文章编号:1005-0523(2026)03-0120-07

基于知识诱导与无用中心驱动的*K*-means算法

王 森,刘青阳,詹小秦,陈 炼

(华东交通大学理学院,江西 南昌 330013)

摘要:*K*-means算法是一种广泛应用的高效无监督聚类算法。然而,研究表明,在处理高维或非球状结构的数据集时,*K*-means算法在确定聚类数目和选择初始质心方面存在显著局限性。为优化*K*-means算法的初始质心选取机制,解决聚类数目确定问题,文章提出了一种基于知识诱导与无用中心驱动的*K*-means算法。该算法首先引入高密度知识点的检测机制,通过识别数据集中的高密度知识点,构建候选质心集合,基于高斯混合模型理论推导最优聚类数;随后,采用无用中心筛选策略对候选质心进行优化选择,最终确定最优初始质心集合。在真实数据集上的实验结果表明,所提算法在聚类性能上总体优于其他对比算法。该算法可有效解决非球状数据分布的聚类问题,并在复杂数据结构场景下展现出较为优越的聚类性能。

关键词:无监督聚类;*K*-means;高密度知识点;高斯混合模型;无用中心

中图分类号:TP391 **文献标志码:**A

本文引用格式:王森,刘青阳,詹小秦,等. 基于知识诱导与无用中心驱动的*K*-means算法[J]. 华东交通大学学报,2026,43(3):120-126.

K-means Algorithm Driven by Knowledge Induction and Useless Center

Wang Sen, Liu Qingyang, Zhan Xiaoqing, Chen Lian

(School of Science, East China Jiaotong University, Nanchang 330013, China)

Abstract: *K*-means is a widely used and efficient unsupervised clustering algorithm. However, studies have shown that when dealing with high-dimensional or non-spherically distributed datasets, the *K*-means algorithm has significant limitations in determining the number of clusters and selecting initial centroids. To thoroughly explore and optimize the initial centroid selection mechanism and the problem of determining the number of clusters in the *K*-means algorithm, a *K*-means algorithm based on knowledge induction and useless center driven is proposed. This algorithm first introduces a detection mechanism for high-density knowledge points, constructs a candidate centroid set by identifying high-density knowledge points in the dataset; then infers the optimal number of clusters based on Gaussian mixture model theory; subsequently adopts a useless center screening strategy to optimally select the candidate centroids, and finally determines the optimal initial centroid set. Experiments on real datasets show that the proposed optimized algorithm generally outperforms other comparison algorithms in clustering performance. This algorithm effectively solves the clustering problem of non-spherical data distribution and exhibits relatively superior clustering performance in scenarios with complex data structures.

收稿日期:2025-03-25
基金项目:国家自然科学基金项目(12361004)

Key words: unsupervised clustering; *K*-means; high-density knowledge points; Gaussian mixture model; useless center

Citation format: WANG S, LIU Q Y, ZHAN X Q, et al. *K*-means algorithm driven by knowledge induction and useless center[J]. Journal of East China Jiaotong University, 2026, 43(3): 120−126.

传统*K*-means算法是一种经典的无监督机器学习聚类算法，其核心思想是将数据集自动划分为 *K* 个不同的簇，其中 *K* 为聚类数目。*K*-means算法的目标是确保簇内数据点相似度较高，而簇间相似度较低，从而最小化簇内平方误差[1]。在应用层面，*K*-means算法可揭示数据分布的结构特征，广泛应用于数据挖掘、机器学习、模式识别、生物信息处理等多个领域[2]。

当前，聚类方法体系已形成多维度的技术分支，主要包括划分式聚类、密度聚类、层次化聚类、网格聚类、谱聚类和概率聚类等。领域知识驱动型发现依赖于现有的领域知识体系，涵盖领域规则、理论框架及结构化知识库，通过融合先验知识与逻辑推理机制，揭示新的知识关联或理论假设；而数据知识驱动型发现则基于对海量数据集的深度挖掘，依托数据结构特征、隐含模式与统计规律，运用机器学习、数据科学及高性能计算方法，自动发现潜在的规律与知识[3]。聚类分析通过有机整合上述两种驱动范式，构建了一种优势互补、动态迭代的知识生成与推理框架，有效克服了单一方法在完备性与可解释性方面的固有局限。

作为经典聚类算法，*K*-means历经持续优化。Hartigan等[4]通过构建损失函数并迭代划分数据，证明了其单调递减性并收敛至局部最优解。Selim等[5]系统分析了算法的收敛性，通过形式化推导不同距离空间下的优化轨迹，证实了其在欧氏空间中的局部收敛特性。Comaniciu等[6]提出的Mean Shift算法将*K*-means视为其特例，通过引入核函数与加权机制，基于密度梯度上升迭代逼近高密度区域，增强了多模态分布建模能力。Chinrungrueng等[7]则通过多核学习构建超球面交集结构，进一步提升该算法对复杂数据的处理能力。

作为划分聚类的基准算法，*K*-means算法的性能高度依赖于初始质心、聚类数目与迭代效率三大要素。在处理复杂结构数据时，初始质心的设置及聚类数目的合理判定会显著影响算法的收敛效果与聚类质量[8]。围绕这一方向对*K*-means进行深入研究，对于实现其高效、准确应用具有重要意义。为此，本文提出一种基于知识诱导与无用中心驱动的*K*-means算法（*K*-KIDUC算法），旨在通过结构化知识驱动机制，优化初始质心生成与聚类数目判定过程。该方法首先提取数据集的高密度区域构建候选质心集，然后采用高斯混合模型拟合全局分布以推断聚类数目，再通过无用中心剔除策略去除冗余质心，最后从优化的候选集中筛选出 *K* 个质心作为初始输入。实验结果表明，该算法在聚类准确性方面表现优异，尤其在非球状分布数据中展现出更强的鲁棒性与泛化能力。

1　*K*-means算法

K-means算法是一种简单的无监督聚类方法，其核心思想是通过迭代优化，将数据集划分为*K*个簇，最小化各个簇内的数据点到质心的距离平方和[9]。

给定数据集 $X=\{x_1,x_2,\cdots,x_i\}$，$(i=1,2,\cdots,N)$，其中 N 为数据集中数据点的总数，x_i 为数据集中第 i 个数据点。质心集合 $C=\{C_1,C_2,\cdots,C_n\}$，簇集合 $S=\{S_1,S_2,\cdots,S_n\}$，$(n=1,2\cdots,K)$。C_n 是第 n 簇的质心，S_n 是第 n 簇中所有数据点的集合，其任意两个簇的数据点不重复。聚类的目标函数 SSEMD(·) 是最小化簇内数据点到其质心的欧氏距离平方和，表达式为

$$\mathrm{SSEMD}(C)=\sum_{n=1}^{K}\sum_{\forall P\in S_n}\mathrm{dis}(P,C_n) \tag{1}$$

式中：dis(·) 为两个数据点的欧氏距离；P 为第 n 簇中的数据点。

传统*K*-means算法的执行流程如下。给定数据集与预设聚类数目 K，算法首先随机初始化 K 个质心，随后迭代执行以下两个步骤直至收敛：① 分配步骤：基于最小欧氏距离将各数据点划分至最近质心所代表的簇；② 更新步骤：重新计算每个簇的

均值向量作为新质心。该过程重复进行，直至满足收敛条件或达到最大迭代次数。算法的时间复杂度为 $O(N{\cdot}K{\cdot}T{\cdot}d)$。其中 d 为数据维度，T 为迭代次数。

2 改进算法 *K*-KIDUC

2.1 基于高密度知识点提取预初始质心

与传统的领域知识驱动不同，高密度知识点是采用非人工处理和约束规则提取的数据内在特征。知识点是指在数据集中一些具有高密度的数据点。与其他数据点相比，知识点具有较高的局部相对密度[10]。现引入平均距离作为数据的密度基准，利用最近邻策略[11]计算数据点的局部密度 ρ_i。

定义 1 任意数据点 x_i 的局部密度为 ρ_i，定义为

$$\rho_i=\frac{\text{average}(X)}{\frac{1}{\varphi}\sum_{P\in B_\varphi}\text{dis}(P,x_i)} \tag{2}$$

式中：B_φ 为距离数据点 x_i 最近的 φ 个数据点的集合，也称为 x_i 最近邻集合[12]；φ 为正整数的阈值；average(X) 是数据集中所有数据点之间欧氏距离的平均值。

如何选择一个合适的 φ 值是确定邻域大小的关键。为此本文提出新的公式。

定义 2 设 N 是数据集 X 中数据点的总数，阈值 φ 定义为

$$\varphi=\frac{\sqrt{N}}{h} \tag{3}$$

式中：h 通常设置为 1 或 2，若数据集规模过大或数据集中存在尺寸特别小的簇时，则需要设置更大的值。

局部密度 ρ_i 提供了更精准的处理策略。average(X) 考虑数据点之间的全局距离信息，基于自然最近邻的思想，定义了自然最近邻权重值，并可计算出数据点 x_i 到更高密度点的超距离 δ_i。

定义 3 数据点 $\forall x_i\in X$ 的超距离为 δ_i，用于衡量数据点 x_i 到其他更高密度点的最小距离，δ_i 定义为

$$\delta_i=\min_{j\in I,\rho_j>\rho_i}\text{dis}(x_i,x_j) \tag{4}$$

式中：I 为 x_i 局部密度大于数据点的所有数据点的下标索引集合；ρ_j 为数据点 x_j 的局部密度。若数据点 x_i 为数据集 X 中局部密度高的点，则定义其超距离 δ_i 为所有数据点间距离的平均值，计算式为

$$\delta_i=\text{average}(X) \tag{5}$$

定义 4 设 W 是数据集 X 中所有数据点的超距离，$W=\{\delta_1,\delta_2,\cdots,\delta_i\}$，$(i=1,2,\cdots,N)$。对 W 进行聚类数目为 2 的 *K*-means 聚类，最终得到 2 个聚类质心，t_h 为这两个聚类质心的均值。

定义 5 设 $x_i,x_j\in X$，$(i\neq j)$。若 x_j 是 x_i 最近的、且密度更高的邻居，则称 x_i 为 x_j 的直联系下级，称 x_j 为 x_i 的直联系上级，记作 $x_i\rightarrow x_j$。将每个数据点直接依赖的数据点记为 η_i，局部密度越大的数据点通常具有更多的依赖数据点，这可用于评估数据点作为局部质心的影响力，并有助于过滤噪声点。数据点 x_i 的直接依赖数量 η_i 的计算式为

$$\eta_i=\sum_{j\in C_\text{O},i\neq j}\zeta(x_i,x_j) \tag{6}$$

式中：C_O 为数据集 X 的下标索引集合；$\zeta(\cdot)$ 定义为

$$\zeta(x_i,x_j)=\begin{cases}1 & (\rho_j<\rho_i,\delta_j<t_\text{h})\\ 0 & (\text{else})\end{cases} \tag{7}$$

定义 6 设 $x_i,x_j\in X,(i\neq j)$，且 D_{η_i} 是数据点 x_i 与 η_i 个最近邻的数据点构成的集合。x_i 的相对密度 ρ'_i 定义为

$$\rho'_i=\sum_{i\in C_\text{O},j\neq i}\xi(x_i,x_j) \tag{8}$$

式中：$\xi(x_i,x_j)$ 的计算式为

$$\xi(x_i,x_j)=\begin{cases}1 & (x_j\rightarrow x_i,x_j\in D_{\eta_i})\\ 0 & (\text{else})\end{cases} \tag{9}$$

整个算法流程如下：首先，根据式(2)和式(3)计算各数据点的局部密度；其次，利用式(4)或式(5)计算数据集中各点的超距离，得到超距离数据集 W，对 W 进行聚类数目为 2 的 *K*-means 聚类，将所得两个聚类质心的均值作为局部密集区域的最大半径；然后，利用式(8)和式(9)计算各数据点的直接依赖数量 η_i 及其 η_i 个最近邻数据点集合 D_{η_i}；最后，根据式(8)和式(9)计算各数据点的相对密度。将相对密度较高的数据点存入候选质心集 G 中。整个算法的时间复杂度为 $O(N^2\cdot d)$。

2.2 基于高斯分布的聚类数目确定方法

传统 *K*-means 算法的聚类数目 K 依赖于先验知识的人工预设。这种方式在处理低维、高可分数据时效果尚可，但在面对复杂数据结构时，由于簇间边界模糊、度量失真以及评估指标敏感性不足，聚类数

目判定准则的鲁棒性会急剧下降[13]。质心分布与数据的多模态密度特征之间存在强耦合关系[14],这是聚类效能随聚类数目设定变化而敏感的核心原因。尽管高斯混合模型(Gaussian mixture model,GMM)可以通过期望最大化(expectation-maximization,EM)算法优化似然函数,但其对数据分布的参数化预设(即假设数据由多个高斯分布混合生成)往往与实际情况不符,这一根本性的假设冲突可能导致聚类质心估计出现偏差。

定义7　设 D_0 为簇内数据服从高斯分布,D_1 为质心周围数据点的分布偏离高斯特性。当簇满足 D_0 状态,无需进行分裂操作;当簇满足 D_1 状态,则执行分裂策略,分割成两个新簇并重新生成质心[15]。

基于高斯混合模型确定聚类数目的策略如下:首先,将初始聚类数目设置为1,并根据传统*K*-means算法更新质心和簇集合 S。检验每个簇是否符合高斯分布(即处于 D_0 或 D_1 状态)。若簇满足 D_0 状态,则保留当前质心;若满足 D_1 状态,则分裂该簇,并更新质心和簇集合 S。重复执行上述步骤,直至聚类质心不再变化,最终确定聚类数目*K*。

2.3　基于无用中心提取初始质心

本文提出一种基于无用中心提取初始质心(UCI)的优化策略。该策略将无用中心的概念由传统*K*-means算法迭代后可能形成的空簇,前置定义为初始化阶段依据数据点间关联性判定的无效候选质心。通过从源头筛选初始质心集,本策略简化了后续计算路径,并显著提升了最终聚类的鲁棒性与准确性。

K-means算法的目标函数是最小化簇内数据点到其质心的距离平方和SSEMD(C),可利用定义1和式(1)计算得出。然而,完全最小化SSEMD(C)并不能保证获得最佳的聚类质量,且该目标函数值对初始质心的选择非常敏感[16]。

定义8(无用中心)　在数据集 X 中,$x_i,x_j,x_t\in X,(i\neq j\neq t)$。若数据点 x_j 对数据点 x_i 无用中心,则需同时满足以下两个条件

$$\mathrm{dis}(x_i,x_t)<\mathrm{dis}(x_i,x_j) \tag{10}$$

$$\mathrm{dis}(x_j,x_t)<\mathrm{dis}(x_i,x_j) \tag{11}$$

如果数据点 x_j 相对于 x_i 不满足上述条件,则它是数据点 x_i 的有用中心。

定义9　在数据集 X 中,数据点 $x_i\in X$ 的平均价值为

$$V(x_i)=\max_{\forall p_o\in C_{\mathrm{UN}}}(\mathrm{dis}(x_i,p_o))\times\sum_{\forall p_o\in C_{\mathrm{UN}}}\ln(\mathrm{dis}(x_i,p_o)) \tag{12}$$

式中:p_o 表示"有用中心"集合 C_{UN} 中的数据点。通常,平均价值 x_i 较高的数据点被选为下一个候选质心 $C_{\mathrm{UN}x}$,并将其存入有用中心的数据点集中。

在整个初始化过程中,首先从数据集中选择某一属性具有最小值的数据点作为第1个质心。随后,在每次迭代中,选取具有最大平均价值的数据点作为下一个质心。该初始化过程持续进行,直至达到预设的聚类数目*K*为止。在初始化过程中,如果最近添加的质心对数据点无用,则直接忽略它;否则,将其加入候选中心集合 C_{UN} 中。在这种情况下,新添有用质心时,也可能会从集合 C_{UN} 删除一些数据点。去除无用中心的时间复杂度表示为 $O(N\cdot K^2\cdot d)$。

2.4　*K*-KIDUC算法流程

考虑到高密度知识点高度依赖数据密度分布的特性,这种依赖关系很容易导致极端情况的出现。通过人工添加约束条件,包括必连约束和勿连约束,来强制某些数据点进行特定分组,以此强化业务逻辑,弥补特征数据存在的不足[17]。

定义10　在数据集 X 中,对于存在必连约束的两个数据点,它们之间直接依赖的数据点所包含的最近邻数据点也必须位于同一个簇中。如果这两个数据点存在勿连约束关系,那么其直接依赖的数据点所包含的最近邻数据点一定不能位于同一个簇中。

以下为*K*-KIDUC算法具体步骤。

输入:必连约束集合 J_{must},勿连约束集合 J_{cannot},数据集 X。

Step 1:对数据集 X 应用高斯混合模型的聚类数目确定方法,得到聚类数目 K;

Step 2:利用基于高密度知识点提取预初始质心算法,计算数据集 X 高密度知识点,构建候选质心集 G;

Step 3:根据无用中心策略,在候选质心集 G 中选取 K 初始质心 $C=\{C_1,C_2,\cdots,C_K\}$;

Step 4:更新分配,计算数据点到各个质心的欧氏距离,将数据点划分到最近邻簇中;

Step 5:对比各个簇以及 J_{must} 和 J_{cannot} 集中的数据点,根据定义10,重新划分数据点;

Step 6:更新质心,根据当前各个簇的情况,计算簇内所有数据点的均值作为新的质心;

Step 7:迭代更新,重复Step4~Step6,不断划分数据点并更新质心,直至达到收敛条件。

输出:最终聚类结果 S 。

K-KIDUC算法的时间复杂度主要分为两个部分。第1部分是确定初始质心的时间复杂度,第2部分是划分数据点部分。综合分析,其时间复杂度表示为 $O(N^2\cdot d+N\cdot K\cdot T\cdot d+K^2\cdot d)$ 。其中 T 为算法迭代次数。

3 实验分析

3.1 实验结果

本部分详细阐述了实验设计与评估框架。在对比实验中,选取了3种基准算法:基于距离硬划分的经典*K*-means算法、通过概率化初始化优化质心选择的*K*-means++算法,以及基于概率生成框架并假设数据来源于有限高斯混合分布的高斯混合模型(GMM)。为验证所提算法组件的有效性,同步设计了消融实验,通过系统性移除或替换关键模块,考察不同聚类数目与约束条件下各算法变体的性能差异。所有实验均在真实的Banknote数据集上进行。

实验结果通过可视化与量化指标相结合的方式呈现。图1展示了数据集聚类结果的可视化对比;图2与图3分别描述了消融实验的设计逻辑与模型变体生成路径;表1,表2则记录了详尽的算法性能指标,包括调整互信息(AMI)、调整兰德指数(ARI)、戴维森堡丁指数(DBI)及运行时间(Time),综合评估各算法在聚类质量与计算效率方面的表现。

AMI是衡量模型生成的簇与真实标签之间一致性的指标,[−1,1],值越接近1表示聚类效果越好。ARI是用于评估聚类分析效果的指标,[−1,1],值越接近1表示聚类结果与真实情况越一致。DBI用于衡量任意两个簇的簇内距离之和与簇间距离之比,其值越小表示聚类效果越好。Time表示聚类运行时间,其值越小,表明算法计算效率越高。

表2中,必连约束和勿连约束列下的0表示未引入该约束,1表示引入该约束。

3.2 分析与评估

在Banknote数据集的对比实验中,该数据集呈现出簇间分布不均衡、噪声干扰显著以及类别边界模糊等复杂特性。实验结果表明,所提出的*K*-KIDUC算法在AMI和ARI指标上优于基准方法,性能提升幅度普遍较高,展现出良好的抗噪声能力和

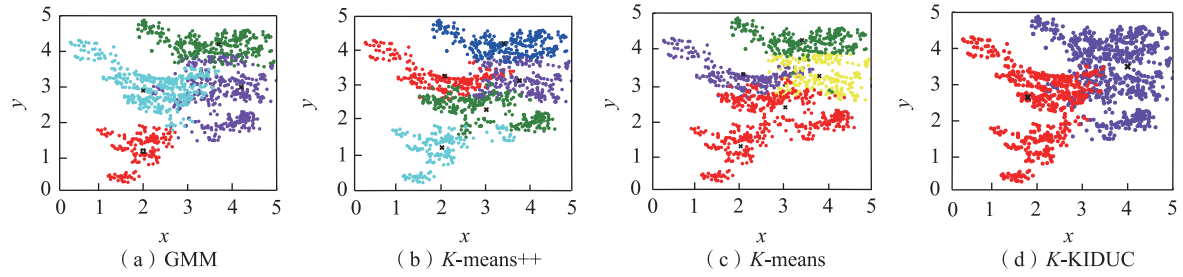


图1 4种算法在Banknote数据集上的聚类结果对比

Fig. 1 Comparison of clustering results of 4 algorithms on the Banknote dataset

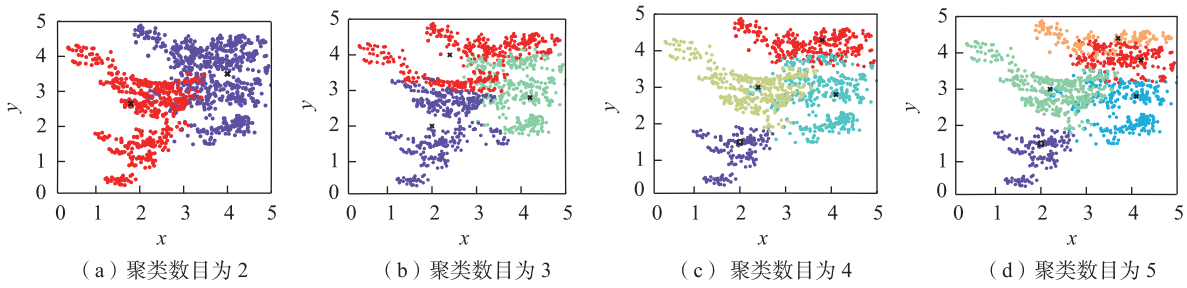
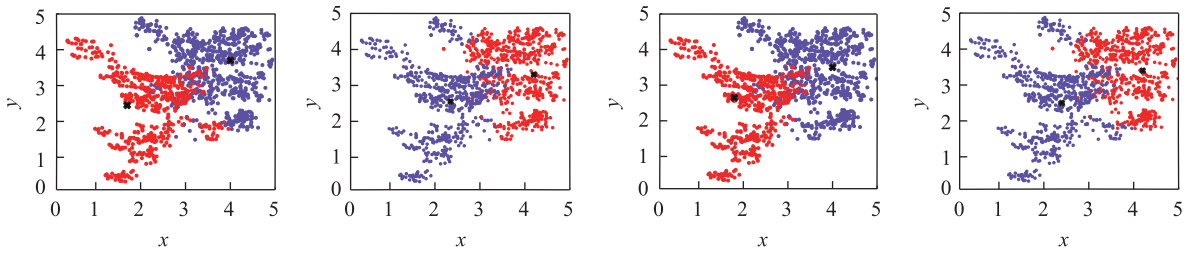


图2 Banknote数据集上4种聚类数目的消融实验图

Fig. 2 Ablation experiment graph of 4 clustering numbers on the Banknote dataset



（a）添加必连约束　（b）添加勿连约束　（c）添加必连和勿连约束　（d）无约束条件

图3　Banknote数据集上约束条件的消融实验图

Fig. 3　Ablation study diagram of constraint conditions on the Banknote dataset

表1　Banknote数据集上4种算法的评价指标对比

Tab. 1　Comparison of evaluation metrics of 4 algorithms on the Banknote dataset

算法	AMI	ARI	DBI	时间/s
K-means	0.769 5	0.768 5	0.735 3	0.563 5
GMM	0.673 4	0.534 2	1.149 7	2.099 6
K-means++	0.332 3	0.239 0	1.071 7	0.679 3
K-KIDUC	0.853 6	0.906 0	0.953 8	8.760 2

表2　Banknote数据集上4种约束条件的消融实验性能分析

Tab. 2　Performance analysis of ablation experiments for 4 constraint conditions in Banknote

必连约束	勿连约束	AMI	ARI	DBI	时间/s
1	0	0.738 8	0.832 6	0.917 3	8.906 1
0	1	0.680 7	0.717 1	2.028 0	9.356 0
0	0	0.638 9	0.663 7	1.885 7	9.722 1
1	1	0.853 6	0.906 0	0.953 8	8.760 2

聚类稳定性。

消融实验进一步揭示，当预设聚类数目与真实类别数一致时，算法性能达到较优状态；而在同时引入必连约束和勿连约束的条件下，聚类效果最佳，这表明约束信息的引入对提升算法的判别能力具有重要作用。

综合来看，*K*-KIDUC算法在聚类质量上显著优于对比算法，但其计算成本较高，表现为运行时间最长，收敛速度较慢。消融分析还表明，聚类数目的增加会进一步延缓算法收敛；而结合使用两类约束则能够在提升聚类性能的同时，有效加速收敛过程，反映出约束机制在权衡算法效率与效果方面的积极价值。

4　结论

为降低*K*-means算法在初始质心选择和聚类数目设定上的局限性，本文提出*K*-KIDUC算法。主要创新工作包括：引入数据内在知识驱动，识别数据集的高密度知识点；利用高斯混合模型预测数据集的趋势走向，获取聚类数目；采用无用中心筛选策略，剔除无用中心，筛选初始质心；添加必连约束和勿连约束，有效减少后续点聚类过程中标签错误分配的情况。

实验结果表明，*K*-KIDUC算法在处理簇间数据点分布不均衡的数据集时更具优势，具有更好的抗噪能力。然而，优化算法虽提升了聚类性能，却导致算法收敛速率降低；同时，初始化阶段的高复杂度致使*K*-KIDUC算法存在可扩展性不足的固有局限。聚类质量与计算效率的平衡是算法实际应用的核心关键，未来研究将聚焦于优化初始化流程或探索轻量化替代策略，以期突破这一瓶颈。

参考文献：

[1] 王森, 邢帅杰, 刘琛. 密度峰值聚类算法研究综述[J]. 华东交通大学学报, 2023, 40(1): 106-116.
WANG S, XING S J, LIU C. Survey of density peak clustering algorithm[J]. Journal of East China Jiaotong University, 2023, 40(1): 106-116.

[2] 周晓东, 董海清, 张昆鹏, 等. 基于几何的*K*-means初始聚类中心优化算法研究[J]. 仪表技术, 2025(2): 66-69.
ZHOU X D, DONG H Q, ZHANG K P, et al. Research on optimization algorithm for initial clustering centers of *K*-means based on geometry[J]. Instrumentation Technology, 2025(2): 66-69.

[3] 姚苏梅, 陆泉. 数据与知识协同驱动的知识发现: 概念、机理与模型[J]. 情报学报, 2025, 44(3): 282-295.
YAO S M, LU Q. Knowledge-discovery method driven by the collaboration of data and knowledge: concept, mechanism, and model[J]. Journal of the China Society for Scientific and Technical Information, 2025, 44(3):

282-295.

[4] HARTIGAN J A, WONG M A. Algorithm AS 136: a *K*-means clustering algorithm[J]. Applied Statistics, 1979, 28(1): 100.

[5] SELIM S Z, ISMAIL M A. *K*-means-type algorithms: a generalized convergence theorem and characterization of local optimality[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1984, 6(1): 81-87.

[6] COMANICIU D, MEER P. Mean shift: a robust approach toward feature space analysis[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(5): 603-619.

[7] CHINRUNGRUENG C, SEQUIN C H. Optimal adaptive *K*-means algorithm with dynamic adjustment of learning rate[J]. IEEE Transactions on Neural Networks, 1995, 6 (1): 157-169.

[8] IKOTUN A M, EZUGWU A E, ABUALIGAH L, et al. *K*-means clustering algorithms: a comprehensive review, variants analysis, and advances in the era of big data[J]. Information Sciences, 2023, 622: 178-210.

[9] 孙林, 刘梦含, 薛占熬. 结合人工蜂群与*K*-means聚类的特征选择[J]. 计算机科学与探索, 2024, 18(1): 93-110.
SUN L, LIU M H, XUE Z A. Feature selection combining artificial bee colony with *K*-means clustering[J]. Journal of Frontiers of Computer Science and Technology, 2024, 18(1): 93-110.

[10] TANG Y M, PAN Z F, HU X H, et al. Knowledge-induced multiple kernel fuzzy clustering[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(12): 14838-14855.

[11] 原泽菲, 张正军, 姜国林. 基于相对邻近度的自适应谱聚类算法[J]. 计算机科学, 2025, 52(10): 79-89.
YUAN Z F, ZHANG Z J, JIANG G L. Adaptive spectral clustering algorithm based on relative proximity[J]. Computer Science, 2025, 52(10): 79-89.

[12] 张清华, 周靖鹏, 代永杨, 等. 基于代表点与*K*近邻的密度峰值聚类算法[J]. 软件学报, 2023, 34(12): 5629-5648.
ZHANG Q H, ZHOU J P, DAI Y Y, et al. Density peaks clustering algorithm based on representative points and *K*-nearest neighbors[J]. Journal of Software, 2023, 34 (12): 5629-5648.

[13] SARMADI H, ENTEZAMI A, MAGALHÃES F. Unsupervised data normalization for continuous dynamic monitoring by an innovative hybrid feature weighting-selection algorithm and natural nearest neighbor searching[J]. Structural Health Monitoring, 2023, 22(6): 4005-4026.

[14] 何选森, 何帆, 徐丽, 等. *K*-means算法最优聚类数量的确定[J]. 电子科技大学学报, 2022, 51(6): 904-912.
HE X S, HE F, XU L, et al. Determination of the optimal number of clusters in *K*-means algorithm[J]. Journal of University of Electronic Science and Technology of China, 2022, 51(6): 904-912.

[15] PATEL E, KUSHWAHA D S. Clustering cloud workloads: *K*-means vs Gaussian mixture model[J]. Procedia Computer Science, 2020, 171: 158-167.

[16] AHMED M, SERAJ R, ISLAM S M S. The *K*-means algorithm: a comprehensive survey and performance evaluation[J]. Electronics, 2020, 9(8): 1295.

[17] 周晨曦, 梁循, 齐金山. 基于约束动态更新的半监督层次聚类算法[J]. 自动化学报, 2015, 41(7): 1253-1263.
ZHOU C X, LIANG X, QI J S. A semi-supervised agglomerative hierarchical clustering method based on dynamically updating constraints[J]. Acta Automatica Sinica, 2015, 41(7): 1253-1263.

第一作者：王森（1969—），男，教授，硕士生导师，研究方向为计算机算法与应用。E-mail：wangsen@ecjtu.edu.cn。

通信作者：刘青阳（2001—），女，硕士研究生，研究方向为无监督聚类分析。E-mail：2023088085410003@ecjtu.edu.cn。

（责任编辑：姜红贵）