**引言部分写作思路和脉络梳理 1/22/2025**

**1.1 数据质量对聚类任务的重要性**

**1. 引出真实场景下的数据质量问题**

- （文献）在许多实际应用（如医疗、金融、工业物联网等）中，数据通常存在缺失值、错误值、噪声等质量问题。

**2. 聚类算法对数据分布变化的敏感性**

- （文献）与分类或回归等有监督学习任务相比，无监督的聚类对数据分布结构依赖更强；

- （文献）当数据中含有较多噪声或不确定性时，聚类算法容易出现簇结构偏离真实分布的现象。这种情况会对后续的模式挖掘、下游决策造成显著影响。

(目的意义：以上说明**“数据质量问题确实会对聚类产生严重影响”**，接下来自然要提到“数据清洗”作为对策。)

**1.2 清洗方法对聚类的影响**

**1. 清洗方法的主要目标**

- 为了应对数据质量问题，常用的清洗策略包括缺失值填充、异常值检测与剔除、错误值纠正等；

- 这些方法的**共同目标**是“修复或减少数据中的噪声和错误”，从而在后续分析中降低不良影响。

**2. 不一定带来纯粹正向提升**

（文献）尽管清洗的初衷是提升数据质量，但有部分文献表明：

- 过度严格的清洗方法可能会修改原本正确的数据；

- 使用常见值填补大量缺失值会破坏原本有价值的分布信息；

- 如果清洗方法改变了数据的原始分布，可能导致聚类算法的性能不升反降。

例如：

* 在有少量噪声的情况下，剔除噪声可能使 k-means 聚类结果更稳定；
* 但对于同一数据集，异常点可能是某些簇的重要特征，剔除后会破坏簇的边界，从而影响密度聚类（如 DBSCAN）。

这意味着清洗与后续的聚类步骤不能简单地“分离”开来，需要结合数据自身特征来评估其影响。

(目的意义：既然清洗方法对聚类有很大影响，就进一步引出需要**协同优化**的问题。)

**1.3 问题与挑战：清洗+聚类的管线协同优化**

**1. 现有研究往往聚焦于单一维度**

1) 机器学习视角

- （文献）大部分关于聚类的研究主要讨论如何改进聚类算法本身（如 k-means 变体、层次聚类、DBSCAN 等）或调优其超参数；

- （文献）通常默认数据已经“足够干净”或简单地用固定的预处理方式（如均值填充）而缺乏深度的清洗评估。

2) 数据质量视角

- （文献）有相当多文献关注提升数据质量，如修补错误值、删除噪声点等，但研究目标往往局限于“让数据更干净”；

- （文献）对于清洗后下游聚类效果的整体提升缺乏全面实验，或仅在单一算法上做了性能评测。

2. 为什么需要“清洗+聚类”协同优化

- 不同清洗策略可能对不同聚类算法表现出明显的差异；

- （文献）不同聚类算法对数据噪声或缺失值的耐受能力不同，也意味着它们对清洗程度的需求不同。

- 关键观点：二者存在耦合关系，若只关注其中一个维度，可能无法获得整体最优方案。

3. 管线式（Pipeline）思维

- 将“清洗策略 + 聚类算法 + 超参数”看作一个整体管线，需要同时优化；

- 但由于可供选择的清洗方法、聚类算法以及超参数组合庞大，人工穷举或简单试验无法在可接受时间内完成搜索。

(过渡：既然需要协同优化且搜索空间巨大，那么自然会想要**自动化**的方法。)

**1.4 自然引出：自动化聚类模型AutoML思路**

**1. 研究空白与本研究切入点**

- 回顾上文：为确保聚类质量并兼顾搜索效率，需要一种能根据数据集特征自动决定清洗策略与聚类算法的一体化方法；

- （文献）目前大多数 AutoML 研究侧重于有监督学习（分类/回归）的自动特征工程与模型选择，对于聚类任务，尤其是“聚类+清洗”的自动化研究仍相对有限。

**2. 在两个理论领域之间搭桥**

- 本研究的思路是在“数据质量（数据清洗理论）”与“自动化机器学习（AutoML理论）”之间寻找结合点；

- 通过将多种清洗方法与聚类算法及其超参数一并纳入搜索空间，并利用模型学习“数据特征 → 优选方案组合”的映射。

**3. 意义与价值**

- 由于聚类在众多无监督场景下应用广泛，一旦实现“可自动化、可适配多种数据特征”的协同优化，将在学术研究与产业应用中都具备重要价值；

这也是一个尚未被深入探索的领域，具有明显的研究新颖性与可拓展性。