

何选森,何帆. 二维块状数据的聚类数量与边界确定[J]. 辽宁工程技术大学学报(自然科学版),2026,45(3):368-377.DOI:10.11956/j.issn.1008-0562.20250537
HE Xuansen, HE Fan. Determination of the cluster number and cluster boundary for two-dimensional blob data[J]. Journal of Liaoning Technical University (Natural Science),2026,45(3):368-377.DOI:10.11956/j.issn.1008-0562.20250537

二维块状数据的聚类数量与边界确定

何选森^{1,2}, 何帆^{3*}

(1. 广州商学院 信息技术与工程学院, 广东 广州 511363; 2. 湖南大学 信息科学与工程学院, 湖南 长沙 410082;
3. 浙江外国语学院 国际商学院, 浙江 杭州 310012)

摘要: 针对经典K-means算法在划分块状(斑点)数据聚类结构时存在不确定性以及易导致聚类结果失真的问题, 提出聚类数量 K 、聚类中心以及聚类边界的确定方法。基于数据特征的样本均值和标准差, 利用先验法则估计数据的最大聚类数 $K_{\max}$, 在 $[1, K_{\max}]$ 范围内计算不同 K 值所对应的各种聚类性能指标以评估聚类质量。通过寻求性能指标的最大值以确定数据所固有的聚类数量 K_{best} ; 将 K_{best} 作为K-means算法的输入参数, 利用期望最大化(E-M)算法迭代估计出相应的聚类中心; 以同一聚类中最远的数据点到该聚类中心的欧氏距离为半径, 给出各个聚类的圆形边界。基于不同样本数量、不同聚类数的二维斑点随机数据的仿真实验, 验证了本文方法的有效性和鲁棒性。

关键词: K-means算法; 聚类性能指标; 聚类数量; 聚类中心; 聚类半径

中图分类号: TP391

文献标志码: A

文章编号: 1008-0562(2026)03-0368-10

Determination of the cluster number and cluster boundary for two-dimensional blob data

HE Xuansen^{1,2}, HE Fan^{3*}

(1. School of Information Technology and Engineering, Guangzhou College of Commerce, Guangzhou 511363, China; 2. College of Information Science and Engineering, Hunan University, Changsha 410082, China;
3. School of International Business, Zhejiang International Studies University, Hangzhou 310012, China)

Abstract: To solve the problem that the classic K-means algorithm divides blob dataset into cluster structure with uncertainty and may cause distortion of clustering result, this paper proposes a set of methods for determining the number of clusters K , the cluster centers and cluster boundaries. Based on the sample mean and standard deviation of the data features, the maximum number of clusters $K_{\max}$ of the dataset is estimated using the prior rule. For different numbers within range of $K \in [1, K_{\max}]$, the multiple clustering performance indicators are calculated, and the inherent number of clusters K_{best} is determined by finding the maximum value of each indicator. With K_{best} as the input parameter of the K-means algorithm, the corresponding cluster centers are iteratively updated using the expectation-maximization (E-M) algorithm. The circular boundaries for each cluster are defined with the Euclidean distance from the farthest data point in the same cluster to the cluster center as the radius. The validity and robustness of the proposed methods are verified based on the simulation experiment of two-dimensional blob random data with different sample sizes and different number of clusters.

Keywords: K-means algorithm; clustering performance indicator; cluster number; cluster center; cluster radius

0 引言

在大数据时代, 数据挖掘^[1]通过抽取数据间关

系, 辨识与提取数据中隐藏的信息, 实现知识发现^[2]并解决应用中的业务问题。数据中隐含的基本信息即为数据分组的结构。对数据结构的辨识技

收稿日期: 2025-12-06 修回日期: 2026-01-09 接受日期: 2026-01-26 责任编辑: 黄琦

基金项目: 广东省普通高校特色创新项目(自然科学)(2024KTSCX131)

作者简介: 何选森(1958-), 男, 河南灵宝人, 硕士, 教授, 主要从事机器学习方面的研究。E-mail: xshe2010@163.com

通信作者: 何帆(1988-), 男, 湖南长沙人, 博士, 讲师, 主要从事经济数据方面的研究。E-mail: fan_he2017@163.com

术包括监督分类^[3]与非监督分类。聚类分析 (clustering)^[4]是探索性数据分析方法,属于非监督学习范畴。聚类分析能够揭示相似数据样本点组成的簇(集群、聚类)结构和模式^[5],数据簇划分的结果可以为监督分类器提供数据的类标签(class label)。

K-means 聚类算法^[6]计算复杂度低且易于执行,已成为经典算法。该算法由 Forgy、MacQueen 分别于 1965 年、1967 年独立开发^[7]。以找到密集型数据集的最优分组策略^[8]为目标,众多学者对其展开研究。STEINLY 等^[9]利用 K-means 聚类的额外信息研究算法的局部以及全局最优解,并对其稳定性进行了分析。SELIM 等^[10]通过将聚类问题转化为非凸数学规划,为 K-means 这类算法的有限收敛性提供了严格证明。在可微的条件下, K-means 算法可收敛至库恩-塔克点 (Kuhn-Tucker point)^[10]。

尽管 K-means 算法具备诸多优良特性,但存在以下局限:一是迭代优化过程不能保证收敛到全局最优^[7],聚类结果对随机初始化的聚类中心(质心)敏感^[8];二是算法执行依赖于预设聚类数量^[8-9],当数据集未知时,主观设定的簇数将导致聚类结果不可靠;三是数据中的离群值(outlier)和噪声(noise)可能会扭曲数据簇形状^[7]。

为解决上述问题,对经典 K-means 算法缺陷的改进方案^[11]和算法变体^[12]应运而生。对于算法的随机初始质心问题,利用最远距离选择原则^[13]确定最初的 2 个聚类中心位置,其余初始质心则采用最小-最大距离规则进行指定与划分^[13],保证数据簇的可分离性。算法的输入参数(即簇数量)与数据结构的适应程度,是决定聚类结果是否有意义以及是否具有可解释性的关键。在综合考虑数据簇内部聚合性以及簇之间分离性的基础上,聚类有效性评价函数^[14]被提出并应用于 K-means 算法中,实现了对数据潜在簇数量的估计。对于高维数据的聚类分析问题,可采用主分量分析(PCA)^[13,15]或线性判别分析(LDA)^[16]对数据进行预处理,并通过轮廓分析法^[13,15-16]实现数据聚类结果的可视化验证。

对 K-means 聚类而言,簇中心可根据期望最大化(expectation-maximization, E-M)算法^[17]逐步迭代获得,因此,对算法的聚类效果进行实时评价具有重要意义。本文从数据特征的统计量出发,采用不同的聚类性能指标来评价聚类质量,通过估计聚类(簇)数量、确定聚类质心和聚类边界,有效解决经典 K-means 算法存在的问题。

1 聚类数量估计

设 p 维数据的样本数量为 n , 则观测数据矩阵(数据集)为 $\mathbf{X}=\{x_{ik}|i=1,2,\cdots,n;k=1,2,\cdots,p\}$, 其中, $\mathbf{X}$ 的每行代表一个样本, 即 $\mathbf{x}_i=[x_{i1}, x_{i2}, \cdots, x_{ip}]$; 每列为一个特征, 即 $\mathbf{x}_k=[x_{1k}, x_{2k}, \cdots, x_{nk}]^T$ 。

在聚类分析中,任意 2 个样本点之间的相异性(dissimilarity)通常采用欧氏距离^[9]来度量,即

$$d(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{k=1}^p |x_{ik} - x_{jk}|^2 \right)^{1/2}, \quad (1)$$

式中, x_{ik} 和 x_{jk} 分别表示数据样本 $\mathbf{x}_i$ 和 $\mathbf{x}_j$ 的第 k 个特征(属性)值。

K-means 聚类算法的目标函数为平方误差和(sum-of-squared-error, SSE)^[7,9], 即

$$I_{\text{SSE}} = \sum_{i=1}^n \sum_{j=1}^K (w_{ij} \|\mathbf{x}_i - \mathbf{m}_j\|_2^2), \quad (2)$$

式中: $\mathbf{m}_j(j=1,2,\cdots,K)$ 为第 j 个聚类 C_j 的质心; K 为聚类(簇)数量; w_{ij} 为权重; $\|\cdot\|_2$ 为 L_2 范数, 即欧氏距离。

式(2)中,若 $\mathbf{x}_i$ 属于聚类 C_j , 权重 w_{ij} 取 1, 否则取 0。这说明 SSE 本质上是数据簇内的 SSE (within cluster SSE)。

K-means 算法通过寻求目标函数 I_{SSE} 的最小值以实现数据的聚类。因此,需确定式(2)中的 2 个重要参数,即 K 和 $\mathbf{m}_j$ 。

任意数据集 $\mathbf{X}$ 的聚类数 K 为正整数^[13,16]。考虑 2 种极端情况:①若整个数据集的所有样本都归属同一簇,则最小簇数 $K_{\min}=1$;②若每个样本独立构成一簇,则最大簇数 $K_{\max}=n$ 。聚类数的范围 $[K_{\min}, K_{\max}]=[1,n]$, 实际应用中通常满足 $K_{\max} \ll n$ 。

在数据分析中,不同特征的量纲与取值范围往往存在显著差异,为此需计算每个特征列的样本统计量,即

$$m_k = \frac{1}{n} \sum_{i=1}^n x_{ik} \quad k=1,2,\cdots,p, \quad (3)$$

$$\sigma_k = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_{ik} - m_k)^2} \quad k=1,2,\cdots,p, \quad (4)$$

式中: m_k 、 σ_k 分别为第 k 个特征列所有样本的均值、标准差。

基于数据的正态假说^[18],以 3σ 准则^[19](即先验法则)推断数据集的聚类结构。对于具有 p 个特征的 $\mathbf{X}$,其最大簇数 $K_{\max}$ 的计算式为

$$K_{\max} = \text{ceil} \left\{ \max_k (m_k + 3\sigma_k) \right\}, \quad (5)$$

式中, $\text{ceil}\{\cdot\}$ 为向上取整函数。

通过式(5)得到数据集的簇数范围 $[1, K_{\max}]$ 。

同理采用 $K_{\min} = \text{ceil} \left\{ \max_k (m_k - 3\sigma_k) \right\}$ 求出数据集 X 的最小簇数, 该公式计算的 $K_{\min}$ 可能为负值, 考虑到簇数最小值为1, 本文统一设定 $K_{\min}=1$ 。

由式(2)可知, I_{SSE} 与簇数 K 具有相关性。在样本容量 n 给定的情况下, 当簇数 K 增大时, 样本被分配至各簇中的间距随之减小, 即 I_{SSE} 随着 K 的增大而减少。在识别聚类数的问题中, 肘部法(elbow method)^[15-16]是一种主流的直观判据, 其核心及逻辑是在 K 从大到小的变化过程中, 找到 I_{SSE} 开始快速增长所对应的 K (即 SSE 曲线的肘部), 从而估计簇数。然而, 肘部法的应用具有较大的局限性, 特别是当簇数较大时, 对 SSE 曲线肘部的判断具有明显的不确定性。因此, 为了客观准确评估真实聚类数, 需引入多维度的统计性能指标体系, 综合考量聚类质量。

1.1 聚类质量评价指标

对不同聚类算法所产生的聚类结果, 分别从算法的运行时间和聚类结果的有效性进行综合评价。常用的性能指标如下。

(1) 兰德指数(Rand index, RI)

兰德指数通过样本对分类一致性进行度量^[20], 定义为

$$I_{\text{RI}} = \frac{a+b}{C_n^2}, \quad (6)$$

式中: C_n^2 为总样本对数; a 、 b 分别为真实标签与聚类结果中同为一簇和不同为一簇的正确样本对数。

I_{RI} 的取值范围为 $[0,1]$, I_{RI} 越接近 1, 表示聚类精度越高。但该指标未剔除随机聚类的影响, 仅适用于简单的分类场景。

(2) 调整兰德指数(adjusted Rand index, ARI)

当数据簇的数量大于 2, 且与样本总数处于相同数量级时, 随机划分策略对应的 I_{RI} 接近 0, 难以区分有效聚类与随机划分。为了抵消这种随机影响, 引入调整兰德指数^[21], 其表达式为

$$I_{\text{ARI}} = \frac{I_{\text{RI}} - E(I_{\text{RI}})}{\max(I_{\text{RI}}) - E(I_{\text{RI}})}, \quad (7)$$

式中, $E(I_{\text{RI}})$ 为随机聚类下兰德指数的数学期望。

I_{ARI} 考虑了机会概率, 其取值范围为 $[-1,1]$ 。 I_{ARI} 越接近 1, 表示聚类结果越准确; I_{ARI} 越接近 0, 表示聚类结果与随机选择相当; I_{ARI} 越接近 -1, 表示聚类结果与真实数据簇完全相反。一般而言, I_{ARI} 取值范围限定于 $[0,1]$ 。

(3) 基于互信息的分数(mutual information-based

score, MI)

对于样本数为 n 的数据集 X , 假设存在两种簇划分方案 U 和 V , 采用信息熵量化簇划分的不确定性, 两类方案的信息熵定义分别为

$$H(U) = -P(i) \log[P(i)], \quad (8)$$

$$H(V) = -P'(j) \log[P'(j)], \quad (9)$$

式中: $P(i) = |U_i|/n$ 为从 U 中随机选取的样本落入 U_i 类的概率; $P'(j) = |V_j|/n$ 为从 V 中随机选取的样本落入 V_j 类的概率。

U 和 V 之间的互信息分数 $MI^{[22]}$ 定义为

$$I_{\text{MI}}(U, V) = \sum_{i=1}^{|U|} \sum_{j=1}^{|V|} P(i, j) \log \left[\frac{P(i, j)}{P(i)P'(j)} \right], \quad (10)$$

式中, $P(i, j) = |U_i \cap V_j|/n$ 为随机选取的样本同时属于 U_i 和 V_j 的概率。

I_{MI} 取值范围为 $[0,1]$ 。 I_{MI} 接近 0, 表示聚类效果较差; I_{MI} 接近 1, 表示聚类效果较好。

(4) 归一化互信息(normalized mutual information, NMI)

在 MI 的基础上, 进一步定义归一化互信息 $NMI^{[22]}$, 即

$$I_{\text{NMI}}(U, V) = \frac{I_{\text{MI}}(U, V)}{\text{mean}[H(U), H(V)]}, \quad (11)$$

式中, $\text{mean}[\cdot]$ 是熵的广义均值。

与 I_{MI} 相比, I_{NMI} 不受数据真实类别数量的影响, 更为稳健。 I_{NMI} 为 0, 表示 U 和 V 没有互信息; I_{NMI} 为 1, 表示 U 和 V 完全相关。

(5) 调整的互信息(adjusted mutual information, AMI)

MI 的期望为

$$E[I_{\text{MI}}(U, V)] = \sum_{i=1}^{|U|} \sum_{j=1}^{|V|} \sum_{n_{ij}=\min(a_i, b_j)}^{\min(a_i, b_j)} \frac{n_{ij}}{n} \log \left(\frac{nn_{ij}}{a_i b_j} \right) \cdot \frac{a_i! b_j! (n-a_i)! (n-b_j)!}{n! n_{ij}! (a_i - n_{ij})! (b_j - n_{ij})! (n - a_i - b_j + n_{ij})!}, \quad (12)$$

式中: a_i 为 U_i 中的元素数量, $a_i = |U_i|$; b_j 为 V_j 中的元素数量, $b_j = |V_j|$; $n_{ij} = (a_i + b_j - n)^+$ 为同时落入 U_i 和 V_j 中的元素数量, 上标 “+” 表示取正值。

通过式(12), 得到调整的互信息 $AMI^{[22]}$ 为

$$I_{\text{AMI}} = \frac{I_{\text{MI}} - E(I_{\text{MI}})}{\text{mean}[H(U), H(V)] - E(I_{\text{MI}})}. \quad (13)$$

I_{NMI} 和 I_{AMI} 的取值依赖于熵的广义均值, 不同均值形式对应不同的计算结果。本文采用算术平

均值作为广义均值,此时, I_{NMI} 和 I_{AMI} 的取值范围均为[0,1]。

(6) 同质性(homogeneity index, HI)、完备性(completeness index, CI)、V测度(V-measure, VM)

在给定数据真实聚类划分的情况下,基于条件熵可定义一些直观的评价指标。参考文献[23],定义2个理想的评价指标^[23],即同质性HI和完备性CI。

HI用于衡量每个簇中是否只包含单一类别的成员(数据样本),其定义为

$$I_{\text{HI}} = 1 - \frac{H(C|K)}{H(C)}, \quad (14)$$

式中: $H(C|K)$ 为在给定簇 K 时类别 C 的条件熵; $H(C)$ 为类别 C 的信息熵。

CI用于衡量给定类别的所有成员是否均被分配至同一簇中,其定义为

$$I_{\text{CI}} = 1 - \frac{H(K|C)}{H(K)}, \quad (15)$$

式中: $H(K|C)$ 为给定类别 C 时聚类 K 的条件熵; $H(K)$ 为聚类 K 的信息熵。

I_{HI} 和 I_{CI} 从不同角度量化聚类质量,取值范围均为[0,1],值越大表示聚类效果越好。在实际应用中,主要观察 I_{HI} 与 I_{CI} 的差值,差值越小聚类效果越好。为了综合考虑 HI 和 CI,采用 VM^[23] 来衡量聚类算法对数据的划分质量,即

$$I_{\text{VM}} = \frac{(1+\beta)I_{\text{HI}}I_{\text{CI}}}{\beta I_{\text{HI}} + I_{\text{CI}}}, \quad (16)$$

式中: β 为权重参数,用于协调指标 I_{HI} 与 I_{CI} 的关系: $\beta < 1$, 则赋予 I_{HI} 更多的权重; $\beta > 1$, 则赋予 I_{CI} 更多的权重。本文取 $\beta=1$, 则

$$I_{\text{VM}} = 2 \frac{I_{\text{HI}}I_{\text{CI}}}{I_{\text{HI}} + I_{\text{CI}}}. \quad (17)$$

实际上, I_{VM} 等同于 I_{MI} 。特别地,当 $\beta=1$ 时, I_{VM} 等同于 I_{NMI} 。

(7) 福克斯-马洛斯指数(Fowlkes-Mallows index, FMI)

FMI最初用于衡量两组聚类结果的相似性,是一种无监督的比较。当已知数据样本的真实聚类时,可将其作为监督自适应指标,本质为聚类精度和召回率的几何平均值^[24],即

$$I_{\text{FMI}} = \frac{\text{TP}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})}}, \quad (18)$$

式中: TP 表示真阳性,即数据集中真实聚类标签和预测标签属于相同簇的数据点对数; FP 表示假

阳性,即真实聚类标签同簇而预测标签不同簇的数据点对数; FN 表示假阴性,即真实标签不同簇、预测标签同簇的数据点对数。

显然, I_{FMI} 基于真实标签和聚类结果的交集、联合集及簇内/簇间数据点对数的比值计算,其取值范围为[0,1],值越大,表示聚类效果越好。

(8) 轮廓系数(silhouette coefficient, SC)

若数据集固有的聚类标签未知,需基于数据自身的相似性开展聚类性能评估,轮廓系数 SC 是该场景下的经典无监督评价指标。SC 依托同一簇中数据样本点间的欧氏距离,以及与下一个邻近簇中所有数据点间的平均欧氏距离,其取值范围为[0,1]。第 i 个数据点 x_i 的轮廓系数^[25]为

$$I_{\text{SC}_i} = \frac{e_i - f_i}{\max\{f_i, e_i\}}, \quad (19)$$

式中: e_i 为数据样本 x_i 与同簇中其他样本点之间的平均欧氏距离; f_i 为样本 x_i 与最邻近簇所有数据点的平均欧氏距离。

实际应用中,单个簇中的轮廓系数常采用该簇中所有样本轮廓系数的平均值。本文采用整个数据集全部样本轮廓系数的平均值。平均轮廓系数越高,算法的聚类性能越好。

1.2 确定聚类数

K-means 算法的应用前提是已知数据的聚类数 K 。聚类有效性(cluster validity)分析^[14]正是建立在准确辨识数据中簇数量 K 的基础之上。前述各聚类性能指标为数据集潜在簇结构的识别提供了客观标准。同时,上述指标的取值范围均为[0,1],为数据集潜在聚类(簇)数量的估计提供了可行思路:在聚类数的可能范围 $[1, K_{\text{max}}]$ 内,分别计算不同 K 对应的聚类性能指标,通过求解各指标的全局最优值,确定数据集的固有簇数量 K_{best} 。

2 聚类中心与聚类边界

K-means 算法的输出结果为指定数量 K_{best} 的聚类划分结果以及聚类中心位置。算法所产生的数据簇满足以下条件^[7]:一是数据簇的中心为该簇内所有数据样本的平均值位置;二是在同一簇中,每个样本点到其簇中心的距离,均小于到其他簇中心的距离。

显然,随着数据集中样本数量的增加,将样本分配至对应数据簇的组合方案数量将急剧增长。若数据中存在噪声和离群值,穷举式搜索将消耗大量时间,导致算法收敛效率显著下降。

对于大样本数据, K-means算法不采用遍历所有可能结果的搜索方法, 而是采用期望最大化(E-M)的迭代算法优化搜索策略。E-M算法是统计学中功能强大的迭代算法^[17], K-means聚类可视为E-M算法的最简单应用形式。在获得簇数量 K_{best} 的前提下, K-means算法的E-M迭代步骤如下。

期望步(E-step): 将数据点划分到距离其最近的聚类中心所对应的簇中。

最大化步(M-step): 将聚类中心更新为当前簇内所有数据样本的平均值。

E-M算法为优化迭代过程, 其目标是不断更新 K_{best} 个聚类中心的位置。一般地, 每一轮E-step和M-step的迭代, 均会获得相较上一轮更优的聚类效果。当算法收敛时, 最终得到 j 个聚类(簇) C_j ($j=1, 2, \dots, K_{\text{best}}$)的质心为

$$m_j = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{ik} \quad j=1, 2, \dots, K_{\text{best}}, \quad (20)$$

式中, n_j 为第 j 个聚类 C_j 中数据样本的数量。

经E-M迭代收敛后得到的聚类质心 m_j , 其位置可能落在某个数据点上, 也可能位于多个数据点之间。

在同一簇 C_j 中, 聚类质心到该簇内最远样本点(可能是噪声)之间的欧氏距离为

$$d_j = \max \{d(x_{ik}, m_j)\} \quad i \in C_j, j=1, 2, \dots, K_{\text{best}} \quad (21)$$

d_j 即为簇 C_j 的半径。以 m_j 为圆心、 d_j 为半径, 即可得到每个簇 C_j 的圆形边界。

3 初始聚类质心改进

在经典K-means算法中, 随机选择 K_{best} 个数据点作为初始聚类质心, 可能引发以下问题: 一是初始质心位置过于接近, 导致在算法后续迭代中2个及以上的聚类被合并, 形成空聚类, 即聚类中无数据点; 二是虽未产生空聚类, 但随机初始化质心选择导致聚类效果较差。为了解决随机初始化问题, 将最小相似度的聚类中心选择原则与最大-最小(max-min)相似性规则相结合, 针对经典K-means算法提出改进方案。

对于数据集 X , 首先计算2个数据样本(假设为 x_1 和 x_2)的最小相似度, 即

$$S_{im}(x_1^0, x_2^0) = \min \{S_{im}(x_i, x_j), i, j=1, 2, \dots, n\}, \quad (22)$$

式中, 上标0表示初始分配。

将 x_1 、 x_2 作为簇 C_1 、 C_2 的初始质心, 即

$$x_1^0 \in C_1^0, x_2^0 \in C_2^0. \quad (23)$$

在此基础上, 利用最大-最小相似性规则找到

第3个数据样本, 作为第3个簇的初始质心, 即

$$\max \{S_{im}(x_3^0, x_r^0), r=1, 2\} = \min \{[S_{im}(x_s^0, x_r^0), r=1, 2], s \neq 1, 2\}, \quad (24)$$

式中: r 为初始质心的序号; s 为第3个质心的序号; x_r^0 为初始质心。

以此类推, 则可依次获得初始 K_{best} 个质心, 满足

$$x_1^0 \in C_1^0, x_2^0 \in C_2^0, \dots, x_{K_{\text{best}}}^0 \in C_{K_{\text{best}}}^0 \quad (25)$$

通过上述改进, 可保证初始聚类质心具有较大的相异性, 从而在K-means算法后续迭代过程中, 聚类不会被合并而形成空聚类。在聚类分析中, 数据样本间的相异性可通过欧式距离进行度量, 也可以利用夹角余弦等来计算^[7]。

4 本文方法的实施步骤

针对给定数据集 X (样本容量 n 、特征维度 p), 本文方法可估计出 X 的簇数量 K_{best} 和聚类中心位置。具体步骤如下。

步骤1 计算数据集 X 各特征维度的样本均值 m_k 和标准差 σ_k ($k=1, 2, \dots, p$)。

步骤2 基于式(5), 利用 m_k 和 σ_k 估计出数据集 X 的最大簇数 K_{max} 。

步骤3 在 $[1, K_{\text{max}}]$ 范围内, 计算不同聚类数 K 对应的各聚类性能指标。

步骤4 定位聚类性能指标在区间 $[1, K_{\text{max}}]$ 内的最大值, 从而确定 X 的最佳聚类(簇)数 K_{best} 。

步骤5 对于确定的 K_{best} , 利用最小相似度规则以及最大-最小相似性规则选择初始聚类中心。

步骤6 采用E-M算法迭代更新聚类中心位置。

步骤7 基于式(21)计算每个簇的半径, 从而获得其圆形边界。

5 仿真实验

由K-means算法的目标函数和迭代过程可知, 算法适用于球形(或超球形)的数据簇划分。对于非球形簇的数据集, K-means的聚类结果与数据本身的簇结构相差较大。因此, 在仿真时, 采用不同样本数量、不同簇数量的球形随机斑点数据。为了聚类结果的可视化, 仅采用二维数据进行仿真, 将经典K-means算法、文献[13]中改进的K-means算法及本文方法的聚类效果进行对比分析。

仿真平台配置: Intel Core i7-8700 CPU、16 GB 内存PC, Windows10系统, Spyder 5运行环境。

5.1 小样本数据的聚类

利用随机数生成器,生成由400个样本点构成的二维 Blobs 数据集,数据散点图见图1,其中,特征1和特征2为数据维度,无量纲。由图1可见,400个样本点大致形成4个数据簇。根据式(5),估计出该数据集的最大簇数 $K_{\max}$ 为12。

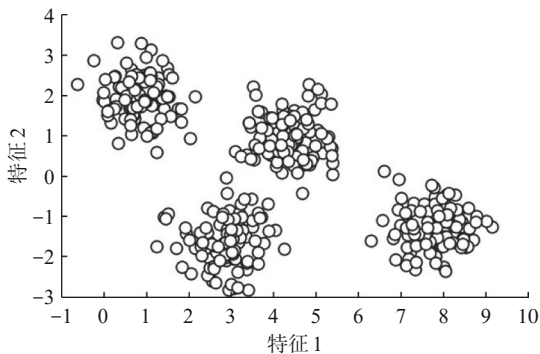


图1 小样本原始数据的散点图

Fig.1 scatter plot of the original small sample datasets

在 $K \in [1, 12]$ 范围内,计算不同 K 对应的各聚类性能指标,提取各指标在该范围内的最大值以估计数据集的固有簇数 K_{best} 。由于 K_{best} 大于2,因此在仿真中采用ARI。另外,由于HI和CI仅从单一维度反映聚类的性能,因此采用VM来评价聚类质量而不采用HI和CI。

小样本数据集聚类性能指标随 K 的变化规律见图2。各性能指标在 $K \in [1, 12]$ 范围内均存在最大值,且最大值均发生在 K 为4处,表明该数据集的最佳簇数 K_{best} 为4。同时,VM、NMI曲线重合,验证了VM等效于NMI且非常接近于AMI的结论。故后续仿真中,不再绘制NMI曲线。

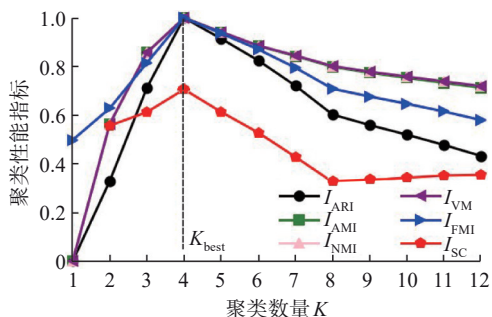


图2 小样本数据集聚类性能指标随 K 的变化

Fig.2 variation of clustering performance indicators with the number of clusters K for datasets

在 K_{best} 为4的条件下,经典K-means算法的聚类结果见图3。与原始数据的散点图相比,经典K-means算法因随机选择数据点作为初始质心,除右下角蓝色数据簇外,其余3个数据簇均出现明显不合理

划分:中部原本属于2个簇的数据(绿色)被算法合并为1个簇;左上角数据簇被算法拆分为2个不同的簇(粉色簇和橙色簇),且橙色簇中混入上部绿色簇的个别样本点。从聚类性能指标来看, I_{AMI} 为0.791 41, I_{VM} 为0.793 26, I_{FMI} 为0.793 50, I_{ARI} 为0.641 10, I_{SC} 仅为0.592 98,与理想值1相差较大,运行时间为0.015 6 s,聚类性能较差。

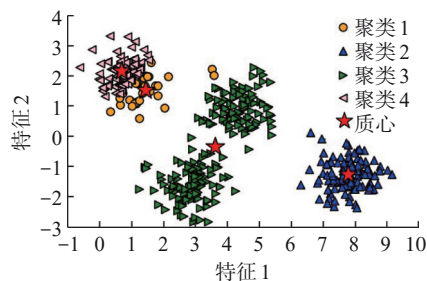


图3 经典K-means算法对小样本数据的聚类结果

Fig.3 clustering results of small sample data by the classic K-means algorithm

在 K_{best} 为4的条件下,分别采用改进K-means算法^[13]及本文方法对该数据集进行聚类分析,聚类结果分别见图4、图5。

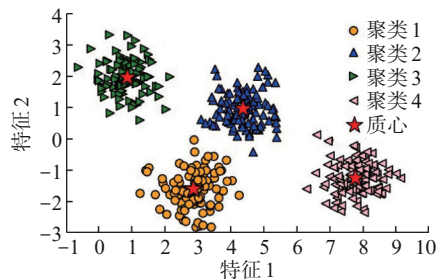


图4 文献[13]算法对小样本数据的聚类结果

Fig.4 clustering results of small sample data by the algorithm in reference [13]

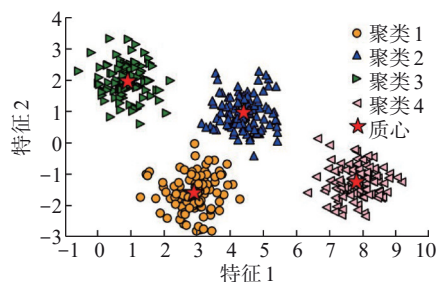


图5 本文方法对小样本数据的聚类结果

Fig.5 clustering results of small sample data by the proposed method in this paper

由图4可以看出,文献[13]算法的聚类结果与原始数据集固有划分结构匹配度较好,其 I_{ARI} 、 I_{AMI} 、 I_{VM} 、 I_{FMI} 、 I_{SC} 分别为0.994 67、0.991 56、0.991 59、0.996 00、0.751 78,运行时间0.015 7 s。

由图5可以看出,本文方法的聚类结果与原始数据簇的划分完全一致。 I_{ARI} 、 I_{AMI} 、 I_{VM} 、 I_{FMI} 均达到理想值1, I_{SC} 为0.885 96, 接近理想值1, 且算法耗时小于0.000 1 s, 聚类质量与迭代效率均实现显著提升。

采用本文方法确定数据集的聚类中心后, 基于式(21)计算各簇的聚类半径, 以聚类中心为圆心, 绘制各簇的分界线, 见图6。

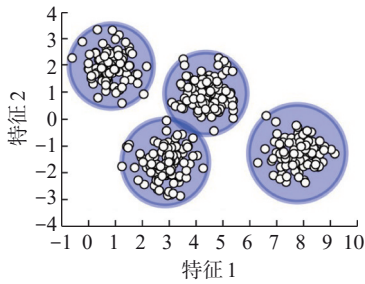


图6 采用本文方法绘制的小样本数据聚类边界线
Fig.6 clustering boundary curves of small sample data drawn by the proposed method in this paper

由图6可以看出, 本文方法绘制的圆形聚类边界线精确度较高, 说明本文方法能够精准估计和正确识别数据集的簇数量, 同时E-M算法为各个聚类中心提供准确定位。

5.2 大样本数据集的验证

为了验证本文算法在“样本容量较大和簇数较多”场景下的聚类性能, 开展不同样本规模 and 不同簇数量的仿真实验。

(1) 样本数为2 100、簇数为7的数据集

随机生成2 100个样本、7个簇的二维数据集, 其散点图见图7。

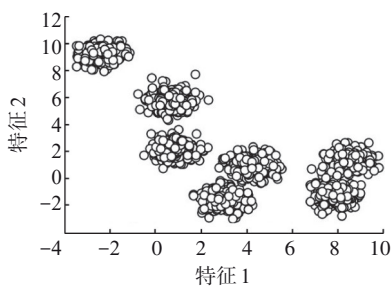


图7 簇数为7的数据散点图
Fig.7 scatter plot of data with 7 clusters

分别计算该数据集2个特征维度的样本均值和标准差, 根据式(5)得出该数据集的最大可能簇数 K_{max} 为14。在 $K \in [1, 14]$ 范围内, 计算不同簇数 K 对应的各聚类性能指标, 并确定各指标的全局最大值, 结果见图8。

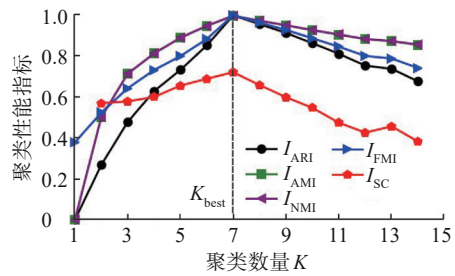


图8 簇数量为7时聚类性能指标随K的变化
Fig.8 performance metrics of clustering with 7 clusters as K varies

由图8可知, 各聚类性能指标均在 K 为7处取得最大值, 即该数据集的最佳簇数 K_{best} 为7。

在 K_{best} 为7的条件下, 经典K-means算法的聚类结果见图9。该算法将中下方蓝色的2个簇合并为1个簇, 将中部上方的1个簇拆分为2个簇(浅蓝色簇和灰色簇), 导致数据集固有聚类结构发生扭曲失真。 I_{ARI} 、 I_{AMI} 、 I_{VM} 、 I_{FMI} 、 I_{SC} 分别为0.805 36、0.916 66、0.917 04、0.839 76、0.769 11, 耗时0.015 7 s, 聚类效果较差。

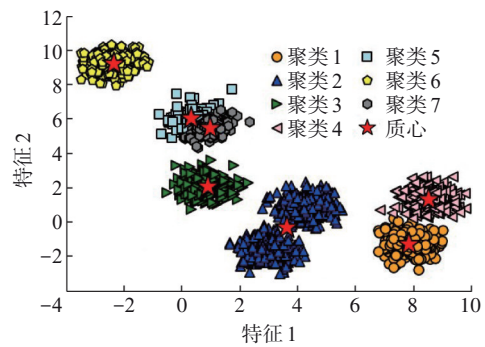


图9 基于经典K-means算法的7簇数据集聚类结果
Fig.9 clustering results of 7 clusters datasets based on classic K-means algorithm

在相同实验条件下, 分别采用文献[13]改进K-means算法和本文方法对该数据集进行聚类分析, 结果分别见图10、图11。

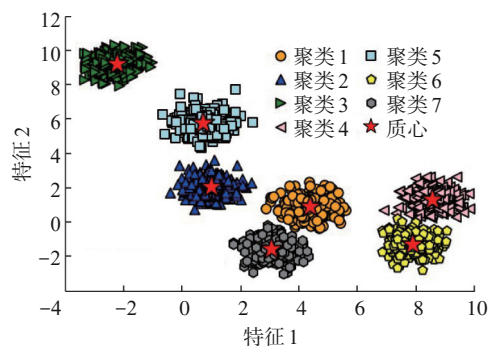


图10 文献[13]算法对簇数为7数据集的聚类结果
Fig.10 clustering results of the datasets with 7 clusters by the algorithm in reference [13]

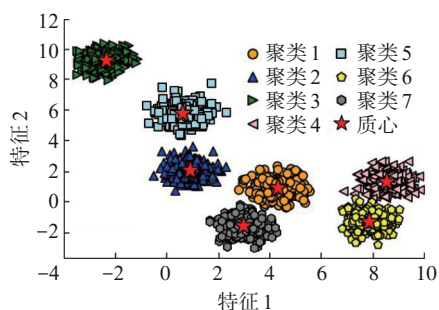


图11 本文方法对簇数为7数据集的聚类结果

Fig.11 clustering results of the datasets with 7 clusters by the proposed method in this paper

由图10、图11可以看出,两种算法的聚类结构与原始数据集固有簇结构一致。文献[13]改进算法的 I_{ARI} 、 I_{AMI} 、 I_{VM} 、 I_{FMI} 、 I_{SC} 分别为 0.949 36、0.969 05、0.969 33、0.958 26、0.813 55, 耗时 0.015 6 s。本文方法的 I_{ARI} 、 I_{AMI} 、 I_{VM} 、 I_{FMI} 、 I_{SC} 分别为 0.995 56、0.993 75、0.993 78、0.996 19、0.895 46, 相对更接近理想值1, 且耗时小于 0.000 1 s, 收敛性能明显提高, 聚类质量显著提升。

根据上述计算结果, 绘制本文方法对该数据集的聚类分界线, 见图12。尽管样本量增加导致相邻簇数据点存在交叉, 本文方法仍能准确识别不同聚类的边界线。

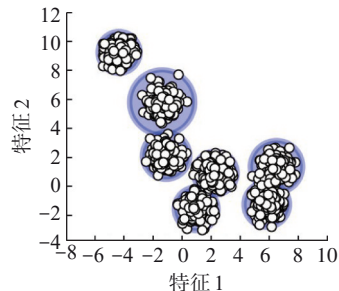


图12 本文方法绘制的7个数据簇边界线

Fig.12 boundary curves of 7 data clusters generated by the proposed method in this paper

(2) 样本数为7 000、簇数为10的数据集

随机生成7 000个样本、10个簇的二维数据集, 其散点图见图13。

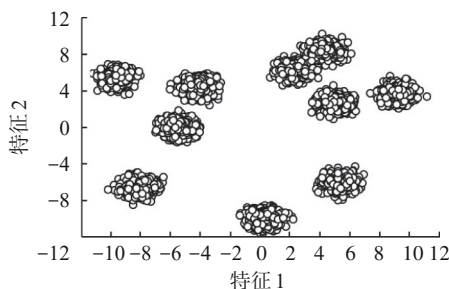


图13 簇数为10的大数据集散点图

Fig.13 scatter plot of big datasets with 10 clusters

根据式(5), 估计该数据集的最大簇数 $K_{\max}$ 为19。在 $K \in [1, 19]$ 范围内, 各类聚类性能指标随 K 的变化规律见图14, 所有指标在 K 为10时取得最大值, 因此确定该数据集的最佳簇数 K_{best} 为10。

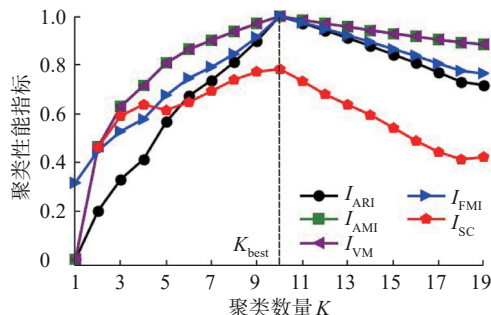


图14 簇数为10时聚类性能指标随K的变化

Fig.14 performance metrics of clustering with 10 clusters as K varies

在 K_{best} 为10的条件下, 经典K-means算法的聚类结果见图15。算法将右上部2个簇合并为1个簇(橙色簇), 将左上部1个簇分裂为2个簇(蓝色簇和玫粉色簇)。 I_{ARI} 、 I_{AMI} 、 I_{VM} 、 I_{FMI} 、 I_{SC} 分别为 0.867 76、0.950 65、0.950 78、0.884 00、0.822 06, 耗时 0.046 9 s。

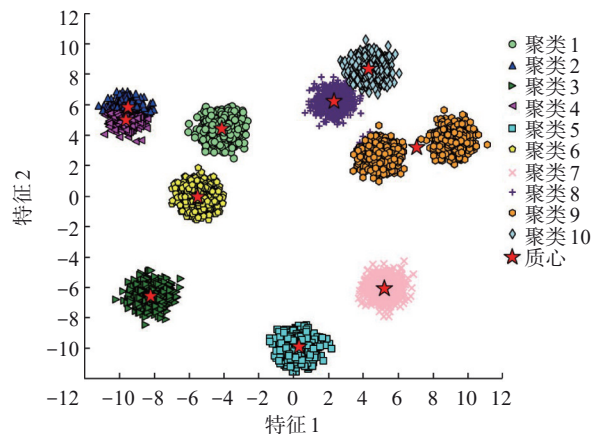


图15 经典K-means算法对大数据集的聚类结果

Fig.15 clustering results of classic K-means algorithm on large datasets

分别采用文献[13]改进算法与本文方法对该数据集进行聚类分析, 结果分别见图16、图17。两种算法的数据划分与数据集固有的聚类结构完全一致。其中, 文献[13]改进算法的 I_{ARI} 、 I_{AMI} 、 I_{VM} 、 I_{FMI} 、 I_{SC} 分别为 0.936 27、0.954 32、0.956 76、0.947 63、0.834 66, 耗时 0.044 4 s; 本文方法的 I_{ARI} 、 I_{AMI} 、 I_{VM} 、 I_{FMI} 、 I_{SC} 分别为 0.999 05、0.998 68、0.998 68、0.999 14、0.834 66, 耗时 0.015 6 s。本文方法的 I_{SC} 超过 0.83, I_{ARI} 、 I_{FMI} 、 I_{AMI} 、 I_{VM} 均高度接近理想值1, 且运行时间仅为经典K-means算法的30%, 收敛性能与聚类质量显著提升。

对比图 11、图 17 的聚类性能指标可知,随着数据集样本数量和聚类簇数量的增加,本文方法的聚类性能不仅未下降,反而显著提升,表明该方法对于数据集的样本容量和聚类(簇)数均具有较强的鲁棒性。

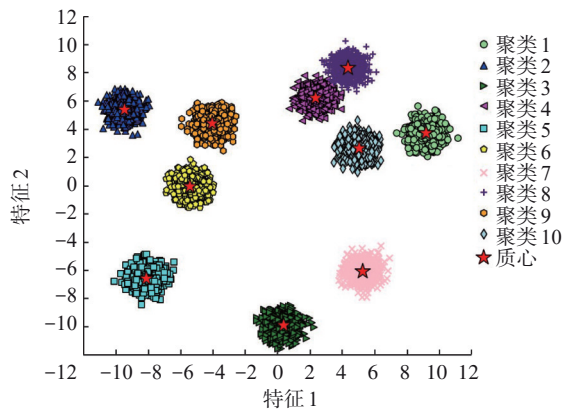


图 16 文献[13]算法对大数据集的聚类结果

Fig.16 clustering results of the algorithm in reference [13] on large dataset

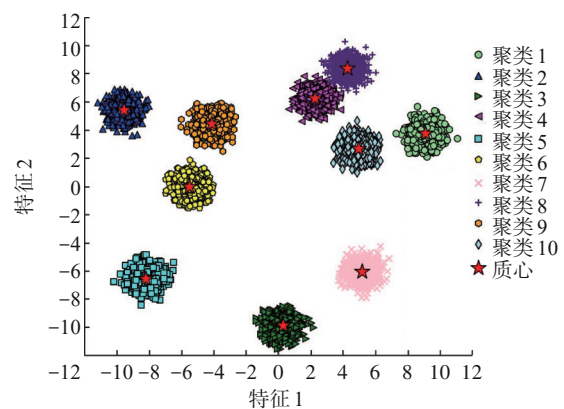


图 17 本文方法对大数据集的聚类结果

Fig.17 clustering results of the proposed method in this paper on large dataset

采用本文方法绘制 10 个数据簇的边界线,结果如图 18 所示。

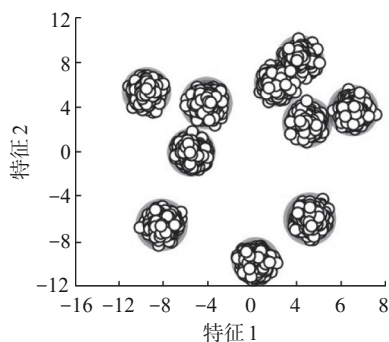


图 18 本文方法绘制的 10 个数据簇的边界线

Fig.18 boundary curves of 10 clusters drawn by the proposed method in this paper

由图 18 可知,尽管样本规模较大导致顶部 2 个簇的数据点存在空间交叉,本文方法仍能较准确地给出各数据簇的边界。

6 结论

针对经典 K-means 算法的固有缺陷,提出一种改进 K-means 方法。采用 3σ 准则估计数据的簇数区间,以区间内多维度聚类性能指标的全局最大值确定数据集潜在的簇数;通过最小相似度原理、最大-最小相似度规则解决经典 K-means 算法的随机初始化问题。以 E-M 算法迭代收敛的簇质心为圆心、簇内距离质心最远样本点的欧氏距离为半径,构建各簇的圆形边界曲线,有效抑制噪声与离群值对聚类结果的干扰。仿真结果表明,本文方法可为球形数据集提供高精度的聚类划分,为监督分类器提供未知数据集的类标签。

本研究仅针对球形结构数据开展验证,存在一定局限。后续将引入数据变换技术,将本文方法的适用范围拓展至更复杂结构数据集的聚类任务。

参考文献(References):

- [1] BANDARU S, NG A H C, DEB K. Data mining methods for knowledge discovery in multi-objective optimization: part A-Survey [J]. Expert Systems with Applications, 2017, 70: 139-159.
- [2] BUI A T. Dimension reduction with prior information for knowledge discovery [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(5): 3625-3636.
- [3] CZARNOWSKI I, JĘDRZEJOWICZ P. Supervised classification problems-taxonomy of dimensions and notation for problems identification [J]. IEEE Access, 2021, 9: 151386-151400.
- [4] SINGH J, SINGH D. A comprehensive review of clustering techniques in artificial intelligence for knowledge discovery: taxonomy, challenges, applications and future prospects [J]. Advanced Engineering Informatics, 2024, 62: 102799.
- [5] MAHNOOR, SHAFI I, CHAUDHRY M, et al. A review of approaches for rapid data clustering: challenges, opportunities, and future directions [J]. IEEE Access, 2024, 12: 138086-138120.
- [6] SUBASI A. Machine learning techniques [M] // Practical Machine Learning for Data Analysis Using Python. Amsterdam: Elsevier, 2020: 91-202.
- [7] XU R, WUNSCH-II D C. Clustering [M]. New Jersey: IEEE Press & WILEY, 2009: 67-73.
- [8] STEINLEY D. K-means clustering: a half-century synthesis [J].

- British Journal of Mathematical and Statistical Psychology,2006,59: 1-34.
- [9] STEINLEY D. Stability analysis in *K*-means clustering[J]. British Journal of Mathematical and Statistical Psychology, 2008, 61: 255-273.
- [10] 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, PAMI-6(1): 81-87.
- [11] HE X S, HE F, FAN Y P, et al. An effective clustering scheme for high-dimensional data[J]. Multimedia Tools and Applications, 2024, 83:45001-45045.
- [12] 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.
- [13] 何选森,何帆,于海澜. *K*-均值算法的初始化改进与聚类质量评估[J]. 西安工程大学学报, 2024, 38(6): 114-123.
- HE Xuansen, HE Fan, YU Hailan. Initialization improvement and clustering quality evaluation of *K*-means algorithm[J]. Journal of Xi'an Polytechnic University, 2024, 38(6): 114-123.
- [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 Xuansen, HE Fan, XU Li, 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] 何选森,何帆,樊跃平,等. 高维数据聚类数量可视化确定模式[J]. 沈阳航空航天大学学报, 2024, 41(3): 71-84.
- HE Xuansen, HE Fan, FAN Yueping, et al. Visualized determination mode for clustering quantity of high-dimensional data[J]. Journal of Shenyang Aerospace University, 2024, 41(3): 71-84.
- [16] 何帆,何选森,刘润宗,等. 数据降维与 *K*-均值聚类的质量评估[J]. 重庆理工大学学报(自然科学), 2024, 38(1): 131-141.
- HE Fan, HE Xuansen, LIU Runzong, et al. Data dimensionality reduction and clustering quality evaluation of *K*-means clustering [J]. Journal of Chongqing University of Technology (Natural Science), 2024, 38(1): 131-141.
- [17] IWASAKI Y, SASAKI Y, NAGATA T, et al. Dynamic mode decomposition based on expectation-maximization algorithm for simultaneous system identification and denoising[J]. Mechanical Systems and Signal Processing, 2025, 223: 111864.
- [18] XIE J R, HONG T, LAING T, et al. On normality assumption in residual simulation for probabilistic load forecasting[J]. IEEE Transactions on Smart Grid, 2017, 8(3): 1046-1053.
- [19] CREIGHTON J H C. A first course in probability models and statistical inference[M]. New York, NY: Springer New York, 1994: 140-156.
- [20] HUBERT L, ARABIE P. Comparing partitions[J]. Journal of Classification, 1985, 2: 193-218.
- [21] STEINLEY D. Properties of the hubert-arabie adjusted rand index [J]. Psychological Methods, 2004, 9(3): 386-396.
- [22] STREHL A, GHOSH J. Cluster ensembles-a knowledge reuse framework for combining multiple partitions[J]. Journal of Machine Learning Research, 2002, 3: 583-617.
- [23] ROSENBERG A, HIRSCHBERG J. V-measure: a conditional entropy-based external cluster evaluation measure[C]// Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL). Prague, Czech Republic Association for Computational Linguistics, 2007: 410-420.
- [24] FOWLKES E B, MALLOWS C L. A method for comparing two hierarchical clusterings[J]. Journal of the American Statistical Association, 1983, 78(383): 553-569.
- [25] BAGIROV A M, ALIGULIYEV R M, SULTANOVA N. Finding compact and well-separated clusters: clustering using silhouette coefficients[J]. Pattern Recognition, 2023, 135: 109144.