [17] Ishaq M, Zahir S, Iftikhar L, et al. Machine learning based missing data imputation in categorical datasets[J]. IEEE Access, 2024, 12: 88332-88344.
- [18] Lee Y, Park S. High-dimensional missing data imputation via undirected graphical model[J]. Statistics and Computing, 2024, 34(5): 160.
- [19] Lin W C, Tsai C F. Missing value imputation: a review and analysis of the literature (2006-2017)[J]. Artificial Intelligence Review, 2020, 53(2): 1487-1509.
- [20] Guo C, Yang W, Liu C, et al. Iterative missing value imputation based on feature importance[J]. Knowledge and Information Systems, 2024, 66(10): 6387-6414.
- [21] Wu R Z, Chaba S Y, Sawlani S, et al. ZeroER: entity resolution using zero labeled examples[C]//Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data. New York: ACM, 2020: 1149-1164.
- [22] Qian Z K, Ding X O, Wang C, et al. DemandClean: a multi-objective learning framework for balancing model tolerance to data authenticity and diversity[J]. Proceedings of the VLDB Endowment, 2025, 18(12): 5339-5342.
- [23] Peng J F, Shen D R, Nie T Z, et al. RL-clean: an unsupervised integrated data cleaning framework based on deep reinforcement learning[J]. Information Sciences, 2024, 682: 121281.
- [24] 苏璇, 王远军. 面向高维不平衡医学数据的特征选择算法[J]. 小型微型计算机系统, 2024, 45(2): 309-318.
- Su X, Wang Y J. Feature selection algorithm for high-dimensional unbalanced medical data[J]. Journal of Chinese Com-

puter Systems, 2024, 45(2): 309–318.

[25] Li P, Rao X, Blase J, et al. CleanML: a study for evaluating the impact of data cleaning on ML classification tasks[C]//Proceedings of the 2021 IEEE 37th International Conference on Data Engineering (ICDE). Piscataway: IEEE Press, 2021: 13–24.

[26] Wang Q L, Guo Y F, Yu L X, et al. Deep Q-network-based feature selection for multisourced data cleaning[J]. IEEE Internet of Things Journal, 2021, 8(21): 16153–16164.

[27] Luo J Y, Wu B H, Luo X, et al. A survey on efficient large language model training: from data-centric perspectives [C]//Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). ACL, 2025: 30904–30920.

[28] Zha D C, Bhat Z P, Lai K H, et al. Data-centric artificial intelligence: a survey[J]. ACM Computing Surveys, 2025, 57(5): 1–42.

作者简介



崔文莹（2004–），女，哈尔滨工业大学计算学部硕士生，主要研究方向为数据清洗、数据治理。



丁小欧（1993–），女，哈尔滨工业大学计算学部副教授、博士生导师，主要研究方向为大数据治理、面向人工智能的数据质量管理。



程海峰（2003–），男，哈尔滨工业大学计算学部硕士生，主要研究方向为数据治理、数据准备。



钱泽凯（2001–），男，哈尔滨工业大学计算学部博士生，主要研究方向为数据质量管理、智能数据治理。



李勇（1986–），男，哈尔滨工业大学学校办公室副主任，主要研究方向为高教管理、信息化建设。



王宏志（1978–），男，哈尔滨工业大学计算学部教授、博士生导师，主要研究方向为数据库系统、大数据管理与分析、数据治理。



王晨（1981–），男，清华大学大数据系统软件国家工程实验室副研究员，主要研究方向为数据库、工业大数据、数据治理与数据标准化。

录用日期: 2026-07-24

通信作者: 丁小欧, dingxiaou_hit@163.com

基金项目: 国家科技重大专项 (No. 2025ZD1607102); 国家自然科学基金面上项目 (No. 62572151); 黑龙江省自然科学基金优秀青年项目 (No. YQ2024F005); 黑龙江省青年科技人才托举工程项目 (No. 2025QNTJ001); 中央高校基本科研业务费专项资金资助 (No. HITAI4SE2026035)

Foundation Items: National Science and Technology Major Project of China (No. 2025ZD1607102), The National Natural Science Foundation of China General Program (No. 62572151), Heilongjiang Provincial Natural Science Foundation Excellent Young Scholars Program (No. YQ2024F005), Heilongjiang Young Scientific and Technological Talent Support Program (No. 2025QNTJ001), The Fundamental Research Funds for the Central Universities (No. HITAI4SE2026035)