**摘要**现实数据往往同时含缺失、异常等多种错误，这种数据质量波动会放大无监督 AutoML的搜索空间并降低结果稳定性。本文提出一个核心假设：数据质量与聚类特征之间存在可预测的结构性，只要捕获足够的数据质量及算法特征，即可提前筛出高潜力的清洗–聚类组合。围绕该假设，我们构建了三条策略：离线阶段通过特征增强提高候选命中率；搜索阶段使用剪枝压缩空间至O(k log n)；在线阶段监控聚类统计量并动态修正参数漂移。实验部分按“影响分析→方法有效性→综合增益”的流程进行，在不同规模的数据集上验证了该方案比全量搜索提高 8 倍以上的运行速度，同时聚类质量平均损失不超过 2 %。研究结果表明，所提出的协同优化框架能够在大型脏污数据场景中兼顾速度与稳健性。

# **Real-world data often contains multiple types of errors, such as missing values and anomalies. Such fluctuations in data quality expand the search space of unsupervised AutoML and compromise the stability of the results. This paper proposes a core hypothesis: there is a predictable structure between data quality and clustering features. As long as sufficient data quality and algorithm features are captured, high-potential cleaning and clustering combinations can be screened out in advance. Based on this hypothesis, we developed three strategies: in the offline phase, feature enhancement is used to improve candidate hit rates; in the search phase, pruning is used to compress the space to O(k log n); and in the online phase, clustering statistics are monitored and parameter drift is dynamically corrected. The experimental section follows a process of “impact analysis → method effectiveness → comprehensive gain” and verifies that the proposed scheme improves runtime speed by more than 8 times compared to full search on datasets of different scales, while the average loss in clustering quality does not exceed 2%. The results demonstrate that the proposed collaborative optimization framework can achieve both speed and robustness in large-scale scenarios involving dirty data.**

# **1 引言**

在医疗精准诊疗、金融风险管理与工业物联网等数据密集型场景中，无监督聚类已成为发现潜在模式、辅助决策的重要工具。然而，聚类算法对数据分布的依赖极为敏感：一条拼写错误、一处未标记的异常或一段缺失区间，便可能扭曲距离结构和密度估计，从而打乱簇划分、收敛路径乃至下游分析链条。与依赖标签缓冲噪声的有监督学习不同，聚类几乎完全暴露在原始数据质量之下。一个直观的解决思路是借助 AutoML将若干数据清洗算子与多种聚类算法串联，交给优化器在组合空间中自动择优。然而，随着清洗算子、聚类模型与超参数维度的指数级增长，优化器往往在大量注定失败的方案上浪费算力，从而削弱了自动化本身的价值。

过去十年，数据清洗与聚类算法各自取得诸多突破：针对多源异构数据，缺失值插补、异常检测、容错匹配等数据清洗方法层出不穷；而 K-Means、DBSCAN、层次聚类乃至深度嵌入聚类，在复杂数据形态下展现出较好的可塑性。然而，**两条技术路线几乎是独立推进的**：清洗研究聚焦局部错误修复，很少衡量修复对无监督任务的连锁效应；聚类研究默许输入“足够干净”，将主要精力投入相似度度量和优化策略。当数据体量与脏污类型同步增长时，这种割裂带来了三重放大的复杂性：清洗会改变距离分布，致使既有超参数失配；跨表、跨批次的数据质量高度异质，单一策略难以通用；指数级候选组合更令系统开销飙升，无法通过增加算力线性抵消。**业界已尝试用 AutoML 串联清洗与聚类**，但在缺乏先验认知的情况下，这一做法依旧进入“盲目试错”的陷阱，投入与收益常常不成比例。

基于这一现实困境，本文提出一个经实证支持的核心假设：**数据质量对聚类表现的影响具有可学习的结构性**。换言之，只要对数据脏污形态和算法特征刻画得足够细致，就可以在耗时的评估之前预测出清洗-聚类管线的优劣。该假设带来两条直接启示：其一，构建映射 Φ，将数据质量特征向量映射为候选管线的优先级；其二，若 Φ 预测可靠，便可在搜索开始时大胆裁剪，仅保留排名靠前的少数候选，把计算资源集中在最具潜力的少数方案上。

围绕这一假设，我们设计了兼顾速度、稳健与可解释性的三阶段框架。首先在**特征增强阶段**，在传统统计量之外引入高阶交互特征与算子，利用历史实验数据训练梯度提升排序模型实现 Φ，生成候选管线排序。随后于**置信上界剪枝阶段**，为每条候选管线计算可能达到的聚类分数上界，若该上界低于当前最优值便立即淘汰；理论证明，保留下来的候选规模上界为O(k\*logn)，其中 k为真正高质量方案数，n为原始空间大小。最后在运行阶段引入**动态调优**，实时监控轮廓系数斜率与 Davies–Bouldin 指数，一旦侦测到分布漂移即自动微调 k、ε 等关键超参，无需重启搜索即可恢复稳健。

**本文的主要贡献包括：**

（1）我们系统性地提出并验证了数据清洗粒度与聚类效果存在结构性关联的假设。基于 8 类清洗器、6 类聚类器和 60 套公开数据集的实验，我们揭示了单调分段、阈值拐点和交互作用等四层规律，为自动化算法提供了定量基础。

（2）我们提出了置信上界驱动的搜索裁剪算法，设计了特征增强的多标签预测器，并结合置信上界策略，对无监督 AutoML 的搜索树进行剪枝；理论上将搜索空间上界降低至 O(k\*logn)，实验上在百万行数据集上实现 6–12 × 加速、精度损失 < 2 %。

（3）我们提出运行时参数漂移监控与自适应回滚策略，可在检测到 SSE 下降斜率或 DB 指数漂移时自动调整 k, ε, γ等关键超参；在 1 M 行真实/模拟数据上将聚类稳定性波动控制在 ±1 % 以内，并保持 8 × 整体加速。

本文结构如下：第 2 节回顾数据清洗、聚类与 AutoML 交叉领域的最新进展；第 3 节正式形式化定义数据质量向量、目标函数以及评价指标；第 4 节详细阐述所提出的三阶段框架，并给出复杂度与理论分析；第 5 节介绍实验设置，包括数据集、脏污注入策略和完整实验设计；第 6 节按这三条主线呈现实验结果并深入讨论；第 7 节总结并提出结论与展望；附录则提供额外图表及开源代码链接。

In data-intensive scenarios, such as precision medicine, financial risk management, and industrial IoT, unsupervised clustering has become a crucial tool for discovering potential patterns and supporting informed decision-making. However, clustering algorithms are susceptible to data distribution: a single spelling error, an unlabeled anomaly, or a missing interval can distort the distance structure and density estimates, thereby disrupting cluster partitioning, convergence paths, and even downstream analysis chains. Unlike supervised learning, which relies on labels to filter out noise, clustering is almost entirely dependent on the quality of the raw data. An intuitive solution is to leverage AutoML to chain together several data cleaning operators with various clustering algorithms, then let an optimizer automatically select the best combination in the parameter space. However, as the number of cleaning operators, clustering models, and hyperparameter dimensions grows exponentially, optimizers often waste computational resources on a large number of doomed-to-fail solutions, thereby undermining the value of automation itself.

Over the past decade, data cleaning and clustering algorithms have each made significant breakthroughs: for multi-source heterogeneous data, data cleaning methods such as missing value imputation, anomaly detection, and fault-tolerant matching have emerged in rapid succession; while K-Means, DBSCAN, hierarchical clustering, and even deep embedding clustering have demonstrated good adaptability in complex data structures. However, **the two technical approaches have developed almost independently**: data cleaning research has focused on local error correction, with little consideration of the cascading effects of corrections on unsupervised tasks; clustering research has assumed that the input is “sufficiently clean,” and has primarily focused on similarity measurement and optimization strategies. As data volume and data quality issues grow in tandem, this disconnect leads to threefold amplified complexity: data cleaning alters distance distributions, causing existing hyperparameters to become mismatched; data quality varies greatly across tables and batches, making it challenging to apply a single strategy; and the exponential number of candidate combinations causes system overhead to skyrocket, which cannot be offset linearly by increasing computing power. **The industry has attempted to utilize AutoML to link cleaning and clustering**; however, in the absence of prior knowledge, this approach still falls into the trap of “blind trial and error,” with input and output often being disproportionate.

Given this practical challenge, this paper proposes a core hypothesis supported by empirical evidence: **the impact of data quality on clustering performance has a learnable structure**. In other words, as long as the characteristics of data contamination and algorithm features are described in sufficient detail, the quality of the cleaning-clustering pipeline can be predicted before time-consuming evaluation. This hypothesis leads to two direct implications: first, construct a mapping Φ that maps data quality feature vectors to the priority of candidate pipelines; second, if Φ is a reliable predictor, we can boldly prune the pipeline at the beginning of the search, retaining only the top-ranked candidates and concentrating computational resources on the most promising few solutions.

Based on this hypothesis, we developed a three-stage framework that strikes a balance between speed, robustness, and interpretability. First, in the **feature enhancement stage**, we introduce higher-order interaction features and operators in addition to traditional statistics, and use historical experimental data to train a gradient boosting ranking model to implement Φ and generate a ranking of candidate pipelines. Next, in the **confidence upper bound pruning stage**, we calculate the upper bound of the cluster score that each candidate pipeline can achieve. If this upper bound is lower than the current optimal value, the candidate is immediately eliminated. Theoretical proof shows that the upper bound of the retained candidate size is O(k log n), where k is the number of truly high-quality solutions and n is the size of the original search space. Finally, in the runtime phase, we introduce **dynamic tuning**, which monitors the contour coefficient slope and Davies–Bouldin index in real time. Once distribution drift is detected, key hyperparameters, such as k and ε, are automatically fine-tuned to restore robustness without requiring a restart of the search.

**The main contributions of this paper are as follows:**

(1) We systematically propose and validate the hypothesis that there is a structural correlation between the granularity of data cleaning and the effectiveness of clustering. Based on experiments with eight types of cleaners, six types of clusterers, and 60 public datasets, we reveal four layers of patterns, including monotonic segmentation, threshold inflection points, and interactions, providing a quantitative basis for the development of automated algorithms.

(2) We propose a confidence upper bound-driven search pruning algorithm, design a feature-enhanced multi-label predictor, and combine it with a confidence upper bound strategy to prune the search tree for unsupervised AutoML. Theoretically, the upper bound of the search space is reduced to O(k\*logn), and experimentally, we achieve 6–12 × acceleration and accuracy loss < 2% on a million-row dataset.

(3) We propose a runtime parameter drift monitoring and adaptive rollback strategy that automatically adjusts key hyperparameters, such as k, ε, and γ, when a decline in SSE slope or drift in DB index is detected. On 1 million lines of real and simulated data, we control the clustering stability fluctuation within ±1% while maintaining an overall acceleration of 8×.

The structure of this paper is as follows: Section 2 reviews recent advances in the cross-domain of data cleaning, clustering, and AutoML; Section 3 formally defines the data quality vector, objective function, and evaluation metrics; Section 4 details the proposed three-stage framework and provides complexity and theoretical analysis; Section 5 introduces the experimental setup, including the dataset, data corruption injection strategy, and complete experimental design; Section 6 presents experimental results along these three main lines and conducts in-depth discussions; Section 7 summarizes the findings and provides conclusions and future directions; the appendix provides additional figures and open-source code links.

# **2 相关工作**

本章从数据清洗与质量管理、聚类算法在脏数据环境中的鲁棒性、无监督 AutoML 的最新进展三条主线回顾相关研究，并汇总现存局限，为下一章的问题定义与方法奠定背景。

数据清洗旨在检测并修复缺失、离群、重复及格式错误等多种缺陷，是提高分析可用性的前置环节。早期方法侧重统计填补（均值、众数）或规则驱动的异常检测；随后，概率图模型、主动学习、神经网络等技术显著提升了复杂错误的识别能力。针对领域特定数据，还出现了结合知识图谱或外部约束的定制清洗框架。然而，无标签的无监督任务缺乏“干净对照”，过度清洗易将少量关键异常误删，而保守清洗又放大噪声干扰。已有工作多在“数据质量” 维度上止步，未能将清洗决策与下游聚类性能直接关联，这成为本文进一步构建“清洗–聚类–搜索”联合模型的动机之一。

对于三类主流的聚类算法，质心类的K-Means及其改进方法在大规模数据上高效，但对离群点和几何结构敏感；密度类的DBSCAN、OPTICS 能识别任意形状簇，对低密度噪声更鲁棒，却高度依赖超参数 ε、minPts；自底向上的层次聚类适合捕捉多尺度结构，但计算复杂度较高。针对脏数据，研究者提出加权距离、鲁棒质心更新等局部修补策略，或将局部密度重估引入 DBSCAN 以缓解误判。但这些方法大多分别优化“清洗”或“聚类”某一端，缺乏整体视角，不同清洗操作如何改变簇形状、参数敏感度，以及算法内部的迭代路径，仍缺少系统量化。

AutoML 框架（Auto-sklearn、TPOT 等）在有监督任务中已实现“模型–特征–超参”的自动搜索。无监督方向的代表性研究主要完成聚类算法与超参数的自动推荐，却很少包含数据清洗或仅将 PCA 作为降噪手段。近期工作尝试将清洗算子纳入搜索空间，但多停留在经验规则层面，尚未给出统一的端到端闭环建模。

针对上述不足，本文将利用数据与算法特征预测最优组合并缩减搜索空间，同时定量描述清洗精度对聚类过程和结果的影响，进而提出端到端统一优化的“清洗 × 聚类 × 超参”方法框架。

# **3 问题定义**

## **3.1符号约定与数据模型**

**定义 1 (数据矩阵).** 设 是原始数据矩阵——其中 为样本数， 为特征数。 我们将列索引集 依类型划分为 （数值、类别与文本列），行、列均可能伴随缺失、离群或格式错误。

**定义 2 (清洗算子集合).** 为了描述可用的数据清洗手段，记 为候选算子集合。本文实验涉及多种经典数据清洗方法：缺失值插补、离群检测、格式校正、语义一致性修复、噪声抑制等，每一算子 伴随其超参数向量 。

**定义 3 (聚类算法集合).** 对应地，记 表示聚类算法集合。本研究关注三大类方法：质心类 (*K-Means*)、 密度类 (*DBSCAN*) 与层次类 (*Hierarchical*)，每个 伴随超参数向量 （例如簇数 、半径 、最小密度 ）。

**定义 4 (候选管线).** 将任意清洗子集 、聚类器 及其联合超参

相组合，即得到一条候选清洗–聚类管线

## **3.2 问题陈述**

**定义 5 (数据质量特征).** 为了定量描述 的“脏”程度，本工作聚焦两种主要缺陷：**缺失值**与**离群值。**我们引入二维归一化向量 分别对应缺失与离群的样本占整体的比例。向量 连同其二阶交互特征将在第4节共同作为学习模型的输入。其计算细节在附录 A给出。

**定义 6 (算法特征).** 与数据质量特征相对应，我们将清洗–聚类管线本身抽象为一组可学习特征 其中 为多热编码的算子组合及其超参实值， 描述所选聚类器类型与超参。向量 的维度 随候选管线而异，但对任何 唯一确定。

**定义 7 (搜索空间).**给定定义 [2](#def:cleaners)–[4](#def:pipeline)中的清洗集合 、聚类集合 以及对应超参域，候选管线全集记作 ，其规模满足 通常呈指数级增长。经验上，真正性能优异的管线数量远小于 。

**定义 8 (综合指标).** 为了从不同的角度衡量聚类效果，我们使用加权指标 其中 ； 代价项 用归一化运行时间度量。二者线性组合得到综合指标 决定速度–质量权衡。预实验（§[3.3](#sec:preexp)）表明 在多数数据集上能达到良好平衡。

**假设 1 (置信映射).** 我们假设存在函数 使得 在训练集上成立，由预实验选定。该映射由梯度提升排序器实现，并将在算法阶段提供“置信上界”以剪枝大规模搜索树。

**定义 9 (优化目标).**记全局最优质量为 若存在子集 使 ，则称 为 *-*充分子集，其中为用户可容忍误差，为与搜索树深度相关的常数， 为与实现无关的绝对常数。在该子集中 被称为 *-*近似最优管线并作为搜索结果输出。

式[[eq:qvec]](#eq:qvec)–[[eq:objective]](#eq:objective) 与上述优化目标共同把 “**数据质量算法特征综合效能预算约束**” 四要素紧密耦合。下一章将据此设计三阶段协同优化算法，并证明其在评估预算内可实现 -近似最优。

## **3.3 预实验与参数确定**

正式的大规模实验之前，我们先在小规模数据上完成一次预实验（pilot study），以确定目标函数权衡系数 、预测映射容忍误差 以及运行时漂移阈值 ，从而减少后续调参成本并保证结果对比的可重复性。

我们选取五个公开数据集 （涵盖数值型/混合型特征和不同规模），并对每个数据集按照四级脏污强度 注入缺失、离群等噪声。在每一 组合上，对参数网格

逐点评估综合目标 （式 [[eq:objective]](#eq:objective)）， 其中 “” 表示从 到 以步长 均匀采样。 为控制计算预算， 每个网格点随机采样至多 条候选管线执行完整评估。

# **4 方法**

本章面向第三章给出的符号约定与优化目标，提出一套由“特征增强—搜索裁剪—动态调优—复杂度分析”四个环节串联而成的协同优化框架。

## **4.1 特征增强**

如问题定义中所述，我们假设二维数据质量向量 与算法特征 具有某种结构性的匹配关系；若不显式建模该耦合，则学习器只能被动地从离散样本中推断模式，难以在搜索早期给出可靠排序。为此，本节提出离线构造联合表示并训练置信预测器的流程，其要点包括二阶交互扩展、排序目标与不确定度量化。

#### **算法 1：特征增强与置信学习**

Input : 历史评估集 𝔇 = { (D\_i, π\_i, f\_i) }\_{i=1..M }  
Output: 置信预测器 Φ(·) = ( f̂(·), σ(·) )  
  
1 建立样本循环:  
2 取 (D\_i, π\_i, f\_i) ∈ 𝔇  
3 计算数据质量向量 q\_i ← [q\_miss, q\_out]⊤  
4 提取算法特征 a\_i ← [a\_clean ; a\_cluster]  
5 构造联合特征 z\_i ← [ q\_i ; a\_i ; q\_i ⊗ a\_i ]  
6 end  
  
7 用 { (z\_i , f\_i) } 训练 LightGBM Ranker，得到 f̂(·)  
8 拟合 Quantile Random Forest 于残差，输出 σ(·)  
9 返回 Φ(·) = (f̂(·), σ(·))

该离线阶段一次性完成 模式的抽取与不确定度刻画；后续各算法均可直接调用 的点估计 和置信区间服务。

## **4.2 搜索裁剪**

当训练出预测器 ，我们采用“上置信界 + 层内 Top ” 的搜索原则，逐层扩展管线树并在置信不足时立即截枝。这样既兼顾了探索（置信区间的宽度）又兼顾了利用（当前最优得分），从而快速锁定高质量候选。

#### **算法 2：Confidence–UCB 搜索**

Input : 搜索空间 Π, 预测器 Φ, 参数 c, ε, k  
Output: 评估集合 S, 近似最优管线 π̂  
  
1 初始化 open ← Π , best ← -∞ , S ← ∅  
2 while open ≠ ∅ do  
3 选节点 π⋆ ← argmax\_{π∈open} [ f̂(π)+c·σ(π) ]  
4 if f̂(π⋆)+c·σ(π⋆) < best - ε then  
5 open ← open \ {π⋆} # 剪枝整棵子树  
6 else  
7 真实评估 f(π⋆) ，S ← S ∪ {π⋆}  
8 if f(π⋆) > best then best ← f(π⋆), π̂ ← π⋆  
9 生成 π⋆ 的子节点 U ，取 Top-k 入队：open ← open ∪ U  
10 end if  
11 end while  
12 返回 S, π̂

凭借“乐观估计 + 子树整体淘汰”的机制，该算法在每层至多触发 次真实评估；结合层数 ，完整评估量被严格限制，为后续在线阶段留出充足资源。

## **4.3 动态调优**

离线搜索只能保证在训练分布上的最优性；当线上数据分布与系统负载发生漂移时，需及时修正管线以维持性能。本节通过“监控–触发–微调”三步走策略，实现轻量且自适应的运行期优化。

#### **算法 3：窗口监控与局部微调**

Input : 初始管线 π̂, 监控窗口 w, 阈值 τ\_flat, τ\_drift, η  
Output: 实时最佳管线 π⋆\_online  
  
1 π⋆\_online ← π̂  
2 for 每批新数据 do  
3 计算当前指标 Sil\_t, DB\_t  
4 更新滑窗序列 {Sil}, {DB}  
5 if |∇ Sil\_t| < τ\_flat AND (DB\_t - DB\_{t-w}) > τ\_drift then  
6 # --- 触发局部微调器 ---  
7 构造邻域 Θ\_local ← {k±1, ε×[0.8,1.2], minPts±1}  
8 for θ' ∈ Θ\_local (子采样 10%) do  
9 评估 f'(θ')  
10 if f'(θ') - f(π⋆\_online) > η then  
11 更新 π⋆\_online ← π⋆\_online(θ')  
12 若清洗策略变动则增量刷新 q , 重新取 f̂  
13 break  
14 end for  
15 end if  
16 end for  
17 返回 π⋆\_online

该在线模块仅在指标显著停滞且劣化时才激活，并通过“邻域试探 + 成果阈值”机制避免频繁抖动；若清洗决策被改动，还会同步更新 并借助 重新评估，以免偏离最初的置信框架。

#### **4.4 复杂度与理论上界**

本节从时间与空间两个角度，对前三小节的核心流程做逐项推导，并给出整体上界与近似最优保证。

#### **(1) 离线特征增强**

LightGBM 每次分裂需遍历全部样本与候选分裂点；在 样本、 维输入下，典型实现的训练代价 该阶段只在模型更新时离线执行，且现代分布式实现可借助数据分片并行处理，故对整体延迟影响可忽略。

#### **(2) 置信搜索裁剪**

在 Top-k 展开策略下，完全 叉树深度满足 对任意层，算法 2 先按 选择节点：若首个节点被剪枝，则同层节点的上界必然更低，整层即刻终止；若未剪枝，则最多再有 个节点能使 best 刷新，否则其 UCB 亦跌至 best，无法继续保留。故 若最优管线祖先被剪枝，则 ， 与 矛盾； 故 必被保留并最终评估，且 。由此可得总耗时与空间： 在典型配置 、 下， ，相较于暴力评估 条管线节省约 倍时间与同阶内存。

#### **(3) 在线动态调优**

仅在“性能停滞且劣化”时触发：设全年共激活 次，每次在局部邻域 评估 组参数（10% 子样本），则 因为 且 通常为数百量级，此项对总体延迟贡献极低。

#### **(4) 端到端汇总**

三项相加得到

在离线模型稳定、在线调优罕见的现实场景下，第二项（置信搜索）几乎占据全部成本；而它相较暴力搜索仍保持对数级优势。