数据清洗与下游聚类的自动化协同优化及其影响机理研究

2025 年 5 月 4 日

摘要

在许多无监督学习任务中，现有的数据清洗与聚类技术已能在一定程度上降低噪声与缺失值带来的影响，但仍难以同时兼顾多样化清洗策略与聚类算法在大规模、高维数据中的协同需求。为进一步提升聚类质量与自动化效率，本文提出了一种清洗-聚类协同优化框架：通过多标签学习模型将数据的特征向量映射到最优或近优的“清洗-聚类”管线组合，从而在大幅减少搜索空间的同时确保较高的聚类性能。在提供解决方案的同时，我们还系统地分析了清洗对聚类算法运行及评价指标的影响，并通过清洗准确度与聚类指标的关联性研究，揭示了在不同数据特征与错误率下清洗策略对聚类收益的关键影响因素。基于对 60 个公开数据集的大规模实验发现，不同数据特征会显著影响清洗与聚类的适配性，例如 Raha-Baran + HC 组合在高维、多特征数据上较为稳健，而 mode + DBSCAN 在低维数值数据上对噪声表现出极端敏感。通过该框架的自动化推荐与筛选，在部分场景下实现了平均 5.83 倍的搜索加速，并在保证聚类质量的同时取得了 19.20% 的平均提升率。研究结果表明，该方法对多样化数据具有一定的稳健性与可扩展性，为噪声较高、规模较大的真实数据环境提供了切实可行的聚类优化方案。

1 引言

在大数据与人工智能的快速发展背景下，无监督学习（如聚类分析）在医疗、金融及工业物联网等众多领域发挥着日益重要的作用 [1–3]。例如，在医疗场景中，通过聚类可从病患数据中挖掘潜在群组，为个性化诊疗提供决策支持 [4]；在金融场景中，聚类方法可以帮助区分用户信用类别、增强客户预期回报的信心等 [5]。已有研究在聚类算法改进和可视化等方面取得了显著成果，常用的方法包括 K-Means 及其变体 [6]、基于密度的 DBSCAN 以及层次聚类、图聚类 [7] 等，这些方法为不同数据形态提供了有效的划分策略。与此同时，数据清洗技术（例如缺失值填补、异常值检测、错误值纠正）也在学术界与工业界获得广泛应用，用以降低噪声影响和提高数据质量 [8–10]。

然而，在无监督学习场景中，数据质量的影响往往更为突出。与分类或回归等有监督学习相比，聚类对于数据分布的依赖更强，一旦噪声、缺失值或错误值的比例较高，就可能破坏簇结构与真实分布之间的对应关系 [11]，从而对模式挖掘和决策支持造成不可忽视的干扰。虽然已有清洗方法在减少噪声方面效果显著，但过度或不当的清洗有时反而扭曲了关键特征 [12]；此外，不同的聚类算法对数据缺陷的敏感度各异 [13]，若仅侧重于数据清洗或聚类算法单方面的优化，往往难以协调两者之间的相互作用，从而难以获得整体最优的策略。

为解决上述问题，研究者逐步认识到“清洗策略 + 聚类算法 + 超参数”一体化管线的重要性 [14]。这种做法能在保证数据分布尽量真实的同时，为不同数据集的特性“量体裁衣”地提供最佳策略。但由于管线搜索空间常呈指数级增长，且无监督场景缺乏显式标签指导，仅靠人工穷举或简单试验往往难以在可接受时间内完成参数寻优。近年来，自动化机器学习（AutoML）在有监督学习领域已呈现出显著优势 [15]，不仅能自动选择模型结构及超参数，还能优化特征工程 [16,17]。然而，大部分 AutoML 研究集中于分类或回归任务 [18]，对无监督学习特别是“清洗 + 聚类”协同自动化的探索仍相对有限 [19]。这为我们带来了新的机遇与挑战：能否将数据质量与无监督聚类的协同优化思路融入 AutoML 框架，并结合更深层次的机理剖析，在大规模及多场景下实现高效且可解释的自动管线搜索。

在此过程中，深度理解清洗操作如何影响下游聚类算法至关重要：只有梳理清楚清洗对聚类影响的机制和环节，才能在自动化管线中有针对性地选择或组合清洗策略与聚类方法。为此，我们将清洗的影响拆分为四个关键层面：**(1)** 清洗对数据集准确度的影响（具体表现在对错误的修复程度），**(2)** 清洗对聚类算法内部过程的干预（如迭代收敛路径、核心点判定），**(3)** 清洗对聚类结果指标（如 Silhouette、DB 指数）的影响，**(4)** 清洗对聚类超参数选择（如 K-Means 的 k 值、DBSCAN 的 ε）的影响。通过从“源数据集—算法运行过程—聚类指标—超参数调优”四步逐层深入地分析，我们不仅能更好地解释清洗策略与聚类性能之间可能存在的因果关联，也为自动化管线的配置与调优提供了更丰富的理论支撑。

基于上述背景与需求，本文针对“数据清洗与下游聚类协同优化”这一交叉方向，提出了一种新的自动化管线模型，并进一步从理论层面对“清洗操作如何影响聚类”展开深度剖析：一方面，借助多标签学习模型将多种清洗策略、聚类算法及其超参数统一纳入搜索空间，在离线阶段学习“数据特征到优选方案”的映射关系；另一方面，系统研究清洗准确度与聚类评价指标之间的关联，为理解清洗操作如何干预聚类运行过程及结果指标提供实证依据。这样，当面对新的数据集时，系统能快速推荐若干最优或近优组合，大幅缩减搜索规模，并根据清洗机理的分析获得更全面的可解释性与稳定性。

本研究的主要贡献包括：

1. **系统性地评估“清洗策略 × 聚类算法”组合的协同表现**

   基于 60 个具备多元质量问题的公开数据集，我们深入研究了不同噪声水平、错误率及规模下的 8 种清洗策略和 6 种聚类算法，对其组合在聚类质量、极端案例和时间开销方面进行了量化与比较。该评估不仅提供了对现有清洗-聚类方案适配度的系统认识，也为后续管线设计提供了实用参考。

2. **提出基于管线思维的协同优化框架**

   将“数据清洗 + 聚类 + 超参数”作为一个整体管线（Pipeline），并结合实验结果总结出多种针对性建议，帮助研究者在实际场景中有的放矢地进行策略选择，避免仅在单一端的优化而忽略全局效果。

3. **构建并验证了一个完整的自动化管线优化模型**

   我们引入多标签学习来捕捉“数据特征与最优清洗-聚类组合”之间的关系，大幅减少了管线搜索空间。在多个数据集上验证表明，自动化模型通常可获得 3× 以上的加速，同时保持较高的聚类准确度，证明了将 AutoML 思路拓展到无监督学习领域的可行性与有效性。

4. **从理论角度剖析清洗对聚类的影响机理**

   通过对清洗准确度（EDR、Precision、Recall、F1）与聚类评价指标（Silhouette、DB、Combined Score）的关联研究，系统揭示了不同数据特征和错误率水平下清洗操作如何影响聚类过程，从而为参数调优与动态适配提供了更深层次的参考。

5. **为多样化数据场景的聚类优化提供可迁移路径**

   通过对损失率、加速比等指标的度量，我们量化了自动化管线在平衡质量与效率方面的潜力，为工业领域部署该思路奠定了实践基础，也为研究者进一步探索清洗与聚类协同优化的动态适配、在线更新等提供了方向。

6. **按错误类型细粒度量化清洗对聚类的收益，并将该洞察用于改进 AutoML 搜索策略**

   本文首次在大规模实验中，将缺失、格式错误、离群值等不同错误类型对聚类流程的影响拆解分析；并把这些类型-级指标注入多标签学习模型，以动态收缩搜索空间，从而进一步提升 AutoML 推荐的精度与效率。

我们的工作不仅加深了无监督场景下“数据清洗—聚类”协同机理的理解，也为自动化机器学习（AutoML）在脏数据环境中的落地探索了新路径。全文结构如下：第 2 章回顾相关工作；第 3 章形式化地定义问题与符号；第 4 章给出清洗与聚类协同的实现框架；第 5 章以 40 个数据集的大规模实验归纳宏观现象；第 6 章则进一步结合各清洗算法原理，对过程指标与错误类型做细粒度机理分析，并将结论嵌入 AutoML 搜索空间以完成强化验证。

2 相关工作

为了更深入理解“清洗策略与聚类算法协同优化”在不同场景下的研究现状，本文从以下三个方面回顾相关工作：首先，探讨数据清洗与数据质量管理的相关方法；其次，分析主要聚类算法的原理及其优化思路；最后，梳理自动化机器学习（AutoML）在无监督学习场景中的研究进展与应用探索。

2.1 数据清洗与数据质量管理

数据清洗旨在识别并修复各种数据缺陷（如缺失值、噪声、重复记录或错误值），是提升数据整体质量的重要途径，已有研究在统计方法和机器学习方法方面均取得了丰富成果。例如，早期工作主要依赖众数/均值填补 [20] 或规则驱动的异常值检测 [21,22]，在处理缺失值和简单错误时比较高效；后续研究则引入高级方法，如概率图模型 [23]，主动学习 [24,25]，

方法	是否包含数据清洗	是否端到端 AutoML 模型
AutoClust [46]	× 无	× 仅聚类优化
cSmartML [47]	× 无	× 仅算法选择 + 超参数优化
MARCO-GE [48]	△ 部分（PCA）	× 主要关注算法推荐
AutoCluster [49]	× 无	△ 部分端到端（集成学习）
TPE-AutoClust [50]	△ 部分（初步聚类）	△ 部分端到端（优化 + 集成）
本文方法	✓ 完整数据清洗	✓ 端到端自动化管线优化

表 1: 当前无监督 AutoML 聚类主要方法对比

神经网络 [26] 等，以应对更复杂的错误类型。部分工作还引入了上下文约束或知识图谱 [27,28]，对特定领域（如医疗、经济数据）的不一致或罕见值进行更有针对性的纠正。

与此同时，研究者也认识到过度或不当清洗可能使原本有价值的异常点被误删或被扭曲 [12]。在有监督学习场景下，数据清洗常可借助标签对比来区分“真正有意义的异常”与“噪声性错误”[29]；然而在无监督场景中，缺乏标签指导，清洗策略一旦过于保守或激进，就会对后续的聚类分析产生不可预测的影响。这些研究进展表明，数据清洗方法的选择与配置应当与下游分析任务（如聚类）紧密结合，而非单独孤立地追求“最干净”的数据 [30]。这也为我们随后探讨的“清洗与聚类协同优化”提供了重要动机。

2.2 聚类算法及其改进

聚类作为典型的无监督学习方法，已在图像识别、文本挖掘、用户分群等领域中得到了广泛应用。现有聚类算法大体可分为基于质心（如 K-Means 及其变体 [6,31,32]）、基于密度（如 DBSCAN [33,34]，OPTICS [35–37]）与层次聚类 [38] 三类。不同算法在簇形状、噪声耐受度、计算复杂度等方面各具优势 [13]。

在面对不完美数据时，上述聚类算法往往对异常值和缺失值表现出不同的敏感度。例如，少量异常点被 K-Means 视为远离中心的“噪声”，可在重新计算均值时抵消 [39]；但若这些点在 DBSCAN 的邻域定义中被错误识别，就可能导致过度分割 [40]。部分工作试图在算法内部引入鲁棒性机制，如改进距离度量或引入加权方案 [41]，但大多仍需事先对数据进行相对独立的预处理，缺乏将“清洗策略”与“聚类算法”放在同一管线中统筹考量，在更复杂的高噪声场景中难以取得较好的聚类结果。

2.3 AutoML 与无监督场景的探索

近年来，AutoML 框架（如 Auto-sklearn [42]、TPOT [43,44]、H2O AutoML）已在有监督学习任务（分类、回归）中展现出卓越的自动化建模与超参数优化能力。典型方法主要依赖贝叶斯优化、遗传算法 [45] 等技术，在预定义的搜索空间内高效探索最优模型配置。然而，这些框架主要针对有监督任务设计，难以直接适用于无监督学习，尤其在聚类任务中面临诸多挑战 [19]。在少量试图探索 AutoML 在无监督学习上应用的研究中（表 1），其方法主要聚焦于聚类算法选择与超参数优化，部分工作结合初步聚类或 PCA 降维以降低特征噪声。

然而，现有研究在“清洗-聚类-AutoML”闭环上仍**缺少系统化建模与量化**：

- 多聚焦于聚类算法推荐 + 超参数搜索，而数据特征如何驱动最优清洗-聚类组合缺乏定量分析；
- 评估侧重最终 Silhouette / DB 等结果指标，而忽略“清洗 → 聚类内部过程”（质心收敛、核心/边界判定等）的干预链路；
- 尚无完整的端到端整合清洗、聚类与调参的无监督 AutoML 框架，难以应对高噪声、多特征、错误类型交叠的数据场景。

2.4 小结与差异

数据质量管理与聚类算法各自已形成成熟体系，AutoML 在有监督学习中也愈趋完善；**但在无监督情形仍存在以下不足**：(i) 缺乏可端到端联动“清洗 + 聚类 + 超参数”的自动化框架；(ii) 现有无监督 AutoML 未将清洗准确度及错误类型特征纳入搜索维度；(iii) 清洗-聚类在多错误类型叠加场景下的协同机理尚无系统量化。

为了改进上述不足，本文提出基于多标签学习的管线化 *AutoML* 框架，核心思路如下：

1. **统一搜索空间**——将“数据清洗策略 × 聚类算法 × 超参数”整体建模，通过离线多标签学习，学习“数据特征 → 优选组合”映射，显著裁剪候选子空间；

2. **过程级记录**——在线阶段细粒度追踪错误类型级清洗准确度（缺失、离群、格式错误等）与聚类运行轨迹（迭代步数、质心位移、核心点变化等），实现清洗收益的因果量化；
3. **原理驱动反馈**——结合各清洗算法原理，分析其在不同错误模式下对聚类的优势与局限，并将所得启示动态反馈至 AutoML 搜索策略，从而持续优化搜索效率与聚类质量。

该框架为回答“清洗何时、如何、在多大程度上提升下游聚类”提供了系统化解决方案，也为无监督 AutoML 的落地实践给出可复制的改进路径。

3 问题定义与挑战

在第 2 节回顾了数据清洗、聚类算法及 AutoML 的研究进展后，我们发现：尽管各自领域已有丰富成果，但在“高维、多源、噪声与缺失并存”的真实场景中，数据清洗与聚类执行并非简单串联——二者在数据分布、算法收敛路径及超参数选择上存在深度耦合，直接影响最终聚类效果。为实现真正的端到端自动化优化，我们需首先构建一个统一的数学模型，将“清洗操作 $\rightarrow$ 数据分布变化 $\rightarrow$ 聚类内部过程 $\rightarrow$ 最终评价”作为一个整体加以刻画。

3.1 数学模型与形式化定义

为在理论与应用中更好地理解并解决“数据清洗与聚类算法”的协同优化，本小节对核心概念进行形式化定义，并建立相应的评价体系。

3.1.1 核心概念与变量定义

设待处理数据集记为 D。其单元格可能同时含有多种脏污类型；本文用 $\mathcal{T}=\{\text{Missing}, \text{Anomaly}, \text{Typo}, \dots\}$ 表示完整的“错误类型”集合，其中的约束条件是：缺失值（*Missing*）与其他类型互斥——因为一旦某个单元格为空值，就不会再出现其他错误信息。

理论特征向量 对任意 $t\in\mathcal{T}$ 记 $r_t(D)\in[0,1]$ 为该类型在 D 中的边际比例；再用对称矩阵 $\mathbf{C}(D)\in\mathbb{R}^{|\mathcal{T}|\times|\mathcal{T}|}$ 描述不同类型在同一单元格中“共现”的二阶比例，其中 $\mathbf{C}_{ij}=P(\text{err}_i\wedge\text{err}_j)$。若 $t=\text{Missing}$ 或 $i=\text{Missing}\vee j=\text{Missing}$ 则 $\mathbf{C}_{ij}=0$。记总观测脏污率 $\text{ErrorRate}(D)=\sum_t r_t-\sum_{i<j}\mathbf{C}_{ij}$。综合得到[1]

$$\mathbf{x}_{\text{gen}}(D)=\Big(\text{ErrorRate}(D)\,;\,r_{t_1}(D),\dots,r_{t_{|\mathcal{T}|}}(D)\,;\,\text{vec}(\mathbf{C}(D))\,;\,m,n\Big), \tag{3.1}$$

其中 m,n 分别为特征维度与样本规模。该向量既统一刻画了**规模**又保留了**错误异质性**，是后续多标签自动化模型 $\Phi:\mathbf{x}(D)\mapsto\Omega'(D)$ 的输入。

记 $\mathcal{C}$ 为数据清洗方法的集合（如缺失值插补、异常值剔除、错误值纠正等），$\mathcal{H}$ 为聚类算法集合（如 K-Means、DBSCAN、层次聚类等），$\mathcal{P}$ 为聚类算法的超参数空间。将一个具体的清洗方法 c、聚类算法 h 及其超参数 $\boldsymbol{\theta}$ 组合成**清洗-聚类策略**：

$$\omega \;=\; \big(c,\,h,\,\boldsymbol{\theta}\big), \tag{3.2}$$

所有可行策略的笛卡尔积构成初始搜索空间：

$$\Omega \;=\; \mathcal{C}\;\times\;\mathcal{H}\;\times\;\mathcal{P}. \tag{3.3}$$

此时，如何在如此庞大的 Ω 中高效找到适配度高的 ω 即是后续的研究重点。

3.1.2 评价系统与最优方案

为了准确衡量任意策略 $\omega\in\Omega$ 在数据集 D 上的聚类质量（或适配性），通常采用若干无监督评价指标加以综合。本文主要使用 Davie-Bouldin（DB）指数 [51] 与轮廓系数（Silhouette）[52] 这两类典型指标，并线性组合为**综合得分**：

$$S(D,\omega)\;=\;\alpha\cdot\big[-DB(D,\omega)\big]\;+\;\beta\cdot\text{Sil}(D,\omega), \tag{3.4}$$

[1] 在实现中可仅存上三角元素 $\text{vec}(\mathbf{C})$，以免维度冗余；此处写成完整形式便于阅读。

其中 $\alpha, \beta > 0$ 为可调权重，$DB(\cdot)$ 越低表明簇内紧凑度与簇间分离度越理想，而 $\mathrm{Sil}(\cdot)$ 越高代表类内相似度越高、类间差异越大。若从聚类精度的角度出发，给定数据集 D 的最优策略可表示为：

$$\omega^*(D) = \arg\max_{\omega \in \Omega} S(D, \omega). \tag{3.5}$$

然而，若要在全量空间 Ω 上评估每个策略 ω，往往需要极高的时间成本。为此我们定义**优化子空间** $\Omega'(D) \subseteq \Omega$，仅在该空间中执行策略评估，以降低计算负担。记评估单个策略的耗时为 $T(D, \omega)$，则完整搜索与缩减搜索的总耗时分别为：

$$T_{\text{original}}(D) = \sum_{\omega \in \Omega} T(D, \omega), \quad T_{\text{reduced}}(D) = \sum_{\omega \in \Omega'(D)} T(D, \omega). \tag{3.6}$$

我们的目标是通过一个合适的 $\Omega'(D)$，在保证较高聚类质量的同时显著减少评估代价。

为量化“性能表现”与“时间加速”之间的平衡，我们引入**损失率**（或提高率）和**综合加速比**两个概念：

$$\eta(D) = 1 - \frac{\bar{S}(\Omega'(D))}{\bar{S}(\Omega)}, \tag{3.7}$$

$$\mathcal{A}(D) = \Big(1 - \eta(D)\Big) \times \frac{T_{\text{original}}(D)}{T_{\text{reduced}}(D)}, \tag{3.8}$$

其中 $\bar{S}(\Omega)$ 表示在完整空间上搜索所得的平均得分，$\bar{S}(\Omega'(D))$ 表示子空间 $\Omega'(D)$ 上的平均得分，$\eta(D)$ 越接近 0 表示缩减空间后带来的聚类性能损失越小，而 $\mathcal{A}(D)$ 越大则表示加速效果越显著。

3.1.3 从数据特征到优选策略的映射

在实际应用场景中，不同数据集 D 往往具有差异显著的质量特征（如 ErrorRate、AnomalyRate、MissingRate 等）。这些特征会显著影响清洗-聚类策略的效果，使得某些组合对特定类型的数据表现更优。若能根据 $\mathbf{x}(D)$（参考式 3.1）提前预测哪些组合更可能获得高分，即可避免对完整搜索空间 Ω 的全量评估。为此，我们引入一个映射函数：

$$\Phi : \mathbf{x}(D) \mapsto \Omega'(D), \tag{3.9}$$

其中 $\Omega'(D) \subseteq \Omega$。通过学习训练集中“特征—策略组合”的关联，再在新数据集上借助该映射快速筛选候选方案，最终只需在子空间 $\Omega'(D)$ 中执行搜索，这种基于历史数据的学习策略可极大降低时间成本 [28]。后续章节将介绍如何具体构建并训练这一映射。

3.2 技术难度

在前文对清洗-聚类协同优化的数学模型与评价体系进行阐释之后，如何在有限的时间与计算预算内高效地找到适配大规模、高噪声数据的组合策略，仍面临多重挑战。为更好地刻画这一过程并揭示潜在挑战，本文聚焦以下五个关键子问题：

(Q_1) **评估不同清洗-聚类组合在多样化数据特征下的适配性**

虽然现有清洗方法和聚类算法选择繁多，但在高维、高错误率（或缺失率）以及复杂噪声场景下，其表现仍难以保证稳定性与最优性。为此，需要系统量化并比较各组合在多种数据特征条件下的优劣，从而为后续策略选择奠定基础。

(Q_2) **构建基于数据特征到优选策略的映射函数**

当数据集特征呈高度异质时，单一清洗或聚类方法往往难以达到稳健性能。本文尝试在离线训练阶段学习 $\Phi(\mathbf{x}(D)) \mapsto \Omega'(D)$，依据数据特征向量自动筛选潜在近优的清洗-聚类组合，以实现快速且精准的策略推荐。

(Q_3) **平衡聚类质量与效率，实现在有限时间内逼近最优**

大规模数据会大幅提升搜索与评估开销，使实时需求难以满足。如何在缩减搜索空间的同时，维持可控的聚类质量损失，并取得显著加速，是本研究所关注的又一关键挑战。

符号	描述
D_{train}	先验数据集（训练集），用于离线评估和学习先验知识
D_{test}	测试数据集，用于实际部署和快速优化
K	Top-K 大小，表示在先验阶段选取的前 K 个最优方案
$\mathbf{M}^{(i)}$	数据集 $D^{(i)}$ 的 Top-K 策略矩阵
ℓ	标签，表示某一优选方案的标识符
$\mathcal{L}$	标签空间，包含所有优选方案的标签集合
$\mathbf{L}^{(i)}$	数据集 $D^{(i)}$ 对应的多标签集合
$\mathcal{M}$	训练集，包含所有先验数据的特征与标签集合
$\mathcal{F}$	多标签分类器，用于预测优选方案标签
$q^{(j)}$	标签 $\ell_{\omega(j)}$ 为优选方案的概率
r	预测阶段保留的最高优选标签数
$\mathbf{L}'$	预测阶段保留的最高优选标签集合
$\Omega'(D)$	数据集 D 的优选子空间，$\Omega'(D) \subseteq \Omega$
G	映射函数，将数据集特征向量映射到优选子空间
$\hat{\omega}$	最优方案，即在 $\Omega'(D_{\text{test}})$ 中得分最高的组合

表 2: 符号与描述

(Q_4) **深度分析清洗对聚类结果的实际影响机理**

本研究将系统考察给定数据集 D 与清洗-聚类策略 ω 时，哪些“有效纠正”决定了聚类得分 S 的形成，并探讨修正更多错误是否必然带来更优聚类表现。此外，还将评估清洗操作对最优超参数选择是否产生系统性偏移，从而为自动化管线的配置和调优提供深入的理论依据。

(Q_5) **利用清洗-聚类机理知识改写 AutoML 的搜索空间与特征工程**

在掌握错误类型对聚类过程与指标的细粒度影响后，如何把这些机理信息反馈到 AutoML：

- 搜索空间收缩——按错误类型显著受益的清洗-聚类组合优先保留，其余策略降权或剔除；
- 特征工程增强——将“错误类型分布、修复难度、过程指标”等高阶特征注入多标签模型，提升优选子空间预测的精度与解释性。

该挑战要求把机理洞察真正融入 AutoML 流程，而非仅作为事后分析，从而进一步降低评估成本并提升推荐质量。

围绕上述五个子问题，后续章节将逐一阐述自动化搜索与映射模型的设计思路，并通过大规模实验证明其在多场景下的可行性与性能优势。特别在 (Q_4) 与 (Q_5) 中，我们将结合清洗准确度指标、聚类算法内部过程的跟踪与错误类型的细粒度分析，深入探讨清洗策略如何从数据集、算法过程、评价指标、超参数，以及 AutoML 搜索空间与特征工程等角度共同影响聚类结果，为自动化管线的优化提供可靠的机理支撑。

4 自动化聚类方法

为进一步提高清洗-聚类策略的搜索效率，并同步分析清洗对聚类内部过程与评价指标的影响机理，本节将在第 3 节所述概念的基础上，介绍将数据划分为先验数据与测试数据、使用多标签学习构建映射函数，以及最终实现自动化聚类优化流程的整体方法。该方法是一个面向数据预处理、清洗、聚类与分析的完整端到端系统，不仅通过离线阶段积累的先验知识来缩减搜索空间、在较短时间内找到近优的清洗-聚类组合，更能对清洗操作的准确度及聚类算法的中间过程进行记录和可视化分析，以揭示“为何”或“何时”清洗能带来显著的聚类性能提升。

以下是本章节所定义的符号与描述：

4.1 先验数据与多标签映射策略

在实际应用中，通常可以从历史任务中获取大量已处理或部分标注的数据集，这些可视为**先验数据**（离线学习）。当面对新任务时，由于需要在较短时间内完成聚类策略的优选与评估，此时的新数据集则称为**测试数据**（在线应用）。通过在先验数据上深入探索并记录“数据特征—策略表现”的关联信息，就能在测试数据上显著减少不必要的搜索开销，从而提升整体效率。

4.1.1 先验数据与测试数据的划分

为便于在实际部署时利用先验知识，本研究将原有数据资源分为以下两类：

- **先验数据集** D_{train}：由多个历史数据集组成，记为 $D^{(1)}, D^{(2)}, \ldots, D^{(N)}$。在离线阶段（训练阶段），这些数据用于对搜索空间 Ω 进行大范围或抽样评估，以收集足够的策略得分信息，为后续自动化优化提供参考。
- **测试数据集** D_{test}：代表实际部署时面临的新数据，需要在线快速找到近优的清洗-聚类组合。此时可借助先验阶段所学知识，显著减少搜索规模并降低评估时间。

在离线评估过程中，对每个先验数据集 $D^{(i)}$ 遍历或随机抽样若干清洗-聚类策略 $\omega \in \Omega$，便可计算各自方案的综合得分 $S(D^{(i)}, \omega)$。为高效记录在 $D^{(i)}$ 上表现最好的候选策略集，我们定义一个 **Top-K 方案矩阵**（式 (4.1)），记为 $\mathbf{M}^{(i)}$，其中每一行是一个评分 S_j 较高的策略组合 $\omega_j^{(i)} = (c_j, h_j, \boldsymbol{\theta}_j)$。该矩阵按照 S_j 降序排列，用于在后续多标签学习中标识“优选”方案。

$$\mathbf{M}^{(i)} = \begin{pmatrix} c_1 & h_1 & \boldsymbol{\theta}_1 & S_1 \\ \vdots & \vdots & \vdots & \vdots \\ c_K & h_K & \boldsymbol{\theta}_K & S_K \end{pmatrix}. \tag{4.1}$$

4.1.2 多标签学习与映射函数构建

在离线阶段，除了得到各数据集 $D^{(i)}$ 的 Top-K 策略外，还要提取其特征向量 $\mathbf{x}(D^{(i)})$。通过**多标签学习**的方法，可将“数据特征”与“优选策略集合”关联起来，从而在面对新数据集 D_{test} 时，根据其特征向量 $\mathbf{x}(D_{\text{test}})$ 预测出最优或近优的策略子空间 $\Omega'(D_{\text{test}})$。

标签空间与多标签样本 离线阶段首先对每个先验数据集 $D^{(i)} \in D_{\text{train}}$ 全量（或抽样）评估策略 $\omega \in \Omega$，得到综合得分 $S(D^{(i)}, \omega)$，并取得分最高的 Top-K 组合 $\mathbf{M}^{(i)} = \{\omega_1^{(i)}, \ldots, \omega_K^{(i)}\}$。令每一条优选策略 $\omega^{(j)}$ 对应唯一标签 $\ell_{\omega^{(j)}}$，则可得到离散标签空间

$$\mathcal{L} = \{\ell_{\omega^{(1)}}, \ell_{\omega^{(2)}}, \ldots, \ell_{\omega^{(m)}}\}. \tag{4.2}$$

对于数据集 $D^{(i)}$，其多标签集合为

$$\mathbf{L}^{(i)} = \{\ell_{\omega_1^{(i)}}, \ell_{\omega_2^{(i)}}, \ldots, \ell_{\omega_K^{(i)}}\}. \tag{4.3}$$

于是可构造多标签训练样本 $\big(\mathbf{x}(D^{(i)}), \mathbf{L}^{(i)}\big)$，汇总为训练集

$$\mathcal{M} = \big\{(\mathbf{x}(D^{(1)}), \mathbf{L}^{(1)}), \ldots, (\mathbf{x}(D^{(N)}), \mathbf{L}^{(N)})\big\}. \tag{4.4}$$

分类器训练与优选子空间映射 在训练集 $\mathcal{M}$ 上训练多标签分类器 $\mathcal{F}$，其对每个标签 $\ell \in \mathcal{L}$ 产生置信度 $q_\ell \in [0,1]$。对新数据集 D_{test}，输入 $\mathbf{x}(D_{\text{test}})$ 得到

$$\mathcal{F}\big(\mathbf{x}(D_{\text{test}})\big) = \big\{(\ell, q_\ell) \,\big|\, \ell \in \mathcal{L}\big\}. \tag{4.5}$$

取置信度最高的 r 个标签形成

$$\mathbf{L}' = \{\ell \,|\, q_\ell \text{ 属于前} r \text{ 大}\}, \tag{4.6}$$

并映射回 **优选子空间**

$$\Omega'(D_{\text{test}}) = \{\omega \,|\, \ell_\omega \in \mathbf{L}'\} \subset \Omega. \tag{4.7}$$

该子空间通常远小于原始搜索空间，因而显著降低评估成本。整体映射可记为

$$\Phi:\ \mathbf{x}(D) \longmapsto \Omega'(D). \tag{4.8}$$

4.1.3 清洗准确度与聚类过程跟踪

在前文 (4.1.1, 4.1.2) 中，我们已经说明了如何在离线阶段获取“清洗-聚类”组合的综合得分 $S(D^{(i)}, \omega)$，并通过多标签学习将其映射到优选子空间。然而，仅仅依赖综合得分并不能充分解释数据清洗对聚类结果的内在影响，也无法揭示

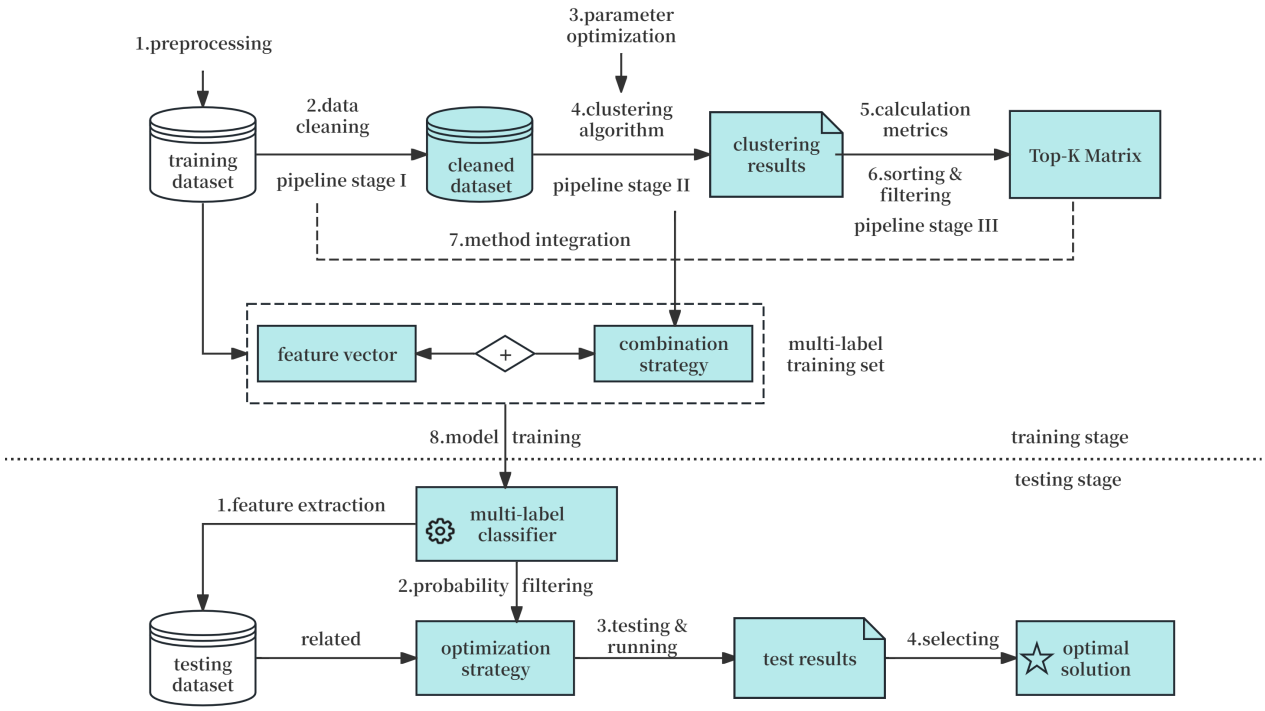


图 1: 自动化聚类优化流程示意图

某些清洗策略为何在高噪声或特定分布下表现优异或失败。为此，我们在离线阶段额外记录清洗准确度和聚类过程数据，以便在后续实验和可视化分析中，深入探究清洗如何改变聚类性能。

记录清洗准确度指标 在离线评估每个策略 ω 时，除了计算 $S(D^{(i)}, \omega)$ 外，我们还会借助对照基准（如 GroundTruth）来衡量清洗操作的有效性，主要包括：

- **EDR (Error Detection Rate)** [12]：清洗阶段检测并成功修复的错误值所占比例；
- **Precision, Recall, F1**：分别表示清洗操作对错误修复的准确率、召回率与综合平衡效果，能够帮助判断“修复了多少”与“修复是否准确”；

这些指标与综合得分一并写入 $\mathbf{M}^{(i)}$ 或其扩展表中，成为后续机理研究的重要依据。

跟踪聚类内部过程 除此之外，我们在聚类算法运行时插入了轻量化的过程跟踪机制，对如下一些中间结果进行记录：

- **K-Means 收敛轨迹**：包括迭代次数、质心坐标变动量、最终 SSE（Sum of Squared Errors）等；
- **DBSCAN 核心点/边界点比例**：了解清洗前后噪声点判定的变化；
- **层次聚类（HC）合并顺序**：可观察高维数据在不同清洗方式下层级拆分的演化差异；

通过这些过程级数据，研究者可在分析中深度观察清洗策略如何影响算法的收敛路径、簇划分形状，以及与数据特点之间的关联。若某些清洗方法在 F1 指标上很高却未能提升最终簇质量，往往能从这些跟踪结果中找到“破坏簇结构”的具体原因。

4.2 自动化聚类优化流程

在第 4.1 节中，我们介绍了如何利用先验数据构建多标签映射策略，以学习数据特征 $\mathbf{x}(D)$ 与优选方案子空间 $\Omega'(D)$ 之间的映射关系。本节将基于这些离线知识，探讨在**新数据集**上的自动化聚类优化流程。其核心思想是：通过多标签分类器在在线阶段快速筛选出若干“候选”清洗-聚类组合，避免大规模穷举搜索，从而在**更短时间**内获得**近优**结果。图 1 展示了该流程的整体示意。

该流程主要包括**离线知识积累**（训练阶段）和**在线优化**（测试阶段）两个核心环节：

Algorithm 1: 离线训练阶段：生成训练数据与训练多标签分类器

Input: 先验数据集 $D_{\text{train}} = \{D^{(1)}, \ldots, D^{(N)}\}$；
搜索空间 Ω；
Top-K 大小 K。
Output: 多标签分类器 $\mathcal{F}$
$\mathcal{M} \leftarrow$ `GenerateTrainingData`$(D_{\text{train}}, \Omega, K)$;
$\mathcal{F} \leftarrow$ `TrainClassifier`$(\mathcal{M})$;
return $\mathcal{F}$

Function `GenerateTrainingData`(D_{train}, Ω, K)**:**
  $\mathcal{M} \leftarrow \emptyset$;
  for $i \leftarrow 1$ **to** $|D_{train}|$ **do**
    foreach $\omega \in \Omega$ *(或采样自 Ω)* **do**
      计算 $S(D^{(i)}, \omega)$;
      记录 EDR/F1 等清洗准确度，以及算法内部过程数据（如质心迭代、核心点等）;
    选出 Top-K 策略 $\mathbf{M}^{(i)} = \{\omega_1^{(i)}, \ldots, \omega_K^{(i)}\}$ 按得分降序;
    映射为多标签集合 $\mathbf{L}^{(i)} = \{\ell_{\omega_1^{(i)}}, \ldots, \ell_{\omega_K^{(i)}}\}$;
    $\mathcal{M} \leftarrow \mathcal{M} \cup \{(\mathbf{x}(D^{(i)}), \mathbf{L}^{(i)})\}$;
  return $\mathcal{M}$
Function `TrainClassifier`$(\mathcal{M})$**:**
  // 可根据具体多标签算法实现
  训练多标签分类器 $\mathcal{F}$;
  return $\mathcal{F}$

1. **训练阶段（离线学习）**：基于先验数据集 D_{train}，计算不同数据特征与清洗-聚类策略的匹配程度，并训练多标签分类器 $\mathcal{F}$，从而建立数据特征到优选方案子空间的映射 $G(\mathbf{x}(D))$；同时收集并记录清洗准确度、聚类过程数据，以备深入机理分析。

2. **测试阶段（在线优化）**：面对新的数据集 D_{test}，利用训练阶段学习到的映射 $G(\mathbf{x}(D_{\text{test}}))$，快速筛选搜索空间 Ω 中的候选策略子集 $\Omega'(D_{\text{test}})$，避免全量穷举，从而在较短时间内获取高质量的清洗-聚类方案。若需进一步探讨其内在机制，可通过与离线记录的分布及过程数据比对来解释为何某些组合在新数据上表现突出或失效。

在接下来的小节中，我们将给出训练与测试环节的关键算法伪代码，并展示如何将多标签训练集 $\mathcal{M}$ 与在线推荐结果有机结合。

4.2.1 训练阶段：离线知识积累

训练阶段的目标是基于先验数据集 D_{train} 生成多标签训练集并学习多标签分类器。算法伪代码如算法 1 所示。

4.2.2 测试阶段：在线预测与最优方案搜索

测试阶段在新数据集 D_{test} 上应用训练好的分类器，快速锁定优选子空间并搜索最优策略。伪代码如算法 2 所示。

4.3 小结

本节系统介绍了自动化聚类方法的整体框架，从离线阶段的多标签训练与清洗准确度/过程数据记录，到在线阶段通过优选子空间快速搜索最优策略。在此基础上，我们不仅能在**大规模搜索空间**中高效获得近优清洗-聚类组合，还能借助**记录下的清洗精度与算法过程**信息，对清洗操作对聚类结果的影响机理进行更深入的剖析。下一章我们将结合具体实验展示该方法在多场景下的适用性与可解释性。

Algorithm 2: 测试阶段：寻找最优方案 $\hat{\omega}$

Input: 测试数据集 D_{test}；
多标签分类器 $\mathcal{F}$；
搜索空间 Ω；
保留标签数 r。
Output: 最优方案 $\hat{\omega}$
计算 $\mathbf{x}(D_{\text{test}})$;
$\mathbf{L}' \leftarrow \{\}$;
foreach $\ell \in \mathcal{L}$ **do**
  $q_\ell \leftarrow$ 置信度$(\mathcal{F}, \mathbf{x}(D_{\text{test}}), \ell)$;
  $\mathbf{L}' \leftarrow \mathbf{L}' \cup \{(\ell, q_\ell)\}$;
选取置信度最高的 r 个标签 $\mathbf{L}'_{\text{top}}$;
映射回优选子空间 $\Omega'(D_{\text{test}})$;
foreach $\omega \in \Omega'(D_{test})$ **do**
  计算 $S(D_{\text{test}}, \omega)$; // 计算综合得分
$\hat{\omega} \leftarrow \arg\max_{\omega \in \Omega'(D_{\text{test}})} S(D_{\text{test}}, \omega)$;
return $\hat{\omega}$

5 宏观实验与现象归纳

本章围绕第 3 节所提出的问题和模型定义（特别是第 3.2 节）展开实验与结果分析，通过对多种数据集和聚类算法的验证，定量评估“数据清洗与聚类协同优化”方案的有效性和适用性，重点回答问题 (Q_1) 到 (Q_3)。

5.1 实验设置

本章实验紧扣第 3 节提出的“多错误特征向量”理论。理论模型允许单元格同时出现多种脏污，但在实验阶段我们做出如下**可控退化**：

1) 仅保留两种最基本的错误类型：**缺失值 (Missing)** 与 **异常值 (Anomaly)**；

2) 同一单元格至多含一种错误，因此在实验中

$$r_{\text{tot}}(D) \;=\; r_{\text{miss}}(D) + r_{\text{anom}}(D).$$

3) **异常值难以在真实数据中精确计数**，故我们将“注入错误比例”当作 $r_{\text{anom}}(D)$；缺失值则在运行时精确统计，得到 $r_{\text{miss}}(D)$。

于是本章用于多标签学习及 AutoML 的实验特征向量为

$$\mathbf{x}_{\text{exp}}(D) = \big(r_{\text{tot}},\, r_{\text{miss}},\, r_{\text{anom}},\, m,\, n\big), \qquad r_{\text{tot}} = r_{\text{miss}} + r_{\text{anom}}.$$

5.1.1 数据集准备

本研究选用 4 个在数据清洗文献中被广泛引用的公开数据集 *beers*, *flights*, *hospital*, *rayyan*。对于每个数据集的**干净副本**，在除主键列外的所有单元格独立注入 $(\text{AnomalyRate}, \text{MissingRate}) \in \{0, 5, 10, 15\}\% \times \{0, 5, 10, 15\}\%$，共产生 $4 \times 4 - 1 = 15$ 份含错文件（排除 0%–0% 组合）。表 3 给出了四个数据集的规模 (n, m)，理论注入搜索网格（两类错误的参数均为 0-15%），以及 15 份带错文件**观测**总错误率 $r_{\text{tot}} = r_{\text{anom}} + r_{\text{miss}}$ 的最小-最大区间。
表中所有任务均满足 $r_{\text{tot}} = r_{\text{anom}} + r_{\text{miss}}$ 后续的多标签学习与 AutoML 管线直接采用

$$\mathbf{x}_{\text{exp}} = \big(r_{\text{tot}},\, r_{\text{miss}},\, r_{\text{anom}},\, m,\, n\big)$$

作为数据特征输入。

数据集	n	m	理论注入 (%)	$r_{\text{tot}}^{\min}$–$r_{\text{tot}}^{\max}$ (%)
beers	2 410	11	0–15 (Anom./Miss.)	9.23–33.10
flights	2 376	7	0–15 (Anom./Miss.)	4.99–29.99
hospital	1 000	20	0–15 (Anom./Miss.)	5.00–30.00
rayyan	1 000	12	0–15 (Anom./Miss.)	18.74–39.85

表 3: 四个数据集的规模、理论注入区间与观测总错误率范围

5.1.2 算法准备

本研究关注两方面算法：(1) **数据清洗策略**；(2) **聚类算法及对应参数**。

数据清洗策略 为便于后续实验复现与比较，表 4 初步汇总了本研究选取的 9 种清洗方法的关键信息。读者可据此快速了解各算法在本章实验中的角色与设置；其具体原理及对聚类过程的影响分析将在第 6 章展开。

算法	针对错误类型	必需配置	模型范式	清洗目标
Mode Impute	MV, FI	—	统计填补	*Repair*
Raha-Baran	MV, FI, Rule viol.	无显式约束	端到端 ML	*Detect + Repair*
HoloClean	MV, FI, Dup, Rule viol.	FD/CF + 外部知识	概率图模型	*Detect + Repair*
BigDansing	Schema viol., Typos	检测规则	规则驱动	*Detect*
BoostClean	Label/Attr Noise	下游模型 (监督)	Boosting Ensemble	Task-Aware Repair
Horizon	MV, Outlier	时序窗口宽度	时序/统计混合	*Repair*
Scared	MV, FI, Outlier	半监督标注预算	主动学习模型	*Detect + Repair*
Unified	MV, Rule viol., Dup	统一约束文件	多策略融合	*Detect + Repair*
GroundTruth	—	—	理想基线	*Upper Bound*

MV: Missing Value；FI: Format Inconsistency；Dup: Duplicate；Rule viol.: 约束违规。

表 4: 实验用 9 种数据清洗方法总览（本章仅作简述，机理详见第 6 章）

聚类算法 表 5 简要列出了本章所使用的 6 类定制化聚类脚本在初始化、超参搜索、过程追踪及复杂度方面相较于 `scikit-learn` 标准实现的主要调整。其设计动机在于为第 5 章实验提供统一且可追踪的运行记录；更深入的原理剖析与过程指标解读将在第 6 章展开。

5.2 实验流程

本节给出大规模统计实验的标准流水线（图 2），覆盖本章后续所有结果所需的输入与输出。整个链条将做数据与指标采集，但并不在此阶段训练或调用 AutoML；步骤 4 产生的中间文件将在第 6 章用作 AutoML 特征与标签。

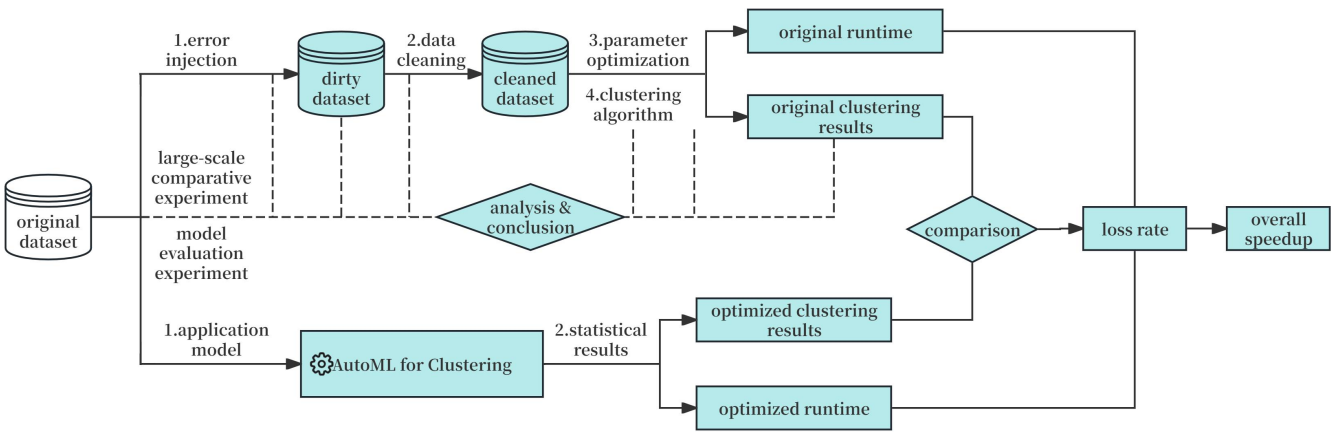


图 2: 本章批量实验的四阶段流水线

算法名称	初始化策略	参数调优	过程指标 (方向)	复杂度 变化
$\text{K-Means}_{\text{base}}$	k-means++	Optuna(k)	$\Delta n_{\text{iter}}\downarrow$, $\text{AUC}_{\Delta}\downarrow$	$\uparrow \mathcal{O}(nkT)$
$\text{K-Means}_{\text{PPS}}$ [53]	K-MC^2 采样	Optuna(k)	同上	$\downarrow \mathcal{O}(n)$ init
$\text{K-Means}_{\text{NF}}$ [54]	随机标签 $\to F$	Optuna(k)	同上	$\uparrow$ (Gram)
GMM-EM (tracking)	k-means++	Optuna+Kneedle(k)	$\Delta n_{\text{iter}}\downarrow$, $\text{AUC}_{\text{LL}}\downarrow$	$\uparrow$ (warm start 循环)
DBSCAN (noise-aware)	—	Optuna(ε, minPts)	$\Delta n_{\text{core}}\uparrow$, $\Delta\rho_{\text{noise}}\downarrow$	$\approx \mathcal{O}(n\log n)$
HC (merge-tree)	—	Optuna(k)+linkage+metric	$\Delta n_{\text{merge}}\downarrow$, $\Delta h_{\max}\downarrow$	$\mathcal{O}(n^2)$

表 5: 6 种定制化聚类脚本的初始化策略、调参方式、过程指标及复杂度概览

Step 1. 可控错误注入 & 特征统计
对每个干净数据集依照 $(\text{AnomalyRate}, \text{MissingRate}) \in \{0, 5, 10, 15\}\,\%$ 的网格注入错误，生成 15 份含错版本。运行时即时统计 $\{r_{\text{tot}}, r_{\text{miss}}, r_{\text{anom}}, m, n\}$，形成特征向量 $\mathbf{x}_{\text{exp}}(D)$。

Step 2. 清洗 → 聚类批处理
对每个含错文件分别执行 9 种清洗算法（表 4）并采集结果；随后用 6 种定制化聚类脚本（表 5）进行聚类与超参搜索，同时记录过程指标（迭代步数、质心位移、核心点等）。

Step 3. 指标计算与分档
对聚类输出计算 Silhouette、Davies–Bouldin 与 Combined Score，并把全部原始指标汇总为三类实验：(i) 得分评估——跨多种清洗-聚类组合的结果求均值/方差；(ii) 错误率梯度——按 r_{tot} 分档绘制曲线；(iii) 错误类型对比——固定 r_{tot} 比较 r_{miss} 与 r_{anom} 的影响。

Step 4. 结果持久化
将每个错误数据集的清洗标签、聚类历史、超参数与评估指标统一记录，同时导出 $\left(\mathbf{x}_{\text{exp}}, \Omega^{\star}, S^{\star}\right)$ 三元组作为 AutoML 训练样本。

完成上述 4 步后，我们即将在第 5.3.1 节到第 5.3.2 节分别开展“得分评估”与“错误敏感性”三组实验，进而回答第 3.2 节中的问题 (Q_1)–(Q_3)。

5.3 实验结果与分析

在完成第 5.2 节所述四阶段实验流水线后，我们对全部 60 份错误数据集 × 9 种清洗 × 6 种聚类共 $60 \times 9 \times 6 = 3\,240$ 组运行结果进行统计分析。本节按照由粗略到细致、由全局到局部的思路拆分为三项对照实验：

1. **得分评估实验**（§5.3.1）——横向比较 9×6 种“清洗 + 聚类”组合在平均分、方差等维度的全局表现；
2. **错误敏感性实验**（§5.3.2）——将 r_{tot} 以 5 % 为步长分档，同时在固定 r_{tot} 的前提下改变 $\left(r_{\text{miss}}, r_{\text{anom}}\right)$ 二元比例，统一考察错误率递增与错误类型差异对聚类指标的综合影响。

下面首先给出 5.3.1 的详细结果。

5.3.1 得分评估实验

实验目的与统计指标 本子实验从“绝对效果 / 相对效果 / 方法稳定性”三个维度，评估每一个 $\langle$清洗, 聚类$\rangle$ 固定组合在 60 份含错数据集中的整体表现。记 S = Combined Score，S_{GT} 为同数据集 *Ground-Truth* 清洗下的得分，则采集三项统计量

$$\left(\overline{S},\ \overline{S}_{\%\text{GT}},\ \sigma_S^2\right), \qquad \overline{S}_{\%\text{GT}} = \frac{1}{N}\sum_{i=1}^{N}\frac{S_i}{S_{i,\text{GT}}} \times 100\%.$$

- $\overline{S}$：绝对平均分；直接对 S 取算术均值。
- $\overline{S}_{\%\text{GT}}$：相对平均分；先对每条记录做“除以 S_{GT} 后乘 100%”的归一化，再取均值，用于跨数据集的横向比较。
- σ_S^2：得分方差；衡量组合在不同数据-错误场景下的波动风险——方差越大，出现“爆分 / 翻车”的概率越高。

结果可视化 针对每个数据集（beers、flights、hospital、rayyan）各生成三幅图：

1) *相对平均分热力图*
9×6 单元；颜色 $= \overline{S}_{\%\mathrm{GT}}$，单元格右下角小灰字 $= \sigma_S^2$。直观看出“最深蓝”块即全局最佳组合。

2) *均值-方差散点图*
X 轴 $= \overline{S}_{\%\mathrm{GT}}$，Y 轴 $= \sigma_S^2$。点颜色代表清洗方法，点形状代表聚类算法，在同一平面呈现“收益-风险”权衡。

3) *Top-10 带误差条形图*
选该数据集上 $\overline{S}_{\%\mathrm{GT}}$ 最高的 10 个组合：横轴 $= \overline{S}_{\%\mathrm{GT}}$，误差条 $= \pm\sqrt{\sigma_S^2}$，并用虚线标出 100% 基准。

结果分析 图 3 汇聚了四个数据集在 54 种 ⟨清洗, 聚类⟩ 组合上的 *相对均值热力图* [(a)(d)(g)(j)], *均值-方差散点* [(b)(e)(h)(k)] 以及 *Top-10 带误差条形图* [(c)(f)(i)(l)]。综合三种视角可归纳出如下共性与差异。

1. **层次聚类 (HC) + 简单填补在纯数值场景最优且稳健。**
在 *beers* 与 *flights* 的热力图 [Fig. 3 (a)(d)]，**mode+HC** 的 $\overline{S}_{\%\mathrm{GT}}$ 分别达到 **1.72** 与 **1.89**，方差仅 $\sigma_S^2 = 0.14, 0.21$。相同组合在散点图 [Fig. 3 (b)(e)] 落于“右端-低纵”的低风险高收益象限，说明对低维、纯数值表格，朴素众数填补已足以配合 HC 的层次划分取得高且稳的综合得分。

2. **深度语义清洗 (baran) 是高维混合表的刚需。**
在 20 维医疗数据 *hospital* [Fig. 3 (g)(h)(i)]，**baran+HC** 以 $\overline{S}_{\%\mathrm{GT}} = 0.88$ 领先所有非语义方法 10% – 25%，且 $\sigma_S^2 \le 0.02$。高维、含规则依赖字段需要“检测 + 修复”一体的语义清洗才能稳定提升 HC 成果；纯统计或局部规则方法 (mode/bigdansing) 散点显著偏左或偏上。

3. **密度聚类 (DBSCAN) 在高离群文本场景爆发但波动最大。**
对标签文本为主的 *rayyan* [Fig. 3 (j)(k)(l)]，**mode+DBSCAN** 取得 $\overline{S}_{\%\mathrm{GT}} = 1.28$，但 $\sigma_S^2 \approx 0.60$，Top-10 误差条宽达 ±21%。同样的“高均值-高方差”也出现在 *beers*/*flights*，反映 DBSCAN 对 ε、minPts 与离群比例极度敏感，需要配合参数重采样或阈值监控方可落地。

4. **滑窗-平滑修复 (horizon) 带来跨领域稳健增益。**
在 *flights* 与 *hospital* 的 Top-10 [Fig. 3 (f)(i)]，horizon+HC/GMM 位列前 2 -4，且误差条均小于 10 %。时序窗口插补减轻了缺失-抖动噪声，使质心型与层次型算法在中高缺失率环境一并受益。

5. **“理想清洗”并非最优——过度修复会削弱层次差异。**
在四张热力图中，GroundTruth+HC 的 $\overline{S}_{\%\mathrm{GT}}$ 仅排第 3-8，例如 *flights* 精确等于 1.00。说明当所有局部噪声被完全抹平后，HC 易产生过细切分、降低 Silhouette。对 AutoML 的启示是：不能简单以“清洗 F1 最高”作为唯一上界，必须同时考虑清洗与聚类的耦合匹配。

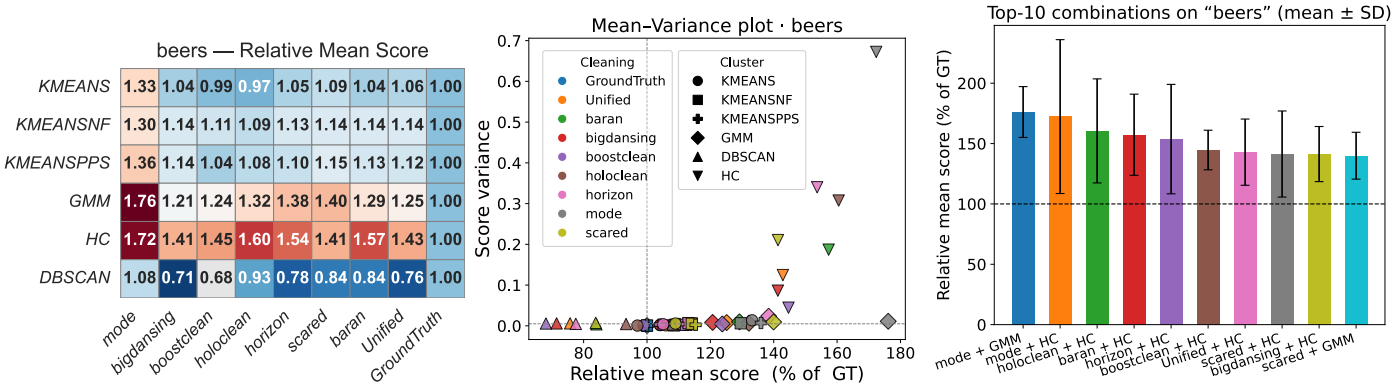
5.3.2 错误敏感性实验

实验目的与统计指标 本实验从两条互补维度定量评估噪声对综合得分 S 的影响：

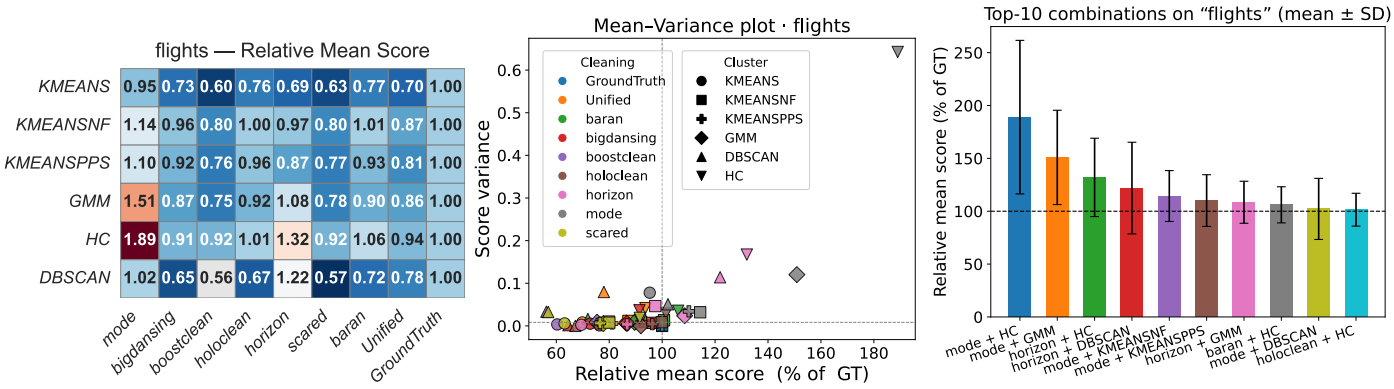
- **总错误率梯度** (error_rate) ——将观测错误率划分为七段 $[0,5), [5,10), \ldots, \ge 30)$ %（记作 error_bin），并在每档仅保留同一清洗 / 聚类方法下的最高 S，绘制各方法的“抗噪极限”曲线。
- **错误类型比例** (*Missing* vs. *Anomaly*) ——在总错误率不超过 15 % 时，枚举 4×4 组 Missing × Anomaly 配比，取每格中非 *GT* 组合的最大 S 作为色值，左上 $(0,0)$ 以 *Ground-Truth* 得分作参考。

结果可视化 图 4 按“清洗曲线 → 聚类曲线 → 类型热力图”的顺序，将 **beers**、**flights**、**hospital** 与 **rayyan** 四个数据集的结果并列排版，便于横向对比：

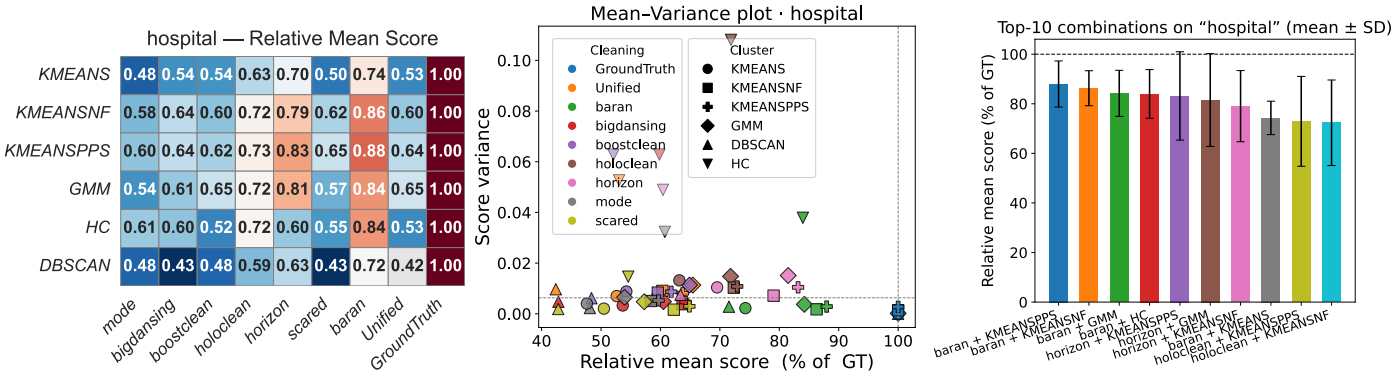
1) *清洗方法梯度曲线*
横轴为七段 error_bin，纵轴对应档内的最高 S，颜色与标记代表不同清洗策略。



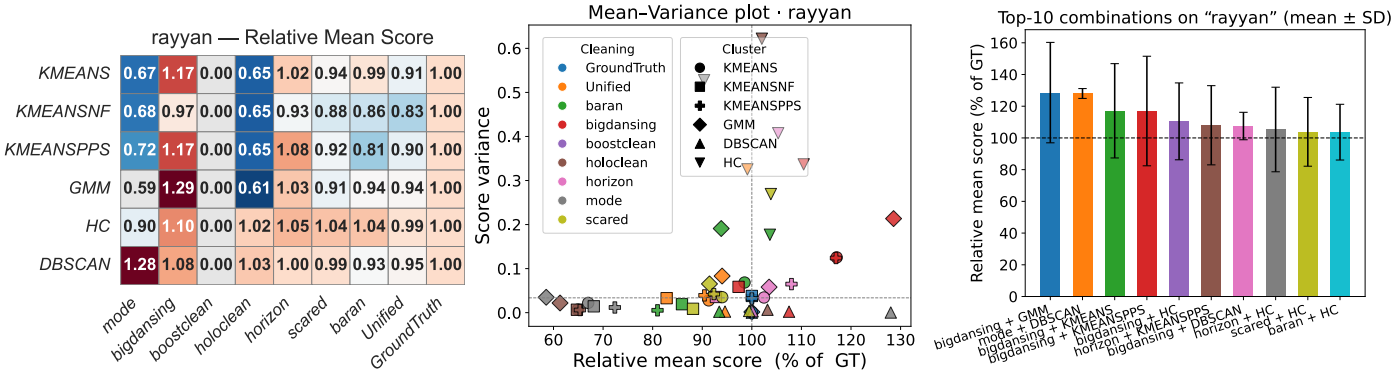
(a) Heat-map • beers (b) Mean–Var • beers (c) Top-10 • beers



(d) Heat-map • flights (e) Mean–Var • flights (f) Top-10 • flights



(g) Heat-map • hospital (h) Mean–Var • hospital (i) Top-10 • hospital



(j) Heat-map • rayyan (k) Mean–Var • rayyan (l) Top-10 • rayyan

图 3: 三种视角的得分评估结果（每行对应一个数据集）。颜色越深蓝表示 $\overline{S}_{\%\mathrm{GT}}$ 越高；散点越靠右/越低表示高收益且低风险；条形误差条体现平均收益的置信区间。详细数值与现象待第 5.3.1 段分析。

2) 聚类算法梯度曲线
坐标系与左图一致，但曲线换为不同聚类算法，用于观察“优清洗 + 优聚类”组合在高噪声区间的稳定性。

3) 错误类型比例热力图
展示 Missing 与 Anomaly 四档比例下的 4×4 网格，色阶范围随本行数据集的 $\min S$ 与 $\max S$ 自适应，直观揭示“哪种错误主导时性能下降最明显”。

结果分析 图 4 (a–l) 将 *Cleaning-curve*、*Cluster-curve* 与 错误类型热力图并列排版，系统呈现错误率递增与 (Missing, Anomaly) 配比对综合得分 S 的联合作用。综合三条视角，可归纳出如下五点共性与差异。

1. **15–25 % 总错误率是性能“峰值带”。**
在 28 个 error_bin 中有 24 个峰值落在此区间。例如 `mode` 在 *beers* 的 20–25 % 档将 S 提升至 **3.76**（+212 % 相对 GT），*flights* 在 10–15 % 档达到 3.68（+248 %）。超过 25 % 后除 *rayyan* 外曲线平均回落 30 – 42 %。

2. **层次聚类 (HC) 搭配鲁棒清洗在高噪末端仍保持领先。**
HC 在全部 28 个档位均列前二；在最高噪声档 (≥ 30 %) 仍维持 $S = 2.31$（*beers*）、2.04（*flights*）、2.07（*hospital*）与 4.12（*rayyan*）。热力图“Missing = 0 %”列出现连续暖带，证明多尺度 linkage 能缓冲 Missing 与 Anomaly 的叠加冲击。

3. **离群点对低维数值和文本表呈“双刃”效应。**
在 *beers* 与 *rayyan*，(Anom = 5%, Miss = 0%) 取得最高 S（3.76 与 4.74）；若两类错误均增至 15 %，S 分别降至 2.35 与 2.83，跌幅 37–40 %。轻度离群拉大簇间距；缺失同步增加时密度被稀释，增益消失。

4. **高维依赖数据对缺失最不敏感。**
hospital 热图 16 格的色阶区间仅 1.42 – 2.01；Missing 从 0% 升至 15% 仅使 S 下降 0.18。Cluster-curve 方差 $\sigma_S^2 = 0.018$ 为四数据集最低，表明层次结构天然兜底缺失扰动。

5. **Ground-Truth 清洗并非聚类最佳起点。**
四张热力图的 $(0, 0)$ 参考分在所在行 / 列均不是极值；其中 *flights* 刚好等于 1.00，被 $(10\%, 0\%)$ 的 3.68 和 $(0\%, 5\%)$ 的 2.94 远超。说明过度修复削弱簇间差异，清洗与聚类需要协同优化而非单端极致。

5.4 本章小结

本章在 4 个基准数据集、15 种错误注入配置和 $9 \times 6 = 54$ 条“清洗–聚类”管线上共计 $60 \times 54 = 3\,240$ 次批量实验，给出了三条核心洞察：

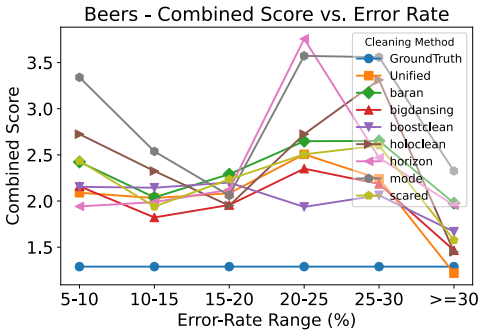
- **全局最优模式**——低维数值表以 `mode`+HC 取得 $\overline{S}_{\%\mathrm{GT}} \approx 1.8$，高维依赖表须用语义清洗 (`baran`) 才能稳定提升 HC，文本 / 时序场景中 DBSCAN 虽偶有爆分却方差最高。
- **错误率与类型**—— 15–25 % 是整体峰值区；轻度 Anomaly 能放大簇间距，但一旦与 Missing 同步升到 15 % 则平均拉低 S 约 40 %。
- **协同大于极致**——多数情况下 Ground-Truth 清洗并未给出最佳聚类，适度保留结构性噪声反而利于层次划分。

这些规律不仅回答了 $(Q_1) \sim (Q_3)$，也引出了下一步的研究问题：清洗究竟通过哪些“微观环节”影响聚类？第六章将顺势深入——逐层剖析清洗准确度、聚类收敛轨迹、评价指标和超参数选择之间的因果链条，为后续 AutoML 搜索空间压缩与动态适配提供理论支撑。

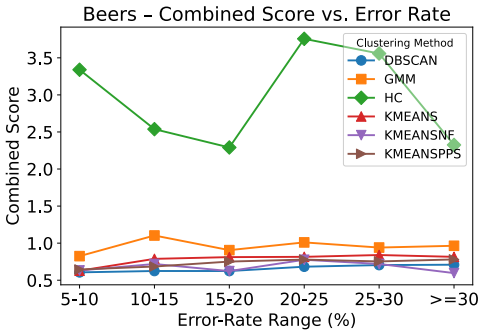
6 机理分析与 AutoML 改进

6.1 研究动机

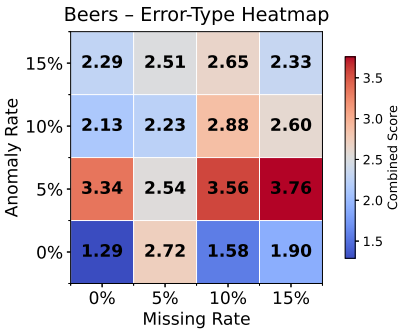
上一章的宏观实验结果表明，“清洗 × 聚类”组合在不同数据特征与错误率下呈现出不同的评分与搜索效率，但这些结果尚未揭示背后的因果机制。清洗操作对哪些错误类型发挥关键作用、又以何种幅度改变质心收敛轨迹、核心点判定以



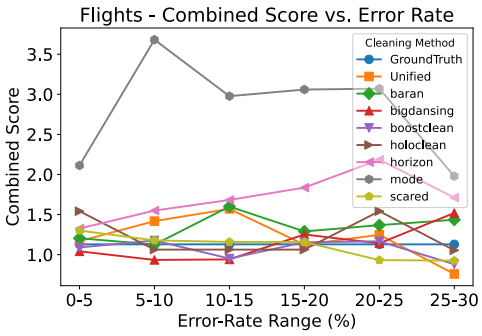
(a) Cleaning-curve • beers



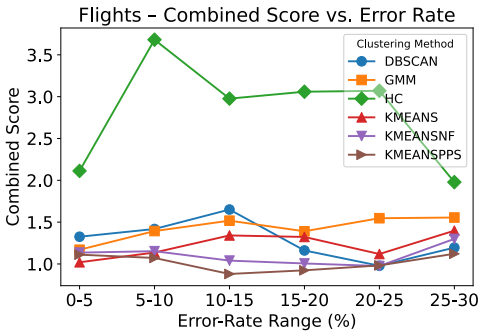
(b) Cluster-curve • beers



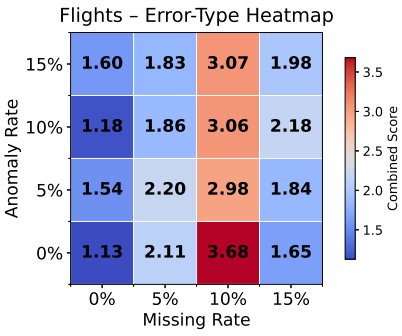
(c) Heat-map • beers



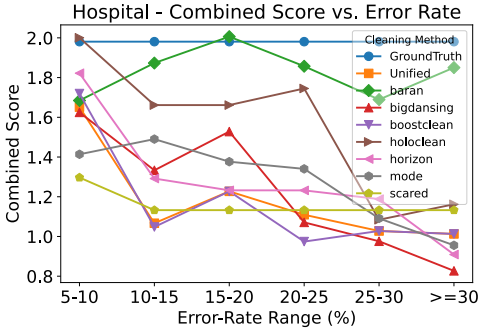
(d) Cleaning-curve • flights



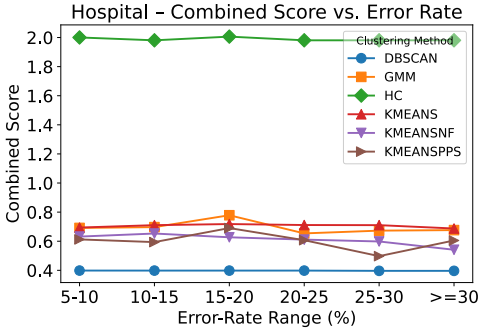
(e) Cluster-curve • flights



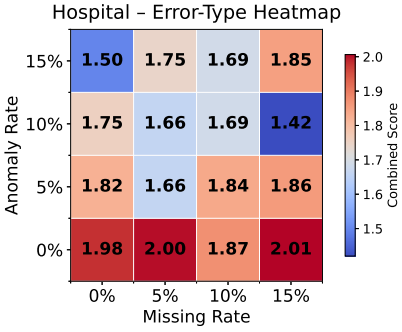
(f) Heat-map • flights



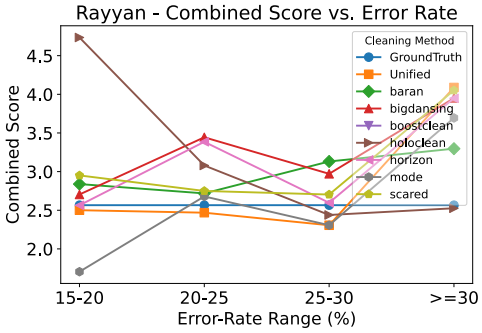
(g) Cleaning-curve • hospital



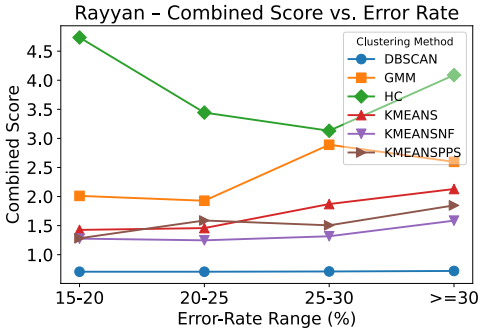
(h) Cluster-curve • hospital



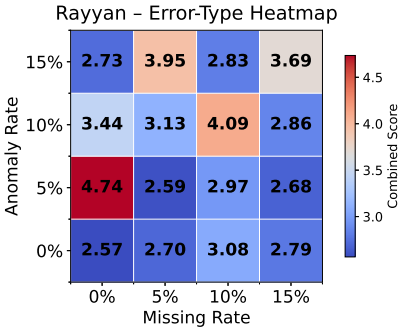
(i) Heat-map • hospital



(j) Cleaning-curve • rayyan



(k) Cluster-curve • rayyan



(l) Heat-map • rayyan

图 4: 错误敏感性可视化：每行对应一个数据集，依次展示清洗策略曲线、聚类算法曲线与类型比例热力图。

及最优超参数区间，仍有待系统阐明。本章按照“错误类型 → 清洗行为 → 算法过程 → 聚类指标 → 超参数偏移”的因果链，依次回答以下四个逐步递进的子问题：

- (Q_1) **清洗准确度——错误类型视角**
  在单元格级真值标签的多错误类型维度上，分别计算每种清洗方法的检测率 EDR_t、精确率 $\mathrm{Precision}_t$、召回率 Recall_t 及综合指标 $\mathrm{F1}_t$（其中下标 t 标识具体错误类型）。这些指标用于量化各方法对不同错误类型的修复覆盖度、修复量及修复精度，为后续机理分析奠定客观基础。

- (Q_2) **清洗干预——过程级视角**
  为判定清洗是否实质改善聚类运行，本研究在三大算法族中各选取衡量收敛速度、噪声判定稳定性与层次结构紧凑度的代表性过程指标，并比较清洗前后的相对改变量。该比较旨在揭示修复率提升是否同步带来运行机理层面的正向变动。

- (Q_3) **清洗收益——聚类评价视角**
  基于 (Q_1) 的分类型修复结果，系统考察错误修复幅度对 $Silhouette_{\mathrm{rel}}$, DB_{rel} 及 $Combined_{\mathrm{rel}}$ 指标的边际影响，识别收益呈现线性、阈值或饱和模式的区间。重点输出“边际收益排序”，以确定最值得优先修复的错误类型并为 AutoML 特征加权提供定量依据。

- (Q_4) **清洗偏移——超参数视角**
  在 (Q_2) 已观测到过程指标显著变化的前提下，进一步检测最优超参数是否出现系统性漂移，例如 $\Delta k = k_{\mathrm{clean}} - k_{\mathrm{raw}}$ 及 $\Delta\varepsilon = \varepsilon_{\mathrm{clean}} - \varepsilon_{\mathrm{raw}}$。该检验旨在判断清洗后是否需要重新缩放或重采样搜索空间，从而指导 AutoML 在效率与性能之间做出更优权衡。

以下内容将给出统一的实验设计与数据采集方案，通过一次性的实验流程来获取涵盖上述四个问题所需的信息，后续小节将基于这些采集结果，对四大问题做深入分析与讨论。

6.2 清洗算法原理与机理假设

下文按照“算法核心思想 → 覆盖错误类型 → 对本章实验指标的预期干预链路”依次介绍 8 种清洗策略。定量幅度留待 §6.4.2 实测，本节给出方向性假设。

Mode Imputation. 该方法把“列内出现次数最多（或均值）即真值”作为唯一假设，一次性用该众数 / 均值回填全部缺失单元，并顺带纠正极少数单字符 *Typo*。实验锚点：我们将在质心类算法的迭代曲线（AUC_Δ）上验证方差是否显著收缩，并留意 DBSCAN 噪声阈值 ε 是否向下微调。

Raha–Baran [28]. 该方法通过大语言模型捕获上下文语义，再辅以轻量依赖约束与三路纠错器，在极少标注条件下端到端修复 Typo 和数值异常。实验锚点：我们将观测噪声点被抑制后 HC 合并层数与 K-Means 迭代是否同步下降，以及 DBSCAN 最优 ε 是否整体偏左。

HoloClean [10]. 该方法将函数依赖、否定约束和外部词典统一表示为概率图模型，对 Missing 和 Rule-Violation 进行高精度推断与修复，Recall 取决于约束覆盖。实验锚点：核心关注 DBSCAN 噪声率 ρ_{noise} 的下降幅度及 Silhouette 方差，以评估误报对簇结构的冲击。

BigDansing [55]. 该方法采用大规模并行规则挖掘与匹配来检测 Schema 违规和格式 Typo，但默认不执行修复，也忽略 Missing 与 Anomaly。实验锚点：我们可以把它当作“检测基线”，主要检查 K-Means 质心摆幅是否因格式统一而轻度缩小，其余过程 / 超参预计保持原状。

BoostClean [56]. 该方法把下游监督模型的预测误差当作目标，用 Boosting 迭代优先修复对任务最有利的脏元。实验锚点：若标签边界与簇结构一致，Combined Score 可能随 Precision 提升；若不一致，则 Comb 波动和参数漂移可能增大。

算法族	过程指标	期望方向
质心型（Baseline / PPS / NF K-Means）	几何衰减率 GeoDecay	↓
	质心位移曲线面积 AUC_{Δ}	↓
	终态 SSE	↓
模型型（GMM）	几何衰减率 GeoDecay	↓
	对数似然收敛面积（AUC）	↓
层次型（HC）	合并层数 $\Delta n_{\mathrm{merge}}$	↓
	最大合并高度 $\Delta h_{\max}$	↓
	簇内 / 簇间距离比 $\Delta R_{\mathrm{intra/inter}}$	↓
密度型（DBSCAN）	核心点计数 Δn_{core}	↑
	平均邻居数 Δn_{avg}	↑
	噪声率 $\Delta \rho_{\mathrm{noise}}$	↓
	CDF-Wasserstein 距离 ΔW_{cdf}	↓

表 6: 四大算法族的过程级指标及其期望方向

Horizon [57]. 该方法针对时序数据的 FD 违规，先用滑窗插补缺失，再通过频繁模式选择修复离群点。实验锚点：我们关注 K-Means SSE 与 HC $h_{\max}$ 是否收缩，以及在 FD 不完备时 Silhouette 方差是否反向上升。

SCARE [58]. 该方法通过主动学习把有限标注预算聚焦于高不确定元组，实现高 Precision、受 δ 约束的 Recall。实验锚点：中等 Missing 场景应看到 DBSCAN 噪声率下降和 K-Means 轻度提速；高缺失时增益预计趋于平坦。

Unified [22]. 该方法在统一代价模型下同时考虑“修改数据”与“松弛约束”，以获得低方差、稳健的整体修复效果。实验锚点：我们预期 SSE、ρ_{noise}、$h_{\max}$ 均有中等幅度改善而超参几乎不变，从而作为 AutoML 搜索的“安全默认”分支。

6.3 实验设计

本节在第 5 章的整体框架上做有限增补，使同一条流水线即可回答 Q_1–Q_4。除特别说明外，数据集、评价指标与硬件配置均沿用上一章。

Step 1. 数据与错误注入
选用 *beers*、*flights*、*hospital*、*rayyan* 共四张公开表；在单元格级别对 **Missing Value (MV)** 与 **Anomaly (Ano)** 分别注入 $0\%, 5\%, 10\%, 15\%$ 四档比例。干净副本保留为真值，用于后续计算 $\mathrm{EDR}_{\{MV,Ano\}}$、Precision、Recall、F1。

Step 2. 清洗执行——回答 Q_1
第 6.2 节列出的 8 种清洗策略逐一作用于 **Step 1.** 生成的所有 (数据集, 错误率) 组合。运行结束后按错误类型记录修复量与修复准确度，得到 $\{\mathrm{F1}_{MV}, \mathrm{F1}_{Ano}\}$。

Step 3. 聚类与过程跟踪——回答 Q_2
清洗前后分别运行四类聚类算法（三变体 K-Means、GMM、HC、DBSCAN），并实时记录其关键过程指标，具体定义及期望方向见表 6。

为便于横向比较，所有过程指标都转换为“相对 *mode*（众数填补）清洗后的改变量”

$$\Delta\% \;=\; \frac{\mathrm{metric}_{\mathrm{CLEAN}} - \mathrm{metric}_{\mathrm{MODE}}}{|\mathrm{metric}_{\mathrm{MODE}}| + 10^{-8}} \times 100\%. \tag{6.1}$$

后续热力图均以 $\Delta\%$ 为色阶，红色表示退化，蓝色代表改善。

Step 4. 聚类结果评估——回答 Q_3
对每份（清洗 $\times$ 聚类）结果计算 Silhouette 与 Davies-Bouldin，并给出加权组合 $\mathrm{Combined} = 0.5\,(1/\mathrm{DB}) + 0.5\,\mathrm{Sil}$。再将这些分数与 $\{\mathrm{F1}_{MV}, \mathrm{F1}_{Ano}\}$ 做相关及边际收益分析，用于判定优先修复的错误类型。

Step 5. 超参数搜索与漂移——回答 Q_4

分别在清洗前后运行 Optuna / KneeLocator，搜索最优 k、ε、*covariance type* 等参数；对比 Δk、$\Delta\varepsilon$ 与 ΔCombined 以量化清洗引起的搜索空间漂移。

Step 6. 汇总输出

单条实验流水线最终同时产出

$$\underbrace{\{\mathrm{F1}_{MV}, \mathrm{F1}_{Ano}\}}_{Q_1} \cup \underbrace{\{\Delta\%\_\mathrm{process}\}}_{Q_2} \cup \underbrace{\{\mathrm{Sil}, \mathrm{DB}, \mathrm{Combined}\}}_{Q_3} \cup \underbrace{\{\Delta k, \Delta\varepsilon\}}_{Q_4},$$

为 § 6.4 的机理验证与 § 6.5 的 AutoML 动态裁剪奠定数据基础。

6.4 实验结果与分析

6.4.1 Q_1：清洗准确度——错误类型视角

目标与方法说明. 在两类单元格错误 $\mathcal{T} = \{\mathrm{Missing}, \mathrm{Typo}\}$ 上，对每个清洗方法 c 计算

$$\mathrm{EDR}_t = \frac{\#\ \mathrm{repaired}_t}{\#\ \mathrm{errors}_t}, \qquad \mathrm{Precision}_t = \frac{\#\ \mathrm{correct}_t}{\#\ \mathrm{repaired}_t}, \qquad \mathrm{Recall}_t = \frac{\#\ \mathrm{correct}_t}{\#\ \mathrm{errors}_t}, \qquad \mathrm{F1}_t = \frac{2\,\mathrm{Prec}_t\,\mathrm{Rec}_t}{\mathrm{Prec}_t + \mathrm{Rec}_t}$$

其中下标 $t \in \mathcal{T}$ 指定错误类型。上述四项分别衡量覆盖度（EDR），修复精度（Precision / Recall），以及综合效果（F1），为后续机理与边际收益分析提供基准。

(a) Dataset: **beers**

Method	Missing				Typo			
	EDR	Prec.	Rec.	F1	EDR	Prec.	Rec.	F1
Mode	-0.008	0.000	0.611	0.661	0.009	0.000	0.018	0.000
Raha-Baran	0.604	0.572	0.845	0.855	0.625	0.575	0.717	0.687
HoloClean	-0.960	-1.215	0.252	0.254	0.033	0.002	0.056	0.004
BigDansing	0.058	0.097	0.635	0.694	0.177	0.177	0.276	0.282
BoostClean	0.045	-0.179	0.631	0.601	0.098	0.109	0.170	0.185
Horizon	-0.037	0.011	0.600	0.666	0.062	0.039	0.111	0.074
SCARE	-0.099	0.062	0.575	0.682	0.103	0.117	0.175	0.200
Unified	0.019	-0.091	0.621	0.631	0.139	0.163	0.227	0.260

(b) Dataset: **flights**

Method	Missing				Typo			
	EDR	Prec.	Rec.	F1	EDR	Prec.	Rec.	F1
Mode	0.014	0.000	0.852	0.914	0.014	0.000	0.028	0.000
Raha-Baran	0.676	0.631	0.953	0.972	0.683	0.651	0.793	0.775
HoloClean	-0.366	-1.321	0.837	0.837	0.809	0.809	0.823	0.823
BigDansing	-0.284	-0.265	0.814	0.897	0.525	0.590	0.638	0.712
BoostClean	0.000	0.000	0.844	0.893	0.000	0.000	0.000	0.000
Horizon	-0.491	-1.226	0.792	0.835	0.400	0.440	0.531	0.577
SCARE	-0.420	0.171	0.798	0.932	0.116	0.197	0.196	0.319
Unified	-0.031	0.758	0.861	0.979	0.245	0.760	0.347	0.855

(c) Dataset: **hospital**

Method	Missing				Typo			
	EDR	Prec.	Rec.	F1	EDR	Prec.	Rec.	F1
Mode	0.103	0.000	0.740	0.770	0.103	0.000	0.179	0.000
Raha-Baran	0.824	0.249	0.946	0.830	0.826	0.252	0.882	0.379
HoloClean	0.241	0.101	0.781	0.793	0.241	0.101	0.363	0.178
BigDansing	-0.015	0.020	0.706	0.775	0.212	0.185	0.324	0.294
BoostClean	-0.000	0.000	0.699	0.770	0.000	0.000	0.000	0.000
Horizon	-0.043	-0.004	0.697	0.769	0.187	0.153	0.292	0.252
SCARE	0.003	-0.128	0.721	0.740	0.285	0.031	0.365	0.060
Unified	0.000	0.000	0.710	0.770	0.000	0.000	0.000	0.000

(d) Dataset: **rayyan**

Method	Missing				Typo			
	EDR	Prec.	Rec.	F1	EDR	Prec.	Rec.	F1
Mode	0.011	0.000	0.528	0.563	0.011	0.000	0.021	0.000
Raha-Baran	0.683	0.298	0.845	0.693	0.686	0.303	0.757	0.422
HoloClean	-0.542	-0.635	0.268	0.288	0.009	0.003	0.017	0.006
BigDansing	0.058	0.048	0.551	0.585	0.173	0.169	0.263	0.263
BoostClean	–	0.045	–	0.571	–	0.259	–	0.356
Horizon	0.028	0.043	0.536	0.582	0.125	0.133	0.203	0.217
SCARE	0.182	0.261	0.609	0.677	0.269	0.282	0.373	0.398
Unified	0.262	0.275	0.647	0.683	0.263	0.276	0.374	0.392

表 7: 各清洗方法在两类错误上的综合指标（四数据集并列展示）

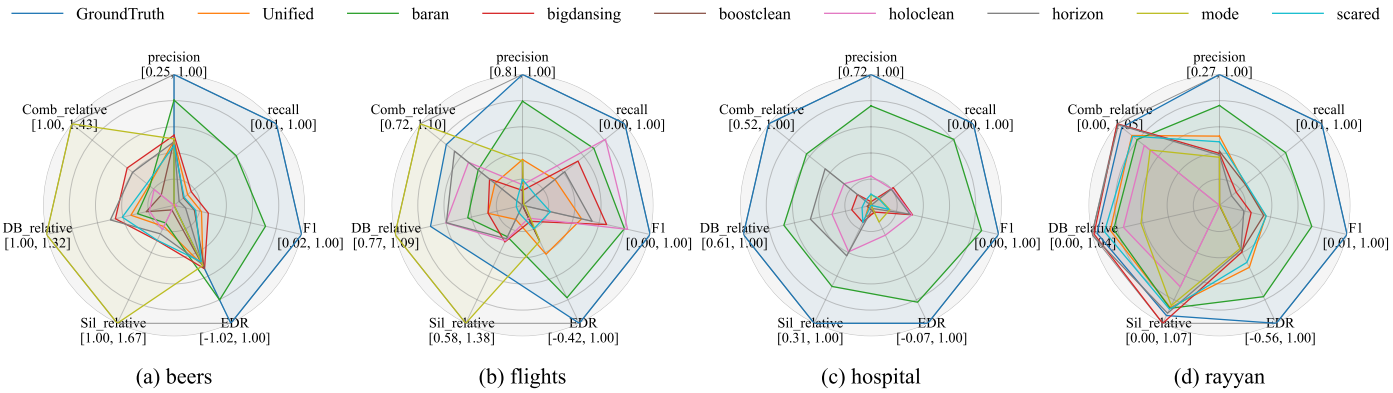
结果与分析

6.4.2 Q_2：清洗干预——过程级视角

本节聚焦“清洗 → 聚类”链条中的运行机理。所有过程级指标均已在 §6.3 的表 6 中统一定义，并通过公式 (6.1) 归一化为“相对 *mode* 填补基线的 $\Delta\%$”。下文不再重复公式或符号，而是直接基于该归一化结果进行比较。

算法族	指标	清洗策略（Δ%，基线 = *mode*）						
		Baran	**Holoclean**	**BigDansing**	**BoostClean**	**Horizon**	**Scared**	**Unified**
质心型	AUC_{Δ}	-19.0	-2.1	-17.8	-10.7	-23.2	-16.8	-24.7
	GeoDecay	0.4	12.0	-2.5	5.2	-3.4	0.1	-2.7
	ΔSSE/NLL	-43.1	-13.0	-36.5	-2.8	-45.3	-27.8	-49.3
密度型	Δn_{core}	-33.3	-1.8	-53.9	-59.5	-55.9	-65.9	-65.7
	Δn_{avg}	4.3	11.9	-1.5	-13.9	7.0	-7.6	-6.7
	$\Delta \rho_{\mathrm{noise}}$	-24.9	-21.5	10.0	39.7	-17.3	16.9	19.1
	ΔW_{cdf}	392.5	363.8	295.5	420.0	356.1	376.3	376.3
层次型	$\Delta n_{\mathrm{merge}}$	-40.4	-3.1	-53.4	-60.2	-58.3	-61.6	-63.5
	$\Delta h_{\max}$	-12.7	-6.0	-7.8	-27.0	-20.2	-31.4	-25.5
	$\Delta R_{\mathrm{intra/inter}}$	13.9	6.5	20.0	46.4	13.9	33.8	24.2

表 8: 清洗策略对过程级指标的相对改变量（Δ%）。数值由公式(6.1)计算；正值表示改善，负值表示退化。



(a) beers (b) flights (c) hospital (d) rayyan

图 5: 归一化七维雷达图——四数据集任务 × 八清洗方法

(a) **跨清洗策略整体分布**

表 8 汇总了 除 *mode* 外的七种清洗方法 在四张数据表上对各过程指标的平均 Δ%。行索引按照“算法族 × 过程指标”展开，列索引为具体清洗策略；正数代表改进、负数表示退化，色块将在最终版本中按同一蓝-红色阶渲染。

(b) **算法族别解析**

- 质心型（*K-Means / GMM*）： ...
- 密度型（*DBSCAN*）： ...
- 层次型（*HC*）： ...

(c) **小结**

将在 §6.4.3 把这些过程层面的发现与聚类结果指标联系起来，验证清洗收益的因果通路，并为后续 AutoML 特征工程提供依据。

6.4.3 Q_3：清洗收益——聚类评价视角

本节按“三层视角”组织实验呈现——综合形状（雷达图）→ 局部相关（散点）→ 全局趋势（CEGR 折线），以回答 Q_3：清洗准确度能否、以及在何种程度上转化为聚类指标收益。

(1) 七维雷达图：同时观察“检测 + 聚类”两类指标

- *beers* 与 *flights.* Baran 在四条检测轴全部外扩，$Comb_{rel}$ 升至 1.35/1.10，并拉高 *beers* 的 Sil_{rel} 至 1.60；BigDansing 仅 Typo 轴小幅提升，其余维度贴近众数填补基线。
- *hospital* 与 *rayyan.* 约束完备的 *hospital* 中，Holoclean 借助 FD/CF 使 Missing / Rule 两轴最外扩，$Comb_{rel} \approx 1.0$；在极高噪声 *rayyan* 里，各方法轮廓趋同，$Comb_{rel}$ 上限仅 1.05——提升 EDR 已难再改善聚类。

小结. Baran 在中低噪声任务展现“检测精度 + 聚类得分”双高外扩；Holoclean 仅在具备完备约束的 *hospital* 发挥优势；当错误率 >30 %（*rayyan*），所有方法雷达面积趋同，修复收益步入饱和区。

(2) 相关散点图：EDR→$\mathbf{Comb_{rel}}$ 与 F1→$\mathbf{Sil_{rel}}$

- *beers* 与 *flights.* F1↛$\mathrm{Sil_{rel}}$ 与 EDR↛$\mathrm{Comb_{rel}}$ 在低 - 中噪声段保持显著正相关：*beers* $r \approx 0.72/0.65$，*flights* 稍弱（$r \approx 0.55$），仅 DBSCAN 出现零星浅红点。
- *hospital* 与 *rayyan. hospital* 呈两极化——K-Means 保持 $r \geq 0.6$，HC / DBSCAN 出现大红圈（负相关），说明规则驱动修复易削弱层次/密度分离度；在噪声最高的 *rayyan*，15% 点深红，F1 与 $\mathrm{Sil_{rel}}$ 可达 $r < -0.3$。

小结. 当错误率 < 20 % 时，清洗准确度与聚类质量整体同涨同跌；噪声攀升或约束失配后，规则型（HC）与密度型（DBSCAN）最先出现负相关，AutoML 应相应下调其权重或收窄搜索空间。

(3) CEGR 折线图：收益随错误率递增的边际变化 CEGR（Clean-Enhanced Gain Ratio）定义为

$$\mathrm{CEGR} \;=\; \frac{\mathrm{Comb}(\mathrm{EDR_{max}}) - \mathrm{Comb}(\mathrm{EDR_{min}})}{\mathrm{EDR_{max}} - \mathrm{EDR_{min}}}, \tag{6.2}$$

并以 5 % 错误率为步长绘制折线。

- *beers* 与 *flights.* 15–25 % 噪声段 CEGR 始终为正：*beers* 的 KMEANS-PPS & KMEANS-NF 曲线由 0.08 升至 0.23（6c），GMM 最终达 0.30；*flights* 多数曲线围绕 0.15–0.20 轻微波动（6f）。说明在“低脏度 + 结构清晰”数据中，只要拉开 EDR，上游收益可线性吸收。
- *hospital* 与 *rayyan. hospital* 在 20–25 % 档出现拐点：K-Means 系列由 0.22 回落至 0.12，HC / DBSCAN 一度跌破 0（6i）。*rayyan* 更严峻——除 GMM 外各算法在 “≥ 30” 档几乎贴近 0，DBSCAN 长期为负（6l）。

小结. CEGR 曲线揭示“清洗投入 → 聚类收益”的边际规律：总体错误率低于约 20 % 时，大多数聚类算法仍能把 EDR 增量转化为正收益；一旦跨过数据依赖阈值（25 % for *hospital*, 30 % for *rayyan*），曲线触顶甚至翻负——此时继续堆叠修复规则性价比低，应转向更鲁棒的聚类策略或先行降维。

直觉一致的发现

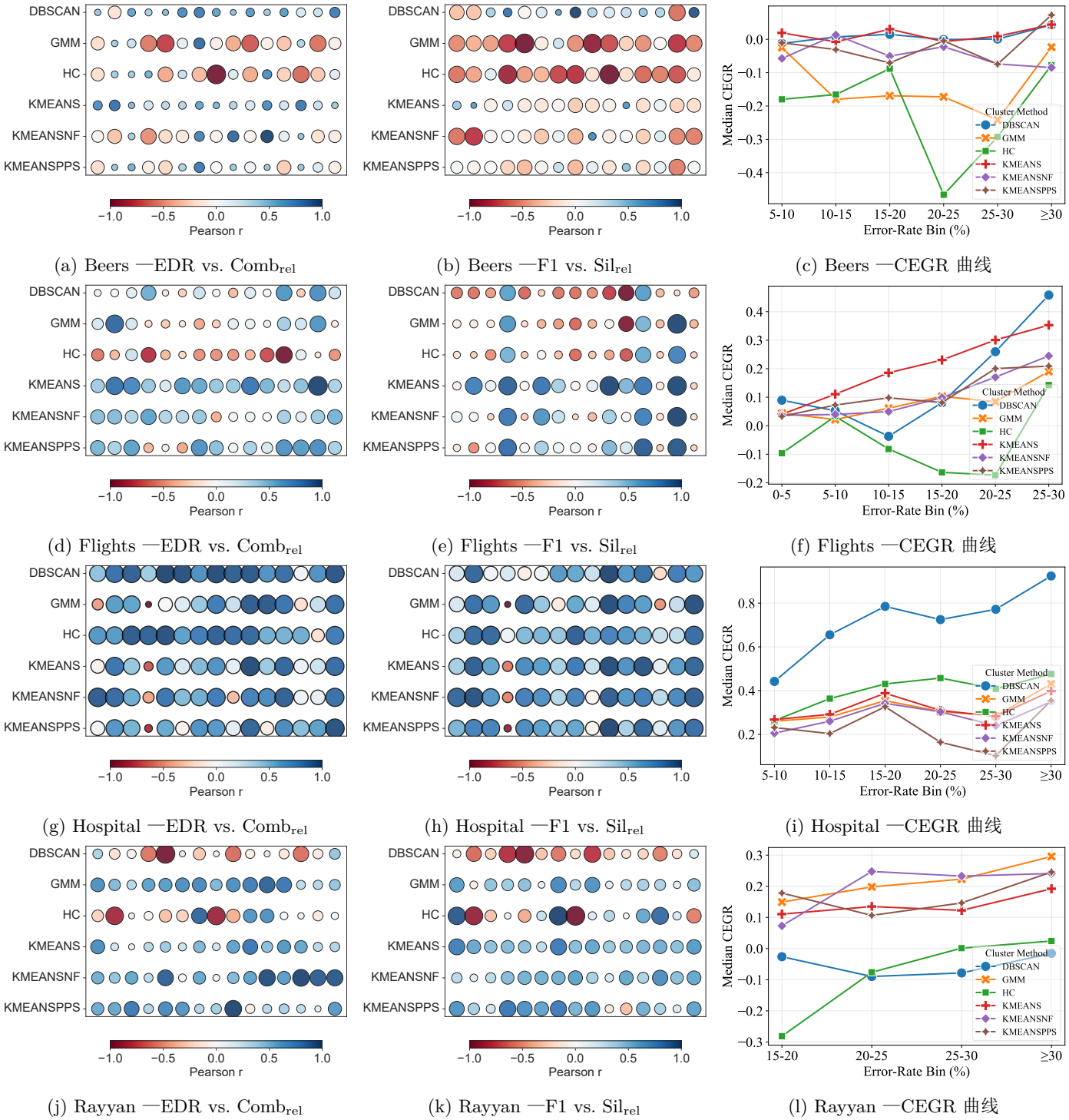
1. 检测-修复越精准，聚类得分越高（低-中噪声段）
在 *beers* 与 *flights*（总体错误率 ≤ 20 %）中，Baran 等高-F1 方法不仅在四条检测轴全部外扩，也同步推高 Silhouette 与 $\mathrm{Comb_{rel}}$；对应散点的 Pearson 系数约 $r \approx 0.7$。说明当脏度可控时，额外的 EDR/F1 增量几乎线性映射到聚类质量提升。

2. 错误率越高，清洗收益越快触顶
CEGR 折线上，当总体错误率突破 ∼ 25 %（*hospital*）或 30 %（*rayyan*）后，各算法曲线先平台而后下滑，甚至转负——再拉大 EDR 上下限已难挽救 Silhouette / Comb。这与“高噪声 ⇒ 边际收益递减”的常识相符。

反直觉的现象

1. 规则驱动清洗可起到反作用
在 *hospital* 的 HC + DBSCAN 组合里，虽然 Holoclean 在 Missing / Rule 轴得分最高，散点相关却出现大红圈（负相关，$r < -0.3$）。过度依赖 FD/CF 的一致化修复压平稀疏离群结构，反而拉低层次/密度聚类质量。

2. 极端高噪声下，高 F1 可能降低分离度
rayyan 的 F1↛$\mathrm{Sil_{rel}}$ 散点中约 15 % 位于深红区。即便检测-修复非常精准（F1 高），过度“平滑”仍会一并消除必要的异常信号，致使簇边界模糊、Silhouette 倒退——与“修得越准越好”的直觉相悖。



(a) Beers —EDR vs. Comb_{rel} (b) Beers —F1 vs. Sil_{rel} (c) Beers —CEGR 曲线

(d) Flights —EDR vs. Comb_{rel} (e) Flights —F1 vs. Sil_{rel} (f) Flights —CEGR 曲线

(g) Hospital —EDR vs. Comb_{rel} (h) Hospital —F1 vs. Sil_{rel} (i) Hospital —CEGR 曲线

(j) Rayyan —EDR vs. Comb_{rel} (k) Rayyan —F1 vs. Sil_{rel} (l) Rayyan —CEGR 曲线

图 6: 四个数据集的“清洗-聚类”相关性与收益趋势：每行同一数据集——左、中的散点图展示 Pearson 相关性；右侧折线图给出随错误率递增的 CEGR 中位数。

6.4.4 超参数选择偏移（正在做）

实验设置 分别在“清洗前后”进行超参数搜索（如 K-Means 中 k，DBSCAN 中 ε 等），记录最优参数及聚类分数，比较其偏移量 Δk 或 $\Delta\varepsilon$。

6.5 讨论：与第五章结果的对照（正在做）

为进一步验证本章实验证据与前述（第 5 章）宏观实验之间的关联，本节从数据集、自动化管线启示两方面展开探讨。

6.5.1 对相同数据集的对照与差异

数据集层面. 将本章结果与第 5 章中相同数据集的聚类表现做对比，检验是否能从本章内部过程或准确度的角度解释某些“爆分”或“收敛异常”现象。若某清洗在第 5 章评测时排名靠前，这里也可展示其收敛曲线或核心点分布更合理。

6.5.2 对自动化管线的启示

自动化搜索层面.

- **清洗准确度可纳入管线特征：**若本章证实 F1/EDR 与聚类指标正相关且稳定，便于日后在自动化过程中更快筛除低准确度的清洗方法。
- **超参数调优的衔接：**若清洗导致显著超参数偏移，提示在自动化工作流程中必须将“清洗-聚类”同步考虑，而非先固定超参数再清洗或反之。

7 结论

本文提出了一种面向数据质量的自动化清洗-聚类优化方法，通过协同优化框架整合数据清洗策略与聚类算法，并利用自动化优化管线缩小搜索空间，以提升聚类效率和质量。研究的主要结论如下：

1. **清洗策略与聚类算法的协同优化是提高聚类质量的关键**。不同清洗-聚类组合在不同数据特征下的适配性差异显著，其中 Raha-Baran + HC 适用于高维、多特征数据，而 mode + DBSCAN 在低维数值数据上可能导致极端分割。
2. **自动化管线有效减少搜索开销，同时保持较高聚类质量**。通过多标签学习建模“数据特征—优选方案子空间”的映射，该方法在平均 5.83 倍加速的情况下，实现了聚类质量 19.20% 的提高，部分数据集在自动化搜索下获得更优结果。
3. **数据特征（如错误率、缺失率、噪声水平）直接影响最优策略的选择**。在高错误率场景下，模式填充（mode）易导致偏差，而 Raha-Baran 在语义受限数据（如医疗、文献分析）中的适配性较优。此外，密度聚类（DBSCAN, OPTICS）对超参数敏感度较高，需要更精细的调优策略。

未来工作 本研究为数据清洗与聚类算法的协同优化提供了理论支持和实验验证，同时为自动化机器学习在无监督场景下的应用拓展了新方向。后续研究可进一步从以下方面优化：

- **数据驱动的自适应清洗策略**

  结合知识图谱、深度学习等方法，提升对复杂数据缺陷（如跨属性错误）的识别与修复能力，确保输入数据的准确性和一致性，为后续聚类优化提供可靠的数据基础。

- **采用更精细的超参数智能调优**

  采用贝叶斯优化、遗传算法等方法，提高聚类算法的稳定性，并增强模型的可解释性。通过智能调优，使密度聚类算法能够适应不同数据分布，减少参数选择对聚类结果的影响。

- **引入更先进的分类模型以优化映射**

  为更准确地捕捉数据特征与最优清洗-聚类组合间的潜在关联，可尝试引入表达能力更强的分类模型（如深度神经网络、集成学习框架等），取代传统多标签或简单判别器。

- **集成最新的聚类算法和评价指标**

  在现有框架中引入近期提出的改进型聚类算法，如自监督聚类、基于图网络的聚类方法等，以提升聚类的泛化能力。同时，结合多种最新的聚类评价指标，如稳定性度量、可解释性分析等，确保模型在不同数据集上的可靠性和适用性。

综上，本文研究表明，清洗-聚类协同优化不仅能够提升数据质量对聚类效果的影响控制能力，还能通过自动化优化方法提升搜索效率，为高噪声、大规模数据环境下的聚类任务提供了可扩展、稳健的解决方案。

参考文献

[1] A. Aljohani, "Optimizing patient stratification in healthcare: A comparative analysis of clustering algorithms for ehr data," *International Journal of Computational Intelligence Systems*, vol. 17, no. 1, p. 173, July 2024. [Online]. Available: https://doi.org/10.1007/s44196-024-00568-8

[2] J. L. Leevy, Z. Salekshahrezaee, and T. M. Khoshgoftaar, "A review of unsupervised anomaly detection techniques for health insurance fraud," in *2024 IEEE 10th International Conference on Big Data Computing Service and Machine Learning Applications (BigDataService)*, 2024, pp. 141–149.

[3] J. Passlick, S. Dreyer, D. Olivotti, L. Grützner, D. Eilers, and M. H. Breitner, "Predictive maintenance as an internet of things enabled business model: A taxonomy," *Electronic Markets*, vol. 31, no. 1, pp. 67–87, March 2021. [Online]. Available: https://doi.org/10.1007/s12525-020-00440-5

[4] A. J. Jabiyeva, "State of the art of big data analytics and clustering algorithms in biomedicine," in *16th International Conference on Applications of Fuzzy Systems, Soft Computing and Artificial Intelligence Tools – ICAFS-2023*, R. A. Aliev, J. Kacprzyk, W. Pedrycz, M. Jamshidi, M. Babanli, and F. M. Sadikoglu, Eds. Cham: Springer Nature Switzerland, 2025, pp. 228–234.

[5] F. Cai, N.-A. Le-Khac, and T. Kechadi, "Clustering approaches for financial data analysis: a survey," *arXiv preprint arXiv:1609.08520*, 2016. [Online]. Available: https://doi.org/10.48550/arXiv.1609.08520

[6] S. Bandyapadhyay, Z. Friggstad, and R. Mousavi, "Parameterized approximation algorithms and lower bounds for k-center clustering and variants," *Algorithmica*, vol. 86, no. 8, pp. 2557–2574, August 2024. [Online]. Available: https://doi.org/10.1007/s00453-024-01236-1

[7] T. Barton, T. Bruna, and P. Kordik, "Chameleon 2: An improved graph-based clustering algorithm," *ACM Trans. Knowl. Discov. Data*, vol. 13, no. 1, Jan. 2019. [Online]. Available: https://doi.org/10.1145/3299876

[8] X. Chu, J. Morcos, I. F. Ilyas, M. Ouzzani, P. Papotti, N. Tang, and Y. Ye, "Katara: A data cleaning system powered by knowledge bases and crowdsourcing," in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 1247–1261. [Online]. Available: https://doi.org/10.1145/2723372.2749431

[9] M. Dallachiesa, A. Ebaid, A. Eldawy, A. Elmagarmid, I. F. Ilyas, M. Ouzzani, and N. Tang, "Nadeef: a commodity data cleaning system," in *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '13. New York, NY, USA: Association for Computing Machinery, 2013, p. 541–552. [Online]. Available: https://doi.org/10.1145/2463676.2465327

[10] T. Rekatsinas, X. Chu, I. F. Ilyas, and C. Ré, "Holoclean: holistic data repairs with probabilistic inference," *Proc. VLDB Endow.*, vol. 10, no. 11, p. 1190–1201, Aug. 2017. [Online]. Available: https://doi.org/10.14778/3137628.3137631

[11] S. Alam, M. S. Ayub, S. Arora, and M. A. Khan, "An investigation of the imputation techniques for missing values in ordinal data enhancing clustering and classification analysis validity," *Decision Analytics Journal*, vol. 9, p. 100341, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2772662223001819

[12] W. Ni, X. Miao, X. Zhao, Y. Wu, S. Liang, and J. Yin, "Automatic data repair: Are we ready to deploy?" *Proc. VLDB Endow.*, vol. 17, no. 10, p. 2617–2630, Jun. 2024. [Online]. Available: https://doi.org/10.14778/3675034.3675051

[13] J. Singh and D. Singh, "A comprehensive review of clustering techniques in artificial intelligence for knowledge discovery: Taxonomy, challenges, applications and future prospects," *Advanced Engineering Informatics*, vol. 62, p. 102799, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1474034624004476

[14] L. Blumenberg and K. V. Ruggles, "Hypercluster: a flexible tool for parallelized unsupervised clustering optimization," *BMC Bioinformatics*, vol. 21, no. 1, p. 428, September 2020. [Online]. Available: https://doi.org/10.1186/s12859-020-03774-1

[15] R. Barbudo, S. Ventura, and J. R. Romero, "Eight years of automl: categorisation, review and trends," *Knowledge and Information Systems*, vol. 65, no. 12, pp. 5097–5149, December 2023. [Online]. Available: https://doi.org/10.1007/s10115-023-01935-1

[16] I. Salehin, M. S. Islam, P. Saha, S. Noman, A. Tuni, M. M. Hasan, and M. A. Baten, "Automl: A systematic review on automated machine learning with neural architecture search," *Journal of Information and Intelligence*, vol. 2, no. 1, pp. 52–81, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2949715923000604

[17] X. He, K. Zhao, and X. Chu, "Automl: A survey of the state-of-the-art," *Knowledge-Based Systems*, vol. 212, p. 106622, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705120307516

[18] P. Li, X. Rao, J. Blase, Y. Zhang, X. Chu, and C. Zhang, "Cleanml: A study for evaluating the impact of data cleaning on ml classification tasks," in *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, 2021, pp. 13–24.

[19] Y. Poulakis, C. Doulkeridis, and D. Kyriazis, "A survey on automl methods and systems for clustering," *ACM Trans. Knowl. Discov. Data*, vol. 18, no. 5, Feb. 2024. [Online]. Available: https://doi.org/10.1145/3643564

[20] D. J. Stekhoven and P. Bühlmann, "Missforest—non-parametric missing value imputation for mixed-type data," *Bioinformatics*, vol. 28, no. 1, pp. 112–118, 10 2011. [Online]. Available: https://doi.org/10.1093/bioinformatics/btr597

[21] G. Beskales, I. F. Ilyas, L. Golab, and A. Galiullin, "On the relative trust between inconsistent data and inaccurate constraints," in *2013 IEEE 29th International Conference on Data Engineering (ICDE)*, 2013, pp. 541–552.

[22] F. Chiang and R. J. Miller, "A unified model for data and constraint repair," in *2011 IEEE 27th International Conference on Data Engineering*, 2011, pp. 446–457.

[23] C. Ge, Y. Gao, X. Miao, B. Yao, and H. Wang, "A hybrid data cleaning framework using markov logic networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 5, pp. 2048–2062, 2022.

[24] S. Krishnan, J. Wang, E. Wu, M. J. Franklin, and K. Goldberg, "Activeclean: interactive data cleaning for statistical modeling," *Proc. VLDB Endow.*, vol. 9, no. 12, p. 948 – 959, Aug. 2016. [Online]. Available: https://doi.org/10.14778/2994509.2994514

[25] F. Neutatz, M. Mahdavi, and Z. Abedjan, "Ed2: A case for active learning in error detection," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, ser. CIKM '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 2249–2252. [Online]. Available: https://doi.org/10.1145/3357384.3358129

[26] S. Krishnan, M. J. Franklin, K. Goldberg, and E. Wu, "Boostclean: Automated error detection and repair for machine learning," *arXiv preprint arXiv:1711.01299*, 2017. [Online]. Available: https://doi.org/10.48550/arXiv.1711.01299

[27] X. Chu, I. F. Ilyas, and P. Papotti, "Holistic data cleaning: Putting violations into context," in *2013 IEEE 29th International Conference on Data Engineering (ICDE)*, 2013, pp. 458–469.

[28] M. Mahdavi and Z. Abedjan, "Baran: effective error correction via a unified context representation and transfer learning," *Proc. VLDB Endow.*, vol. 13, no. 12, p. 1948–1961, Jul. 2020. [Online]. Available: https://doi.org/10.14778/3407790.3407801

[29] M. Bernhardt, D. C. Castro, R. Tanno *et al.*, "Active label cleaning for improved dataset quality under resource constraints," *Nature Communications*, vol. 13, p. 1161, March 2022. [Online]. Available: https://doi.org/10.1038/s41467-022-28818-3

[30] W. Hu, A. Zaveri, H. Qiu *et al.*, "Cleaning by clustering: methodology for addressing data quality issues in biomedical metadata," *BMC Bioinformatics*, vol. 18, p. 415, September 2017. [Online]. Available: https://doi.org/10.1186/s12859-017-1832-4

[31] S. Huang, Z. Kang, Z. Xu, and Q. Liu, "Robust deep k-means: An effective and simple method for data clustering," *Pattern Recognition*, vol. 117, p. 107996, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320321001837

[32] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," *Information Sciences*, vol. 622, pp. 178–210, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025522014633

[33] T. Zaki Abdulhameed, S. A. Yousif, V. W. Samawi, and H. Imad Al-Shaikhli, "Ss-dbscan: Semi-supervised density-based spatial clustering of applications with noise for meaningful clustering in diverse density data," *IEEE Access*, vol. 12, pp. 131 507–131 520, 2024.

[34] D. Cheng, C. Zhang, Y. Li, S. Xia, G. Wang, J. Huang, S. Zhang, and J. Xie, "Gb-dbscan: A fast granular-ball based dbscan clustering algorithm," *Information Sciences*, vol. 674, p. 120731, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025524006443

[35] M. Hajihosseinlou, A. Maghsoudi, and R. Ghezelbash, "A comprehensive evaluation of optics, gmm and k-means clustering methodologies for geochemical anomaly detection connected with sample catchment basins," *Geochemistry*, vol. 84, no. 2, p. 126094, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0009281924000187

[36] C. Tang, H. Wang, Z. Wang, X. Zeng, H. Yan, and Y. Xiao, "An improved optics clustering algorithm for discovering clusters with uneven densities," *Intell. Data Anal.*, vol. 25, no. 6, p. 1453–1471, Jan. 2021. [Online]. Available: https://doi.org/10.3233/IDA-205497

[37] I. S. Kamil and S. O. Al-Mamory, "Enhancement of optics' time complexity by using fuzzy clusters," *Materials Today: Proceedings*, vol. 80, pp. 2625–2630, 2023, sI:5 NANO 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2214785321048355

[38] Z. Chen, J. Feng, D. Yang, and F. Cai, "Hierarchical clustering algorithm based on crystallized neighborhood graph for identifying complex structured datasets," *Expert Systems with Applications*, vol. 265, p. 125714, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417424025818

[39] M. Atif, M. Farooq, M. Shafiq *et al.*, "Uncovering the impact of outliers on clusters' evolution in temporal data-sets: an empirical analysis," *Scientific Reports*, vol. 14, p. 30674, 2024. [Online]. Available: https://doi.org/10.1038/s41598-024-75928-7

[40] H. Guo, H. Yin, S. Song *et al.*, "Application of density clustering with noise combined with particle swarm optimization in uwb indoor positioning," *Scientific Reports*, vol. 14, p. 13121, 2024. [Online]. Available: https://doi.org/10.1038/s41598-024-63358-4

[41] Y. Lai, S. He, Z. Lin, F. Yang, Q. Zhou, and X. Zhou, "An adaptive robust semi-supervised clustering framework using weighted consensus of random kk-means ensemble," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 5, pp. 1877–1890, 2021.

[42] M. Feurer, K. Eggensperger, S. Falkner, M. Lindauer, and F. Hutter, "Auto-sklearn 2.0: hands-free automl via meta-learning," *J. Mach. Learn. Res.*, vol. 23, no. 1, Jan. 2022.

[43] R. S. Olson and J. H. Moore, *TPOT: A Tree-Based Pipeline Optimization Tool for Automating Machine Learning*. Cham: Springer International Publishing, 2019, pp. 151–160. [Online]. Available: https://doi.org/10.1007/978-3-030-05318-5_8

[44] J. D. Romano, T. T. Le, W. Fu, and J. H. Moore, "Tpot-nn: augmenting tree-based automated machine learning with neural network estimators," *Genetic Programming and Evolvable Machines*, vol. 22, no. 2, pp. 207–227, June 2021. [Online]. Available: https://doi.org/10.1007/s10710-021-09401-z

[45] A. Singh, N. Prakash, and A. Jain, "Chronic diseases prediction using two different pipelines tpot and genetic algorithm based models: A comparative analysis," in *Proceedings of the 2024 9th International Conference on Machine Learning Technologies*, ser. ICMLT '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 175–180. [Online]. Available: https://doi.org/10.1145/3674029.3674058

[46] Y. Poulakis, C. Doulkeridis, and D. Kyriazis, "Autoclust: A framework for automated clustering based on cluster validity indices," in *2020 IEEE International Conference on Data Mining (ICDM)*, 2020, pp. 1220–1225.

[47] R. ElShawi, H. Lekunze, and S. Sakr, "csmartml: A meta learning-based framework for automated selection and hyperparameter tuning for clustering," in *2021 IEEE International Conference on Big Data (Big Data)*, 2021, pp. 1119–1126.

[48] R. ElShawi and S. Sakr, "csmartml-glassbox: Increasing transparency and controllability in automated clustering," in *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, 2022, pp. 47–54.

[49] Y. Liu, S. Li, and W. Tian, "Autocluster: Meta-learning based ensemble method for automated unsupervised clustering," in *Advances in Knowledge Discovery and Data Mining*, K. Karlapalem, H. Cheng, N. Ramakrishnan, R. K. Agrawal, P. K. Reddy, J. Srivastava, and T. Chakraborty, Eds. Cham: Springer International Publishing, 2021, pp. 246–258.

[50] R. ElShawi and S. Sakr, "Tpe-autoclust: A tree-based pipline ensemble framework for automated clustering," in *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, 2022, pp. 1144–1153.

[51] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, 1979.

[52] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0377042787901257

[53] O. Bachem, M. Lucic, S. H. Hassani, and A. Krause, "Approximate k-means++ in sublinear time," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, ser. AAAI'16. AAAI Press, 2016, p. 1459–1467.

[54] F. Nie, Z. Li, R. Wang, and X. Li, "An effective and efficient algorithm for k-means clustering with new formulation," *IEEE Trans. on Knowl. and Data Eng.*, vol. 35, no. 4, p. 3433–3443, Apr. 2023. [Online]. Available: https://doi.org/10.1109/TKDE.2022.3155450

[55] Z. Khayyat, I. F. Ilyas, A. Jindal, S. Madden, M. Ouzzani, P. Papotti, J. Quiané-Ruiz, N. Tang, and S. Yin, "BigDansing: A system for big data cleansing," in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. Melbourne, Australia: ACM, 2015, pp. 1215–1230. [Online]. Available: https://doi.org/10.1145/2723372.2747646

[56] S. Krishnan, M. J. Franklin, K. Goldberg, and E. Wu, "BoostClean: Automated error detection and repair for machine learning," *CoRR*, vol. abs/1711.01299, 2017. [Online]. Available: https://arxiv.org/abs/1711.01299

[57] E. K. Rezig, M. Ouzzani, W. G. Aref, A. K. Elmagarmid, A. R. Mahmood, and M. Stonebraker, "Horizon: scalable dependency-driven data cleaning," *Proc. VLDB Endow.*, vol. 14, no. 11, p. 2546–2554, Jul. 2021. [Online]. Available: https://doi.org/10.14778/3476249.3476301

[58] M. Yakout, L. Berti-Équille, and A. K. Elmagarmid, "Don't be scared: use scalable automatic repairing with maximal likelihood and bounded changes," in *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '13. New York, NY, USA: Association for Computing Machinery, 2013, p. 553–564. [Online]. Available: https://doi.org/10.1145/2463676.2463706