文章编号:1671-9352(2026)01-0049-16 DOI:10.6040/j.issn.1671-9352.4.2025.004

基于知识度量的模糊粗糙 *c*-均值算法

李文焱[1],李丽红[1,2,3*],王洪欣[1]
(1.华北理工大学理学院, 河北 唐山 063210; 2.河北省数据科学与应用重点实验室, 河北 唐山 063210; 3.唐山市数据科学重点实验室, 河北 唐山 063210)

摘要:提出基于知识度量的模糊粗糙 *c*-均值聚类(fuzzy rough *c*-means based on the knowledge measure, KFRCM)算法。传统聚类算法在处理具有模糊边界的数据时存在一定的局限性,表现为对初始聚类中心较为敏感且在高维空间中效率较低。为解决上述问题,引入特征加权的知识度量,结合模糊隶属度函数与粗糙集近似算子,采用高斯核相似度以增强边界特性。实验采用 14 个数据集,实验结果表明,KFRCM 算法的聚类准确性、稳定性和计算效率均优于 6 种主流聚类算法。该研究首次将知识度量与模糊粗糙聚类相结合,为开发更为可靠和适应性更强的聚类算法提供了新的思路和算法。

关键词:模糊粗糙集;知识度量;聚类分析;高斯核函数;上下近似集
中图分类号:TP391 **文献标志码**:A
引用格式:李文焱,李丽红,王洪欣. 基于知识度量的模糊粗糙 *c*-均值算法[J]. 山东大学学报(理学版),2026,61(1):49-64.

Fuzzy rough *c*-means based on the knowledge measure

LI Wenyan[1], LI Lihong[1,2,3*], WANG Hongxin[1]
(1. School of Science, North China University of Science and Technology, Tangshan 063210, Hebei, China; 2. Hebei Key Laboratory of Data Science and Application, Tangshan 063210, Hebei, China; 3. Tangshan Key Laboratory of Data Science, Tangshan 063210, Hebei, China)

Abstract: A knowledge-based fuzzy rough c-means clustering method (KFRCM) is introduced. Traditional clustering methods have limitations in handling data with fuzzy boundaries, which are sensitive to the initial cluster centers, and exhibit low efficiency in high-dimensional spaces. To address these issues, the KFRCM is proposed. a feature-weighted knowledge measure is incorporated, fuzzy membership functions are integrated with rough set approximation operators, and Gaussian kernel similarity is utilized to enhance boundary characterization. Experimental results on 14 datasets demonstrate that the proposed KFRCM algorithm outperforms 6 mainstream clustering algorithms in terms of accuracy, stability, and computational efficiency. This study is recognized as the first integration of knowledge measurement with fuzzy rough clustering, offering a new perspective and an advanced algorithmic framework for developing more reliable and adaptable clustering techniques.

Key words: fuzzy rough sets; knowledge measurement; clustering analysis; gaussian kernel function; upper and lower approximation sets

0 引言

数据聚类是一种无监督学习技术,旨在通过相似的特征将样本进行分组,揭示隐藏的模式和结构[1]。聚类算法已被广泛应用于数据分析[2]、模式识别[3]、图像处理[4]、自然语言处理[5]和生物信息学[6]等领域。在数据科学和统计学中,聚类不仅是数据预处理[7]、特征提取[8]和模型构建[9]的基础,而且通过最大化簇内相似性并最小化簇间差异性的原则,显著提升数据分析的精度与有效性[10]。聚类算法主要分为硬聚类和软

收稿日期:2025-05-15; **网络出版时间**:2025-12-11
基金项目:唐山市基础研究项目(24130202C)
第一作者:李文焱(1999—),女,硕士研究生,研究方向为数据挖掘和三支决策研究. E-mail:1375339465@qq.com
*通信作者:李丽红(1979—),女,教授,硕士,研究方向为数据挖掘和三支决策研究. E-mail:22687426@qq.com

聚类。硬聚类算法适用于数据点具有清晰边界的情形,将每个数据点分配到唯一的簇中[11]。例如,k 均值聚类算法(k-means clustering algorithm, k-means)[12]、k 中心点算法(partitioning around medoids algorithm, k-medoids)算法[13]和聚集层次聚类算法(agglomerative hierarchical clustering algorithm, AHC)[14]等算法已在多个领域取得了广泛应用。然而,当面对模糊边界或高度不确定的数据时,硬聚类算法的效果会显著下降。在这种情况下,软聚类算法更具优势,它允许每个数据点具有不同的隶属度函数,并为每个数据点分配多个簇。常见的软聚类算法包括模糊聚类算法(fuzzy clustering method, FCM)、直觉模糊聚类算法(intuitionistic fuzzy clustering method, IFCM)[15]、高斯混合模型(gaussian mixture model, GMM)算法[16]和粗糙 c-均值聚类算法(rough c-means clustering algorithm, RCM)[17]等。

近年来,针对不确定性聚类的研究取得显著进展,逐步解决模糊边界特征化和高维数据表示的问题。其中一个关键方向是将知识量化与软聚类相结合。Szmidt 等[18]率先提出了直觉模糊熵的概念,建立框架衡量不确定数据中固有的模糊性。在此基础上,Guo 等[19]提出一种知识度量模型,明确区分信息量与信息清晰度之间的差异。此研究通过双参数模型得到扩展[20],该模型自适应地平衡信息的具体性与模糊性,为动态数据集中量化模糊知识提供新的方向。知识驱动的度量算法能够有效提高噪声环境下的数据集聚类的稳定性,尤其对于在噪声环境下的数据集。

尽管取得了这些进展,但当前算法中仍存在 2 个关键性的问题:特征可辨识性,传统的模糊聚类算法对所有特征赋予相同权重,忽略了在实际场景中的不同重要性。虽然 Patel 等[21]提出基于相似度的属性加权算法,但该算法缺乏与知识度量框架的系统性结合;边界适应性,基于粗糙集和核技术的算法[22]在一定程度上解决了边界不确定性的问题,但仍然难以调和局部与全局数据的关系。Zhang 等[23]的多粒度粗糙集研究强调分层特征交互的重要性,但他们只研究静态矩阵表示,无法适应不断变化的聚类结构。这些研究揭示直觉模糊熵和知识度量算法在解决复杂数据分析问题中的潜力。基于以上研究基础,本文提出基于知识度量的模糊粗糙 c-均值聚类算法(fuzzy rough c-means based on the knowledge measure, KFRCM)。KFRCM 算法旨在通过引入知识度量的概念克服传统软聚类算法在处理模糊边界、初始聚类中心选择和高维度数据时的局限性,旨在推动聚类技术在复杂数据集中的应用和发展。

本文提出基于知识加权的初始聚类中心优化算法,通过构建知识度量指标,量化数据特征的分布差异性,为各维度特征分配自适应权重,从而提高初始聚类中心选择的准确性,解决传统随机初始化导致的收敛不稳定问题。设计新型高斯核模糊粗糙隶属度函数算子,将高斯核相似度函数引入粗糙集近似空间,建立基于全局数据结构的隶属关系计算模型,并通过核函数调节样本间相似性度量,提高模糊边界样本的聚类精度。实现知识加权机制、模糊粗糙集理论与传统 FCM 算法的三重融合,通过知识加权机制保障特征选择合理性,利用模糊粗糙集处理不确定性数据,并基于模糊聚类算法优化聚类迭代过程,在保持 FCM 算法计算效率的同时提升算法对复杂数据的处理能力。

1 相关工作

近年来,粗糙集聚类、模糊聚类及知识度量的研究取得显著进展,这些算法在处理不确定性和模糊性数据方面具有优势。粗糙集理论主要用于知识发现和规则生成,其核心思想是通过上下近似集对数据进行分类,能够有效处理不确定性和不完整性数据。近年来,粗糙集聚类在理论研究和应用领域取得了重要突破。例如,多粒度粗糙集引入多层次的上下近似集,显著提高分类精度,特别是在医疗诊断和金融风险评估中,利用粗糙集聚类处理高维和不完整数据取得了较好的效果[24]。此外,结合深度学习技术,基于神经网络的粗糙集聚类算法进一步提高计算效率和分类性能[25]。

模糊聚类分析则是基于模糊集理论的一种软分类算法,能够处理类属中介性问题。模糊聚类通过隶属度函数函数描述样本的类属关系,近年来在目标函数优化、算法实现和应用领域取得了重要进展。例如,基于熵的模糊聚类算法通过引入信息熵,显著提高对噪声数据的鲁棒性。在算法实现方面,结合分布式计算技术,研究者提出适用于大规模数据集的模糊聚类算法,显著提高计算效率。在应用领域,模糊聚类被广泛应用于图像分割和模式识别中,特别是在医学图像分割中,模糊聚类能够有效区分不同组织区域[26]。

知识度量是评估数据分类和信息提取效果的重要指标。提出了基于信息熵和模糊集的知识度量算法,

能够更准确地评估分类效果。例如,基于模糊集的知识度量算法通过引入隶属度函数函数,能够更精确地评估分类结果的不确定性。在知识发现和数据挖掘中,知识度量被用于评估规则生成和分类模型的性能,为决策支持提供重要依据[27]。

粗糙集与模糊聚类的结合研究也逐渐成为热点。基于粗糙集和模糊集的混合聚类算法能够同时处理不确定性和模糊性问题。例如,混合聚类算法通过结合粗糙集的上下近似和模糊集的隶属度函数函数,显著提高分类精度[28]。在复杂系统建模和智能决策中,混合聚类算法被用于处理多源异构数据,取得较好的效果。例如,在智能交通系统中,混合聚类算法能够有效处理多源传感器数据[29]。

尽管粗糙集聚类、模糊聚类及知识度量在理论研究和实际应用中取得显著进展,但仍面临一些挑战。例如,需要进一步探索粗糙集与模糊聚类的深度融合算法,处理更复杂的数据[30]。此外,结合人工智能技术,开发更高效的聚类算法也是未来的研究方向。例如,基于深度学习的混合聚类算法有望进一步提高计算效率和分类性能。在应用领域,这些算法有望在智能制造和智慧城市中得到更广泛的应用。例如,在智能制造中,粗糙集聚类和模糊聚类用于设备故障诊断和生产优化[31]。未来的研究应重点关注多数据融合、实时计算和大规模数据处理等问题,推动粗糙集聚类和模糊聚类在实际应用中的进一步发展。

2　预备知识

2.1　直觉模糊知识度量

直觉模糊集(intuitionistic fuzzy set, IFS)概念[32]扩展了模糊集理论,每个元素通过隶属度函数、非隶属度函数和犹豫度表征。

设 X 是一个非空集合。X 上的直觉模糊集 U 定义为 $U=\{\langle x,\mu_U(x),v_U(x)\rangle \mid x\in X\}$。元素 x 在 U 中的隶属度函数由函数 μ_U 表示,而非隶属度函数由函数 v_U 表示。隶属度函数与非隶属度函数必须满足以下条件 $0\leqslant\mu_U(x)+v_U(x)\leqslant 1,\ \forall x\in X$,犹豫度函数为

$$\pi_U(x)=1-\mu_U(x)-v_U(x), \tag{1}$$

式中:当 $\pi_U(x)=0$ 时,直觉模糊集简化为经典的模糊集, $\mu_U(x)+v_U(x)=1$。

Guo 等[33]提出基于公理的直觉模糊知识度量算法,不依赖于熵。知识量从 2 个角度量化的信息的多少,分别是信息量 $A=\mu_U(x_i)+v_U(x_i)$ 和信息清晰度 $E=\mid\mu_U(x_i)-v_U(x_i)\mid$, A 越大,表明知识的量越大。Guo 等提出双参数直觉模糊知识度量模型[34]为

$$K_1(\boldsymbol{U};\alpha,\beta)=\frac{1}{n}\sum_{i=1}^{n}(\alpha\pi_U(x_i)E+\beta A(1-E))+AE, \tag{2}$$

式中: $\boldsymbol{U}$ 的知识度量由该集合传达的特定知识量和蕴含的非特定潜在知识量构成。参数 $\alpha\in[0,1]$ 和 $\beta\in(0,1)$ 分别反映了主体对未知信息的态度以及对信息数量或清晰度的偏好。α 平衡信息数量和清晰度之间的权重,当 $\alpha\in[0,0.5]$ 时,评估趋于保守;当 $\alpha\in(0.5,1]$ 时,评估则呈现出乐观的态度。参数 β 描述主体在清晰度和信息数量之间的偏好。当 $\beta\in(0,0.5)$ 时,主体倾向于优先考虑信息的清晰度,导致较小的知识度量;而当 $\beta\in(0.5,1)$ 时,主体则更注重信息的数量,导致较大的知识度量。若 $\beta=0.5$,则表示主体没有明确的偏好。

在完全接受潜在知识($\alpha=1$)且没有明确偏好($\beta=0.5$)的情况下,模型将进一步简化为

$$K_2(\boldsymbol{U};\ \alpha=1;\ \beta=\frac{1}{2})=1-\frac{1}{2n}\sum_{i=1}^{n}(1-\mid\mu_U(x_i)-v_U(x_i)\mid)(1+\mu_U(x_i)), \tag{3}$$

式(3)更为简洁地呈现了原始公式的核心思想,并且提供了一个在直觉模糊逻辑背景下直接衡量 $\boldsymbol{U}$ 中所包含知识的度量算法。

2.2　模糊相似度度量和核密度

高斯核函数是常用的模糊相似度度量函数,广泛应用于评估 2 个模糊集之间的相似性[35]。在低维空间中使用核算法计算数据点在高维特征空间中的内积,避免显式计算高维空间中的特征映射。方法避免复杂的显式特征映射计算,节省计算时间。在高维空间中执行点积运算用核函数表示,即

$$h(U_i,U_j)=\varphi(U_i)^T\varphi(U_j), \tag{4}$$

式中:U_i 和 U_j 是直觉模糊集 U 中的子集

为了进一步描述数据在高维空间中的关系,并进行非线性变换,数据从 U_i 映射到高维的特征空间 $\varphi(U_i)$,U_j 映射到高维的特征空间 $\varphi(U_j)$,映射结果为

$$\varphi(U_i)=\langle\varphi(\mu_{U_i}),\varphi(v_{U_i}),\varphi(\pi_{U_i})\rangle, \tag{5}$$

$$\varphi(U_j)=\langle\varphi(\mu_{U_j}),\varphi(v_{U_j}),\varphi(\pi_{U_j})\rangle, \tag{6}$$

式中:$\varphi(\mu_{U_i})$、$\varphi(v_{U_i})$、$\varphi(\pi_{U_i})$表示 U_i 转化后的隶属度、犹豫度和非隶属度函数,$\varphi(\mu_{U_j})$、$\varphi(v_{U_j})$、$\varphi(\pi_{U_j})$表示 U_j 转化后的隶属度、犹豫度和非隶属度函数。核距离定义为衡量 2 个集合在高斯核映射后的高维特征空间中的相似性,即

$$\delta_\varphi^2(U_i,U_j)=\|\varphi(U_i)-\varphi(U_j)\|^2=2\left(1-\exp\left(-\frac{\|U_i-U_j\|^2}{\sigma^2}\right)\right), \tag{7}$$

式中:参数 σ 控制高斯核的宽度,决定了高斯核函数对距离的敏感度,$\|U_i-U_j\|$ 为 U_i 到 U_j 的欧氏距离。

设 $A_i\subseteq A\subseteq \boldsymbol{U}$,基于高斯核距离定义,高斯密度函数为

$$\rho(A_i,A_j)=\left(1+\sum_{j=1}^{n}\delta_\varphi^2(A_i,A_j)\right)^{-1}。 \tag{8}$$

通过上述公式的推导,核距离的计算通过高斯核映射后的高维特征空间中的相似性进行量化,并结合高斯密度函数反映数据点与中心点之间的紧密程度。

2.3 模糊粗糙算子

模糊等价关系在传递性方面具有独特的特征,建立模糊相似关系扩展二元模糊关系的传递性范围[36]。设 $R:\boldsymbol{U}\times\boldsymbol{U}\to[0,1]$ 为定义在域 $\boldsymbol{U}$ 上的一个模糊二元关系,$R(s,t)$ 表示元素 s 与 t 之间的相关程度,若关系 R 满足以下条件,则称 R 为一个模糊等价关系:

(1) 自反性,对于任意元素 $s\in\boldsymbol{U}$,有 $R(s,s)=1$。

(2) 对称性,对于任意元素 s, $t\in\boldsymbol{U}$,有 $R(s,t)=R(t,s)$。

(3) 最大-最小传递性,对于任意元素 $s,t,m\in\boldsymbol{U}$,有 $R(s,t)\geqslant\sup\limits_{m\in U}\{\min[R(s,m),R(m,t)]\}$。

模糊粗糙集理论是模糊集理论与粗糙集理论的结合,为处理复杂数据中的不确定性和不完全性提供有效的解决方案,既能提高决策的准确性和效率,也能够有效应对实际问题中的各种挑战。

2.4 FCM 算法及其演化

FCM 算法是一种模糊聚类技术,通过评估对象与聚类之间的隶属关系对数据划分。在 FCM 算法中,通过最小化目标函数实现聚类。对于具有 n 个数据点和 k 个聚类的数据集,目标函数定义为

$$L=\sum_{i=1}^{n}\sum_{j=1}^{k}\mu_{ij}^m\delta(x_i,c_j), \tag{9}$$

式中:模糊指数为 m, $\delta(x_i,c_j)$ 为数据点 x_i 与聚类中心 c_j 之间的欧几里得距离,聚类 j 的中心为 c_j,数据点 x_i 在聚类 j 中的隶属度函数为 μ_{ij}。

由此,得到相应的隶属度函数和聚类中心函数为

$$\mu_{ij}=\left(\sum_{l=1}^{k}\left(\frac{\delta(x_i,c_j)}{\delta(x_i,c_l)}\right)^{\frac{2}{m-1}}\right)^{-1}, \tag{10}$$

$$c_j=\frac{\sum\limits_{i=1}^{n}\mu_{ij}^m x_i}{\sum\limits_{i=1}^{n}\mu_{ij}^m}。 \tag{11}$$

FCM 算法的步骤:先初始化聚类中心矩阵 $\boldsymbol{C}$ 和隶属度函数矩阵 $\boldsymbol{M}$,再根据式(10)、(11)更新隶属度函数和聚类中心。评估目标函数的变化,判断其是否低于指定的阈值 ε。如果满足该条件,则过程终止,输出最终的隶属度函数矩阵;否则,返回步骤(2),重复计算。最后根据最高隶属度函数原则将对象分配到各自的聚类中。

算法 1 FCM 算法。

输入　数据集 $\boldsymbol{X}=\{x_1,x_2,\cdots,x_n\}$;聚类数目 k;模糊指数 m;收敛阈值 ε;最大迭代次数 $t_{\max}$

输出　$\boldsymbol{M};\boldsymbol{C}$

(1) 初始化 $\boldsymbol{C}$、$\boldsymbol{M}$,迭代次数 $t=0$
(2)　当 $t<t_{\max}$时执行
(3)　　对 i 执行 1 到 n
(4)　　　对 j 执行从 1 到 c
(5)　　　　计算 μ_{ij}
(6)　　　结束循环
(7)　　结束循环
(8)　对 j 执行从 1 到 k
(9)　　　计算 c_j
(10)　结束循环
(11)　根据式(9)计算新目标函数值 $L(t)$
(12)　$t\leftarrow t+1$
(13)　　如果 $\|L(t)-L(t-1)\|\leqslant\varepsilon$ 返回 $\boldsymbol{M}$、$\boldsymbol{C}$
(14)　　结束条件
(15) 结束循环
(16) 返回 $\boldsymbol{M},\boldsymbol{C}$

FCM 算法将模糊概念引入聚类分析,解决数据集的模糊性和不确定性。然而,FCM 算法也存在局限性,包括收敛速度慢、对噪声敏感以及易受局部最优解的影响。为了解决这些问题,提出几种改进算法。初始中心选择优化,传统的 FCM 算法采用随机选择初始中心,导致局部最优解。为了解决这个问题,须开发各种改进算法优化初始中心选择,增强算法性能。常见的策略包括基于聚类有效性指标、密度峰值和遗传算法的策略。这些策略不仅提高初始中心的选择的质量,还增强最终聚类的效率和有效性。增强空间结构的表示,在传统 FCM 算法中,通常用数据点表示欧几里得距离空间中的向量,处理高维数据时可能导致“维度灾难”。为了解决这一问题,提出基于核算法和子空间投影的 FCM 算法。这些算法能够更有效地处理高维数据, 提高算法的准确性和效率。增强目标函数,传统的 FCM 算法基于欧几里得距离误差平方和的目标函数,使得算法容易受到噪声和离群点的影响。为了解决这一问题,提出改进的目标函数,其中包括基于信息熵和模糊熵的目标函数,更好地处理数据中的不确定性,从而提高算法的鲁棒性和稳定性。

总之,对 FCM 算法的改进主要集中在提高迭代效率、增强聚类稳定性以及提高对异常数据的鲁棒性。本文提出的基于知识度量的模糊粗糙 c-均值算法对这 3 方面优化,不仅加快收敛速度,增强聚类结果的稳定性,还提高算法在复杂数据集上的适应性,进一步提高了其整体性能和实际应用价值。

3　基于知识度量的模糊粗糙 c-均值算法

本章介绍 KFRCM 算法,旨在解决经典 FCM 算法对初始聚类中心选择的敏感性以及模糊边界描述的不明确性。KFRCM 算法基于知识度量加权。聚类过程包括 3 个部分:(a)数据集首先进行直觉模糊化,再基于直觉模糊知识量、核空间密度与核距离计算特征权重,最终选择初始聚类中心;(b)模糊粗糙度算子替代聚类中心的迭代优化过程;(c)聚类迭代继续进行,直到满足收敛条件,得到最优的聚类划分迭代算法。

3.1　模糊聚类中心初始化

设直觉模糊集 $\boldsymbol{U}$ 采用一种具有偏好性的策略生成隶属度函数,$\boldsymbol{U}$ 中第 i 个样本的第 j 个特征对应的隶属度函数为 μ_{ij},而非隶属度函数为 ν_{ij},满足 $\mu_{ij}+\nu_{ij}\leqslant 1$,即

$$\begin{cases}\mu_{U_{ij}}=\dfrac{\max\limits_{1\leqslant i\leqslant n}\{x_{ij}\}-x_{ij}}{\max\limits_{1\leqslant i\leqslant n}\{x_{ij}\}-\min\limits_{1\leqslant i\leqslant n}\{x_{ij}\}},\\ \nu_{U_{ij}}=(1-\mu_{U_{ij}})^{\frac{1}{a}},\\ \pi_{U_{ij}}=1-\mu_{U_{ij}}-\nu_{U_{ij}},\end{cases}\tag{12}$$

式中:α 是 Yager 运算符的可调参数[37], $\alpha\in[0.8,1)$。α 用于调节隶属度函数的模糊程度,控制模糊集的生成和转换过程。

通过直觉模糊化算法得到

$$\boldsymbol{U}=\begin{pmatrix}\langle\mu_{U11},v_{U11},\pi_{U11}\rangle & \langle\mu_{U12},v_{U12},\pi_{U12}\rangle & \cdots & \langle\mu_{U1s},v_{U1s},\pi_{U1s}\rangle\\ \langle\mu_{U21},v_{U21},\pi_{U21}\rangle & \langle\mu_{U22},v_{U22},\pi_{U22}\rangle & \cdots & \langle\mu_{U2s},v_{U2s},\pi_{U2s}\rangle\\ \vdots & \vdots & \vdots & \vdots\\ \langle\mu_{Un1},v_{Un1},\pi_{Un1}\rangle & \langle\mu_{Un2},v_{Un2},\pi_{Un2}\rangle & \cdots & \langle\mu_{Uns},v_{Uns},\pi_{Uns}\rangle\end{pmatrix}。\tag{13}$$

数据集加权是通过直觉模糊知识度量实现的。根据知识度量公理框架,知识度量包含信息量$\mu_U(x_{ij})+v_U(x_{ij})$和信息的清晰度$|\mu_U(x_{ij})-v_U(x_{ij})|$。由于优化模型的非线性特性,清晰度度量中的绝对值导致模糊划分矩阵的不稳定性,因此知识度量模型将清晰度度量替换为平方形式$|\mu_U(x_{ij})-v_U(x_{ij})|^2$。为了确保加权过程的客观性,设 $\alpha=0$。因此,修订后的知识度量模型为

$$k(U_{ij})=\frac{1}{2n}\sum_{i=1}^{n}(\mu_U(x_{ij})+v_U(x_{ij}))(1+|\mu_U(x_{ij})-v_U(x_{ij})|^2),\tag{14}$$

式中 $w=\{w_1,w_2,\cdots,w_s\}$ 表示 $\boldsymbol{U}$ 中各特征的权重决策变量。每个特征 w_j 在 $\boldsymbol{U}$ 中的权重为

$$w_j=\frac{\sum_{i=1}^{n}k(U_{ij})}{\sum_{i=1}^{n}\sum_{j=1}^{k}k(U_{ij})},\tag{15}$$

式中:$k(A_{ij})$表示与特征值 A_{ij}相关的函数,且权重满足 $\sum_{j=1}^{s}w_j=1$, $j=1,2,\cdots,s$。

令 $\boldsymbol{W}$ 为模糊加权数据集,$\boldsymbol{W}$ 中第 i 个元素的第 j 个特征值为 W_{ij}。W_{ij}的隶属度函数、非隶属度函数和犹豫度函数共同构成了直觉模糊集,且满足

$$\begin{cases}\mu_{W_{ij}}=1-(1-\mu_{U_{ij}})^{w_j},\\ v_{W_{ij}}=v_{w_j}\cdot v_{U_{ij}},\\ \pi_{W_{ij}}=1-\mu_{W_{ij}}-v_{W_{ij}},\end{cases}\tag{16}$$

式中:$i=1,2,\cdots,n$; $j=1,2,\cdots,s$。

$\boldsymbol{W}$ 为

$$\boldsymbol{W}=\begin{pmatrix}\langle\mu_{W11},v_{W11},\pi_{W11}\rangle & \langle\mu_{W12},v_{W12},\pi_{W12}\rangle & \cdots & \langle\mu_{W1s},v_{W1s},\pi_{W1s}\rangle\\ \langle\mu_{W21},v_{W21},\pi_{W21}\rangle & \langle\mu_{W22},v_{W22},\pi_{W22}\rangle & \cdots & \langle\mu_{W2s},v_{W2s},\pi_{W2s}\rangle\\ \vdots & \vdots & \vdots & \vdots\\ \langle\mu_{Wn1},v_{Wn1},\pi_{Wn1}\rangle & \langle\mu_{Wn2},v_{Wn2},\pi_{Wn2}\rangle & \cdots & \langle\mu_{Wns},v_{Wns},\pi_{Wns}\rangle\end{pmatrix}。\tag{17}$$

令$W_i\in\boldsymbol{W}$,在核空间中的样本密度定义为

$$\rho(W_i)=\left(1+\sum_{j=1}^{n}\delta^2(W_i,W_j)\right)^{-1},\tag{18}$$

式中:$\delta(W_i,W_j)$为核距离函数,即

$$\delta(W_i,W_j)=2\left(1-\exp\left(-\frac{\|W_i-W_j\|^2}{2\sigma^2}\right)\right)。\tag{19}$$

利用核距离公式,核空间的样本密度表示为

$$\rho(W_i)=\left(1+2\sum_{j=1}^{n}\left(1-\exp\left(-\frac{\|W_i-W_j\|^2}{\sigma^2}\right)\right)^2\right)^{-1},\tag{20}$$

式中:$\boldsymbol{W}$ 中的样本点距离接近时,样本密度 $\rho(W_i)$较高。$\rho(W_i)$从大到小排序,得到核空间密度排序

$$\rho(W_1^*)\geqslant\rho(W_2^*)\geqslant\cdots\geqslant\rho(W_l^*)\geqslant\cdots\geqslant\rho(W_n^*),\tag{21}$$

式中:l 是一个密度控制参数,$l\in\left[\frac{n}{2},\frac{3n}{2}\right]$。$\boldsymbol{W}$ 中的前 l 个决策变量组成一个高密度的决策变量集合为

$H=\{W_1^*,W_2^*,\cdots,W_l^*\}$。在选择初始聚类中心时,应同时考虑聚类中心的代表性和聚类间的差异,从 H 中选择 c 个聚类中心($c<l$)。每次选择一个聚类中心后,相应的决策变量将从 H 中移除,导致 H 的元素个数从 l 减少到 $l-1$。选择过程:

(1) 选择 H 中密度最大的决策变量作为第 1 个初始聚类中心 V_1。

(2) 根据式(19),计算在 H_{l-1} 中离 V_1 最远的决策变量作为第 2 个聚类中心 V_2。

(3) 通过最大最小法,首先计算 $d(W_j^*,V_{i-1})$ 的最小值,然后选择最大值。

(4) 利用 $V_i=\arg\max\limits_{W_j^*}[\min\limits_{W_j^*\in H-\{V_1,\cdots,V_{i-1}\}}(d(W_j^*,V_1),\cdots,d(W_j^*,V_{i-1}))]$ 得到满足要求的 c 个初始聚类中心。

3.2　模糊粗糙集算子

在传统的模糊粗糙集理论中,上近似算子确定一个对象是否可能属于目标聚类,而下近似算子则确定一个对象是否一定不属于目标聚类。然而,单独使用上近似或下近似算子可能会得出过于乐观或过于保守的判断,无法充分反映对象之间的相互关系及其隶属度函数。引入核相似度的概念,量化对象之间的相似性,利用相似度度量精确地定义对象对目标聚类的隶属度函数。

设 x, $y\in U$,模糊等价关系 R 的上近似和下近似算子分别为

$$\overline{F}(C_j)(x)=\max_{y\in U}\{\min[R(x,y),C_j(y)]\},\tag{22}$$

$$\underline{F}(C_j)(x)=\min_{y\in U}\{\max[1-R(x,y),C_j(y)]\}。\tag{23}$$

$F(x,C_j)$ 表示项 x 与聚类 C_j 之间的模糊粗糙度,$F(x,C_j)$ 定义为

$$F(x,C_j)=\frac{\overline{F}(C_j)(x)}{\underline{F}(C_j)(x)}。\tag{24}$$

模糊粗糙度 $F(x,C_j)$ 反映对象 x 与聚类 C_j 之间关系的确定性程度。需要注意的是,$F(x,C_j)$ 越小,表示确定性越高。较小的 $F(x,C_j)$ 表明 x 属于聚类 C_j 的可能性更大,应赋予较高的隶属度;相反,较大的 $F(x,C_j)$ 则表明存在更多的模糊性或不确定性,导致较低的隶属度。将具有最小模糊粗糙度 $F(x,C_j)$ 的聚类 C_j 分配给对象 x,确保每个对象都被分配到与其最相似、关系最确定的聚类中,提供更准确和可靠的分类结果。

3.3　算法的通用公式与逻辑

本文提出的 KFRCM 算法是改进的 FCM 聚类算法,利用模糊粗糙算子和知识度量优化聚类中心。计算特征权重和核相似度,结合模糊粗糙集的近似能力,提升算法对不确定和模糊数据的处理能力,提高聚类精度。本文算法改进了传统 FCM 算法的中心更新机制,更适用于复杂数据集。

数据预处理阶段。通过模糊知识度量计算特征的权重 w_{ij},将计算得到的权重应用于数据集的各个特征中;初始化聚类中心。计算核空间密度,确定加权数据集矩阵 w 和初始聚类中心 v;聚类分配。根据 $F(x,C_j)$ 的最小值,将加权数据集矩阵 w 中的数据点 x_i 分配到对应的聚类 C_j;获得新的聚类中心。获取每个聚类 C_j 中的所有数据点,计算它们的平均位置,并将其作为新的聚类中心;基于新的聚类中心重新计算隶属度函数矩阵;直到满足 $|\mu_i-\mu_{i-1}|\leqslant\varepsilon$ 或达到一定的迭代次数 m。

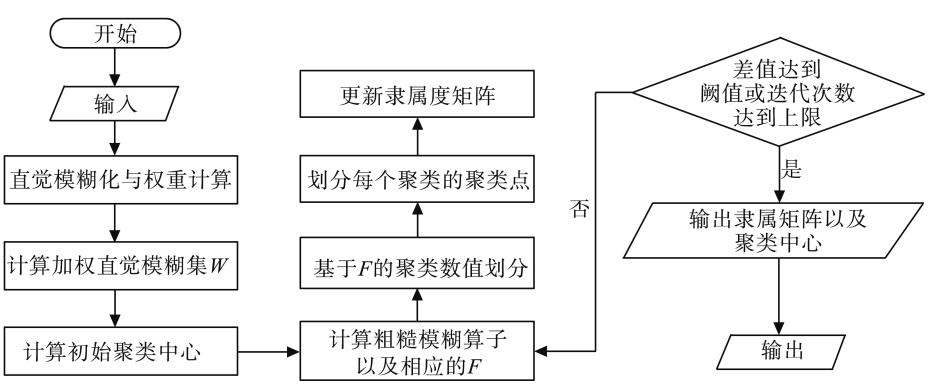


图 1　KFRCM 算法流程图

Fig.1　Flowchart of KFRCM algorithm

算法 2 KFRCM 算法。

输入 $\boldsymbol{U}$, c, m, ϵ, t, σ

输出 隶属度函数矩阵 $\boldsymbol{M}$ 和聚类中心矩阵 $\boldsymbol{C}$

(1) 利用 $\boldsymbol{U}$ 计算 μ 和 v
(2) 计算 $k(\boldsymbol{U})$ 和权重 w
(3) 利用 μ、v 和 w 转换 $\boldsymbol{W}$
(4) 计算 $\boldsymbol{W}$ 的密度并选择高密度点
(5) 利用高密度点初始化 $\boldsymbol{C}$
(6) 重复
(7) 对每个 $x_i \in U$ 执行
(8) 将 x_i 划分到 F 最小的簇中
(9) 结束循环
(10) 执行对每个簇 $j=1$ 到 c
(11) 利用式(11)计算 $\boldsymbol{C}_j$
(12) 结束循环
(13) 利用式(10)计算 μ_{ij}
(14) 直到满足阈值要求或达到最大迭代次数 t
(15) 返回 $\boldsymbol{M}$, $\boldsymbol{C}$

结合模糊粗糙集的上近似和下近似算子,能够有效捕捉对象与聚类中心之间的模糊关系,同时保留 FCM 算法在处理模糊数据集方面的优势。KFRCM 算法解决边界模糊的问题,增强隶属度函数计算的精确性,加速迭代收敛过程。引入模糊粗糙近似算子,进一步优化 FCM 算法的聚类效果。此外,σ 在聚类过程中具有重要意义,它决定高斯核的宽度以及相似度度量的敏感性,对聚类结果产生显著影响。

4 实验结果

4.1 比较算法和数据集

为了验证 FKRCM 算法的有效性,实验采用合成数据集和机器学习库数据集,各数据集参数见表 1 所示,其中基于模糊粗糙集的 FCM 算法在 14 个基准数据集上进行性能测试,并与其他六种聚类算法进行比较。选取 14 组常见的公开数据集,数据集均为经典的机器学习基准数据集,涵盖不同维度、规模和分类任务的真实与合成数据,广泛应用于分类、聚类等算法的性能测试与比较。所选的聚类算法包括 k-means 算法、FCM 算法、GMM 算法、KFCM 算法[38]、RCM 算法[39]、IFCM 算法[40]和 AHC[41]。本文实验深入分析 KFRCM 算法在不同数据集上的表现,并通过统计分析评估其相对竞争力。

表 1 数据集
Table 1 Dataset information

数据集	数量	维度	分类
Iris	150	4	3
Wine	178	13	3
Breast Cancer	569	30	2
Make Moons	1 000	2	2
Aggregation	788	2	7
WDBC	569	30	2
Cancer	286	9	2
Haberman	306	3	2
WBC	699	9	2
Seed	210	7	3
Thyroid	215	5	3
Dermatology	366	34	6
Banknote Authentication	1 372	4	2
Flame	240	2	2

其中,威斯康星乳腺癌(Wisconsin breast cancer, WBC)数据集为 9 个直接形态特征且存在缺失值,而威斯康星诊断乳腺癌(Wisconsin diagnostic breast cancer, WDBC)数据集扩展为 30 个统计衍生特征且无缺失值,具有更高维度和更丰富的诊断信息。本文采用 3 种指标评估聚类效果:聚类准确率 A_{ACC}[42]、福尔克斯-马洛斯指数 F_{FMI}[43]、兰德指数 R_{RI}[44]。A_{ACC}是通过将分配的聚类标签与真实标签进行比较,A_{ACC}是衡量正确匹配的样本数量。F_{FMI}用于计算召回率和精确度的几何平均值,F_{FMI}全面评估聚类结果的准确性和完整性。R_{RI}用于衡量聚类结果中样本对之间的相似度、同意度和

准确性,这 3 个指标的取值范围均为 0~1,指标越大表示聚类性能越优。

4.2　比较分析与数据可视化

实验结果表明,KFRCM 算法的准确率、福尔克斯-马洛斯指数和兰德指数均大于其他聚类算法,具有更好的聚类效果,见表 2—4 所示。

表 2　不同数据集不同算法的准确率

Table 2　The accuracy of different algorithms on different datasets

数据集	KMEANS 算法	FCM 算法	AHC 算法	RCM 算法	IFCM 算法	KFCM 算法	KFRCM 算法
Iris	0.705 2±0.077 7	0.840 0	0.826 7	0.840 0	0.920 0	0.798 1	**0.960 0**
Wine	0.966 3±0.006 9	0.965 0	0.927 0	0.966 3	0.819 6	0.601 1	**0.971 9**
Breast Cancer	0.879 1±0.014 5	0.896 3	0.880 5	0.896 3	0.895 9	**0.920 9**	0.910 4
Make Moons	0.840 7±0.020 4	**0.854 1**	0.850 0	0.870 0	0.838 5	0.853 2	0.810 0
Aggregation	0.727 2±0.000 6	0.727 2	0.756 2	0.815 3	0.611 0	0.732 8	**0.855 3**
WDBC	0.872 7±0.008 7	0.896 7	0.688 9	0.919 2	0.894 9	**0.920 9**	0.915 6
Cancer	0.963 6±0.009 4	0.970 6	0.824 3	0.824 3	0.824 3	0.796 3	**0.972 2**
Haberman	0.741 8±0.009 5	0.735 7	0.575 2	0.516 2	0.735 3	0.522 9	**0.758 2**
WBC	0.963 6±0.009 4	0.970 4	0.824 3	0.824 3	0.969 3	0.910 2	**0.972 2**
Seed	0.915 9±0.011 7	0.919 6	0.842 9	0.690 5	0.857 1	0.900 0	**0.914 3**
Thyroid	0.863 6±0.032 0	0.893 0	0.844 4	0.816 3	0.893 6	0.853 9	**0.902 3**
Dermatology	0.712 0±0.102 8	0.762 6	0.586 6	0.782 1	0.508 4	0.508 4	**0.854 7**
Banknote Authenticatio	0.574 9±0.001 4	0.607 1	0.668 4	0.662 5	0.598 4	0.598 4	**0.617 3**
Flame	0.850 4±0.011 2	0.850 0	0.516 7	0.770 8	0.841 7	0.841 7	**0.862 5**

从表 2 可以看出,采用的 14 个数据集中的 9 个数据集的 KFRCM 算法的准确率较大,充分验证了 KFRCM 算法效果优良。例如数据集 Iris,KFRCM 算法的准确率为 0.96,相较于 FCM 算法提高了 14%,相较于 k-means 算法提高了 25.48%。表明 KFRCM 算法在处理经典分类问题时具有更高的有效性。类似地,数据集 Wine,KFRCM 算法的准确率为 0.971 9,不仅高于 FCM 算法的准确率(0.965),也优于 RCM 算法的准确率(0.966 3),验证该算法在多元分类任务中的优势。在处理更具挑战性的复杂数据集时,KFRCM 算法的准确率较大。例如,高维数据集 Dermatology 上,KFRCM 算法的准确率 0.854 7,相比 FCM 算法的 0.762 6 提高了 9.21%;在具有复杂分布特性的 Aggregation 数据集上,KFRCM 算法的准确率为 0.855 3,相比 FCM 算法的准确率提高了 12.81%。这些结果表明,KFRCM 算法在处理高维特征和非线性可分数据时具有更强的鲁棒性。此外,KFRCM 算法对噪声干扰的数据集的准确率也较高。以 Haberman 数据集为例,尽管该数据集存在明显的类别不平衡问题,KFRCM 算法的准确率为 0.7582,高于 FCM 算法的准确率(0.735 7)和 KFCM 算法的准确率(0.522 9)。说明 KFRCM 算法对噪声数据具有较好的适应能力。

表 3　不同数据集不同算法的 F_{FMI}

Table 3　The F_{FMI}-index of different algorithms on different datasets

数据集	KMEANS 算法	FCM 算法	AHC 算法	RCM 算法	IFCM 算法	KFCM 算法	KFRCM 算法
Iris	0.893 3±0.042 8	0.751 7	0.749 8	0.752 0	0.856 3	0.798 0	**0.923 3**
Wine	0.932 2±0.013 5	0.929 3	0.860 2	0.931 9	0.691 8	0.631 1	**0.941 7**
Breast Cancer	0.772 0±0.027 2	0.689 8	0.778 5	0.689 6	0.773 1	**0.861 2**	0.847 2
Make Moons	0.657 9±0.015 3	0.658 6	**0.694 0**	0.663 0	0.681 8	0.683 6	0.691 0
Aggregation	0.785 9±0.001 0	0.758 8	0.714 9	0.734 3	0.563 9	0.798 2	**0.822 7**
WDBC	0.760 4±0.010 8	0.689 7	0.719 4	0.692 4	0.771 7	**0.861 2**	0.854 7
Cancer	0.843 4±0.080 1	0.895 7	0.722 8	0.722 1	0.893 8	0.823 9	**0.950 1**
Haberman	0.540 8±0.035 8	0.458 4	0.570 3	0.552 7	0.543 1	0.550 9	**0.728 3**
WBC	0.843 4±0.080 1	0.895 5	0.722 8	0.722 1	0.893 8	0.774 1	**0.950 2**
Seed	0.843 0±0.019 6	**0.848 2**	0.717 1	0.826 3	0.759 1	0.822 9	0.842 7
Thyroid	0.834 9±0.022 4	0.859 5	0.845 5	**0.868 4**	0.859 7	0.853 6	0.867 9
Dermatology	0.748 1±0.084 0	0.766 6	0.608 1	0.796 4	0.585 5	0.585 5	**0.835 1**
Banknote Authentication	0.513 6±0.000 4	0.525 4	**0.666 2**	0.552 1	0.521 6	0.521 6	0.546 4
Flame	0.753 8±0.015 1	0.753 0	0.623 0	0.766 6	0.741 7	0.741 7	**0.770 8**

表 4 不同数据集不同算法的 R_{RI}

Table 4 The R_{RI}-index of different algorithms on different datasets

数据集	KMEANS 算法	FCM 算法	AHC 算法	RCM 算法	IFCM 算法	KFCM 算法	KFRCM 算法
Iris	0.785 4±0.089 3	0.630 1	0.615 3	0.630 3	0.786 0	0.694 8	**0.885 7**
Wine	0.897 9±0.020 2	0.893 6	0.789 9	0.897 5	**0.936 3**	0.682 1	0.912 1
Breast Cancer	0.542 0±0.046 5	0.837 2	0.636 3	0.856 8	0.862 1	0.854 1	**0.871 4**
Make Moons	0.397 9±0.026 2	0.408 4	**0.467 5**	0.410 6	0.446 0	0.415 2	0.382 3
Aggregation	0.678 4±0.001 5	0.640 2	0.662 7	0.671 7	0.688 2	0.705 1	**0.774 0**
WDBC	0.823 5±0.022 0	0.837 1	0.570 6	0.849 3	0.859 4	0.854 1	**0.889 2**
Cancer	0.694 9±0.135 7	0.783 4	0.731 6	0.739 8	0.780 2	0.693 1	**0.890 9**
Haberman	0.528 9±0.028 4	0.495 6	0.598 5	0.493 1	0.503 7	0.499 4	**0.619 8**
WBC	0.794 9±0.135 7	0.782 9	0.732 4	0.813 7	0.780 2	0.802 5	**0.890 8**
Seed	0.765 7±0.029 3	0.773 3	0.803 4	**0.885 8**	0.638 3	0.884 1	0.864 9
Thyroid	0.653 8±0.091 1	0.643 5	0.678 5	0.659 0	0.644 4	0.675 0	**0.702 2**
Dermatology	0.886 1±0.050 1	0.886 9	0.806 1	0.916 9	0.657 6	0.657 6	**0.935 2**
Banknote Authentication	0.510 3±0.000 4	0.522 6	**0.556 4**	0.552 5	0.519 0	0.519 0	0.527 2
Flame	0.744 7±0.015 7	0.743 9	0.498 5	0.645 2	0.732 4	0.732 4	**0.761 8**

从表 3 可以看出,KFRCM 算法的 F_{FMI} 较大,尤其在类别边界较为模糊的数据集上表现出明显优势。具体而言,在 Iris 数据集上,KFRCM 算法的 F_{FMI} 为 0.923 3,显著高于 FCM 算法的 F_{FMI}(0.751 7)和 IFCM 算法的 F_{FMI}(0.856 3);在 Wine 数据集上,KFRCM 算法的 F_{FMI} 为 0.941 7,相比 FCM 算法的 F_{FMI} 提高了1.24%,同时也高于 IFCM 算法的 F_{FMI}(0.691 8)。此外,在含有噪声的 Haberman 数据集上,KFRCM 算法的 F_{FMI} 为 0.728 3,较 FCM 算法的 F_{FMI} 提高了 58.9%,进一步验证了该算法对噪声和异常值具有较强的抗干扰能力。

从表 4 可以看出,KFRCM 算法的 R_{RI} 指数较大,KFRCM 算法在存在显著类别重叠的数据集上表现良好。在数据集 Cancer 上,KFRCM 算法的 R_{RI} 为 0.890 9,相比 FCM 算法的 R_{RI} 提高了 13.72%;在 Dermatology 数据集上,KFRCM 算法的 R_{RI} 为 0.935 2,较 RCM 算法的 R_{RI} 提高了 1.83%。此外,在低可分性数据集 Make Moons 上,KFRCM 算法的 R_{RI} 为 0.382 3,虽略低于 AHC 算法的 R_{RI}(0.467 5),但仍高于 FCM 算法的 R_{RI}(0.408 4),表明该算法在处理复杂分布数据时仍具备一定竞争力,并且在该类场景下存在进一步优化的空间。

实验结果表明,相较于传统 FCM 算法及其改进算法,KFRCM 算法融合模糊粗糙集理论和核相似度计算,在处理高维数据、噪声干扰和模糊信息时具有优势,其中 14 个测试数据集中的 9 个数据集的准确率较大,8 个 F_{FMI} 较大和 7 个 R_{RI} 较大。特别是在 Iris、Wine 等经典数据集上,KFRCM 算法的准确率分别达到0.96 和 0.971 9,较传统算法有所提高。可视化分析进一步验证了该算法的有效性,如图 2 所示,Wine 数据集经降维处理后,KFRCM 算法能更准确地划分聚类边界。总体而言,KFRCM 算法通过创新的聚类中心优化机制,在保证算法鲁棒性的同时显著提高聚类精度。

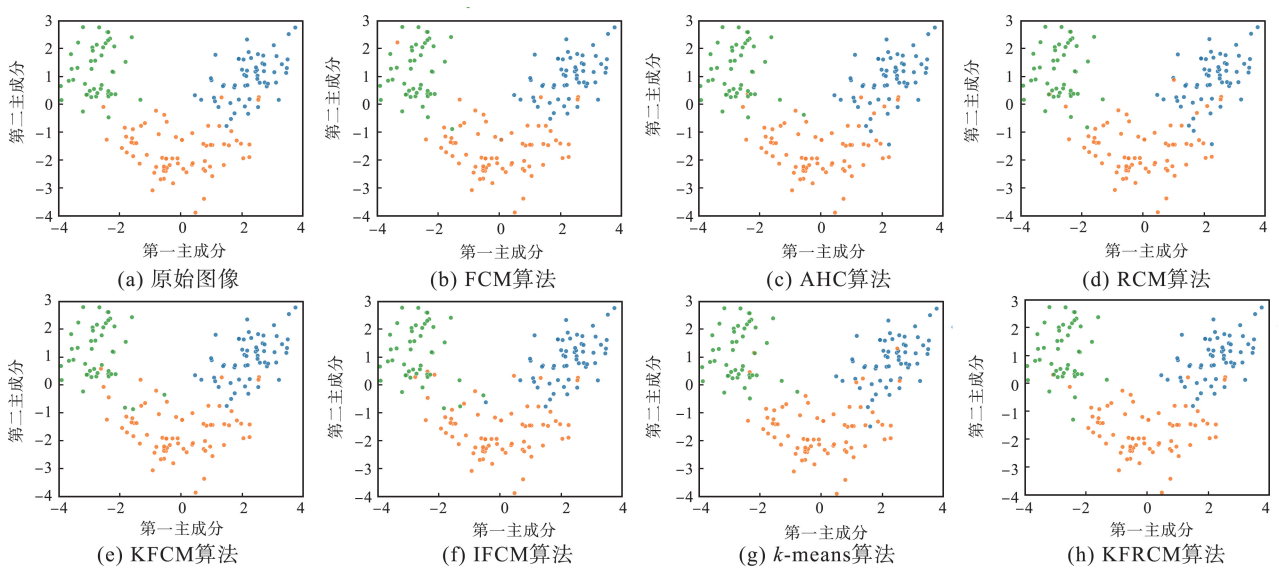


图 2 葡萄酒数据聚类图

Fig.2 Wine data clustering plot

重复实验表明,KFRCM 算法在同一数据集上聚类准确率的标准差趋近于 0,显示出稳定性和可重复性。从算法和数据处理两个角度来进行解释。从算法设计层面来看,KFRCM 算法通过融合模糊集与粗糙集理论构建了双重约束。首先,模糊隶属度函数约束通过知识度量实现了样本对聚类中心的初始点进行选择。其次,粗糙近似空间约束则通过上近似集与下近似集的对比降低了聚类的敏感性。初始点接近正确分类点的情况下,样本 x 对于第 j 个簇的模糊粗糙度 $F(x,C_j)$ 由核矩阵唯一确定。当 σ 取值恰当时,对数据局部结构的稳定表征使得算法能够稳定收敛于确定解。

从数据特性角度分析,在类间分离度显著且噪声水平较低的情况下,KFRCM 算法的隶属度函数有较强的收敛性。实验观察发现,不同初始条件下的聚类中心最终稳定在相同位置,从而进一步解释标准差趋近于零的现象。

4.3　迭代分析

聚类过程中会存在聚类准确率与效率矛盾问题。增加迭代次数提高精度,但也显著增加计算复杂度。不充分的迭代可能降低聚类效果,且划分不准确。本文提出一种融合知识度量机制和模糊粗糙算子的改进算法。通过建立科学的评估体系优化初始聚类中心选择,并利用核相似度量化样本与中心的模糊程度,KFRCM 算法在保证聚类精度的同时显著提高了计算效率。

图 3 的迭代次数与准确率关系曲线直观展示算法改进效果。通过融合模糊粗糙集理论和优化初始中心选择策略,KFRCM 算法在收敛速度、聚类精度和稳定性等方面均获得显著提高。一方面,模糊粗糙算子与高斯核函数的结合有效提高了单次迭代效率;另一方面,针对样本规模、特征维度、聚类数目等关键参数的计算优化,控制算法复杂度的同时增强处理能力。这种改进算法适合处理边界模糊、数据重叠的复杂聚类问题,在维持计算效率的基础上,提高传统 FCM 算法性能。

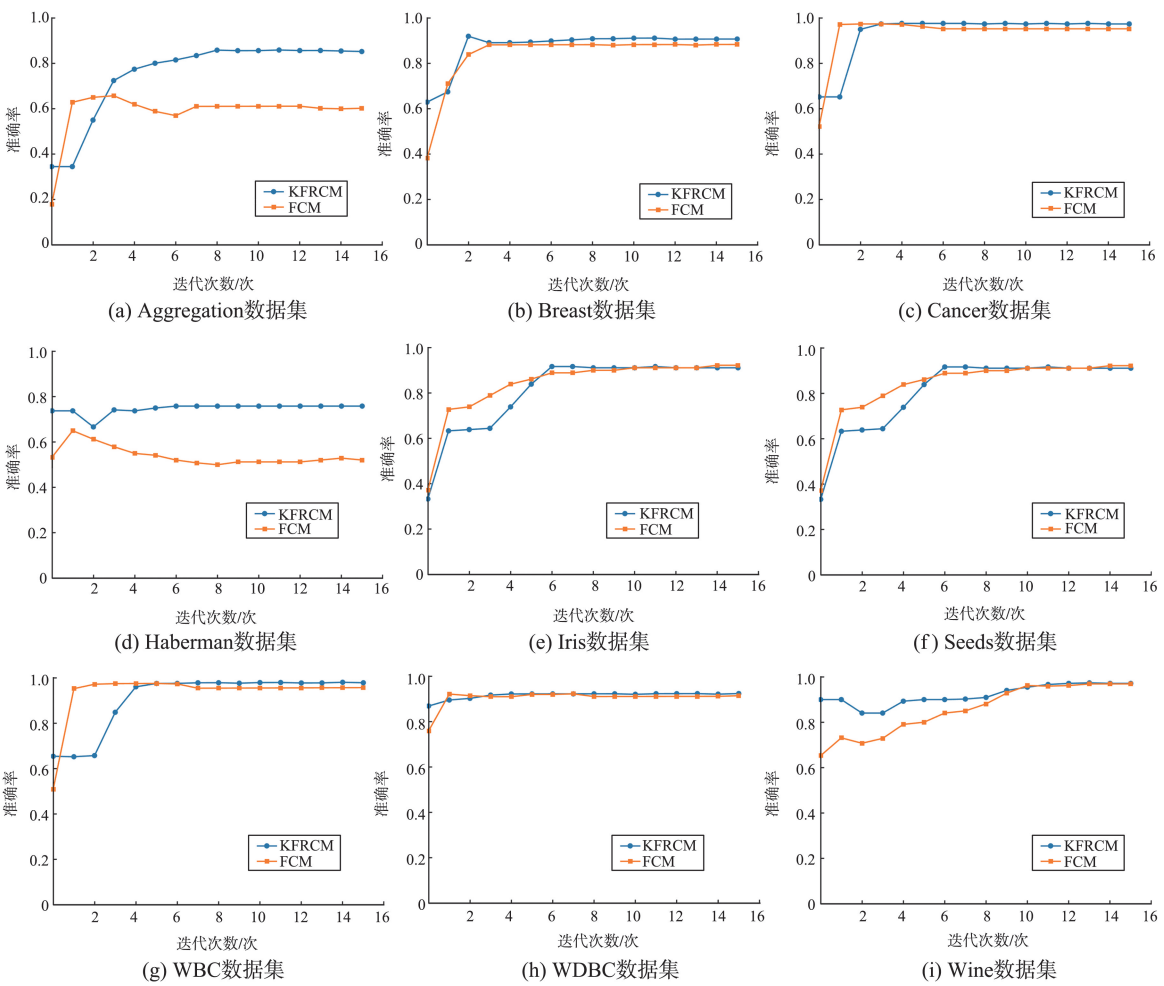


图 3　迭代分析

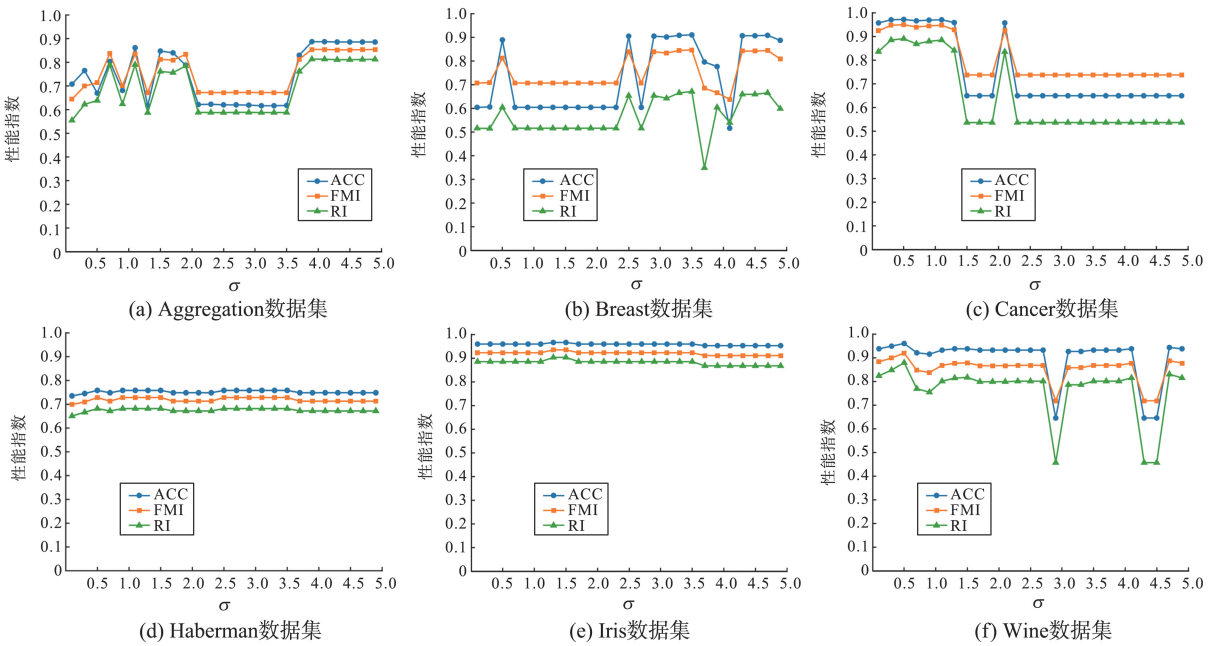
Fig.3　Iteration analysis

4.4 参数敏感性分析

4.4.1 σ 的变化

在软聚类中,通过数据点与聚类中心之间的关系度量相似性,这直接影响聚类过程中权重的分配。若将 σ 设置为较小的值,核的作用范围减小,导致无法充分捕捉数据的全局结构。相反,当 σ 设置过大时,核的作用范围过宽,导致数据点间的相似性过于泛化,进而模糊聚类边界,降低聚类效果。

如图 4 所示。随着 σ 的变化(其中 $\sigma \in [0,5]$,步长为 0.1),不同数据集的聚类性能不同。对于 Haberman 和 Iris 数据集,A_{ACC}、F_{FMI}、R_{RI}保持相对稳定。然而,对于 Aggregation、Breast 和 Wine 数据集,当 σ 较高时,聚类性能下降。突显选择合适的 σ 对优化聚类性能的重要性,尤其是在那些对 σ 变化较为敏感的数据集。因此,在实际应用中,合理调整 σ 是提高聚类效果的关键。



(a) Aggregation数据集 (b) Breast数据集 (c) Cancer数据集

(d) Haberman数据集 (e) Iris数据集 (f) Wine数据集

图 4 不同参数 σ 对聚类性能的影响

Fig.4 Impact of varying σ on clustering performance

4.4.2 m 的变化

m 在确定隶属度函数的模糊程度方面起着至关重要的作用。一方面,增大 m 降低聚类结果,导致聚类边界模糊;另一方面,较小的 m 使聚类边界更加明确,可能会提高聚类结果,但拟合程度不高。

当 $m \in [0 \sim 3]$,长为 0.1,不同数据集上的聚类性能不同。如图 5(a)所示,m 从 1.0 增加到 3.0,Aggregation 数据集的准确率呈现先上升后下降的趋势,F_{FMI}和 R_{RI}也表现出相似的波动特征。如图 5(i)所示,当 $m \geqslant 2.25$ 时,Wine 数据集的准确率逐渐减小且减小速率逐渐增加,F_{FMI}和 R_{RI}也同步减小;如图 5(c)所示,Cancer 数据集的准确率对 m 变化不敏感,各评估指标随 m 增大仅缓慢波动。实验结果表明,在实际应用中要根据具体决策需求和数据特性,选择适当的 m 而获得最优的聚类效果。敏感性分析凸显了根据数据集特征选择 σ 和 m 的重要性。对于具有重叠或噪声聚类的数据集,如 Wine 数据集和 Aggregation 数据集,较小的 m 和适中的 σ 具有好的聚类性能。在实际应用中,通过交叉验证算法或在参数值范围内确定网格搜索确定参数的最优值。未来的研究可以着重于开发自适应策略,在聚类过程中动态调整 σ 和 m,增强算法在多样化数据环境中的处理能力。

4.5 KFRCM 算法的理论意义和局限性

KFRCM 算法将知识度量与模糊粗糙集聚类过程相结合,通过知识度量在迭代过程中优化初始聚类中心选择。引入模糊粗糙度信息,KFRCM 算法显著提高对模糊边界和数据点重叠的识别准确率。实验结果表明,KFRCM 算法的迭代效率和聚类性能均优于其他 5 种传统聚类算法。KFRCM 算法不仅提高了聚类模型的准确率,减少收敛所需的迭代次数,显著提高算法的整体效率。此外,引入知识度量的概念,KFRCM 算法能够根据数据集中的数据点的权重客观地选择初始聚类中心,缩短迭代时间并提高聚类效率,使聚类过程更加高效和准确。KFRCM 算法适用于处理具有模糊边界和不确定性的复杂数据集,将模糊集理论与粗糙

集理论相结合,KFRCM 算法能够有效地处理数据集中的不确定性和模糊性,提高复杂数据的处理能力。

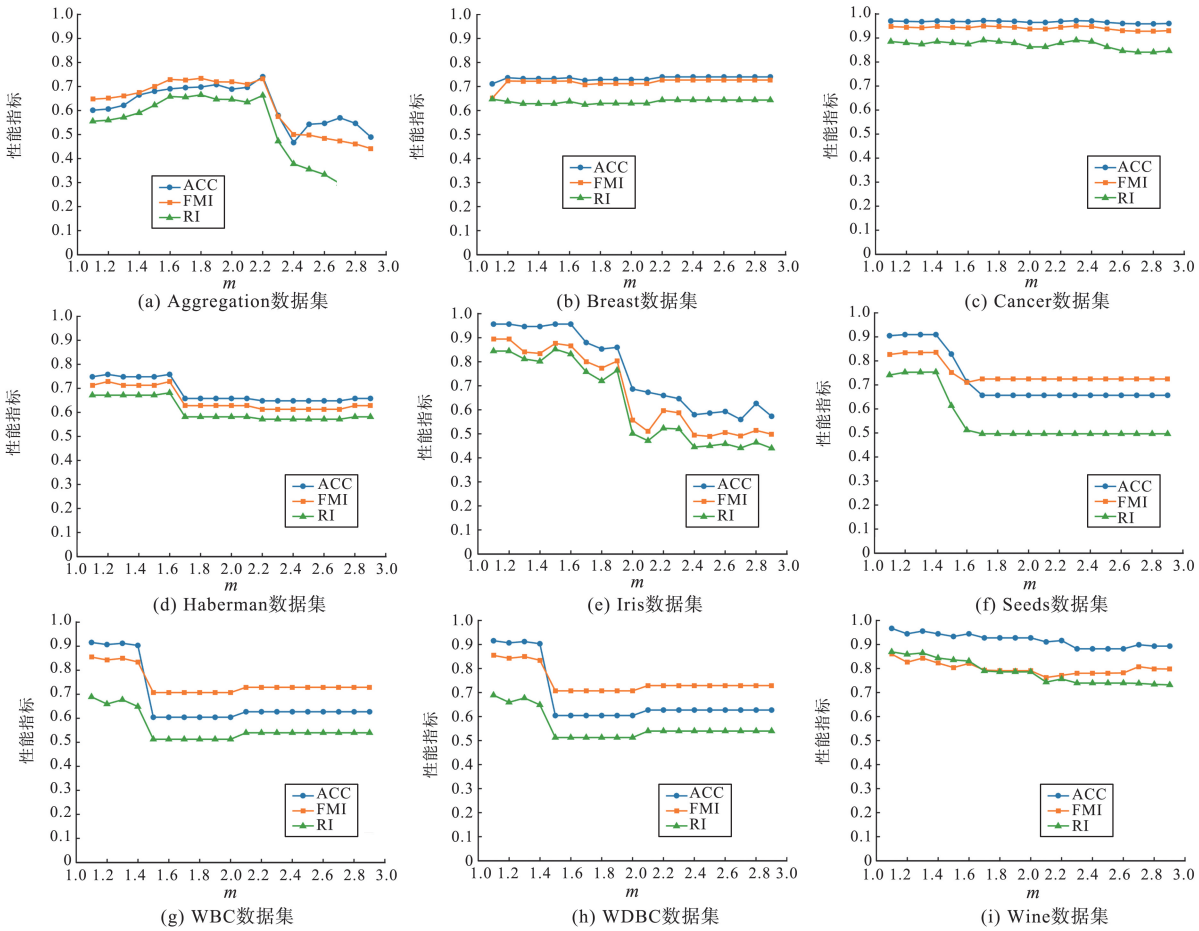


图 5　不同参数 m 对集群性能的影响

Fig.5　Impact of varying m on clustering performance

在引入噪声的 2d-4c-no4 数据集上,KFRCM 算法的标准差(s=0.002 8)仍显著低于传统 FCM 算法(s=0.034 1)和 k-means 算法(s=0.049 7)。参数敏感性分析表明,当 m 稳定且 σ 接近数据平均距离时,KFRCM 算法能够同时保持较高的聚类准确率与稳定性。说明模糊粗糙理论框架能提高算法稳定性。

表 5 中针对合成数据集的不平衡、密度不均匀、异形、不对称、中高维度和高维度特征,采用 2d-4c、2d-4c-no4、Flame、Square、HyperBlob-135 和 HyperBlob-300 数据集,用于评估 KFRCM 聚类算法的性能。该分析揭示了算法的一些局限性。由图 6(a)、(b)可知,2d-4c、2d-4c-no4 数据集为大规模数据集,且存在不对称性。KFRCM 算法在处理大规模、不平衡数据时,计算复杂性较高,随着数据量的增加,运行时间呈指数级增长,限制在大数据集中的应用和可扩展性。尽管 KFRCM 算法在处理密集和稀疏聚类时存在一定偏差,但图 6(c)中的月牙形聚类显示出 KFRCM 算法在特定数据模式下的有效性。KFRCM 算法能够成功区分不同密度和不均匀的聚类形状,但在仍存在识别不准确的情况。

表 5　具有不同数据特征的合成数据集

Table 5　Synthetic datasets with different data features

数据集	特征	数量	维度	分类
2d-4c	不平衡	714	3	3
2d-4c-no4	密度不均匀	862	6	4
Flame	异形	240	2	2
Square	不对称	800	2	4
HyperBlob-135	中高维度	1 500	135	15
HyperBlob-300	高维度	1 500	300	30

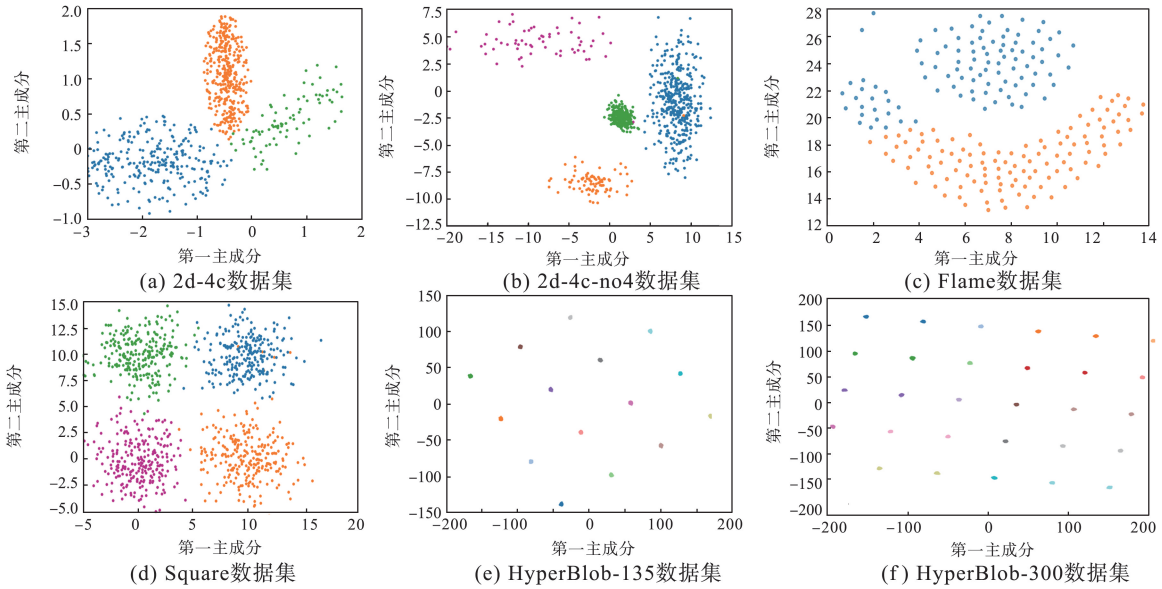


图 6 合成数据集的聚类图像

Fig.6 Cluster images of the synthetic dataset

图 6(a)中 2d-4c 的数据集的样本量较大,数据分布不平衡,KFRCM 算法聚类效果好。随着数据量增加,算法的运行时间将呈指数级增长,使得 KFRCM 算法处理大型数据集时的实用性受限。图 6(b)、(c)和(d)分别展示 2d-4c-no4、Flame 和 Square 数据集的聚类效果。结果表明,KFRCM 算法能够有效区分不同聚类,由于基于中心的聚类机制依赖中心点数据分配,在处理非球形分布的 Flame 数据集时出现部分数据分配错误的情况。此外,当采用图 6(e)、(f)的 HyperBlob-135 和 HyperBlob-300 数据集时,能验证 KFRCM 算法在中高维数据上的有效性。采用高斯核函数度量样本相似性,在一定程度上缓解维度问题,关注数据局部关系能提升高维空间中的计算效果。

总的来说,KFRCM 算法虽然在许多聚类任务中表现优异,但仍存在一些局限性。时间复杂度较高,尤其在处理大规模数据集时效率较低。在面对形状复杂或密度不均的数据集时,聚类效果较差。未来的研究可以着重于优化时间复杂度,引入更加灵活的聚类中心选取方式,并开发自适应机制动态调整参数,以提高算法的效率、适应性和稳定性。

5 结语

将知识度量与模糊粗糙集理论结合提出 KFRCM 算法,得到改进的直觉模糊知识度量公式,通过知识度量指标化数据特征的分布差异性,为各维度分配自适应权重,提高初始聚类中心选择的准确性。设计一种新型高斯核模糊粗糙隶属度函数算子,首次将高斯核相似度函数引入粗糙集近似空间,建立基于全局数据结构的隶属关系计算模型,有效解决模糊边界样本的聚类难题。实现知识加权机制、模糊粗糙集理论与传统 FCM 算法的三重融合,通过知识加权保障特征选择合理性,利用模糊粗糙集处理不确定性数据,在保持 FCM 算法计算效率的同时提高对复杂数据的处理能力。实验结果表明,KFRCM 算法的指标均优于 6 种主流算法,如在 Wine 数据集上准确率达到 97%,验证了 KFRCM 算法的优越性,为创新更可靠和适应性更强的聚类技术提供了新的理论框架和实践路径。KFRCM 算法的迭代效率和收敛速度也较高。首次系统地将知识度量与模糊粗糙聚类相结合,不仅拓展软聚类理论的应用范围,也为处理高维不确定数据提供有效工具,具有重要的理论意义和广泛的应用前景。

参考文献:

[1] 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.

[2] D′ANDRADE R G. U-statistic hierarchical clustering[J]. Psychometrika, 1978, 43(1):59-67.

[3] BEZDEK J C. Patternrecognition with fuzzy objective function algorithms[J]. Advanced Applications in Pattern Recognition, 1981, 22(1171):203-239.

[4] MIRKIN B. Mathematical classification and clustering[M]. New York: Springer, 2013:3-7.

[5] HORNG Y J, CHEN S M, CHANG Y C, et al. A new method for fuzzy information retrieval based on fuzzy hierarchical clustering and fuzzy inference techniques[J]. IEEE Transactions on Fuzzy Systems, 2005, 13(2):216-228

[6] QUINTANA F J, GETZ G, HED G, et al. Cluster analysis of human autoantibody reactivities in health and in type 1 diabetes mellitus: a bioinformatic approach to immune complexity[J]. Autoimmunity, 2003, 21(1):65-75.

[7] GENTHER H, GLESNER M. Advanced data preprocessing using fuzzy clustering techniques[J]. Fuzzy Sets and Systems, 1997, 85(2):155-164.

[8] QIAO Xiaoguang, CHEN Caikou, WANG Weiye. Efficient subspace clustering and feature extraction vial ℓ -norm and ℓ -norm minimization[J]. Neurocomputing, 2024, 595:127813.

[9]CHOI S, YOON S. Change-point model-based clustering for urban building energy analysis[J]. Renewable and Sustainable Energy Reviews, 2024, 199:114514.

[10] SCHAFFER M, VERA-VALDÉS J E, MARSZAL-POMIANOWSKA A, et al. Exploring smart heat meter data: a co-clustering driven approach to analyse the energy use of single-family houses[J]. Applied Energy, 2024, 371:123586.

[11] GÓMEZ-FLORES W, HERNÁNDEZ-LÓPEZ J. Automatic adjustment of the pulse-coupled neural network hyperparameters based on differential evolution and cluster validity index for image segmentation[J]. Applied Soft Computing, 2020, 97: 105547.

[12] LIAO Jiyong, WU Xingjiao, WU Yaxin, et al. K-NNDP: k-means algorithm based on nearest neighbor density peak optimization and outlier removal[J]. Knowledge-Based Systems, 2024, 294:111742.

[13] HEIDARI J, DANESHPOUR N, ZANGENEH A. A novel k-means and k-medoids algorithms for clustering non-spherical-shape clusters non-sensitive to outliers[J]. Pattern Recognition, 2024, 155:110639.

[14] CAO Xinyu, YU Min, ZHANG Shuming, et al. Hierarchical clustering evolutionary tree-support for SLA[J]. Journal of Manufacturing Processes, 2024, 125:189-201.

[15] XU Zeshui, WU Junjie. Intuitionistic fuzzy c-means clustering algorithms[J]. Journal of Systems Engineering and Electronics, 2010, 21(4):580-590.

[16] SHANG Taotao, HUANG Qianwen, WANG Yongyi. Vibration reduction and energy harvesting on the ship thrust bearing unit excited by a measured shaft longitudinal vibration using NES-GMM[J]. Ocean Engineering, 2024, 294:116914.

[17] LIN Rongfu, GUO Weizhong, CHENG Shing Shin. Type synthesis of novel 1R, 2R, 1R1T, and 2R1T hybrid RCM mechanisms based on topological arrangement and modular design method[J]. Mechanism and Machine Theory, 2024, 200: 105692.

[18] SZMIDT E, KACPRZYK J. Entropy for intuitionistic fuzzy sets[J]. Fuzzy Sets and Systems, 2001, 118(3):467-477.

[19] GUO Kaihong, ZANG Jie. Knowledge measure for interval-valued intuitionistic fuzzy sets and its application to decision making under uncertainty[J]. Soft Computing, 2019, 23:6967-6978.

[20] GUO Kaihong, XU Hao. Knowledge measure for intuitionistic fuzzy sets with attitude towards non-specificity[J]. International Journal of Machine Learning and Cybernetics, 2019, 10:1657-1669.

[21] PATEL A, JANA S, MAHANTA J. Construction of similarity measure for intuitionistic fuzzy sets and its application in face recognition and software quality evaluation[J]. Expert Systems with Applications, 2024, 237:121491.

[22] GUO Kaihong, XU Hao. A unified framework for knowledge measure with application: from fuzzy sets through interval-valued intuitionistic fuzzy sets[J]. Applied Soft Computing, 2021, 109:107539.

[23] ZHANG Xiaoyan, WANG Jinghong, HOU Jianglong. Matrix-based approximation dynamic update approach to multi-granulation neighborhood rough sets for intuitionistic fuzzy ordered datasets[J]. Applied Soft Computing, 2024, 163: 111915.

[24] 陈宝国,邓明. 基于对象更新的邻域多粒度粗糙集模型增量式算法[J]. 智能系统学报,2023,18(3):562-576.
CHEN Baoguo, DENG Ming. Incremental algorithm for neighborhood multi-granular rough set model based on object update [J]. Journal of Intelligent Systems, 2023, 18(3):562-576.

[25] 莫子孟,尹立平. 基于模糊粗糙集的大型汽轮机组设备故障识别算法[J]. 能源科技,2024,22(3):44-48.
MO Zimeng, YIN Liping. Fault diagnosis method for large turbine units based on fuzzy rough set[J]. Energy Technology, 2024, 22(3):44-48.

[26] 刘以,张小峰,孙玉娟,等. 基于加权滤波与核度量的鲁棒图像分割算法[J]. 激光与光电子学进展,2024,61(8):380-392.

LIU Yi, ZHANG Xiaofeng, SUN Yujuan, et al. Robust image segmentation algorithm based on weighted filtering and kernel metrics[J]. Progress in Laser and Optoelectronics, 2024, 61(8):380-392.

[27] 刘子源,马占有,李霞,等.基于模糊测度的最大可能性互模拟等价研究[J/OL]. 郑州大学学报(理学版). (2025-01-13)[2025-11-16]. https://doi.org/10.13705/j.issn.1671-6841.2024144.7.
LIU Ziyuan, MA Zhanyou, LI Xia, et al. Study on maximum likelihood mutual simulation equivalence based on fuzzy measures[J/OL]. Journal of Zhengzhou University (Natural Science Edition). (2025-01-13) [2025-11-16]. https://doi.org/10.13705/j.issn.1671-6841.2024144.7.

[28] 李繁,张晓宇,刘林东. 基于广义粒度自编码器的模糊粗糙聚类算法[J]. 计算机应用与软件,2024,41(3):266-275.
LI Fan, ZHANG Xiaoyu, LIU Lindong. Fuzzy rough clustering method based on generalized granular self-encoder[J]. Computer Applications and Software, 2024, 41(3):266-275.

[29] 朱世超,王骋程,王超,等. 基于支持向量聚类和模糊粗糙集的交通流数据修复算法[J]. 森林工程,2023,39(1):157-165.
ZHU Sichao, WANG Chengcheng, WANG Chao, et al. Traffic flow data repair method based on support vector clustering and fuzzy rough sets[J]. Forest Engineering, 2023, 39(1):157-165.

[30] 任浩伟,王青海,张巧珍. 区间值直觉模糊β覆盖粗糙集模型[J]. 陕西科技大学学报,2024,42(5):214-224.
REN Haowei, WANG Qinghai, ZHANG Qiaozhen. Interval-valued intuitionistic fuzzy β-covering rough set model[J]. Journal of Shaanxi University of Science and Technology, 2024, 42(5):214-224.

[31] 商钰玲,李鹏,朱枫,等. 基于模糊逻辑的物联网流量攻击检测技术综述[J]. 计算机科学,2024,51(3):3-13.
SHANG Yuling, LI Peng, ZHU Feng, et al.Overview of IoT traffic attack detection technology based on fuzzy logic[J]. Computer Science, 2024, 51(3):3-13.

[32] ATANASSOV K T. Intuitionistic fuzzy sets[J]. Fuzzy Sets and Systems, 1986, 20(1):87-96.

[33] GUO Kaihong, XU Hao. Preference and attitude in parameterized knowledge measure for decision making under uncertainty[J]. Applied Intelligence, 2021, 51(10):7484-7493.

[34] MÜLLER K R, MIKA S, RATSCH G, et al. An introduction to kernel-based learning algorithms[J]. IEEE Transactions on Neural Networks, 2001, 12(2):181-201.

[35] WAN Jihong, CHEN Hongmei, LI Tianrui, et al. Dynamic interaction feature selection based on fuzzy rough set[J]. Information Sciences, 2021, 581:891-911.

[36] MERCER J. Functions of positive and negative type, and their connection with the theory of integral equations[J]. Philosophical Transactions of the Royal Society of London, 1909, 209:415-446.

[37] YAGER R R. Some aspects of intuitionistic fuzzy sets[J]. Fuzzy Optimization and Decision Making, 2009, 8(1):67-90.

[38] YANG Xiaowei, ZHANG Guangquan, LU Jie, et al. A kernel fuzzyc-means clustering-based fuzzy support vector machine algorithm for classification problems with outliers or noises[J]. IEEE Transactions on Fuzzy Systems, 2010, 19(1):105-115.

[39] KASTRITSI T, DOULGERI Z. A controller to impose a RCM for hands-on robotic-assisted minimally invasive surgery[J]. IEEE Transactions on Medical Robotics and Bionics, 2021, 3(2):392-401.

[40] PU Yue, YAO Wenbin, LI Xiaoyong. EM-IFCM: fuzzy *c*-means clustering algorithm based on edge modification for imbalanced data[J]. Information Sciences, 2024, 659:120029.

[41] RAN Xingcheng, XI Yue, LU Yonggang, et al. Comprehensive survey on hierarchical clustering algorithms and the recent developments[J]. Artificial Intelligence Review, 2023, 56:8219-8264.

[42] YU Hong, WANG Xincheng, WANG Guoyin, et al. An active three-way clustering method via low-rank matrices for multi-view data[J]. Information Sciences, 2020, 507:823-839.

[43] LIU Rui, WANG Hong, YU Xiaomei. Shared-nearest-neighbor-based clustering by fast search and find of density peaks[J]. Information Sciences, 2018, 450:200-226.

[44] WANG Wei, XIA Feng, NIE Hansong, et al. Vehicle trajectory clustering based on dynamic representation learning of internet of vehicles[J]. IEEE Transactions on Intelligent Transportation Systems, 2020, 22(6):3567-3576.

(编辑:陈丽萍)