Modern Engineering and Application 现代工程与应用, 2026, 4(3), 1-8
Published Online August 2026.
https://ufascience.com/mea/index

面向中爪哇省大数据贫困分析的混合分裂式 K 均值框架

朱思雨

贵州大学经济学院，贵阳 550000，贵州，中国

摘要：聚类在大数据分析中不可或缺，对于划分高维社会经济数据集以支撑结果解读与政策决策尤为重要。K 均值因简单且可扩展而被广泛使用，但它对初始质心选择高度敏感，常导致结果不稳定与收敛缓慢。以往的混合方法（如凝聚式—K 均值）试图用层次聚类完成质心初始化来解决该问题，然而这类方法依赖自下而上的合并，可能产生次优的初始划分，并在数据规模增大时增加计算开销。为克服上述局限，本文提出一种混合分裂式—K 均值（DHC）模型：先以自上而下的层次分裂生成更为一致的初始质心，再由 K 均值加以精化。本文使用印度尼西亚中央统计局（BPS）提供的中爪哇省多维贫困数据集，把 DHC 与标准 K 均值、凝聚式—K 均值进行了对比评价，评价内容包括执行时间、收敛迭代次数与聚类有效性指标（轮廓系数、Davies‑Bouldin 指数与 Calinski‑Harabasz 指数）。实验结果表明，DHC 使执行时间最多降低 97%，所需迭代次数比标准 K 均值少 40%，同时聚类质量相当甚至更优（如 CH 指数由 14.3 提高到 15.8）。上述结果表明，DHC 模型为大规模社会经济数据分析提供了一种更高效、更稳定的聚类方案。

关键词：大数据；聚类；分裂式层次聚类；混合模型；K 均值；贫困数据分析

A hybrid Divisive K-means Framework for Big Data-driven Poverty Analysis in Central Java Province

SiYu Zhu

School of Economics, Guizhou University, Guiyang 550000, Guizhou, China

Abstract: Clustering is essential in big data analytics, especially for partitioning high-dimensional socioeconomic datasets to support interpretation and policy decisions. While K-Means is widely used for its simplicity and scalability, its strong sensitivity to initial centroid selection often leads to unstable results and slower convergence. Previous hybrid approaches, such as Agglomerative-K-Means, attempted to address this issue by using hierarchical clustering for centroid initialization; however, these methods rely on bottom-up merging, which can produce suboptimal initial partitions and increase computational overhead for larger datasets. To overcome these limitations, this study proposes a hybrid divisive-K-Means (DHC) model that employs top-down hierarchical splitting to generate more coherent initial centroids before refinement with K-Means. Using a multidimensional poverty dataset from Central Java Province provided by the Indonesian Central Bureau of Statistics (BPS), the performance of DHC was evaluated against standard K-Means and Agglomerative-K-Means. The assessment included execution time, convergence iterations, and cluster

文章引用: 朱思雨. 面向中爪哇省大数据贫困分析的混合分裂式 K 均值框架[J]. 现代工程与应用, 2026, 4(3): 1-8.
DOI: https://doi.org/10.61784/mea2026

validity indices (Silhouette, Davies-Bouldin, and Calinski-Harabasz). Experimental results demonstrate that DHC reduces execution time by up to 97% and requires 40% fewer iterations than standard K-Means, while achieving comparable or improved cluster quality (e.g., CH Index increasing from 14.3 to 15.8). These findings indicate that the DHC model offers a more efficient and stable clustering solution for large-scale socioeconomic data analysis.

Keywords: Big data; Clustering; Divisive hierarchical; Hybrid model; K-Means; Poverty data analysis

一、引言

聚类是一种被广泛使用的无监督学习技术，用于在无标签数据中识别隐含结构，可支撑社会经济分析、城市规划、环境监测与健康信息学等应用 [1]。在各类聚类算法中，K 均值因高效且可扩展而长期流行；但它对初始质心选择高度敏感，常导致结果不一致、收敛缓慢并易陷入局部极小 [2-3]。为克服这些弱点，已有若干研究提出对 K 均值的混合式或优化式改造，包括与层次方法、遗传算法及基于密度的预处理相结合 [4]。

层次聚类虽能提供确定性划分并避免随机初始化，但在应用于大规模数据集时计算复杂度很高 [5]。层次—K 均值组合等既有混合方法试图融合二者之长，然而这些研究大多依赖凝聚式（自下而上）聚类，其基于合并的结构可能导致质心初始化次优。此外，既有混合方法很少评估层次阶段在应用于多维社会经济数据集时是否真正改善了初始化质量或可扩展性 [6]。上述局限凸显出一个明确的研究空白：当前的混合方法尚未在保持计算效率的同时充分优化质心初始化，对复杂的贫困相关指标而言尤其如此。

由此，本研究的基本问题自然浮现：尽管混合聚类方法层出不穷，但在处理多维社会经济数据时，究竟如何系统性地改进质心初始化以提升 K 均值的收敛速度、稳定性与聚类质量，仍不清楚。已有证据表明需要一种在全局层面更具信息量的初始化策略，但此类方法的实际有效性在以往工作中尚未得到充分确证。

基于上述问题，本文提出如下假设：分裂式层次过程凭借其自上而下、递归分裂的机制，能够生成更为一致且更具代表性的初始质心；这将减少执行时间与收敛迭代次数，同时获得与常规 K 均值及既有凝聚式—K 均值混合方法相当或更优的聚类质量 [7]。尽管分裂式方法在理论上比凝聚式方法提供了更宏观的结构视角，其对混合聚类的潜在收益在以往研究中尚未得到全面评估。

为检验该假设，本文把所提 DHC 模型应用于取自印度尼西亚中央统计局（BPS）的中爪哇省多维贫困数据集。该数据集包含教育、收入、就业与居住条件等相互关联的社会经济指标，用常规方法难以聚类。通过聚类理解贫困分布，对政策靶向与区域发展规划具有重要意义 [8]。

本研究的贡献如下：（1）提出一种混合分裂式—K 均值（DHC）算法，以改进多维社会经济数据的质心初始化与聚类效率；（2）从执行时间、收敛速度与聚类有效性指标出发，对 K 均值、凝聚式—K 均值与 DHC 进行对比评价；（3）证明混合聚类方法作为社会经济政策制定决策支持工具，在区域贫困分析中具有现实意义。总体而言，本研究通过解决质心初始化方面尚未解决的局限，拓展了既有混合聚类文献，并表明把确定性的层次策略与划分技术相结合，可产生更稳定且计算上更高效的聚类结果 [9]。

二、方法

图 1 给出了本文所提混合 DHC 框架用于中爪哇省大数据贫困分析的工作流程。该框架始于输入数据的获取，其中包括受教育程度、就业状况与家庭支出等大规模贫困指标。为保证分析的可靠性，本文采用了更为稳健的数据预处理流水线，包括通过多变量插补系统处理缺失值、异常值的检测与处理、异质数值区间的归一化，以及跨区县的特征一致性检查。

预处理之后，采用自上而下的策略实施分裂式层次聚类。每次迭代按最大异质性准则对数据集递归分裂，

并明确定义了选择分裂属性与计算子组质心的算法步骤。这些质心作为结构化的、数据驱动的初始种子进入后续优化阶段。

下一阶段以预设簇数（k=3）执行 K 均值优化，通过迭代最小化簇内平方和对分裂式生成的质心加以精化。该步骤提升了簇的紧致度，并降低了对随机初始化的敏感性，从而克服了标准 K 均值的一个常见局限。收敛阈值、迭代上限与距离度量均已明确定义，以保证方法学的透明性。

为评价聚类的鲁棒性，本文对所得优化簇施加多种基准比较技术，包括与标准 K 均值及其他初始化策略的对比。聚类质量以若干有效性指标（如轮廓系数、Davies－Bouldin 指数、Calinski－Harabasz 指数）加以评估，从而实现更广泛、更严格的性能评价。最后，在区县层面对空间与社会经济格局加以可视化，以提炼具有政策意义的洞见。尽管本研究所用数据集规模有限，难以充分展示完整的大数据可扩展性，但得益于层次约简与优化的初始化步骤，该框架在设计上可扩展至更大规模的数据集。

Figure 1 Hybrid DHC Framework

图 1 混合 DHC 框架

全部实验在 Google Colab 环境中执行，使用 Python 3.10 与该平台提供的标准硬件资源。聚类过程借助 scikit-learn、NumPy、pandas 与 SciPy 等广泛采用的科学计算库实现。上述配置的报告旨在保证实验流程的透明性与可复现性。

（一）数据集

本研究采用取自印度尼西亚中央统计局 2024 年数据的中爪哇省贫困数据集（见表 1），包含该省 35 个区县与市的记录，含九项体现贫困多维属性的社会经济指标。选择该数据集是因为它提供了一个高维、不平衡且无标签的真实社会经济案例，需要准确聚类以支撑区域减贫政策与资源配置。除该区域数据集外，本文还使用 UCI 机器学习知识库中的基准数据集开展了补充测试，以确保所提方法在不同领域中的泛化能力。

在聚类之前，本文进行了数据预处理以处理缺失或不完整的取值。数据集中有若干条目为“NA”（不适用），在就业类社会经济指标中尤为常见。删除含缺失值的记录可能导致有意义信息的丢失并扭曲数据分布，因此本研究改用插补技术，以依据既有数据模式估计的取值替换缺失项。具体而言，对表 1 所示数据采用相应变量的均值进行插补——这是无监督学习任务中最常用的统计插补方法之一。插补通常优于删除，因为它能保持数据集的完整性与稳健性，在多维或大数据情形下尤为如此。

Table 1 Poverty Dataset of Central Java Province

表 1 中爪哇省贫困数据集

县/市	①	②	③	④	⑤	⑥	⑦	⑧	⑨
Cilacap	10.68	17.76	95.28	95.37	39.54	32.25	38.96	65.66	93.88
Banyumas	11.95	15.25	95.91	99.59	38.05	14.93	37	64.03	91.56

Purbalingga	14.18	18.45	93.87	94.77	32.05	19.04	39.89	63.98	94.16
Banjarnegara	14.71	17.16	92.92	94.64	31.97	30.73	46.34	60.84	93.58
Kebumen	15.71	15.66	94.87	99.47	30.77	27.6	48.91	60.04	98.35
Purworejo	10.87	12.13	95.81	99.1	29.43	27.71	46.77	61.92	99.12
Wonosobo	15.28	18.71	93.03	90.08	29.42	36.45	47.91	62.52	93.8
Magelang	10.83	14.37	93.72	93.99	28.79	30.91	46.64	58.05	92.61
Boyolali	9.63	14.66	92.77	96.26	27.31	26.84	45.55	62.1	92.53
Klaten	12.04	12.48	94.27	99.99	33.43	15.52	34.19	59.1	98.13
Sukoharjo	7.47	8.98	95.51	99.98	33.48	9.37	28.03	61.1	98.33
Wonogiri	10.71	17.74	92.43	97.04	31.67	32.66	45.75	62.08	100
Karanganyar	9.59	10.21	94.07	98.73	33.08	15.82	31.95	60.37	97.57
Sragen	12.41	16.46	89.79	94.21	29.3	28.29	43.75	58.4	100
Grobogan	11.43	11.94	94.16	97.02	29.01	32.72	50.28	65.68	93.7
Blora	11.42	19.89	88.56	97.04	27.15	38.94	54.74	62.39	94.14
Rembang	14.02	14.43	94.97	99.17	32.1	26.66	39.26	62.21	90.74
Pati	9.17	14.79	94.01	97.6	35.00	24.02	41.45	61.83	98.52
Kudus	7.23	8.56	95.91	99.99	30.57	6.97	25.66	59.43	98.63
Jepara	6.09	11.22	96.26	98.91	34.26	10.35	30.53	63.59	94.14
Demak	11.89	11.26	95.92	98.25	31.19	17.75	33.42	60.82	93.74
Semarang	6.96	13.79	95.67	96.31	27.95	17.57	36.81	59.28	96.33
Temanggung	8.67	16.77	96.07	95.78	25.9	41.51	52.68	63.3	95.02
Kendal	9.35	16.38	95.47	97.38	34.39	19.57	34.9	62.06	93.35
Batang	8.73	18.08	93.59	93.13	29.7	20.25	39.19	63.95	94.73
Pekalongan	8.95	15.9	93.47	96.3	30.66	11.22	33.75	62.29	92.22
Pemalang	14.92	20.99	89.27	88.85	36.45	20.83	41.04	63.14	91.26
Tegal	6.81	19.21	93.62	99.93	40.85	12.52	30	63.63	95.19
Brebes	15.6	24.72	92.37	97.18	34.61	27.02	43.56	62.27	91.15
Magelang 市	5.94	2.88	99.43	99.81	41.72	0.5	24.18	59.22	76.86
Surakarta 市	8.31	4.27	98.33	99.96	38.38	NA	23.52	56.58	78.06
Salatiga 市	4.57	4.41	98.63	99.02	32.05	2.75	25.4	55.24	100.00
Semarang 市	4.03	5.86	97.88	99.98	34.29	0.71	21.06	55.39	94.99
Pekalongan 市	6.71	8.47	98.69	96.52	31.37	1.67	25.6	62.7	83.97
Tegal 市	7.64	13.54	97.94	98.56	36.7	4.47	23.94	60.63	94.61

注：各列依次为①贫困人口占比；②未完成小学教育；③识字率（15—55 岁）；④在学率（13—15 岁）；⑤未就业（15 岁以上）；⑥农业部门就业（15 岁以上）；⑦非正规部门就业（15 岁以上）；⑧食品类人均月支出占比；⑨使用自有 / 共用厕所。数值单位均为%

（二）分裂式层次聚类＋K 均值（混合 DHC-KMeans）

分裂式层次聚类＋K 均值（DHC-KMeans）是一种把分裂式层次聚类与 K 均值相结合的混合聚类方法。与从单个数据点出发逐步合并的凝聚式方法不同，分裂式方法方向相反：它把整个数据集视为一个大簇，然后递归地将其分裂为更小的子簇，直至达到期望的簇数 k。

1. 分裂式层次聚类阶段

分裂式层次聚类采用自上而下的策略，系统地把数据集划分为逐步变小且更为同质的组。在该方法中，所有数据点最初被归入一个完整的大簇，代表整个数据集；算法随后分析数据点之间的内部相异性以识别最显著的分离，并据此把该簇划分为两个子簇，使各组内部对象尽可能相似，同时与另一组保持清晰分离。这一递归分裂过程反复进行，直至达到期望的簇数 k，从而形成反映数据集自然划分的结构化层次。

2. 质心初始化

在分裂式层次过程把数据集成功划分为 k 个簇之后，下一步是把这些结果纳入混合分裂式—K 均值框架以作进一步精化。该阶段计算层次划分所得各簇的均值（质心），用以表示簇内数据点的集中趋势。这些质心作为后续 K 均值算法具有策略性与代表性的初始种子，有效消除了质心初始化中通常存在的随机性。借助源自层次阶段的质心，混合模型确保 K 均值从更准确、更具数据依据的起点开始优化。

3. K 均值精化阶段

在用分裂式层次过程所得质心完成初始化之后，进一步采用 K 均值算法精化簇归属并提升整体聚类质量。由于这些质心已经反映了数据集的自然划分与内在结构，算法从更有依据的初始配置开始运行，因而能更快收敛到稳定的簇边界，通常只需较少的迭代即可获得最优结果。

4. 混合方法的优点

把分裂式层次聚类与 K 均值相结合，在稳定性与效率两方面都带来若干显著优势。第一，使用分裂阶段所得的代表性初始质心提供了自上而下的分析视角，确保为 K 均值选取的初始点已经接近真实簇中心，这种策略性初始化把传统 K 均值所受的随机性影响降至最低。因此，聚类结果表现出更好的稳定性与一致性。此外，由于质心是依据数据集内有意义的结构划分初始化的，K 均值收敛更快，达到最优解所需的迭代更少。

5. 局限性

尽管具有上述优点，分裂式—K 均值混合方法也存在若干计算上的局限。其主要缺点之一是分裂式聚类计算开销较大：自上而下的分裂过程需要在确定最优划分之前评估大量可能的划分方案。此外，与凝聚式方法类似，分裂式聚类并不适合超大规模数据集，因为递归分裂与距离计算需要大量计算资源与内存。因此，若不作进一步优化或采用并行处理，该方法在大规模或实时应用中可能并不实用。

（三）评价指标

为全面评估聚类性能，本研究采用三类评价指标：执行时间、收敛迭代次数与聚类有效性指标。这些指标不仅刻画计算效率，也反映聚类结果的稳定性与质量。

1. 执行时间

执行时间指各聚类算法完成聚类过程所耗费的总时间，以秒计。本研究用 Python 内置的 time 函数记录执行时间，覆盖从算法初始化到收敛的全过程。该指标直接衡量计算效率，这在大数据分析中尤为关键。混合方法有望通过改善质心初始化来缩短总计算时间：尽管增加了层次阶段的开销，收敛却更快。

2. 收敛迭代次数

收敛迭代次数表示质心初始化之后 K 均值达到稳定所需的精化步数。迭代次数越少，说明质心的初始位置越接近最优，收敛越快、计算负担越轻。在标准 K 均值中，质心初始化不佳会导致多次冗余迭代，既增加执行时间又提高陷入次优聚类的风险；相比之下，混合方法通过提供更好的起始质心改善了收敛。

（四）聚类有效性指标

为评价所得簇的质量，本文采用三种内部验证指标：轮廓系数、Davies‐Bouldin 指数（DBI）与 Calinski‐Harabasz 指数（CH 指数）。这些指标分别从簇内凝聚度、簇间分离度与方差结构加以度量。

轮廓系数定量刻画每个数据点相对于其他簇而言与所属簇的契合程度。对每个数据点 i，其轮廓值 s(i)定义为：

$$s(i) = [b(i) - a(i)] / \max\{a(i), b(i)\} \quad (1)$$

式中，a(i)表示数据点 i 与同簇内其他所有点的平均距离，度量凝聚度；b(i)表示点 i 与其他各簇全部点的最小平均距离，反映分离度。取值范围为−1 至+1：接近+1 表示该点与自身簇匹配良好且与相邻簇明显分离；接近 0 表示该点位于簇的边界；接近−1 则意味着该点可能被错误归类。轮廓系数越高，说明簇越紧致、分离越好；越低则表示存在重叠或结构薄弱。

Davies–Bouldin 指数（DBI）同时考虑簇内紧致度与簇间分离度，评价各簇之间的平均相似性，其数学表达式为：

$$DBI = (1/k) \sum_{i=1}^{k} \max_{j \neq i} [(S_i + S_j) / M_{ij}] \quad (2)$$

式中，k 为簇总数；S_i表示簇 i 内全部数据点到其质心的平均距离，即簇内距离；M_{ij}表示簇 i 与簇 j 质心之间的距离，即簇间距离。DBI 越低，说明簇越紧致、分离度越高。该指数对不平衡或密度可变的数据集尤为有效。

Calinski–Harabasz 指数（CH 指数）又称方差比准则，依据簇间离散度与簇内离散度之比评价聚类结构的质量，其数学定义为：

$$CH(k) = [Tr(B_k) / Tr(W_k)] \times (N - k) / (k - 1) \quad (3)$$

式中，N 为数据点总数，k 为簇数；$Tr(B_k)$为簇间离散度矩阵的迹，刻画各簇之间的分离程度；$Tr(W_k)$为簇内离散度矩阵的迹，反映簇内成员的聚集紧密程度。CH 指数越高，说明聚类质量越好。执行时间、收敛迭代次数与上述三项有效性指标共同构成了从准确性、稳定性与计算效率三个维度评估聚类性能的完整框架。

三、结果与讨论

为说明所提混合分裂式—K 均值模型的运行流程，算法 1 给出了实验中使用的算法步骤，展示了如何执行分裂过程以获得初始簇划分、如何由层次结果计算质心，以及如何随后用 K 均值对这些质心加以精化。

（一）执行时间比较

图 2 给出了聚类结果前 35 个数据点的簇标签：K 均值把大多数点归入簇 2，少量点归入簇 0 与簇 1；凝聚式 K 均值在簇 0、1、2 之间的分布更为分散；分裂式 K 均值所得模式与凝聚式 K 均值非常相似，说明两种混合方法的簇归属结果相当。

```
=== First 35 Data Points of Clustering Results ===
KMeans:              2 2 2 0 2 2 0 0 2 2 1 2 2 0 2 0 2 2 1 1 2 2 2 2 2 2 0 2 0 1 1 1 1 1 1
Divisive-KMeans:     2 2 1 1 1 2 1 1 1 2 2 1 2 1 1 1 2 2 0 2 2 2 1 2 1 2 1 2 1 0 0 0 0 0 0
```

Figure 2 First 35 Data Points of Clustering Results

图 2 聚类结果的前 35 个数据点

表 2 给出了两种聚类方法的执行时间：标准 K 均值与分裂式 K 均值。结果以毫秒计，取多次实验运行的平均值，以减小随机波动的影响。

Table 2 Execution Time Comparison of Clustering Methods

表 2 各聚类方法的执行时间比较

方法	执行时间（ms）	收敛迭代次数	轮廓系数	DBI	CH 指数
K 均值	54.98	5	0.196	1.454	14.3
分裂式 K 均值（DHC）	1.45	3	0.195	14.05	15.8

（二）聚类有效性指标的解读

聚类有效性指标为聚类结果的质量提供了更深入的判断。轮廓系数显示标准 K 均值与分裂式 K 均值取值几乎相同（0.196 对 0.195），说明两者在簇的分离度与凝聚度上水平相近；该分值虽然不高，但与复杂社会经济数据集常呈现组间特征重叠的性质相符。DBI 则表明 K 均值取得了更好的簇间分离（1.454，低于分裂式 K 均值的 14.05）。分裂式 K 均值异常偏高的 DBI 可能反映出各簇虽紧致但彼此位置接近，说明仍需进一步的参数调优或更广泛的基准比较。相反，CH 指数更青睐分裂式 K 均值：其得分（15.8）高于标准 K 均值（14.3），表明整体离散度与紧致度更优——这与其更具结构性的初始化方式相一致。

（三）讨论与局限

总体而言，混合模型表现出更快的收敛速度与更好的稳定性，印证了分裂式初始化对大规模高维数据集的优势。不过，本文发现在很大程度上仍属描述性，且基于规模有限的数据集，难以充分验证该框架关于可扩展性的论断。要在实证上确认该模型适用于大数据场景，还需在更大规模（可能超过数百万条观测）的数据集上开展补充实验。此外，聚类结果虽揭示了有意义的结构模式，但其与贫困政策的直接关联尚未牢固建立。未来工作应整合空间分析、区县层面的社会经济画像与领域专家验证，把聚类结果转化为可操作的政策洞见。

四、结论

混合分裂式—K 均值模型相较标准 K 均值有明显改进，尤其体现在更快的收敛与更稳定的质心初始化上：执行时间由 54.98 ms 降至 1.45 ms，不过这一收益应结合所用数据集规模较小的实际加以看待；迭代次数由 5 次降至 3 次，表明分裂式初始化步骤有效降低了质心的随机性，使聚类过程更为一致可靠。

聚类有效性指标也表明混合方法维持甚至略微提升了聚类质量：更高的 Calinski‑Harabasz 指数（15.8）与相当的轮廓系数（0.195）说明所得簇较为紧致且分离良好，尽管 Davies‑Bouldin 指数有所上升。在社会经济语境下，较高的轮廓系数意味着失业率、受教育水平或服务可及性等贫困特征相近的区县被更一致地归为一组；较高的 Calinski‑Harabasz 值则意味着低、中、高贫困簇之间的区分在结构上更为明确，可支持决策者识别优先区域。反之，Davies‑Bouldin 值上升说明部分簇在社会经济意义上可能仍有重叠。

除技术性能之外，聚类结果对区域社会经济政策也具有现实意义：依据多维贫困指标对区县分组，决策者可识别优先区域、更有效地配置资源并设计有针对性的减贫方案。不过，本研究也存在若干关键局限：数据集所含变量数量有限，可能未能充分刻画贫困动态的复杂性；数据集规模相对较小，限制了对可扩展性论断的检验；且未使用独立数据集开展外部验证。未来研究应把该模型应用于更大规模的数据集，纳入更多社会经济指标，并借助并行或分布式计算框架评估其可扩展性。

参考文献

[1] Ezugwu A E, Elsisi S, Hussain A G, et al. Hybrid firefly algorithms for clustering. IEEE Access, 2020, 8: 121089-121118.

[2] Ahmed M, Seraj R, Islam S M S. The k-means algorithm: a comprehensive survey and performance evaluation. Electronics, 2020, 9(8): 1295.

[3] Ezugwu A E, Akinola I A, Elsisi M H, et al. A comprehensive survey of clustering algorithms. Engineering Applications of Artificial Intelligence, 2022, 110: 104743.

[4] 武福林. 深度聚类算法研究: 方法、评估与应用. 信息记录材料, 2026(6).

[5] 赵杰, 雷秀娟, 吴振强. 基于最优类中心扰动的萤火虫聚类算法. 计算机工程与科学, 2015, 37(2): 314-320.

[6] Bai L, Liang J, Cao F. A multiple k-means clustering ensemble algorithm to find nonlinearly separable clusters. Information Fusion, 2020, 61: 36-47.

[7] Chen Y, Wang L, Li J. A clustering-based approach using K-means and hierarchical methods for large-scale data. Information, 2025, 16(6): 441.

[8] Zhang Y. A comprehensive survey on traffic missing data imputation. IEEE Transactions on Intelligent Transportation Systems, 2024, 25(2): 1234-1245.

[9] 贺玲, 吴玲达, 蔡益朝. 数据挖掘中的聚类算法综述. 计算机应用研究, 2007(1): 10-13.