DOI:10. 13382/j. jemi. B2508433

种群优化联合鲁棒距离度量的公平性 K-means 算法*

谢一涵[1] 毕鹏飞[1] 王爱萍[2]
(1. 南京信息工程大学人工智能学院(未来技术学院) 南京 210044;2. 南京信息工程大学软件学院 南京 210044)

摘 要:随着聚类算法在智能测量系统、多源传感数据分析与嵌入式状态识别等场景中的广泛应用,如何在保证聚类质量的同时兼顾敏感属性的公平性,已成为制约聚类算法在关键测量任务中应用效果的瓶颈问题。为解决上述问题,提出了一种创新的种群优化联合鲁棒距离度量的公平性 K-means 聚类算法(PODM-Kmeans)。该方法在构建过程中,充分考虑到敏感属性的公平性与聚类质量之间的平衡性,引入改进的布谷鸟搜索算法以实现初始聚类中心选择过程中的全局搜索能力和局部搜索能力的平衡,有效增强了聚类效果的稳定性。在此基础上,在聚类迭代目标函数的构建上,该方法有效采用了公平性约束和簇大小约束机制,并融合了灵活的加权欧氏范数作为距离度量方法,合理抑制了异常值所带来的消极影响,助力了公平性的提升。通过在 5 个合成数据集和 5 个真实数据集上进行的大量实验结果表明,PODM-Kmeans 在同类方法中具有较优的性能表现,尤其在 Adult、Bank、Census1990 和 CreditCard 4 个数据集上,在维持一定的聚类效果的同时,PODM-Kmeans 的公平性比率(FR)指标均超过 0. 95。

关键词: K-means 聚类;公平性;种群优化;鲁棒距离度量;布谷鸟搜索算法;欧式距离
中图分类号: TN911. 3 **文献标识码**: A **国家标准学科分类代码**: 510. 40

Population optimization combined with robust distance metric for fair K-means clustering algorithm

Xie Yihan[1] Bi Pengfei[1] Wang Aiping[2]
(1. School of Artificial Intelligence (School of Future Technologies), Nanjing University of Information Science & Technology, Nanjing 210044, China; 2. School of Software, Nanjing University of Information Science & Technology, Nanjing 210044, China)

Abstract: With the widespread application of clustering algorithms in intelligent measurement systems, multi-source sensor data analysis, and embedded state recognition, ensuring fairness with respect to sensitive attributes while maintaining clustering quality has become a key challenge that limits their effectiveness in critical measurement tasks. To address this issue, we propose a population optimization combined with robust distance metric for fair K-means clustering method (PODM-Kmeans). The proposed method balances clustering quality and fairness by incorporating an enhanced Cuckoo Search algorithm to achieve a trade-off between global and local search capabilities during the initialization of cluster centers, thereby improving clustering stability. Furthermore, fairness constraints and cluster size constraints are effectively integrated into the iterative clustering objective function. A flexible weighted Euclidean norm is adopted as the distance metric to mitigate the negative impact of outliers, contributing to improved fairness. Extensive experiments conducted on five synthetic and five real-world datasets demonstrate the superior performance of PODM-Kmeans compared to existing methods. Notably, on the Adult, Bank, Census1990, and CreditCard datasets, PODM-Kmeans achieves a fairness ratio (FR) exceeding 0. 95 while maintaining high clustering quality.

Keywords: K-means flustering; fairness; population-based optimization; robust distance metric; cuckoo search algorithm; Euclidean distance

收稿日期:2025-06-03 Received Date: 2025-06-03
*基金项目:江苏省基础研究计划基金(BK20220452)项目资助

0 引　言

随着机器学习在银行、医疗、招聘、教育和刑事司法等关键领域的广泛应用,确保其决策公正性和无偏性变得尤为重要[1]。在机器学习中,聚类算法通过分析数据的内在结构,自动识别数据中的模式或类别,广泛应用于数据分析、智能测试、识别检测等应用领域[2-4]。群体公平性与平衡性是衡量聚类算法公平性的两个关键维度。群体公平性要求聚类结果中每个群体的成员分布比例与该群体在总体中的比例尽可能一致,聚类的平衡性则要求敏感群体中成员数量与另一个敏感群体中成员数量之间的最小比例尽可能接近[5]。聚类分析的公平性在多个领域应用中对决策质量和用户体验有着重要影响[6]。例如,在招聘与人才筛选中,聚类分析有助于识别潜在的候选人,通过引入公平性约束,可减少种族、性别、年龄等敏感属性对决策结果的干扰,提升算法的社会可信度与可接受性。在金融信贷评估中,聚类技术被用于客户分群与信用风险评估,聚类结果的公平性对于构建公正透明的评估系统至关重要,尤其在面对结构复杂或历史偏见数据时更显关键。随着各类智能测量系统与嵌入式决策设备在招聘评估、金融风控、医疗监测等领域的广泛部署,聚类分析作为底层数据处理与特征提取的重要手段,其算法的稳定性与公平性正成为电子测量与仪器系统亟需关注的核心问题之一。

K-means 算法因其计算效率高、实现简便而成为最常用的聚类方法之一。在实际应用中,K-means 也展现出良好的适应性[7]。韦子辉等[8]结合 ISODATA 思想改进了 K-means 算法,有效提高了非视距环境下识别技术的精确性与稳定性,拓展了该算法在无线通信领域的实际应用能力。Lee 等[9]将其用于标准单元的寄生电容预测问题,进一步展示了其在工业信息领域的良好适用性。然而,随着 K-means 被应用于涉及敏感群体的数据分析任务中,算法在公平性方面的局限性逐渐显现。传统的聚类方法往往忽视了群体代表性分布,可能在无意中加剧原有数据中的偏差,损害特定群体的利益。为了提升聚类算法的公平性,近期许多学者提出了很多改进办法。Suman 等[10]提出了一种可适用于多种目标函数的通用公平聚类框架,能够在确保群体代表性约束的同时维持良好的聚类性能。Ziko 等[11]进一步引入变分推理机制,通过将群体分布偏差建模为正则项,实现了公平性与聚类质量之间的灵活权衡。除了群体层面的公平性,Xu 等[12]在经典的 K-means 算法中加入了群体比例约束,以保证各簇中各群体的代表性与整体分布相符。在此基础上,Yang 等[13]则关注个体层面的公平性,提出在谱聚类中引入成本保证机制,以限制相似样本被分配至不同簇的风险。然而,在满足某些数据集的公平性和群体平衡性约束时,这些算法可能需要对群体分布进行过度调整,从而导致簇被划分得过于细致,进而生成小簇。针对这一问题,Pan 等[14]进一步优化了公平性与聚类效果的融合,融合了簇大小不均的惩罚项在公平性目标函数中,有效避免了小簇的生成。然而,它的聚类效果仍然受限于初始聚类中心的选择,若初始簇中心选取不当,算法可能会陷入局部最优解,可能倾向于聚合某些特定的群体或样本,忽视其他群体的平衡,从而影响聚类公平性。同时,文献[15-19]提出了优化初始聚类中心选取以提升聚类效果的算法,然而,这些方法都并未考虑到聚类过程中的公平性问题。

综上所述,现有公平性研究虽在公平性建模方面取得一定进展,但普遍存在初始聚类不稳定、距离度量缺乏鲁棒性以及优化策略效率不足等问题。因此,本文提出了一个种群优化联合鲁棒距离度量的公平性 K-means 聚类算法(PODM-Kmeans),旨在高效处理包含单一敏感多值属性的数据集,实现聚类效果与公平性之间的更优平衡。

PODM-Kmeans 算法采用了一种新颖的种群优化算法来选取初始聚类中心,其不仅能提升聚类的稳定性,而且可在全局范围内寻找到最优解,实现对于各类别数据点平等均衡的对待。构建了同时兼顾聚类效果与公平性的目标函数,并引入了加权欧氏范数作为距离度量,能够保证聚类质量的同时有效提升结果的公平性。针对离散型目标函数的求解问题设计了一种迭代算法,该算法能够快速收敛到最优解,实现计算效率的有效提升并确保所获取解的准确性。

1 相关方法

1.1 欧氏距离(Euclidean distance)

欧氏距离是最常见的距离度量之一,用于计算空间中两点之间的直线距离[20]。其计算方法简单、含义直观,因此被广泛应用于各类聚类算法中。对于 $X=(x_1,x_2,\cdots,x_n)$,$Y=(y_1,y_2,\cdots,y_n)$,欧氏距离的计算公式为:

$$d(X,Y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2} \tag{1}$$

欧氏距离假设数据点在特征空间中均匀分布,但在许多实际应用中,数据往往具有异质性或包含敏感属性,这使得欧氏距离在计算数据点之间的相似度时可能无法准确反映不同群体之间的实际差异。

1.2 K-means 算法

K-means 算法基于欧氏距离对样本进行聚类,通过最小化簇内样本点到簇中心的平方距离和来实现数据的有效划分。K-means 算法旨在最小化各个聚类簇内样本

点到其中心的距离平方和,从而使同一簇内数据相似性最大,聚类效果最优,其公式如下:

$$J = \sum_{k=1}^{K} \sum_{i=1}^{n_k} \| x_i^{(k)} - \mu_k \|^2 \tag{2}$$

式中:J 为是总代价函数;K 是聚类的数量;$x_i^{(k)}$ 是 k 类中的第 i 个数据点;μ_i 是 k 类的聚类中心;n_k 是 k 类中的样本数。

然而,传统的 K-means 算法的效果往往依赖于初始聚类中心的选择,并未考虑到聚类过程中的公平性问题。由于初始聚类中心的随机性,算法容易陷入局部最优解,导致聚类结果的不稳定性。同时,该算法对数据的分布和密度变化较为敏感,尤其是在处理具有复杂结构或不均匀分布的数据时,往往无法得到理想的聚类效果和公平性。

1.3　平衡公平 K 均值聚类算法(balanced fair K-means,BFKM)

针对传统 K-means 在公平性和簇平衡方面的不足,Pan 等[14]提出了 BFKM 算法,在 K-means 框架下引入了平衡性约束与群体代表性约束,从而在一定程度上提升了聚类结果对不同敏感群体的公平性与均衡性。

BFKM 算法通过定义指示矩阵 $\boldsymbol{F} = [f_1, f_2, \cdots, f_n]^{\mathrm{T}} \in \boldsymbol{R}^{n \times m}$ 和 $\boldsymbol{Y} = [y_1, y_2, \cdots, y_n]^{\mathrm{T}} \in \boldsymbol{R}^{n \times k}$ 分别表示每个样本所属敏感属性和所属簇,其公平性约束可以表示为:

$$\frac{\boldsymbol{Y}_{:j}^{\mathrm{T}} \boldsymbol{F}_{:l}}{\boldsymbol{F}_{:l}^{\mathrm{T}} F_{:l}} = \frac{\boldsymbol{Y}_{:j}^{\mathrm{T}} 1}{n} \tag{3}$$

$$\frac{\boldsymbol{Y}^{\mathrm{T}} \boldsymbol{F}}{\boldsymbol{F}^{\mathrm{T}} \boldsymbol{F}} - \frac{\boldsymbol{Y}\boldsymbol{Y}}{n} = 0 \tag{4}$$

此算法为了避免在进行公平约束聚类过程中出现极小簇影响聚类效果的现象,应尽可能保证每个簇的大小一致。因此,通过引入惩罚项来优化目标函数,其表达式为:

$$\min_{M,Y} (d(i,c) + \rho \| \frac{\boldsymbol{Y}^{\mathrm{T}} \boldsymbol{F}}{\boldsymbol{F}^{\mathrm{T}} \boldsymbol{F}} - \frac{\boldsymbol{Y}\boldsymbol{Y}}{n} \| + \lambda tr((\boldsymbol{Y}^{\mathrm{T}} \boldsymbol{Y})^{-1})) \tag{5}$$

相比仅依赖欧氏距离和无约束分配的 K-means 算法,BFKM 有效缓解了簇大小不均或某些群体被忽视的问题。然而,BFKM 仍然存在一些稳定性和公平性方面的缺点。在稳定性方面,BFKM 对初始中心较为敏感,可能导致局部最优解。在公平性方面,尽管 BFKM 引入了平衡性约束,但由于采用欧氏距离进行迭代,算法可能忽略数据点之间的特征差异及其内部结构,这可能导致部分簇的质量下降,并降低了算法的鲁棒性。

综上所述,这 3 种方法在聚类分析中体现了从基础距离度量到聚类策略设计,再到公平性优化的逐步演化过程。欧氏距离作为最基本的相似性度量方法,计算简便且易于理解,但在处理属性差异较大或涉及敏感信息的数据时,区分能力相对有限。在此基础上,K-means 算法利用欧氏距离对数据进行快速划分,在聚类效率和簇内紧凑性方面表现较好,然而未对群体代表性加以约束,可能引发结果对部分群体的偏倚。进一步地,BFKM 在 K-means 框架之上引入了平衡性与公平性约束,从而在一定程度上改善了聚类分析中的簇规模及群体分布的合理性。

2　算法介绍

2.1　算法总体框架

由于聚类分析在聚类中心初始化的稳定性、距离度量的适用性以及聚类结果的公平性方面尚存优化空间。对此,本文提出了 PODM-Kmeans 算法。其中包括了初始化聚类中心模块,鲁棒距离度量模块,聚类公平性模块以及算法求解策略。其流程结构如图 1 所示。PODM-Kmeans 算法在初始化聚类中心阶段引入种群优化算法进行全局搜索,随后构造加权欧氏距离与公平性和聚类大小平衡约束相融合的目标函数,最后结合坐标下降法求解非连续优化问题。

2.2　初始化聚类中心

在初始化聚类中心部分,本文融合了改进的布谷鸟搜索算法,如图 2 所示,最佳初始聚类中心搜索过程模拟布谷鸟寄生繁殖的行为。本文应用其强大的全局搜索能力初始化聚类中心,旨在克服 K-means 算法中初始化聚类中心随机性带来的不稳定性问题。

首先,PODM-Kmeans 中布谷鸟通过立方混沌映射选择 K 个聚类中心作为培养孩子的候选巢。立方混沌映射是一种改进型的混沌映射方法,它通过非线性递归方程生成伪随机数列,能够提供高度遍历性和不可预测性[21]。相比于传统布谷鸟搜索算法的随机初始化,立方混沌映射可提供更加均匀和分散的初始解,有助于提升搜索效率,其数学表达式如下:

$$r_{t+1} = \rho r_t (1 - r_t^3) \tag{6}$$

式中:r_t 表示混沌序列的当前值;ρ 为控制参数。

通过多次实验,发现最佳初始候选巢数和数据集大小成正相关。因此,为了兼顾时间复杂度和搜索效果,本文将初始候选巢数 num_nest 设置为:

$$num\_nest = \lfloor n \rfloor \tag{7}$$

式中:n 为数据集 X 样本数量。

然后,根据环境是否适宜孩子的成长,布谷鸟选择合适的巢穴来培育后代。为了有效评估不同初始聚类中心的质量,本文采用了平方误差和(sum of squared errors,SSE)作为巢穴的适应度。SSE 是一种简单且计算高效的

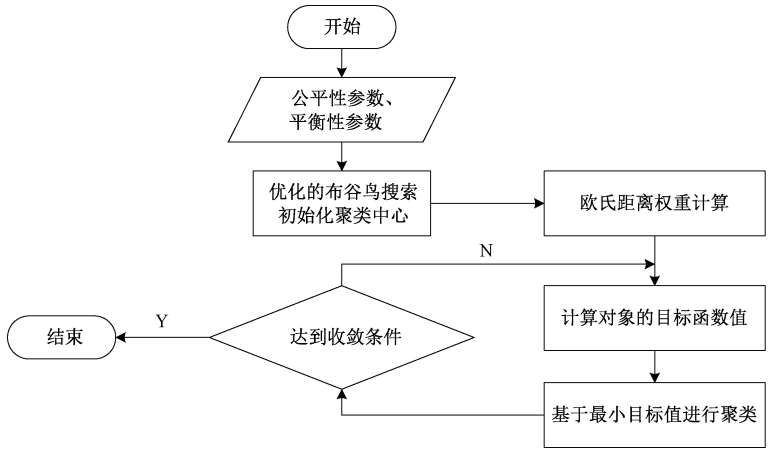


图 1　算法的流程

Fig. 1　Flowchart of the proposed algorithm

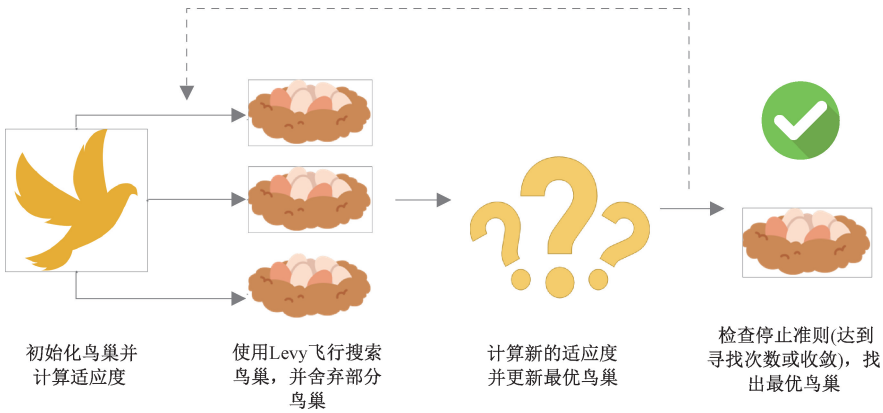


图 2　布谷鸟搜索巢示意图

Fig. 2　Illustration of nest search in cuckoo search algorithm

评估指标[22]。在每次搜索中,布谷鸟通过最小化 SSE 值,能够在全局范围内选出最优的聚类中心。

$$SSE = \sum_{i=1}^{k} \sum_{x_j \in C_i} \| x_j - \mu_i \|^2 \tag{8}$$

式中:x_j 为在第 i 个簇里的样本点;μ_i 为第 i 个簇的聚类中心;C_i 为第 i 个簇。

接着,在每一次迭代中引入随机立方混沌映射生成巢穴和 Levy 飞行来模拟布谷鸟飞行寻找和丢弃巢穴。Levy 飞行是一种随机游走,它是由许多小的移动和少量大的移动组成[23]。在飞行过程中,布谷鸟每次迭代每个巢以小于 PA 的概率进行 Levy 随机搜索新的巢以增加多样性。通过这种方式,布谷鸟能够自适应探索更广阔的搜索空间,避免陷入局部最优解。此外,Levy 飞行的特性使得布谷鸟能够在较少的迭代中实现较大范围的搜索,有效加速聚类中心的优化过程。其中 Levy 飞行的过程为:

$$C_{new} = C_{old} + L(\beta) \tag{9}$$

$$L(\beta) = \frac{u}{|v|^{1/\beta}} \tag{10}$$

$$\delta = \left(\frac{\Gamma(1+\beta)\sin(\pi\beta/2)}{\Gamma((1+\beta)/2) \cdot \beta \cdot 2^{(\beta-1)/2}}\right)^{1/\beta} \tag{11}$$

式中:$u \sim N(0,\delta^2)$;$v \sim N(0,1)$;β 是 Levy 分布参数。

传统布谷鸟算法中的 Levy 分布参数 β 是固定值,不能动态的调整布谷鸟飞行的步长,β 过大,在算法前期全局搜索能力增强,而到算法后期局部搜索能力较弱,算法的寻优精度不高。β 偏小,算法全局搜索能力弱,局部搜索能力强,算法很容易陷入局部最优解。因此,为了平衡全局搜索力和局部搜素力进而提升搜索性能,本文引入一个自适应步长公式对 Levy 分布参数 β 进行动态调整:

$$\beta(t) = m \times \beta_{\max} \exp(c) \tag{12}$$

$$c = \ln\left(\frac{\beta_{\min}}{\beta_{\max}}\right) \times \frac{t}{max\_iter} \tag{13}$$

式中:t 表示当前迭代次数;max_iter 表示最大迭代次数;$\beta_{\min}$ 和 $\beta_{\max}$ 表示 Levy 分布参数 β 的下界和上界;m 代表

动态调整的全局参数。分析式(12)和(13)可知 Levy 分布参数 β 随迭代次数的增加而减小,从而进一步动态调整 Levy 飞行距离,提高了算法搜索效果和效率。

PA 表示布谷鸟发现新巢的概率。在传统布谷鸟算法中,PA 是一个固定值。如果 PA 的选取过大,会导致抛弃最优解的概率变大,从而很难收敛到一个最优解。如果 PA 选取过小,会导致对于一个较差解收敛过慢,从而算法会陷入局部最优解的漩涡中。基于以上问题,本文提出一个动态调整的 PA,具体公式如下:

$$PA(t) = \frac{m \times 2PA}{1 + \exp(T \times \frac{t}{\max\_iter})} \tag{14}$$

式中:T 为下降参数。

由式(14)可知,在搜索前期,算法有较大的全局搜索能力,在搜索后期,算法有较大的局部搜索能力,故而保证了算法对于最优解的搜索性能。

在式(12)和(14)中,本文使用了动态调整的全局参数 m。全局参数 m 是为了避免算法一直在局部不停迭代并无法达到预期效果而设置的。其计算公式为:

$$m = \begin{cases} 1, & f_{best}^{t} > f_{best}^{t+1} \\ m + 0.001, & \text{其他} \end{cases} \tag{15}$$

首先,初始化 count 为 0。此后,如果本次迭代的最佳适应度小于上一次的最佳适应度,表示布谷鸟找到了一个更优的巢,将动态调整的全局参数 m 重置为 1,以便继续有效探索并避免陷入局部最优。如果未能找到更好的解,则会对 m 进行相应的上调。若算法经过多次精确搜索仍未找到更优解,则 m 会持续增加。然而,若此时 m 值已超过 2,算法会进行进一步的判断。如果当前的迭代次数小于总迭代次数的一半,且 count = 0,则将当前的最佳适应度值保存,并将 count 设为 1,同时重置所有解并进行新一轮的迭代。如果迭代次数超过总次数的 1/2,则会继续增加 m,并动态调整 Levy 分布参数 β 和布谷鸟发现新巢的概率 PA,直到发现更优的适应度值。

最后,重复以上步骤进行迭代,直至达到收敛条件。改进后的种群优化算法会输出一个初始聚类中心集合,这些聚类中心将作为公平性 K-means 算法的初始聚类中心。通过这种方式,改进后的种群优化搜索能够有效避免 K-means 算法中由于随机初始化聚类中心而带来的不稳定性,提升聚类结果的质量和公平性。

2.3　鲁棒距离度量

传统的欧氏距离难以应对数据中的异质性和敏感属性差异,尤其在特征尺度、类别或属性差异显著时,可能无法准确刻画数据点间的实际相似性,影响聚类结果的公平性与准确性。为此,本文在聚类过程中引入加权欧氏度量,根据各特征的标准差 σ_m 赋予特征不同的重要性,有效提升聚类算法在面对多样化和异质性数据时的表现。其中加权欧氏度量 D_{ij} 的计算方法为:

$$\sigma_m = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_{im} - \mu_m)^2} \tag{16}$$

$$w_m = 1/\sigma_m \tag{17}$$

$$D_{ij} = \sqrt{\sum_{m=1}^{M} w_m (X_{im} - C_{jm})^2} \tag{18}$$

式中:X_{im} 是第 i 个样本的第 m 个特征值;μ_m 表示第 m 个特征所有样本的均值;C_{jm} 是第 j 个聚类中心的第 m 个特征。

2.4　聚类公平性

在传统聚类算法中,目标通常是将数据点根据相似性划分到不同的簇中。然而,这些算法往往忽视了簇的大小和公平性问题,即每个簇的规模可能存在较大差异,或者簇的划分可能对某些敏感群体不公平。为了克服这些问题,本文在聚类过程中引入了公平性和簇大小平衡的正则项约束,旨在确保聚类结果既具有合理的均衡性,又符合不同群体的公平性要求。本文更新聚类过程中的目标代价函数为:

$$\min_{Y,M}(D_{ic} + p_1 \| \frac{\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{F}}{\boldsymbol{F}^{\mathrm{T}}\boldsymbol{F}} - \frac{\boldsymbol{Y}\boldsymbol{Y}}{n} \| + p_2 tr((\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y})^{-1})) \tag{19}$$

2.5　算法求解

本文构建的目标函数属于非连续组优化问题,难以直接通过求导等方式解决。故采用坐标下降策略,在每轮迭代中交替更新聚类中心与样本分配矩阵,逐步逼近最优解。其核心思想是,固定一个变量,优化另一个变量,交替进行,直到收敛。因此,本文将迭代分为两个阶段进行。首先,固定 $\boldsymbol{Y}$,更新中心 $\boldsymbol{M}$。此时目标函数可以简化为关于 $\boldsymbol{M}$ 的最小二乘问题:

$$\min_{M} \| \boldsymbol{X} - \boldsymbol{M}\boldsymbol{Y}^{\mathrm{T}} \|_F^2 \tag{20}$$

得到闭式解 $\boldsymbol{M}$ 的推导如下:

$$\| \boldsymbol{X} - \boldsymbol{M}\boldsymbol{Y}^{\mathrm{T}} \|_F^2 = \mathrm{tr}((\boldsymbol{X} - \boldsymbol{M}\boldsymbol{Y}^{\mathrm{T}})(\boldsymbol{X} - \boldsymbol{M}\boldsymbol{Y}^{\mathrm{T}})^{\mathrm{T}}) \tag{21}$$

$$\frac{\partial}{\partial \boldsymbol{M}} \| \boldsymbol{X} - \boldsymbol{M}\boldsymbol{Y}^{\mathrm{T}} \|_F^2 = 2(\boldsymbol{M}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{Y}) = 0 \tag{22}$$

$$\boldsymbol{M} = \boldsymbol{X}\boldsymbol{Y}(\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y})^{-1} \tag{23}$$

此后,固定 $\boldsymbol{M}$,更新标签 $\boldsymbol{Y}$。由于标签矩阵 $\boldsymbol{Y}$ 是离散矩阵,不能使用常规的梯度办法。本文采用逐样本更新的策略,即对每个样本 x_i 独立进行如下优化:

$$y_i = \arg\min_{j=1,\cdots,k} \mathcal{L}_{\mathrm{local}}(x_i, j) \tag{24}$$

考虑到式(19)的构造,其局部函数能够定义为:

$$\mathcal{L}_{local}(x_i, j) = \mathcal{L}_1(x_i, j) + \mathcal{L}_2(x_i, j) + \mathcal{L}_3(x_i, j) \tag{25}$$

$$\mathcal{L}_2(x_i, j) = p_1\left(\frac{(\boldsymbol{Y}^{(i\to j)})^{\mathrm{T}}\boldsymbol{F}}{\boldsymbol{F}^{\mathrm{T}}\boldsymbol{F}} - \frac{\boldsymbol{Y}^{(i\to j)}\boldsymbol{Y}^{(i\to j)}}{n} - \frac{\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{F}}{\boldsymbol{F}^{\mathrm{T}}\boldsymbol{F}} - \frac{\boldsymbol{Y}\boldsymbol{Y}}{n}\right) \tag{26}$$

$$\mathcal{L}_3(x_i,j)=p_2(\mathrm{tr}((\boldsymbol{Y}^{(i\to j)})^{\mathrm{T}}\boldsymbol{Y}^{(i\to j)})^{-1}-\mathrm{tr}((\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y})^{-1})) \tag{27}$$

在每次迭代过程中,目标函数值均呈现单调递减趋势且存在非负下界,故整个优化过程能够快速收敛到最优解。收敛曲线如图3和4所示,可以看出所提算法在小规模数据集2d-4c-no0与大规模数据集adult上均具有良好的收敛特性,有效验证了其在实际应用中的可行性。

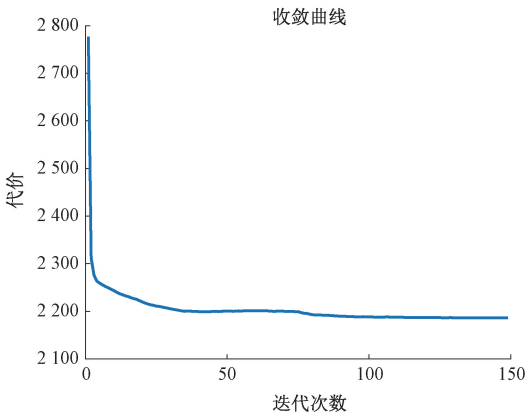


图3　在2d-4c-no0数据集上目标函数收敛

Fig. 3　Convergence plot of the objective function on the 2d-4c-no0 dataset

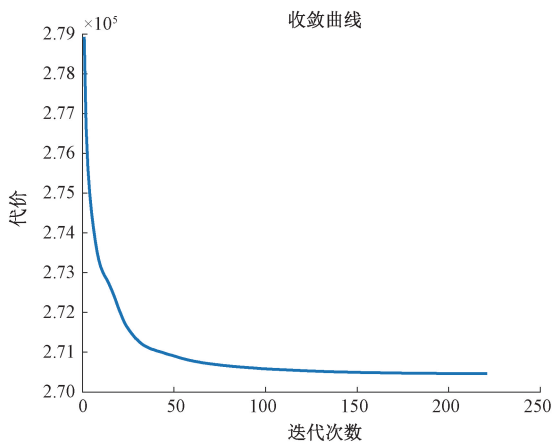


图4　在adult数据集上目标函数收敛

Fig. 4　Convergence plot of the objective function on the adult dataset

2.6　时间复杂度分析

本文提出的PODM-Kmeans算法主要包括初始聚类中心优化、加权欧氏距离计算、公平性约束构建以及坐标下降优化求解过程。设数据集样本数量为n,特征维数为d,聚类数量为K。算法初始阶段利用改进的布谷鸟搜索算法确定聚类中心,复杂度为$O(TNKd)$,其中N为初始巢穴数,T为最大迭代次数;在随后的聚类过程中,引入加权欧氏距离和公平性约束,通过坐标下降法迭代优化,每轮更新复杂度均为$O(nKd)$,设迭代次数为L,整体复杂度为$O(LnKd)$。综合来看,算法总的时间复杂度为$O(TNKd+LnKd)$。由于实际应用中一般满足$T\ll L$,算法的主导复杂度可近似为$O(LnKd)$。与传统K-means的复杂度$O(nKd)$相比,PODM-Kmeans仅在常数因子上有所增加,仍具有良好的计算效率和可扩展性。

3　实验

3.1　数据和实验背景

如表1所示,实验为了更好的评估算法的性能和公平性,采用了公平性聚类研究常用的5个合成数据集和5个真实世界数据集进行测试与验证。其中,5个合成数据集包括Elliptical、DS-577、2d-4c-no0、2d-4c-nol和2d-4c-no4,这些数据集广泛应用于聚类算法的研究,具有一定代表性。5个真实数据集包括Adult、Bank、Census1990、CreditCard和Diabetic,它们来自于UCI机器学习库。这些真实世界数据集涵盖广泛的领域,具有多样化的特征维度和群体结构,反映出不同的数据分布特性。实验惩罚项P_1和P_2的设定基于数据集的实际分布特性,通过多次实验的超参数优化和模型验证过程得出,其具体数值参考表1所示。由于本文在聚类过程中引入了公平性考虑,如图5所示,本文在聚类可视化时采用颜色区分不同的聚类类别,并用形状标识不同的敏感属性,以更好地可视化公平性与聚类效果。

表1　数据集及参数

Table 1　Datasets and parameters used in the experiments

数据集	P_1	P_2	规模	特征
Elliptical($k=2$)	4 000	7 700	500	2
DS-577($k=3$)	2 100	20 000	577	2
2d-4c-no0($k=4$)	6 500	7 700	1 572	2
2d-4c-no1($k=4$)	9 000	170 000	1 623	2
2d-4c-no4($k=4$)	3 900	100 000	863	2
Adult($k=10$)	8 000 000	80 000 000	32 561	5
Bank($k=6$)	56 000	9000	4 000	6
Census1990 ($k=5$)	150 000	20 000	3 000	25
CreditCard ($k=10$)	2 000 000	200 000	30 000	14
Diabetic($k=10$)	110 000	14 000	10 000	2

本文实验PODM-Kmeans代码在MATLAB中实现。对每个数据集,本实验采用了归一化处理,包括行归一化和列归一化。本文采用的对比方法包括传统的K-means聚类算法(Lloyd's K-means,Lloyd)、公平的非归一化谱聚类算法(fair unnormalized spectral clustering,FSCUN)、公平的归一化谱聚类算法(fair normalized spectral

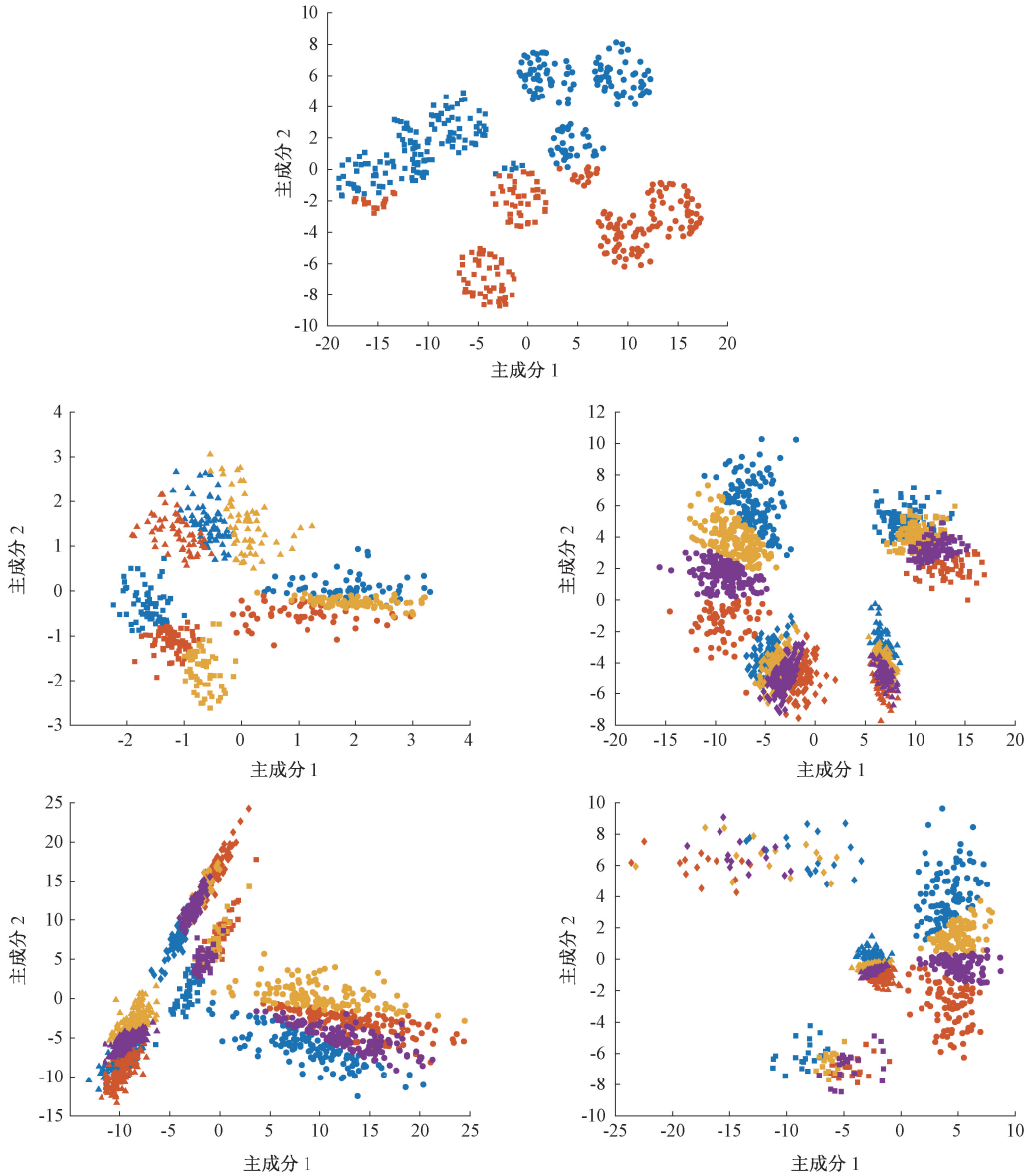


图 5　在合成数据集上 PODM-Kmeans 聚类可视化

Fig. 5　Clustering visualization of PODM-Kmeans on the synthetic dataset

clustering, FSCN)、公平 K 均值聚类算法(fair K-means, FrKM)和 BFKM 算法。为了实验对比的公正性,实验对比方法参数的设定参考了文献[14]。

3.2　评估标准

为了兼顾聚类效果和公平性的评估,本文采用了公平性聚类中常用的 4 种指标,实验的评估指标包括以下 4 个方面。

1)公平性比率(fairness ratio, FR)最早是在 2017 年被提出,而后在研究算法公平性中被广泛应用[14,24]。该指标评估各个簇中不同群体的比例是否与其在整体数据集中的比例一致。

$$FR(C_j)=\min_{1\leqslant l\leqslant m}\left(\frac{r(j,l)}{r(l)},\frac{r(l)}{r(j,l)}\right) \tag{28}$$

$$FR(C)=\min_{C_j\in C}FR(C_j) \tag{29}$$

2)平均瓦瑟斯坦距离(average wasserstein distance, AWD)[25]是用于评估不同敏感属性群体在各个簇中的分布偏差,通过计算概率向量之间的瓦瑟斯坦距离来量化不公平性。

$$AWD=\frac{\sum_i^k|C_i|\times WD(P_i,P_x)}{n} \tag{30}$$

3)Dunn 指数(dunn index, DI)用于评估聚类结果的紧凑性和可分离性[26]。DI 的计算基于簇间最小距离与

簇内最大直径之比。

$$DI = \frac{\min\limits_{1\leqslant i<j\leqslant k}\left(\min\limits_{x\in C_i, y\in C_j} dist(x,y)\right)}{\max\limits_{1\leqslant l\leqslant k}\left(\max\limits_{x,y\in C_l} dist(x,y)\right)} \tag{31}$$

4)SSE 用于评估数据点相对于其所属簇中心的紧密程度[21]。是一种常用的聚类性能评价指标,用于衡量数据点与其所属簇中心之间的紧密程度,具体公式如式(8)。

3.3　对比实验

合成数据集与真实数据集的实验结果如表 2~5 所示,最优结果已加粗标出。本文将结合聚类公平性和聚类效果两方面来分析实验结果。

1)聚类公平性对比

如表 2 所示,PODM-Kmeans 在 FR 上对所有数据集均表现出一定的优势。在 4 组高维大规模数据集 Adult、Bank、Census1990 和 CreditCard 上,PODM-Kmeans 的 FR 指标均超过 0.95,这表明 PODM-Kmeans 在高维数据聚类任务中能够更好地保持聚类的公平性。相比之下,方法 Lloyd、FSCUN 以及 FSCN 在多个数据集上出现 FR 值为 0,这意味着它们在这些数据集上未能保证敏感属

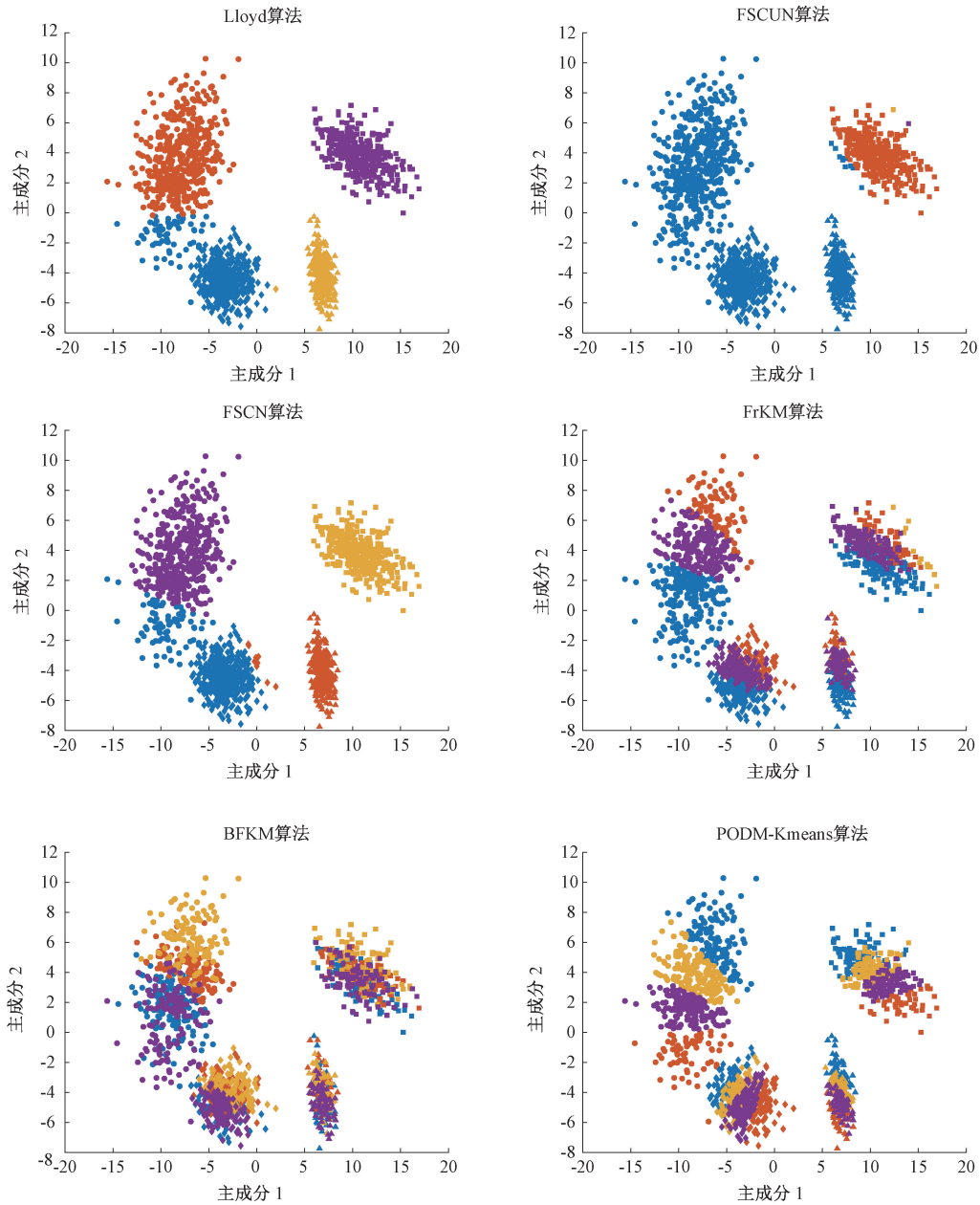


图 6　在 2d-4c-no0 数据集上本文算法与对比算法聚类可视化

Fig. 6　Clustering visualization of the proposed algorithm and comparison algorithms on the 2d-4c-no0 dataset

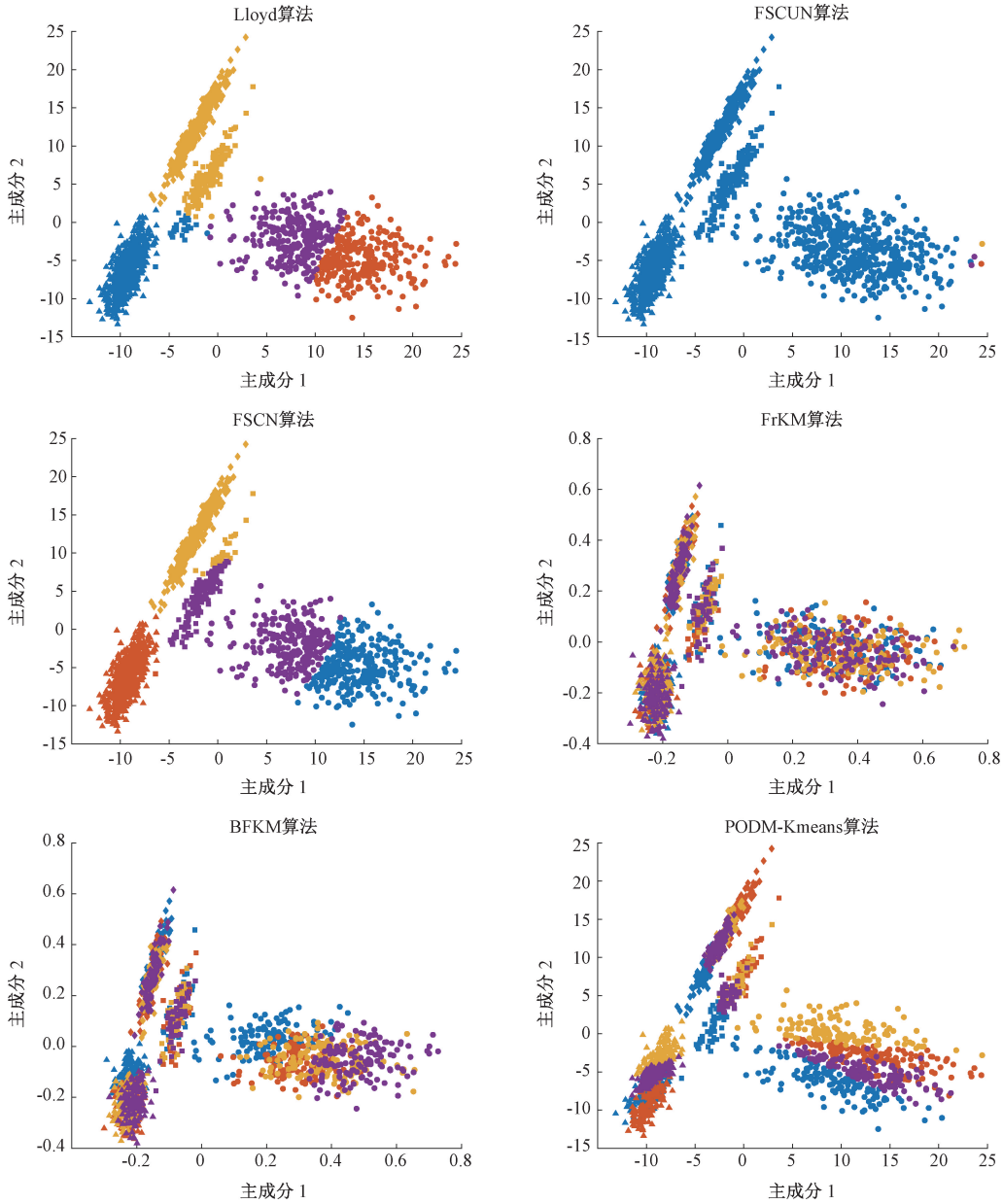


图 7　在 2d-4c-no1 数据集上本文算法与对比算法聚类可视化

Fig. 7　Clustering visualization of the proposed algorithm and comparison algorithms on the 2d-4c-no1 dataset

表 2　在不同数据集上本文算法与对比算法 FR 值

Table 2　FR values of the proposed algorithm and comparison algorithms on different datasets

数据集	Lloyd	FSCUN	FSCN	FrKM	BFKM	PODM-Kmeans
Elliptical(k=2)	0	0. 887 8	0. 883 5	0. 883 5	0. 900 5	**0. 977 5**
DS-577(k=3)	0	0	0	0. 801 9	0. 804 2	**0. 844 1**
2d-4c-no0(k=4)	0	0	0	0. 811 4	0. 814 1	**0. 822 8**
2d-4c-no1(k=4)	0	0	0	0. 801 8	0. 801 9	**0. 892 7**
2d-4c-no4(k=4)	0	0	0	0. 738 8	0. 756 6	**0. 870 6**
Adult(k=10)	0. 436 2	0	0. 559 9	0. 905 1	0. 918 0	**0. 979 8**
Bank(k=6)	0. 292 9	0. 336 8	0. 530 6	0. 809 0	0. 813 7	**0. 870 3**
Census1990(k=5)	0. 512 9	0. 696 4	0. 741 8	0. 915 1	0. 916 9	**0. 954 2**
CreditCard(k=10)	0. 739 0	0	0. 885 1	0. 890 0	0. 890 0	**0. 954 9**
Diabetic(k=10)	0. 837 6	0	0. 823 9	0. 874 4	0. 874 4	**0. 880 7**

表 3 在不同数据集上本文算法与对比算法 AWD 值

Table 3 AWD values of the proposed algorithm and comparison algorithms on different datasets

数据集	Lloyd	FSCUN	FSCN	FrKM	BFKM	PODM-Kmeans
Elliptical(k=2)	0.494 0	0.046 0	0.048 0	0.048 0	0.047 2	**0.010 0**
DS-577(k=3)	0.418 3	0.099 5	0.149 6	**0.024 2**	0.031 9	0.029 3
2d-4c-no0(k=4)	0.312 7	**0.004 9**	0.064 9	0.005 2	0.010 3	0.010 1
2d-4c-no1(k=4)	0.270 5	**0.001 2**	0.070 1	0.001 4	0.012 5	0.011 5
2d-4c-no4(k=4)	0.172 3	0.075 9	0.079 9	**0.007 6**	0.013 9	0.014 3
Adult(k=10)	0.057 3	0.042 8	0.059 9	0.008 2	0.007 4	**0.002 1**
Bank(k=6)	0.069 1	0.035 7	0.035 8	0.020 3	0.021 0	**0.014 3**
Census1990(k=5)	0.067 1	0.051 1	0.047 0	0.014 1	0.014 1	**0.012 0**
CreditCard(k=10)	0.036 3	0.003 4	0.018 8	0.017 0	0.019 4	**0.003 3**
Diabetic(k=10)	0.035 7	0.023 5	0.028 5	0.024 2	0.026 0	0.024 5

表 4 在不同数据集上本文算法与对比算法 DI 值

Table 4 DI values of the proposed algorithm and comparison algorithms on different datasets

数据集	Lloyd	FSCUN	FSCN	FrKM	BFKM	PODM-Kmeans
Elliptical(k=2)	0.064 4	0.041 8	**0.071 6**	**0.071 6**	0.004 0	0.005 8
DS-577(k=3)	**0.007 9**	0.000 5	0.006 8	0	0	0
2d-4c-no0(k=4)	**0.007 6**	0.003 2	0.000 2	0	0	0
2d-4c-no1(k=4)	0.001 7	**0.011 2**	0.000 1	0	0	0
2d-4c-no4(k=4)	**0.005 7**	0.000 2	0.000 1	0	0	0
Adult(k=10)	**0.000 6**	0	0.000 4	0	0	0
Bank(k=6)	**0.019 0**	0	0.001 8	0.000 1	0	0
Census1990(k=5)	0.050 7	0.058 0	0.025 8	0.043 3	0.064 0	**0.093 0**
CreditCard(k=10)	0.009 4	**0.027 5**	0	0	0	0
Diabetic(k=10)	**0.040 6**	0	0	0	0	0

表 5 在不同数据集上本文算法与对比算法 SSE 值

Table 5 SSE values of the proposed algorithm and comparison algorithms on different datasets

数据集	Lloyd	FSCUN	FSCN	FrKM	BFKM	PODM-Kmeans
Elliptical(k=2)	**206.298 2**	344.421 6	343.964 0	343.964 0	351.109 0	351.048
DS-577(k=3)	**71.013 4**	449.029 2	361.429 9	518.153 6	516.065 5	516.105 4
2d-4c-no0(k=4)	**114.483 4**	$1.528\ 3\times10^3$	$1.355\ 8\times10^3$	$1.478\ 9\times10^3$	$1.455\ 5\times10^3$	$1.451\ 2\times10^3$
2d-4c-no1(k=4)	**82.312 2**	$1.615\ 0\times10^3$	$1.271\ 8\times10^3$	$1.583\ 5\times10^3$	$1.539\ 7\times10^3$	$1.536\ 2\times10^3$
2d-4c-no4(k=4)	**104.002 3**	705.999 1	666.311 7	714.986 6	704.524 2	706.287 5
Adult(k=10)	$\mathbf{9.508\ 3\times10^3}$	$1.438\ 0\times10^4$	$1.025\ 1\times10^4$	$1.027\ 7\times10^4$	$1.058\ 3\times10^4$	$1.066\ 5\times10^4$
Bank(k=6)	$\mathbf{1.231\ 2\times10^3}$	$1.785\ 8\times10^3$	$1.254\ 3\times10^3$	$1.369\ 8\times10^3$	$1.329\ 8\times10^3$	$1.331\ 8\times10^3$
Census1990(k=5)	$\mathbf{1.760\ 4\times10^3}$	$1.820\ 7\times10^3$	$1.821\ 9\times10^3$	$1.852\ 6\times10^3$	$1.852\ 0\times10^3$	$1.841\ 4\times10^3$
CreditCard(k=10)	$\mathbf{8.199\ 8\times10^3}$	$1.842\ 0\times10^4$	$9.344\ 1\times10^3$	$8.282\ 7\times10^3$	$8.226\ 7\times10^3$	8.406×10^3
Diabetic(k=10)	243.263 4	$3.261\ 6\times10^3$	**235.258 3**	327.115 8	298.483 8	261.0071

性的公平性,导致聚类结果高度偏向特定类别或歧视某个敏感群体。分析表 3 可知,PODM-Kmeans 在 AWD 评估指标上整体优于其他算法,表明该方法能够最大程度上减少不同群体在各个簇中的分布失衡情况。尤其是在 Elliptical、Adult、Bank、Census1990 和 CreditCard 这 5 个数据集上,PODM-Kmeans 都是最小值。在所有算法中,PODM-Kmeans 实现最小 AWD 值的次数最多,进一步证明了其在不同数据分布下的广泛适用性和优越性。

在 FR 和 AWD 指标的评估中,PODM-Kmeans 展现出了良好的性能。深入分析,PODM-Kmeans 在聚类公平性方面的突出表现主要归因于两大核心设计机制。(1)在初始聚类中心选取阶段,本文引入了改进的布谷鸟搜索算法,通过全局搜索与局部搜索的协同平衡,有效提升了初始聚类中心的质量,从而为后续的聚类过程奠定了良好的基础;(2)在聚类迭代过程中融合了加权欧氏范数、公平约束和簇大小平衡约束,增强了模型对数据处理的鲁棒性。这使其在面对复杂高维数据时仍能保持稳定的聚类公平性,减少了对敏感属性的歧视或偏倚,尽量确保所有群体在聚类过程中得到合理分配。这些优化共同提升了 PODM-Kmeans 的适应性和泛化能力,使其在多种

数据分布下 FR 和 AWD 表现出持续的优势。

2)聚类效果对比

由于公平性强调不同敏感群体间的均衡分布,这必然与传统聚类效果追求簇内紧密性和簇间区分度的目标存在一定的冲突,导致在提升公平性的同时,聚类结构的固有质量难以避免地受到一定程度的牺牲。结合表 4 可知,包括 PODM-Kmeans 在内的大多数公平性聚类算法在簇间分离性指标(DI 值)上普遍偏低,这一现象也验证了公平性约束对聚类可分性带来的必然影响。然而,即使在这种情况下,PODM-Kmeans 在聚类可视化上依然在可视化簇形状、边界连续性及群体混合程度上均展现了良好的聚类结构。本文采用颜色区分不同聚类类别,并以点的形状标识不同的敏感属性。如图 6、7 所示,从可视化结果可以看出,相较于 Lloyd 算法和 FSCN 算法,通过点的形状(敏感属性)分布可得,PODM-Kmeans 能有效保证同一簇中包含多种敏感属性样本,且各敏感群体在各簇中的占比更均衡,这与表 2 的较高 FR 及表 3 的较低 AWD 相一致。相比之下,Lloyd 和 FSCN 算法的聚类结果虽然在簇整体形状上较为紧凑,但同一颜色中往往部分形状占比过高,表明同一簇内敏感属性呈现同质化,无法满足聚类分析中对于敏感属性公平性和均衡性的要求。同时,PODM-Kmeans 在维持整体聚类结构稳定性方面上优于其他公平性算法,有效避免了类似于 FrKM 方法出现的极端簇合并和聚类边界难以清晰体现的现象。此外,相较于 BFKM,PODM-Kmeans 所生成的聚类簇在二维主成分空间中形态更紧凑自然,边界更加平滑且簇间过渡连续,未出现簇形状拉伸或边界模糊等问题。

另一方面,从表 5 可以看出,在簇内紧密性方面,PODM-Kmeans 在 SSE 指标上的表现与其他公平性聚类算法(FSCUN、FSCN、FrKM、BFKM)基本相当,仅略高于传统未考虑公平性约束的 Lloyd 算法所获得的最小 SSE。需要指出的是,公平性聚类方法的核心目标是在提升簇内敏感属性分布均衡性的同时,尽量保持聚类结果的紧密性和稳定性。因此,SSE 值的适度增加是引入公平性约束所带来的可接受的代价,这也是当前公平聚类研究中的普遍现象。结合图 6 和 7 的可视化结果可进一步验证,PODM-Kmeans 在保证较高公平性的同时,依然能够维持良好的簇结构可解释性和空间分布连续性,未因公平性约束而出现簇分布异常或结构失真。与其他公平性聚类方法相比,PODM-Kmeans 在实现较高公平性指标的同时,有效抑制了 SSE 的过度增大,显示出一定的聚类质量保护能力,体现了本文方法在实际场景下公平性与聚类质量之间的兼顾能力。

进一步分析,这种兼顾能力的取得主要源于 PODM-Kmeans 机制本身:初始阶段的全局优化策略确保了敏感属性的分布均衡,减少了后续聚类调整的剧烈程度;同时,鲁棒的距离度量方式进一步有效抑制了异常数据点带来的聚类质量损失。这种设计策略使得 PODM-Kmeans 能够在提升公平性的同时,更好地保护聚类结构的稳定性与合理性,展现出比现有公平聚类方法更加良好的综合性能,尤其适用于对公平性与聚类精度同时具有严格需求的实际数据分析场景。

4　结　论

本文提出了一种 PODM-Kmeans 方法,合理平衡了聚类质量与公平性需求。该方法以公平性为核心驱动,融合改进的种群优化策略与鲁棒距离度量准则,在提升聚类质量的同时,有效缓解了聚类过程中的群体偏差问题。其核心思路在于通过优化聚类初始中心的选取机制,引入全局与局部协同演化的种群搜索算法,从源头提升聚类稳定性,并结合加权欧氏范数实现对数据噪声的有效抑制。在此基础上,构建了集公平性约束与簇大小平衡为一体的目标函数,从结构上保障了敏感群体在聚类过程中的合理分布,体现了算法在公平性建模方面的系统性与前瞻性。本研究的关键技术难点在于如何协调优化策略与公平性约束之间的冲突关系,确保全局搜索效率与聚类精度不因引入公平机制而受损。PODM-Kmeans 在多个公开数据集上展现出的稳定性能,充分验证了其在公平性与聚类质量间达成有效折中的能力,体现出方法的先进性与实用价值。未来的研究可进一步聚焦于算法的计算复杂度控制与簇间分离能力提升,拓展其在嵌入式测量系统中的部署潜力,并探索其在金融风控、医疗诊断、推荐系统等对实时性与公平性要求极高的关键场景中的应用前景,从而推动聚类算法在智能决策中的实用性边界。

参考文献

[1] ATABEK A, ERALP E, GURSOY M E. Trust, privacy and security aspects of bias and fairness in machine learning [C]. Proceedings of the 5th IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA), 2023: 111-121.

[2] 范馨月,张阔,张干,等. 细微特征增强的多级联合聚类跨模态行人重识别算法[J]. 电子测量与仪器学报, 2024,38(3):94-103.
FAN X Y, ZHANG K, ZHANG G, et al. Multi-level joint clustering cross-modal person re-identification algorithm with enhanced fine-grained features [J]. Journal of Electronic Measurement and Instrumentation, 2024, 38(3): 94-103.

[3] 邓子文,段勇. 基于深度聚类学习的无监督行人重识别[J]. 电子测量与仪器学报,2025,39(3):1-9.

DENG Z W, DUAN Y. Unsupervised person re-identification based on deep clustering learning [J]. Journal of Electronic Measurement and Instrumentation, 2025, 39(3): 1-9.

[4] 韩莹,朱宏宇,李琨. 融合聚类及随机配置网络的短期光伏功率预测[J]. 电子测量与仪器学报, 2023, 37(11):205-216.
HAN Y, ZHU H Y, LI K. Short-term photovoltaic power prediction based on clustering and random configuration network [J]. Journal of Electronic Measurement and Instrumentation, 2023, 37(11): 205-216.

[5] CHHABRA A, MASALKOVAITĖ K, MOHAPATRA P. An overview of fairness in clustering [J]. IEEE Access, 2021, 9: 130698-130720.

[6] ABRAHAM S S, DEEPAK P, SUNDARAM S S. Fairness in clustering with multiple sensitive attributes [C]. Proceedings of the 23rd International Conference on Extending Database Technology (EDBT), 2020: 287-298.

[7] SINAGA K P, YANG M S. Unsupervised K-means clustering algorithm [J]. IEEE Access, 2020, 8: 80716-80727.

[8] 韦子辉,廖戈,李明轩,等. 基于 ISODATA 改进 K 均值聚类算法的 NLOS 识别技术[J]. 电子测量技术, 2024,47(4):172-180.
WEI Z H, LIAO G, LI M X, et al. NLOS identification technology based on ISODATA improved K-means clustering algorithm [J]. Electronic Measurement Technology, 2024, 47(4): 172-180.

[9] LEE J, KIM S. Parasitic capacitance prediction for standard cells using machine learning and K-means clustering algorithm [C]. Proceedings of the 2025 International Conference on Electronics, Information, and Communication (ICEIC), 2025: 1-4.

[10] SUMAN B K, CHAKRABARTY D, FLORES N J, et al. Fair algorithm for clustering [C]. Proceedings of the 33rd Annual Conference on Neural Information Processing Systems (NeurIPS), 2019: 4955-4966.

[11] ZIKO I M, GRANGER E, YUAN J, et al. Variational fair clustering [C]. Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI), 2021: 11202-11209.

[12] XU W, HU J, DU S, et al. K-means clustering with fairness constraints [C]. Proceedings of the 16th International Conference on Intelligent Systems and Knowledge Engineering (ISKE), 2021: 215-222.

[13] YANG Z, ZHANG H, YANG C, et al. Cost guarantee for individual fairness on spectral clustering [C]. Proceedings of the 29th IEEE International Conference on Parallel and Distributed Systems (ICPADS), 2023: 1546-1553.

[14] PAN R, ZHONG C, QIAN J. Balanced fair K-means clustering [J]. IEEE Transactions on Industrial Informatics, 2024, 20(4): 5914-5923.

[15] ALAM A, MUQEEM M. Automatic clustering for selection of optimal number of clusters by K-means integrated with enhanced firefly algorithms [C]. Proceedings of the 2nd International Conference on Technological Advancements in Computational Sciences (ICTACS), 2022: 343-347.

[16] YANG M S, HUSSAIN I. Unsupervised multi-view K-means clustering algorithm [J]. IEEE Access, 2023, 11: 13574-13593.

[17] 刘熹,陈晨,双丰. 基于改进 YOLOv7-tiny 的多种类绝缘子检测算法[J]. 仪器仪表学报, 2024, 45(9): 101-110.
LIU X, CHEN CH, SHUANG F. Multi-class insulator detection algorithm based on improved YOLOv7-tiny [J]. Chinese Journal of Scientific Instrument, 2024, 45(9): 101-110.

[18] CHEN X, MIAO P, BU Q. Image segmentation algorithm based on particle swarm optimization with K-means optimization [C]. Proceedings of the 2019 IEEE International Conference on Power, Intelligent Computing and Systems (ICPICS), 2019: 156-159.

[19] MARDI M, KEYVANPOUR M R. GBKM: A new genetic based K-Means clustering algorithm [C]. Proceedings of the 7th International Conference on Web Research (ICWR), 2021: 222-226.

[20] MALKAUTHEKAR M D. Analysis of Euclidean distance and Manhattan distance measure in face recognition [C]. Proceedings of the Third International Conference on Computational Intelligence and Information Technology (CIIT 2013), 2013: 503-507.

[21] 洪丽啦,莫愿斌,鲍冬雪. 立方混沌非线性哈里斯鹰优化算法在无线传感器节点部署分析研究[J]. 现代电子技术,2023,46(6):161-168.
HONG L L, MO Y B, BAO D X. Research on deployment analysis of wireless sensor nodes based on cubic chaotic nonlinear harris hawk optimization algorithm [J]. Modern Electronics Technique, 2023, 46(6): 161-168.

[22] CHARISMA A, HIDAYAT M R, ZAINAL Y B. Speaker recognition using mel-frequency cepstrum coefficients and sum square error [C]. Proceedings of the 3rd International Conference on Wireless and Telematics (ICWT), Palembang, Indonesia, 2017: 160-163.

[23] ICHINOSE G, MIYAGAWA D, CHIBA E, et al. How Lévy flights triggered by the presence of defectors affect evolution of cooperation in spatial games [J]. Artificial Life, 2023, 29(2): 187-197.

[24] CHHABRA A, MASALKOVAIT K, MOHAPATRA P. An overview of fairness in clustering [J]. IEEE Access, 2021, 9: 130698-130720.

[25] MURALI N, MISHRA D. Wasserstein distance for attention based cross modality person re-identification [C]. Proceedings of the 2022 IEEE 19th India Council International Conference (INDICON), 2022: 1-6.

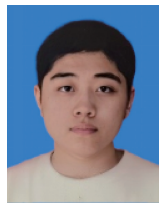
[26] BHADANA A, SINGH M. Fusion of K-means algorithm with Dunn's index for improved clustering [C]. Proceedings of the 2nd International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS), 2017: 1-5.

作者简介

谢一涵,现为南京信息工程大学人工智能学院(未来技术学院)本科生,主要研究方向为机器学习、聚类分析、深度学习、目标检测。

E-mail: 202383290356@ nuist. edu. cn

Xie Yihan is now a B. Sc. candidate at the School of Artificial Intelligence (Future Technology School), Nanjing University of Information Science and Technology. Her main research interests include machine learning, clustering analysis, deep learning, and object detection.



毕鹏飞(通信作者),2021 年于哈尔滨工程大学获得博士学位,现为南京信息工程大学讲师,主要研究方向为机器学习、深度学习、模式识别、水下目标感知。

E-mail: pfcx@ nuist. edu. cn

Bi Pengfei (Corresponding author) received his Ph. D. degree from Harbin Engineering University in 2021 and is now a lecturer at Nanjing University of Information Science and Technology. His main research interests include machine learning, deep learning, pattern recognition, and underwater target perception.

王爱萍,现为南京信息工程大学软件学院本科生,主要研究方向为边缘计算、数据分析。

E-mail: 1816463803@ qq. com

Wang Aiping is now a B. Sc. candidate at the School of Software, Nanjing University of Information Science and Technology. Her main research interests include edge computing and data analysis.