结合自动参数调优与数据清洗策略的聚类性能优化：跨多数据集的对比研究

摘要

真实世界数据常包含噪声、缺失值以及其他数据质量问题，这些因素显著影响了聚类分析的效果。为改善这一情况，本研究对不同数据清洗策略（包括众数填补、Raha-Baran 算法及理想状态下的 GroundTruth）与 6 种常见聚类算法的组合表现进行深入探索，并利用 Optuna 方法对聚类过程实施自动化超参数调优。我们在 40 个具备不同数据类型与错误比率的真实数据集上对每种方案进行系统性的测试与评估。实验结果表明，以 Raha-Baran 或 Mode 策略进行数据清洗后搭配层次聚类（HC）算法，往往能在轻度噪声或缺失值环境中取得接近甚至优于参考基准的聚类效果；同时，该组合方案在多样化的数据条件下展现出了较强的稳定性与适应性。本研究结果为在不完美数据条件下提升聚类性能提供了实用参考和有效策略。

关键词：数据质量，数据清洗，聚类优化，AutoML，缺失值处理

1　引言

背景。在数据挖掘与机器学习领域，聚类分析作为一种重要的无监督学习方法，被广泛应用于客户细分、图像分割、生物信息学等多个领域。然而，现实数据通常包含噪声、缺失值等质量缺陷，这些问题会干扰数据的整体分布，进而影响聚类算法对簇结构的正确识别。此外，不同数据集的特性（如特征类型、噪声水平、缺失比率），聚类算法参数的设置，检测指标的不确定性也进一步增加了聚类任务的复杂程度。因此，为了在数据质量受损的情况下提升聚类性能，选择合适的数据清洗策略和优化聚类算法的参数显得尤为关键。

研究动机。尽管已有研究分别关注了数据清洗和聚类算法性能的优化，但大多数工作将二者分开进行分析和比较，缺乏对多数据集、多清洗策略与自动化参数优化方法的系统性研究。此外，目前针对自动化参数调优（如 AutoML 工具 Optuna）与数据清洗策略的协同效应尚缺乏深入探讨。因此，需要一个全面的研究框架来评估数据清洗与聚类算法优化结合的性能，并验证其在多种数据条件下的适用性。

研究目标。本研究的目标是通过结合多种数据清洗策略（如众数填补、Raha-Baran 算法、GroundTruth 基准）与自动化超参数调优技术（Optuna），系统探索它们对聚类性能的影响。我们通过跨多个真实数据集的实验分析，研究这些清洗与聚类优化组合策略在不同数据质量条件下的表现，从而为数据质量受损情况下的高效聚类提供可靠的

参考和解决方案。

研究贡献

本文的主要贡献包括：

- 系统评估多种数据清洗与聚类算法组合策略，涵盖从简单到复杂的清洗方法与多种常用聚类算法。
- 引入 Optuna 框架，实现聚类算法的自动化超参数调优，适应不同数据条件的最优配置。
- 在 40 个具有不同数据类型、噪声比例和缺失比率的真实数据集上开展实验，验证清洗与优化组合策略的适用性与有效性，为数据质量受损条件下的聚类提供实用指导。

文章结构

本文其余部分安排如下：

- 第二部分定义数据质量问题、清洗策略及评价指标；
- 第三部分回顾数据清洗方法、聚类算法及自动化调优技术；
- 第四部分介绍实验设计、测试流程及结果分析；
- 第五部分探讨实验发现及其理论依据；
- 第六部分总结相关研究并指出创新点；
- 第七部分总结研究结论并讨论未来研究方向。

2 问题定义

2.1 数据质量问题

本研究基于 40 个具有不同特征类型和数据质量问题的真实数据集开展实验，这些数据集涵盖数值型、字符型、混合型等多种类型，具有广泛的代表性。每个数据集的基本统计特征包括样本数和特征数，分别反映数据的规模与复杂程度；扩展统计特征包括缺失率、噪声比例和错误率，这些特征的形式化定义如下：

错误率：衡量数据集中错误单元占总单元的比例，其定义如下：

$$\text{Error Rate} = \frac{\text{错误单元数量}}{\text{总单元数量}} \times 100\%$$

错误类型包括：

- 缺失值：单元值为空。
- 语法错误：单元值不符合格式规则（如日期为“2023-32-01”）。
- 语义错误：单元值与上下文逻辑矛盾（如“性别”为“男”，但职业为“女演员”）。
- 知识库错误：单元值违反领域知识约束（如国家“美国”但货币为“欧元”）。

缺失率：表示数据集中缺失值的比例，量化缺失数据的严重性，其定义如下：

$$\text{Missing Value Ratio} = \frac{\text{缺失单元数量}}{\text{总单元数量}} \times 100\%$$

虽然缺失值在严格意义上可以归类为错误值的一种，但由于其“N/A”特性可能对聚类性能造成特殊影响（例如影响距离计算、相似度度量或簇中心的确定），因此本研究将缺失值单独列出并进行独立讨论。

噪声比例：噪声比例表示数据集中异常值占总数据单元的比例，用以模拟数据污染问题，其定义如下：

$$\text{Noise Ratio} = \frac{\text{噪声单元数量}}{\text{总单元数量}} \times 100\%$$

通过这些指标，能够系统地量化数据质量问题并为后续实验设计提供明确的依据。

2.2 数据生成与预处理

为了模拟真实世界中的数据质量问题，本研究以一组干净数据集为基础，注入多种错误类型生成不同质量条件的数据集。错误注入的过程包括以下几步：

1. 缺失值注入：按照预设的缺失率随机选择数据单元，将其值设置为空，以模拟数据丢失问题。
2. 语法错误注入：随机选择数据单元更改其格式，使其不符合预期规则（如“100.00”改为“$USD100”）。
3. 语义错误注入：更改单元格内容，使其与上下文不一致（如“年龄”为“-25”）。
4. 知识库错误注入：将单元值更改为不符合领域知识的内容（如“国家”为“火星”）。
5. 噪声注入：对数值型字段随机添加偏移或高斯噪声，生成异常值。

注入的错误比例由错误率、缺失率和噪声比例共同控制，生成后的数据集具备多样化的质量问题，为后续数据清洗与聚类优化策略的评估提供了真实可信的测试对象。

2.3 评价指标

为了全面评估聚类算法的性能，本研究引入了以下几个评价指标

Davies-Bouldin Score：衡量簇内紧致度和簇间分离度，值越低表示聚类效果越好。定义如下：

$$\text{DB\_score} = \frac{1}{N}\sum_{i=1}^{N}\max_{j\neq i}\left(\frac{s_i + s_j}{d_{ij}}\right)$$

其中 N 是簇的总数。s_i 表示簇 i 的平均直径。d_{ij} 表示簇 i 和簇 j 的中心之间的距离。

Silhouette_score：衡量每个点聚类结果的合理性，取值范围为 $[-1,1]$，值越高表示聚类效果越好。定义如下：

$$\text{Silhouette\_score} = \frac{1}{N}\sum_{i=1}^{N}\frac{b_i - a_i}{\max(a_i, b_i)}$$

其中 a_i 表示样本点 i 到其所属簇内其他样本点的平均距离。b_i 表示样本点 i 到最近簇的样本点平均距离。

综合评分：根据以下公式计算综合得分，全面衡量聚类性能，
α 和 β 可以根据需要进行选择。

$$\text{Combined\_Score} = \frac{1}{\text{DB\_score}} \times \alpha + \text{Silhouette\_score} \times \beta$$

百分比综合得分：将综合评分标准化为百分比形式，用来衡量某个算法组合相对于基准值效果，其定义如下：

$$\text{Normalized Score} = \frac{\text{Combined\_Score}}{\text{Reference Score}} \times 100\%$$

2.4 优化目标

本研究的优化目标是通过数据清洗策略与聚类算法的组合实现聚类性能的全面提升，并探索其适用性与协同作用。具体目标如下：

自动化超参数调优。聚类算法的性能高度依赖于关键超参数的合理设置，例如

KMeans 的簇数量 k、DBSCAN 的 ϵ 与 min_samples。本研究利用 Optuna 框架对各聚类算法进行自动化超参数调优，以适应不同数据集和清洗策略的特性，从而提升局部性能，为后续选择最佳方案奠定基础。

提升综合性能。在完成超参数调优后，本研究对多种数据清洗策略（如众数填补、Raha-Baran 算法和 GroundTruth 基准）与聚类算法的组合进行全局性能评估。综合评分基于 DB_score 和 Silhouette_score 两个核心指标，衡量每种组合在不同数据质量条件下的适应性和稳定性，进而筛选出性能最优的方案。

多条件下的适用性验证。在获得最佳组合方案后，本研究进一步探索其在多样化数据质量条件下的适用性与协同性。通过分析最佳方案在不同数据质量问题（如高噪声、高缺失率）下的表现，研究数据清洗与聚类优化的协同作用机制，筛选出在特定数据条件下实现聚类性能最大化的算法组合策略。

3 算法回顾（还没有具体写，以下是内容安排）

- **数据清洗策略：**
    1. **Mode 填补：**基于众数填充方法的原理与应用场景。
    2. **Raha-baran 算法：**清洗异常值与缺失值的技术细节与优缺点。
    3. **GroundTruth：**作为理想参考清洗方法的描述。
- **自动化参数优化（自行修改后的 Optuna，附上参考文献和源代码）：**
    - Optuna 的基本工作原理（贝叶斯优化/TPE）。
    - 搜索空间定义与调优流程示意。
- **聚类算法：** 简要介绍使用的聚类算法及其调优的关键参数（具体怎么优化的，结合程序）：
    - **层次聚类（HC）：**连结方式与性能特点。
    - **K 均值（KMeans）：**k 值选择的影响。
    - **高斯混合模型（GMM）：**成分数与协方差类型。
    - **DBSCAN 与 OPTICS：**密度参数（epsilon、min_samples）的敏感性。
    - **AffinityPropagation：**对初始参数的依赖。

4 实验评估

4.1 实验设置

4.1.1 数据集准备与清洗策略

本研究实验基于 40 个具有多样化特征的真实且公开的数据集，这些数据集涵盖数值型、字符型以及混合型数据，规模从小样本到中等规模数据集不等。通过选择具有不同缺失率、噪声比例和错误类型的多样化数据集，确保实验结果的广泛适用性。

以下是本次实验的数据集详细信息：

Dataset Name	Error Rate (%)	Samples	Features	Missing Value Ratio	Data Types Distribution
beers	**6.1700**	2410.0000	11.0000	**0.0626**	{dtype('float64'): 8, dtype('int64'): 3}
beers	**12.4800**	2410.0000	11.0000	**0.0965**	
beers	**20.3400**	2410.0000	11.0000	**0.0961**	
beers	**26.6000**	2410.0000	11.0000	**0.0737**	
beers	**29.7800**	2410.0000	11.0000	**0.1239**	
beers	31.7300	2410.0000	11.0000	0.2070	
beers	**40.2300**	2410.0000	11.0000	**0.1788**	
beers	**45.0800**	2410.0000	11.0000	**0.3183**	
beers	**46.7300**	2410.0000	11.0000	**0.2344**	
beers	**52.0000**	2410.0000	11.0000	**0.2900**	
beers	0.0000	2410.0000	11.0000	0.0404	
flights	8.6900	2376.0000	7.0000	0.0499	{dtype('float64'): 6, dtype('int64'): 1}
flights	16.9700	2376.0000	7.0000	0.0998	
flights	**24.5600**	2376.0000	7.0000	**0.1497**	
flights	**30.6300**	2376.0000	7.0000	**0.1750**	
flights	**38.9200**	2376.0000	7.0000	**0.2498**	
flights	**40.7600**	2376.0000	7.0000	**0.2447**	
flights	45.4400	2376.0000	7.0000	0.2997	
flights	**51.5900**	2376.0000	7.0000	**0.3499**	
flights	62.8700	2376.0000	7.0000	0.4497	
flights	**67.7400**	2376.0000	7.0000	**0.4999**	
flights	0.0000	2376.0000	7.0000	0.0000	
hospital	**8.5300**	1000.0000	20.0000	**0.0190**	{dtype('float64'): 19, dtype('int64'): 1}
hospital	**11.9600**	1000.0000	20.0000	**0.0380**	
hospital	**15.3400**	1000.0000	20.0000	**0.0570**	
hospital	**21.6500**	1000.0000	20.0000	**0.0950**	
hospital	**24.8300**	1000.0000	20.0000	**0.1140**	
hospital	**27.9600**	1000.0000	20.0000	**0.1330**	
hospital	**33.6800**	1000.0000	20.0000	**0.1710**	
hospital	36.5200	1000.0000	20.0000	0.1900	
hospital	**46.5200**	1000.0000	20.0000	**0.1900**	
hospital	49.2900	1000.0000	20.0000	0.2850	
hospital	0.0000	1000.0000	20.0000	0.0000	
rayyan	**10.7500**	1000.0000	11.0000	**0.1544**	{dtype('float64'): 10, dtype('int64'): 1}
rayyan	**13.7900**	1000.0000	11.0000	**0.1719**	
rayyan	**16.8800**	1000.0000	11.0000	**0.1888**	
rayyan	**19.7100**	1000.0000	11.0000	**0.2029**	
rayyan	**22.7700**	1000.0000	11.0000	**0.2214**	
rayyan	**24.3500**	1000.0000	11.0000	**0.2313**	
rayyan	29.2500	1000.0000	11.0000	0.2630	
rayyan	**40.2400**	1000.0000	11.0000	**0.2935**	
rayyan	**47.8800**	1000.0000	11.0000	**0.3993**	
rayyan	**52.7300**	1000.0000	11.0000	**0.3935**	
rayyan	0.0000	1000.0000	11.0000	0.1461	

表 4.1 本实验所用数据集的详细质量问题统计信息

4.1.2 清洗策略和聚类算法的准备 [1]

实验中采用了三种数据清洗策略：

- **Mode 填补**：通过对缺失值使用众数填补来模拟最简单的数据修复情形，同时利用众数修复少量易修正的错误。在实验中，Mode 填补作为基础清洗策略，用于评估低复杂度清洗方法在轻度数据问题下的适用性。
- **Raha-Baran**：结合上下文推理与规则匹配的高级数据清洗方法，能够处理复杂错误（如语义冲突和知识库错误）。在实验中，Raha-Baran 用于模拟深度优化的清洗策略，以评估高性能清洗算法在复杂数据质量条件下的表现。
- **GroundTruth**：假设对数据进行完美清洗，所有噪声、缺失值和错误均已修复，提供理论上的最优基准作为实验参考，用于评估其他清洗策略与聚类组合的性能差距。

实验中包含以下 6 种常见聚类算法：HC，KMeans，GMM，DBSCAN，OPTICS，AffinityPropagation，每种清洗策略与聚类算法的组合运行后，记录其生成的聚类簇数和对应评价指标的计算结果。

本实验评估了 **3 种清洗策略 × 6 种聚类算法** 的 18 种组合策略，并在 40 个具备不同数据特性和质量问题的数据集上进行测试。通过这一系统性设计，实验能够全面评估清洗策略与聚类算法在多样化数据条件下的适配性。

4.1.3 实验指标与得分标准

簇数量。簇数量是判断聚类结果合理性的重要标准。过多或过少的簇数可能导致结果失真或难以解释，因此本研究对簇数量设置了以下范围限制：

- **标准范围**：簇数量一般不小于 5 个且一般不大于样本数的算数平方根。
- **筛选规则**：若实验结果中的簇数量明显偏离此范围（例如簇数过少导致过度聚合，或簇数过多导致过度分散），则将其标记为不合法结果并从分析中排除。

综合得分。在本实验中，综合得分的权重设置为 $\alpha=0.75, \beta=0.25$. 这一权重的选择是根据抽样检测与预实验的结果：当轮廓系数的权重过高时，结果倾向于生成极少的簇（如 2000 个元组仅分为 2 个簇），与实际场景需求不符。因此，降低 β 的占比，有助于生成更接近实际应用需求且较为稳定的簇数量。

百分比得分。为了便于比较不同算法组合之间的性能，本研究对综合得分进行百分比

[1] 清洗策略和聚类算法的详细工作原理及适用场景见第三章

归一化处理，以 GroundTruth 清洗策略修复后的最高综合得分为基准，将其定义为 100%。对于部分得分超过 100% 的结果（如数据清洗策略特别适合某些数据集的特殊分布），实验中予以保留并单独分析[2]。不合理的实验结果（如算法运行超时、不收敛或簇数明显偏离标准范围）被标记为 0%，以避免对整体结果分析的干扰。

4.2 实验步骤

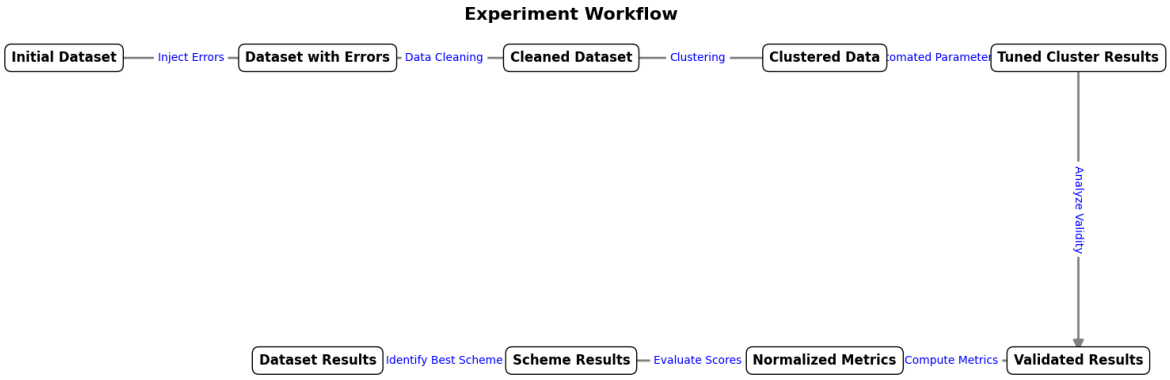

图 4.2.1 实验完整流程图

// 这里可以具体写一下每个步骤的任务，后续补上。

4.3 实验统计结果

图 4.3.1 展示了 *beers*、*flights*、*rayyan* 和 *hospital* 数据集在不同错误率条件下的得分与缺失值比例的对比情况，图表结构如下：

[2] 本分析详见 5.1

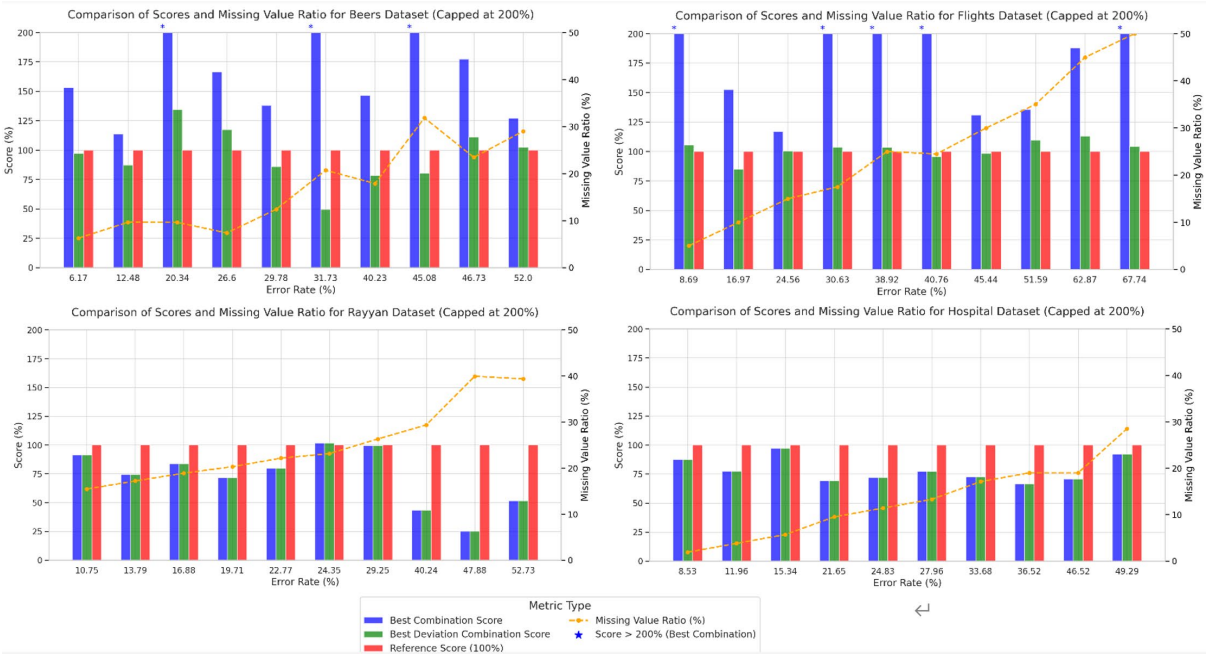

图 4.3.1 各数据集聚类性能得分与缺失值比例随错误率变化图像

表 4.3.2 列出了不同数据集在多种错误率 (Error Rate) 条件下获得的“最佳组合 (Best Combination)”及其综合得分（Best Combination Score %，简记为 Score_1），同时给出“最接近基准 (Best Deviation Combination)”及对应分数（Best Deviation Combination Score %，简记为 Score_2）。所有结果都以 GroundTruth + HC（100%）作为参考基准。

Dataset Name	Error Rate (%)	Missing Value Ratio	Best Combination	Score_1 (%)	Best Deviation Combination	Score_2 (%)
beers	**6.1700**	**0.0626**	**mode + HC**	153.1105	raha-baran + HC	97.2143
beers	**12.4800**	**0.0965**	**raha-baran + HC**	113.6862	mode + HC	87.3677
beers	**20.3400**	**0.0961**	**mode + HC**	259.8521	raha-baran + HC	134.4963
beers	**26.6000**	**0.0737**	**mode + HC**	166.5351	mode + AffinityPropagation	117.4489
beers	**29.7800**	**0.1239**	**raha-baran + HC**	138.0655	mode + HC	86.1352
beers	31.7300	0.2070	mode + AffinityPropagation	232.9467	mode + DBSCAN	49.4929
beers	**40.2300**	**0.1788**	**mode + HC**	146.5448	mode + KMeans	78.3762
beers	**45.0800**	**0.3183**	**mode + HC**	208.1525	raha-baran + HC	80.1914
beers	**46.7300**	**0.2344**	**mode + HC**	177.3738	mode + AffinityPropagation	111.1102
beers	**52.0000**	**0.2900**	**mode + HC**	127.0622	mode + DBSCAN	102.5086
flights	**8.6900**	**0.0499**	**mode + HC**	290.4399	mode + KMeans	105.3636
flights	16.9700	0.0998	mode + GMM	152.6073	raha-baran + HC	84.8866
flights	**24.5600**	**0.1497**	**raha-baran + HC**	117.0251	raha-baran + AffinityPropagation	100.5333
flights	**30.6300**	**0.1750**	**mode + DBSCAN**	242.2105	mode + GMM	103.6810
flights	**38.9200**	**0.2498**	**mode + DBSCAN**	1817.8023	mode + GMM	103.5511
flights	**40.7600**	**0.2447**	**mode + DBSCAN**	2543.5122	mode + AffinityPropagation	95.7994
flights	45.4400	0.2997	mode + AffinityPropagation	130.9496	mode + DBSCAN	98.4787
flights	**51.5900**	**0.3499**	**mode + HC**	135.6610	raha-baran + HC	109.6794
flights	62.8700	0.4497	mode + KMeans	187.7741	raha-baran + HC	113.1024
flights	**67.7400**	**0.4999**	**mode + HC**	204.9410	raha-baran + HC	104.2668
hospital	**8.5300**	**0.0190**	**raha-baran + HC**	87.6306	raha-baran + HC	87.6306
hospital	**11.9600**	**0.0380**	**raha-baran + HC**	77.4230	raha-baran + HC	77.4230
hospital	**15.3400**	**0.0570**	**raha-baran + HC**	97.1416	raha-baran + HC	97.1416
hospital	**21.6500**	**0.0950**	**raha-baran + HC**	69.2354	raha-baran + HC	69.2354
hospital	**24.8300**	**0.1140**	**raha-baran + HC**	72.1247	raha-baran + HC	72.1247
hospital	**27.9600**	**0.1330**	**raha-baran + HC**	77.1589	raha-baran + HC	77.1589
hospital	**33.6800**	**0.1710**	**raha-baran + HC**	72.4997	raha-baran + HC	72.4997
hospital	36.5200	0.1900	mode + AffinityPropagation	66.4276	mode + AffinityPropagation	66.4276
hospital	**46.5200**	**0.1900**	**mode + HC**	70.7028	mode + HC	70.7028
hospital	49.2900	0.2850	mode + AffinityPropagation	92.1726	mode + AffinityPropagation	92.1726
rayyan	**10.7500**	**0.1544**	**raha-baran + HC**	91.3802	raha-baran + HC	91.3802
rayyan	**13.7900**	**0.1719**	**raha-baran + HC**	74.3516	raha-baran + HC	74.3516
rayyan	**16.8800**	**0.1888**	**raha-baran + HC**	83.6298	raha-baran + HC	83.6298
rayyan	**19.7100**	**0.2029**	**mode + HC**	71.6916	mode + HC	71.6916
rayyan	**22.7700**	**0.2214**	**raha-baran + HC**	79.8734	raha-baran + HC	79.8734
rayyan	**24.3500**	**0.2313**	**raha-baran + HC**	101.7058	raha-baran + HC	101.7058
rayyan	29.2500	0.2630	raha-baran + AffinityPropagation	99.4943	raha-baran + AffinityPropagation	99.4943
rayyan	**40.2400**	**0.2935**	**raha-baran + HC**	43.5159	raha-baran + HC	43.5159
rayyan	**47.8800**	**0.3993**	**mode + HC**	25.2223	mode + HC	25.2223
rayyan	**52.7300**	**0.3935**	**mode + HC**	51.7525	mode + HC	51.7525

表 4.3.2 不同数据集在多错误率条件下的最佳组合与最接近基准组合得分对比表

图 4.3.3 展示了 18 种算法组合在不同聚类性能指标上的表现。该图分为 5 个部分，按指标类型与对性能的意义分类：

- *Average Score (%)* - 各算法组合的平均得分百分比，表明它们在所有实验场景中的总体表现。得分范围从 0% 到 150%，算法组合按得分从高到低排序。
- *Average Combined Score* - 各算法在所有实验场景中综合分数的均值。综合分数体现了聚类紧致性和分离度的平衡，范围从 0 到 2，越高越优。
- *Standard Deviation of Percentage Score* - 各算法得分的波动范围（标准差）。波动范围越小，表示算法在不同场景中的表现越稳定。
- *Standard Deviation of Combined Score* - 综合分数的标准差，评估算法在多场景下的一致性。
- *Average Deviation from Reference (100%)* - 各算法得分与参考基准（100%）的平均偏离程度。偏差越小，说明算法组合与基准方案的表现越接近。

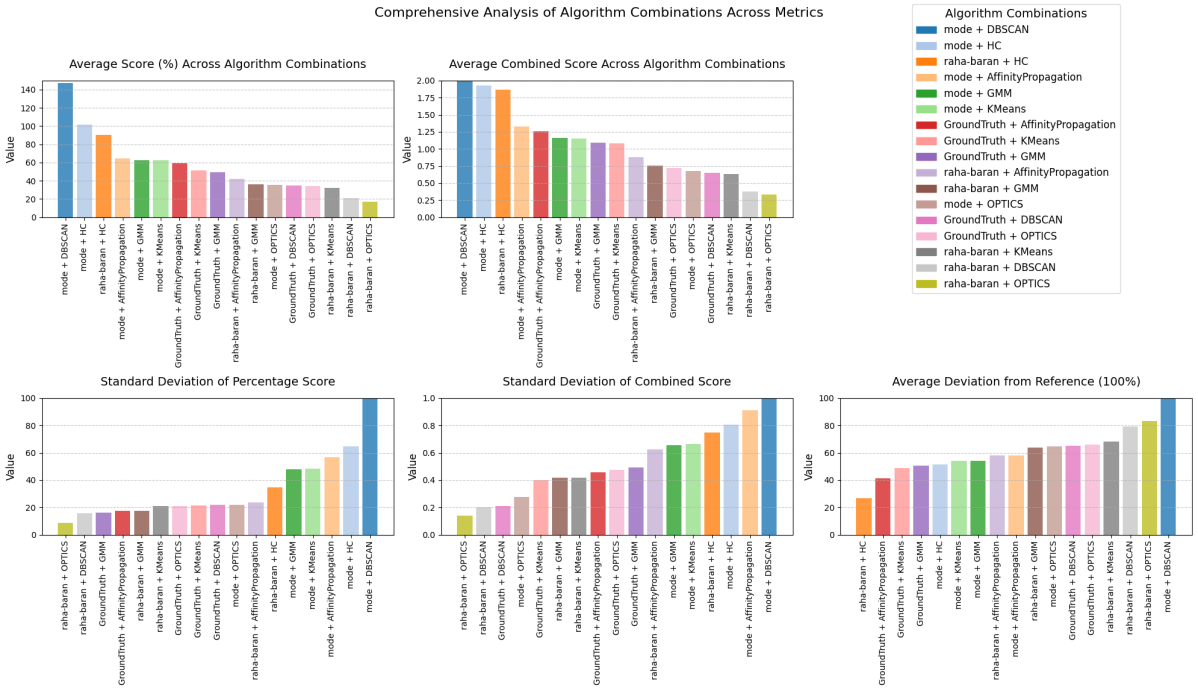

图 4.3.3 基于多种指标的算法组合综合性能分析图（全体数据集）

图 4.3.4 在图 4.3.3 的基础上选择了错误率低于 25% 的数据，用来分析不同算法组合方案在非极端错误率情况下的聚类性能与方案稳定性：

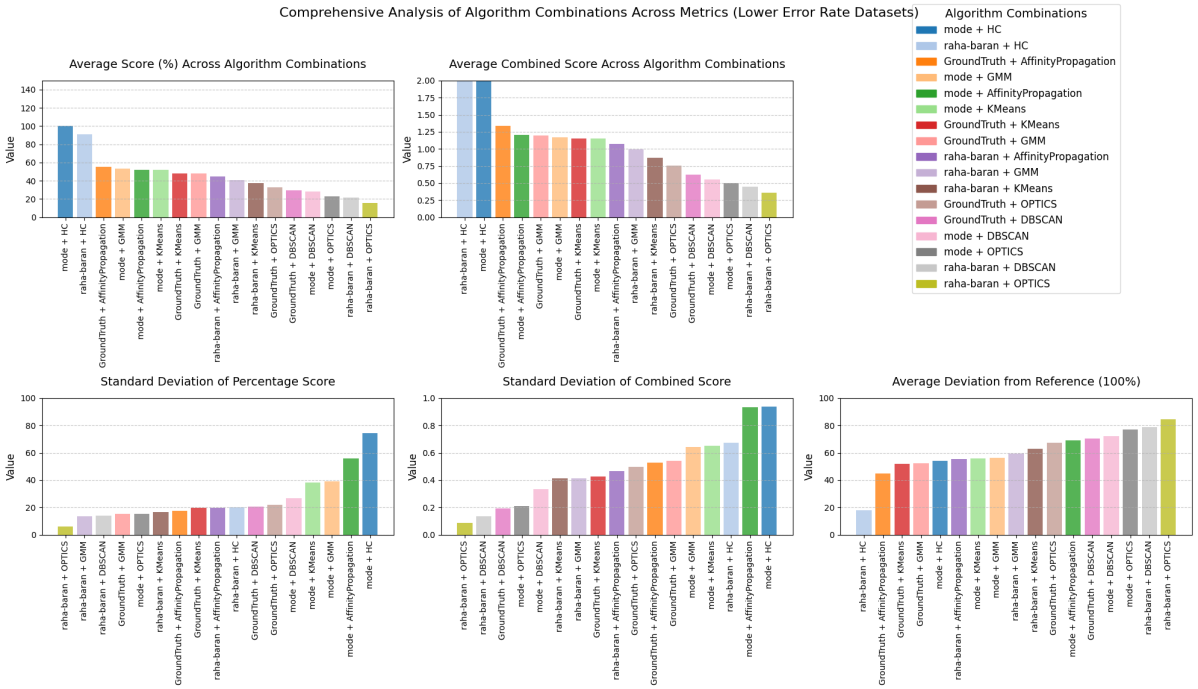

图 4.3.4 基于多种指标的算法组合综合性能分析图（低错误率数据集）

4.4 实验结果分析

4.4.1 基于单数据集纵向分析

通过观察各数据集（beers、flights、hospital、rayyan）在不同错误率下的排名与偏差（表 4.3.2），可以得到以下几点认识：

“最佳组合”与“最接近基准”经常不一致。在 *beers* 与 *flights* 等样本量相对较大（分别 2410 与 2376），数值型特征占主导的数据集中，mode + HC 或 mode + DBSCAN 有时出现高达 200%～300%、甚至 1000%+ 的“爆分”。这说明当“简单填补 (mode)”遇上合适的超参数 (HC 或 DBSCAN) 时，可能在内部指标（Davies-Bouldin 与 Silhouette）上得到极端的划分效果。然而，这些高分结果往往与基准存在较大偏差，原因或在于 DBSCAN 对噪声与缺失值极度敏感，或 HC 采取了与理想簇数相差较远的分割方式。此时，“最接近基准”的方案通常从 raha-baran + HC 或 mode + KMeans 中产生，其分值虽不及“爆分”方案，但与 GroundTruth 的聚类结构更为接近。

不同数据集的极端现象分布并不均衡。在 *flights*（7 个特征、全部数值型）数据集中，mode + DBSCAN 出现过 1800%～2500% 的极端情况，表明在中等规模、低维且噪声成分比较明显的数据里，密度聚类对填补方式与超参数尤为敏感。而在 *hospital*（1000 条样本、20 个特征、多为数值型但更高维度）与 *rayyan*（1000 条样本、11 个特征），聚类分数整体较为温和，极少越过 200%。这些数据往往在特征和数据类型上更具一定的“语义约束”或“上下文逻辑”，使得 raha-baran 在修复后为 HC 或其他算法保留了更合理的分布结构，减少了极端划分。

Raha-Baran + HC 在中低错误率时表现稳健。在 *hospital* 和 *rayyan* 中，当错误率低于 25% 左右时，raha-baran + HC 往往能同时获得“最高分”和“与基准最贴近”这两项佳绩。这说明对于具备一定语义约束、特征相对丰富（20 维或 11 维）且数据规模中等（1000 条）的数据集而言，Raha-Baran 的精细化修复能大幅降低噪声干扰；而 HC 的层次聚类方式在质量较好的数据环境下更容易逼近基准结构。当错误率升至 40% 以上，虽然这组方案的分数也会波动下降，但对比其他组合依然较为稳定，表明在高维场景下（如 20 特征的 *hospital* 数据），上下文修复 + 分层聚类的协同作用相对优于其他算法组合。

4.4.2 全局方案表现与离散度分析

图 4.3.3 将所有数据集（含不同错误率）的结果综合起来，有助于评估每种方案在“全局”角度的稳定性和整体适应性，以下是从该图得出的全局分析结果：

部分组合平均分明显高于 100%，但方差极大。典型例子是 mode + DBSCAN：其平均分约 147.02%，最高能到 2000%+，标准差却达 481.75，表明极端聚类结果频发，实际使用中需谨慎对待。这类“高均值+高方差”组合多见于数据集以数值特征为主、尺

寸中大型（>2000 条），且噪声或缺失较多的场景，比如 *flights*、*beers*。若目标是“可解释、贴近基准”，则不建议此路线；若希望在特定内部指标上挖掘极端方案，可以考虑尝试但要严格控制超参数搜索。

Raha-Baran + HC、Mode + HC 的平均分相对较高且方差适中。mode + HC 平均分约 101.53%，标准差 64.62，整体分数分布在 50%～300% 区间；而 raha-baran + HC 平均分 90.03%，标准差 34.80，极端情况更少。这与表 4.3.2 中的观察一致，说明层次聚类对不同规模与特征数都有一定鲁棒性。同时，Raha-Baran 的高精度清洗对于有语义或规则错误的数据（如 *hospital*, *rayyan*）尤其有效。相较之下，Mode 填补在数值型为主的 *beers*, *flights* 中更易“爆分”。

其余组合多数集中在 30%～70% 区间。例如 raha-baran + OPTICS 均值仅 16.75%，意味着在当前参数搜索及数据特征下，OPTICS 并未与 Raha-Baran 产生良好协同。这些结果并非表明此类方法“不可用”，而是暗示对于更高维或存在较多噪声的数据，密度聚类 (DBSCAN, OPTICS) 若参数无法匹配数据分布，往往难以取得理想效果。

综上，图 4.3.3 在宏观层面验证了 DBSCAN/OPTICS 类方法虽有潜力但波动极大，而 HC + 适宜的清洗（Mode 或 Raha-Baran）往往兼具较高均值和更低方差。在不同规模与类型的数据下，HC 都能为聚类结果提供相对稳定的保障。

4.4.3 低错误率情形的表现

图 4.3.4 在图 4.3.3 的基础上仅保留了错误率 ≤25% 的数据，进一步考察当数据质量相对较好时各方案的表现差异。以下是对低错误率表现图的分析：

Mode + HC 在相对干净数据条件下均值可达 100.14%。但标准差约为 74.20，表明该方法在某些情况下能够显著超越基准，但在特定场景（如高维度或噪声分布不均的数据）下表现波动较大。这种现象在数值型数据集（如 *beers* 和 *flights*）上尤为明显：当错误率较低时，简单的填补方法可以较好地保留数据分布的完整性，而 HC 聚类通过合适的层次切分，能够在内部指标上实现优异表现。

Raha-baran + HC 仍具有稳定性。其平均分在低错误率子集中达 91.04%，标准差降至 20.04，相比全部数据时（均值 90.03%，标准差 34.80）更加稳定；对于具备语义或知识库错误的场景（如 *hospital*, *rayyan*），在错误率较低时，Raha-Baran 修复带来的好处更突出，HC 聚类结果基本可与基准相近甚至出现略超 100% 的情况。

其余组合虽略有改善，但相比于 mode + HC 和 raha-baran + HC 仍有明显不足。例如 mode + GMM、mode + KMeans 等平均分通常介于 50%～60%，有一定提升但相对有限。说明在部分中等维度的数据中（如 11~20 个特征），KMeans 或 GMM 聚类若不

能适配具体数据分布，其生成的内部指标可能难以占据优势。

4.4.4 综合结果与启示

结合实验结果和多种清洗策略与聚类算法的表现，针对不同数据集的规模、特征类型及错误率水平，提出以下建议以指导清洗与聚类方法组合的选择：

小至中规模数据集、高维/多元特征且可能存在语义错误：优先选择 raha-baran + HC 组合。Raha-Baran 的上下文修复能力能够有效处理复杂语义错误，而 HC（层次聚类）的分层聚类特性适合高维、多特征数据。若错误率较高，需进一步结合更先进的修复策略，或通过特征选择与降维手段减少特征复杂性。

中至大规模数据集、以数值型特征为主、对内部指标最优化有较强需求：考虑使用 mode + HC 或 mode + DBSCAN。简单填补策略（Mode）在数值型数据上的效率较高，而 HC 和 DBSCAN 均能对中大规模数据集进行合理划分，但需谨慎对待 mode + DBSCAN 的高方差风险。若应用场景要求聚类结果与基准结构更贴近且解释性更强，应该选择更稳健的组合（如 mode + HC 或 mode + KMeans）。

强噪声分布、不规则簇形状的数据集可使用 DBSCAN 或 OPTICS 等密度聚类方法。这些方法在处理不规则簇形状和高噪声数据时具有明显优势。密度聚类方法对超参数（如 ϵ 和 min_samples）敏感，因此需要针对不同错误率和特征分布进行精细化超参数调优，以避免出现极端分割（过多单点簇）或过度聚合的情况。自动调优框架（如 Optuna）在此类场景中尤为重要。

5 讨论（还没开始写，以下是写作思路）

5.1 理论解释

1. Mode 与 Raha-Baran 的差异机理
    - 在前文（第 4 节）对比了这两种清洗策略的性能差异，这里可进一步从理论层面探讨：
    - 将这些结果与已有文献进行对比，说明本研究如何证实了“适度复杂的清洗策略在错误率不高时效果最佳，而在错误率极端时也可能受限”等结论；同时可引用其他学者的类似或相反发现，与之形成呼应或对照。
2. 层次聚类 (HC) 的稳健性与适用性
    - 结合聚类原理：HC 逐步合并或拆分簇，不依赖随机初始化；因此在轻度噪声和缺失值场景下通常更稳健。

- 与 KMeans、DBSCAN 等对比：
  - KMeans 假设球形分布，对噪声更敏感；
  - DBSCAN/OPTICS 在噪声识别和 ϵ\epsilonϵ 参数选择不当时容易出现极端聚类形态。
- 理论视角可再次佐证：HC 在适当的规模和噪声水平下能更好地逼近基准结果，尤其当清洗策略能减少无效或混乱的数据点时，HC 的逐步分层更能体现出稳定优势。

3. 特殊现象的机理解释

- 高于 100%的得分：可从“内部指标的放大效应”切入，说明在距离度量或簇间离散度被极端拉大的场合，DB_score、Silhouette_score 等会呈现过高评价。
- 不收敛或超时：可将失败场景归因于参数初始值过于极端、数据分布被错误修复扭曲，以及部分算法复杂度（如层次聚类在大规模数据下的 O(n2)甚至 O(n3)开销过高等。

5.2 方法与评估指标的深入分析

1. 自动化调优（Optuna）对性能与稳定性的提升
   - 在第 4 节已经提到 Optuna 搜参的成果，这里可更深入讨论：
     - Optuna 的 TPE / 贝叶斯优化在多大程度上缩小了盲目搜索空间？
     - 是否有场景（如超高维度或数据规模非常大）需要新的采样策略或并行化搜索？
   - 可以结合已有文献或 AutoML 相关研究，说明本研究的选择如何符合实际聚类任务在无监督场景中的需求。
2. 内部聚类指标（DB_score 与 Silhouette_score）的局限
   - 在第 4 节多次提到这些指标导致部分组合“爆分”。在本节可更系统地指出：
     - 高分不一定代表与真实业务意义吻合；
     - 当簇数过多或初始分区极端时，DB_score 与 Silhouette_score 可能

被高估；

- 同时，这些指标对噪声敏感度较高，当清洗策略改变数据分布后，很容易产生过于理想化的聚类效果。

- 建议在实际应用中，若需要更可解释或稳健的聚类结构，可在内部指标之外结合外部指标或专家评估。

5.3 清洗与聚类的协同作用

1. 协同提升聚类性能的机制
   - 在结果分析（第 4 节）中已指出“清洗能提升聚类效果”，此处可结合更多参考文献与理论模型，阐述“噪声/缺失值的减少如何降低簇内离散度、增强簇间分离度”。
   - 如果有可能，可插入一个示意性流程图，展示“清洗 -> 数据分布优化 -> 超参数调优 -> 聚类模型收敛更好”的框架。
2. 清洗策略对协同作用的影响
   - 对比 Mode 与 Raha-Baran 对于协同提升的不同贡献路径：
     - Mode 更简单，但若数据本身多语义冲突，协同效果有限；
     - Raha-Baran 能深度修复语义与逻辑冲突，在高维、具有复杂错误类型的数据上对聚类帮助更大。

5.4 局限性

1. 数据集与清洗策略的限制
   - 第 4 节聚焦在 40 个数据集，这里可从通用性与外部有效性等学术角度指出：
     - 数据规模是否足够广泛？
     - 数据类型以数值或可转化为数值居多，是否对文本/图像/多模态数据也能适用？
     - Mode 与 Raha-Baran 都属于已知的典型清洗策略，但对更复杂的知识库约束或跨表数据一致性，是否需要新的清洗范式？

2. 超参数优化范围的局限性
    - Optuna 在聚类中虽有助于自动找优，但依赖于用户定义的搜索空间。若搜索范围不合理，仍可能错过全局最优。
    - 在大规模或高维数据中，优化时长与内存开销均需进一步平衡。
3. 聚类算法种类的局限
    - 本研究聚焦于 6 种经典算法，暂未涉及深度学习聚类（如 Deep Embedded Clustering 等）。
    - 未来可研究对高维度（>100 特征）、大量非线性特征或图结构数据的聚类策略适用性；或与自监督学习、表示学习相结合。

6 相关工作

数据清洗方法。Mode 填补是一种简单高效的清洗方法，广泛用于处理缺失值问题，特别是在类别型数据中表现出良好适用性。然而，该方法难以应对复杂错误（如语义冲突或语法错误），而且可能引入偏差或丢失语义信息。相比之下，Raha-Baran 策略结合上下文推理和规则匹配技术，可修复多种类型的复杂错误，在多样化数据场景中表现更优，但计算复杂度较高，限制了其在大规模数据上的应用。

经典与新兴聚类算法。KMeans 是最常用的聚类算法，因其计算简单、实现容易而广泛应用，但对数据分布的敏感性限制了其在复杂场景中的表现。层次聚类（HC）通过分层构建簇结构，不依赖初始参数，对轻度噪声和缺失数据具有较强鲁棒性，但计算复杂度较高。DBSCAN 和 OPTICS 等密度聚类方法适用于任意形状簇且耐受噪声，但对参数设置较为敏感。近年来，深度聚类（如 Deep Embedded Clustering）结合特征学习和聚类优化，在高维和复杂分布数据场景下表现出显著优势，扩展了聚类算法的应用范围。

AutoML 在聚类中的应用。自动化机器学习（AutoML）在近年来逐渐成为提升聚类算法性能的有效工具。相比于监督学习，AutoML 在无监督学习中的研究较少，主要困难在于缺乏明确的目标函数和评价指标。然而，Optuna 作为一种高效的自动化调优框架，依赖贝叶斯优化和树结构搜索算法，能够在高维参数空间中快速找到性能最优的配置，实验表明其显著提升了聚类算法的适应性和性能。

本研究的创新点

1. 清洗策略对聚类优化支持的验证
   验证数据清洗策略如何通过改善数据分布特性，降低聚类目标空间的复杂性，

为后续的聚类参数优化提供更优的初始条件。这一分析明确了清洗策略在提升聚类性能中的关键作用。

2. 多样化数据集的广泛实验验证
   基于 40 个具备多样化数据质量问题的真实数据集进行实验，涵盖不同类型的数据质量问题（如高缺失率和高噪声）。这一设计显著增强了研究结果的普适性，为清洗与优化策略在不同场景中的适用性提供了丰富的实验支持。

3. 清洗与优化协同作用的深入分析
   深入探讨数据清洗与聚类优化的协同工作机制，结合实验数据，系统分析清洗策略如何影响参数调优的效果。这为实际应用中的数据质量提升与聚类性能优化提供了明确的理论依据和实践指导。

7 结论

本研究围绕数据质量问题对聚类性能的影响，系统地评估了数据清洗策略与聚类优化的结合方案。通过对 Raha-Baran、Mode 填补及 GroundTruth 等清洗方法与 6 种常见聚类算法的组合评估，发现 Raha-Baran + HC 组合在轻度噪声和缺失值条件下展现了稳定的优越性，部分情况下超越参考基准。同时，利用 Optuna 进行自动化超参数调优，大幅提升了聚类算法对多样化数据集的适应性和性能表现。

与现有研究相比，本研究以广泛的数据实验验证了清洗与优化协同作用的有效性，为复杂数据质量问题提供了系统化的解决思路。在实践中，这一研究框架可为客户分群、异常检测等应用场景提供可靠的技术参考。未来的研究将重点扩展清洗方法与优化策略的多样性，并探索在更高维度与多模态数据条件下的适用性。