基于局部模糊粒球的高效模糊聚类方法研究

Jiang Xie、Qiao Deng、Shuyin Xia*、Yangzhou Zhao、Guoyin Wang、Xinbo Gao

arXiv:2303.03590v2 [cs.LG] 8 Mar 2023

摘要——**近年来，模糊聚类问题受到广泛关注。现有方法大多从全局角度进行隶属度迭代，在噪声环境中存在诸多问题；对于样本量差异较大的簇，迭代计算也不够准确和高效。本文从大规模优先策略出发，利用粒球对数据进行模糊迭代，且计算数据的隶属度时仅考虑其所在的两个粒球，从而提高迭代效率。所形成的模糊粒球集能够针对不同数据场景采用更多处理方法，从而增强模糊聚类计算的实用性。**

索引词——**聚类、模糊聚类、粒球、模糊C均值**

一、引言

近年来，随着互联网的快速发展，大数据时代已经到来。数据分析在理解各种现象方面发挥着不可或缺的作用 [1]。因此，机器学习研究取得了长足进展。聚类是根据组（簇）对模式（观测值、数据项或特征向量）进行的无监督分类。[2] 数据聚类是许多领域中的重要问题，包括数据挖掘、模式识别和生物信息学。 [3]解决这一问题的算法有很多，例如经典的模糊C均值聚类，常用于分割医学图像 [4]。

将一组物理或抽象对象划分为多个由相似对象组成的类别，这一过程称为聚类。一个簇是由与同一簇中的对象相似、而与其他簇中的对象不同的数据对象组成的集合。自然科学和社会科学中存在许多分类问题。聚类分析又称群分析，是一种研究分类问题（样本或指标）的统计分析方法。根据应用类型和所考虑的数据类型，各个领域已发展出许多聚类方法，并广泛应用于实际场景。 [5] 聚类可能是探索性数据分析和数据处理应用中的首要任务。聚类分析的主要任务是将数据集划分为不同的对象组。在一定程度上，同一组中的对象，也就是所谓的簇中的对象，比起其他组中的对象 [6]彼此更加相似。

聚类分析源自分类学，但聚类不等同于分类。聚类与分类的区别在于，聚类所需的类别是未知的。聚类分析内容十分丰富，包括系统聚类、有序样本聚类、动态聚类、模糊聚类、图论聚类、聚类预测等。聚类也是数据挖掘中的一个重要概念。聚类作为知识提取的工具，可用作预处理步骤，将数据划分为易于处理的部分。聚类是一种探索性数据分析形式，它将数据分组成若干组或子集，使每组中的对象具有一定的相似性。传统的聚类分析计算方法如下：1. 划分法；2. 层次法；3. 基于密度的方法；4. 基于网格的方法；5. 基于模型的方法。

根据不确定性的不同，聚类可分为硬聚类和软聚类两类。硬聚类将每个待识别对象严格划分到某个类别中，使其只能属于其中一个类别。模糊聚类通过样本建立对类别的不确定描述，能够更客观地反映客观世界，因此成为聚类分析的主流。模糊聚类算法是一种基于函数优化方法的聚类算法。它使用微积分计算技术来寻找最优代价函数。基于概率的聚类方法将使用概率密度函数。为解决这一问题，在假设存在适当模型的前提下，模糊聚类算法中的向量可以同时属于多个簇。近年来，K均值及其衍生方法一直是聚类研究的热点。硬C均值聚类（HCM）属于硬聚类，而模糊C均值聚类（FCM）则是K均值领域中的软聚类。 [7]

FCM [8] 已被证明比HCM具有更好的性能，并已成为聚类分析中最知名、最有力的方法。与硬聚类不同，FCM算法根据数据点属于某一类的程度来划分簇。 [9] 然而，这些FCM算法在含噪环境中存在诸多问题，在样本规模差异较大的情况下进行聚类迭代计算时准确性也不高。优秀的聚类算法应具有鲁棒性，能够容忍应用系统中经常出现的这些情况。 [10] 为了解决FCM中的上述问题，我们提出了一种将粒球与FCM相结合的方法来确定初始迭代中心，该方法不仅提高了迭代效率，也提升了准确性和聚类结果。因此，粒度模糊聚类的主要贡献如下：

1. 使用粒球进行局部隶属度迭代，不仅能够保证数据的模糊属性，还能大幅提高隶属度迭代的效率。

2. 所得到的模糊粒度集可针对不同的数据集场景采用多种方法完成最终聚类，

例如，使用连接方法可自适应地获得复杂形状的聚类结果；K均值可更快速地获得聚类结果。

3. 所提出方法的迭代过程完全自适应，无需输入任何参数。

II. 相关工作

A. 模糊集

模糊集由Lotfi Zadeh在1965年发表的一篇论文中提出。1969年，Ruspini发表了一篇开创性论文，该论文已成为大多数模糊聚类算法的基础。他的思想确立了模糊划分的基本结构，并描述和举例说明了首批实现模糊划分的算法。Bezdek于1973年提出了模糊C均值模型的一般情形。 [11]。Senapati等人提出了费马模糊集，将费马模糊集与勾股模糊集和直觉模糊集进行比较，并找到了费马模糊集的基本运算集。此外，还定义了用于对费马模糊集排序的得分函数和精确函数，并研究了两个费马模糊集之间的欧几里得距离。文献还提出采用费马模糊TOPSIS方法解决多准则决策问题 [12]。费马模糊集是模糊集在处理信息不确定性方面最有用的扩展之一 [13]。区间值模糊集、直觉模糊集、二型模糊集、N型模糊集、犹豫模糊集（ HFS）、对偶模糊集和中性粒细胞集等，都是模糊集的广义扩展，在解决决策问题方面颇有用处，且相比经典集成具有许多优势 [14]。

B. 模糊聚类

随着模糊集理论的形成、发展和深化，Ruspini 首次提出了模糊划分的概念。以此为出发点和基础，模糊聚类的理论与方法迅速发展并日益成熟。通常，模糊聚类是根据对象自身的属性构造模糊矩阵，并依据一定的隶属度确定聚类关系。通过模糊数学，定量确定样本之间的模糊关系，从而实现客观、准确的聚类 [15]。根据聚类过程的不同，模糊聚类分析大致可分为三类：基于模糊关系的模糊聚类 [16], 基于目标函数的模糊聚类算法，以及基于神经网络的模糊聚类算法。模糊聚类分析是一种为数据提供软划分的经典方法 [17]。当事物之间存在模糊边界时，模糊聚类分析是一种按照一定要求对事物进行分类的数学方法。在模糊聚类中，所有对象都隶属于所有簇，其隶属度与对象和相应簇表示之间的距离成反比 [18]。模糊聚类方法可以发现模糊划分，使观测值能够以软划分方式归属于多个簇。 [19]聚类分析是数理统计中的一种多元分析方法。它运用数学方法定量确定样本之间的关系，从而客观地划分类型。事物之间的边界有些是明确的，有些则是模糊的。样本群体中面部相似程度的边界是模糊的，阴晴天气之间的边界也同样模糊。当聚类涉及事物之间的模糊边界时，会采用模糊聚类分析方法。

C. 模糊*C*均值（*FCM*）方法

模糊C均值（FCM）是首个将模糊集理论引入簇群、从而量化簇隶属度不确定性的方法 [20]。由于具有很强的模糊属性，模糊C均值（FCM）算法是模糊聚类中最经典的方法。FCM 是一种基于中心点模糊距离的聚类方法。在 FCM 中，一个数据点可以按一定的概率归属于多个簇。 [21]。FCM 算法根据最小化的目标函数输出隶属度矩阵，并将每个数据点隶属度最大的类别作为该数据点的分类结果。在 FCM 中，数据点的隶属度通过迭代更新，簇数量预先确定，且每个簇都包含一个具有相应隶属度的数据点。 [22] 为最大化簇间差异，可以将分数阶熵作为正则化函数加入 FCM 算法的目标函数中 [23]。模糊C均值算法是对普通C均值算法的改进。普通C均值算法对数据划分采取刚性处理，而 FCM 则采用灵活的模糊划分。FCM 算法简单，因此在大多数情况下具有很强的鲁棒性，并且在聚类算法中具有强大的生命力。 [24] 由于 FCM 以模糊理论为基础，因此主要通过模糊划分矩阵进行描述，突破了硬聚类中每个数据点只能被划分到一个类别的限制。最终输出的模糊矩阵反映了数据集的所有信息，因此能够准确考察数据集的分布和整体特征 [25]。尽管传统 FCM 算法存在一些不足，例如簇中心初始化、最优簇数量的确定以及对噪声敏感；此外，FCM 的实现依赖于模糊指数的取值，使用不同指数可能会得到不同的结果[26]。研究最多的是基于目标函数的模糊聚类算法，它将聚类问题描述为一个约束优化问题，并通过求解该优化问题来确定数据集的模糊划分和聚类结果。

尽管传统的FCM算法存在一些缺点，例如簇中心初始化、确定最优簇数量等问题，并且容易受到噪声和离群值的影响，这是因为该算法以相同的方式为所有点分配权重，无法从实际数据点中有效学习如何区分噪声和离群值 [27] ，但它仍是应用最广泛的聚类算法。以往研究提出了许多适用于FCM算法的有效性函数，然而，这些传统的模糊集有效性

指标存在一些明显的不足，例如，当簇数量变得很大并接近数据点数量时，会呈现单调递减趋势；模糊加权指标可能导致数值不稳定性、簇数量不足等问题 [28]。

隶属函数表示对象 x 属于集合 A 的程度，通常记为 $\mu_{A(x)}$。自变量的取值范围是所有可能属于集合 A 的对象，函数值的取值范围为 [0,1]，即 $0 \leq \mu_{A(x)} \leq 1$。$\mu_{A(x)} = 1$表示 x 完全属于集合 A，这相当于在传统集合概念中 $x \in A$。K均值聚类也称为 C均值聚类。其核心思想是，算法将 n 个向量 X_j （i =1,2,..., n）划分为 c 个类别 V_i （i = 1,2,..., c），并求出每组的簇中心，使非相似性（或距离）指标的值函数（或目标函数）最小化。目标函数可定义为:

$$J_m(U,V) = \sum_{i=1}^{c}\sum_{j=1}^{n} \|X_j - V_i\|_A^2 \tag{1}$$

算法实现的主要步骤：步骤 1：随机确定 K 个初始点作为质心。步骤 2：为数据集中的每个数据点找到最近的簇。步骤 3：对每个簇，计算簇内所有点的均值，并将该均值作为质心。步骤 4：重复步骤 2，直到任意数据点的簇分配结果都不再变化。

FCM算法的目标函数为

$$J_m(U,V) = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^m \|X_j - V_i\|_A^2 \tag{2}$$

其中，U = $[u_{ij}]$ 是隶属度矩阵， u_{ij} 是第j个样本对第i类的隶属度，M是一个模糊常数。FCM算法的约束条件：对于任意样本，其对各个簇的隶属度之和为1。

$$\sum_{i=1}^{c} u_{ij} = 1 \tag{3}$$

求解U、V：为了在约束条件下求目标函数的极值，我们使用拉格朗日乘子法构造一个新函数。

$$F = \sum_{i=1}^{c} u_{ij}^m \|X_j - V_i\|^2 + \lambda\left(\sum_{i=1}^{c} u_{ij} - 1\right) \tag{4}$$

在 λ 成为拉格朗日乘子，并在以下条件下求出函数 F 的最优值:

$$\frac{\partial F}{\partial \lambda} = \left(\sum_{i=1}^{c} u_{ij} - 1\right) = 0 \tag{5}$$

$$\frac{\partial F}{\partial u_{ij}} = \left[m\left(u_{ij}\right)^{m-1}\|X_j - V_i\|^2 - \lambda\right] = 0 \tag{6}$$

$$\frac{\partial F}{\partial v_i} = \sum_{j=1}^{n}\left(u_{ij}\right)^m x_j - v_i\sum_{i=1}^{c} u_{ij}^m = 0 \tag{7}$$

由极值条件求解:

$$u_{ij} = \left[\sum_{k=1}^{c}\left(\frac{\|X_j - V_i\|^2}{\|X_j - V_k\|^2}\right)^{\frac{2}{m-1}}\right] \tag{8}$$

$$v_i = \frac{\sum_{j=1}^{n} u_{ij}^m X_j}{\sum_{j=1}^{n}\left(u_{ij}^m\right)} \tag{9}$$

D. Granular Computing

Chen [29] 指出，人类认知具有全局优先的特征，这一点在1982年得到了体现。它不同于现有的主流人工智能算法，后者以最细粒度的点作为输入。Wang [30]首次将大规模认知规则引入粒球计算，并提出了多粒度认知计算。Xia和Wang [31]-[33] 进一步使用不同大小的超球体来表示“粒”，并提出了粒球计算，其中，大粒球表示粗粒度，而小粒球表示细粒度。

给定数据集 $D = x_i(i = 1, 2, \ldots\ldots, n)$，其中 n 是 D 上样本的数量。使用粒球 GB_1、GB_2、……、GB_m 覆盖并表示数据集 D。假设 j^{th} 粒球 GB_i 中的样本数量表示为 $|GB_j|$，则其覆盖度可表示为$\sum_{i=1}^{m}\left(|GB_j|\right)/n$。粒球覆盖的基本模型可表示为

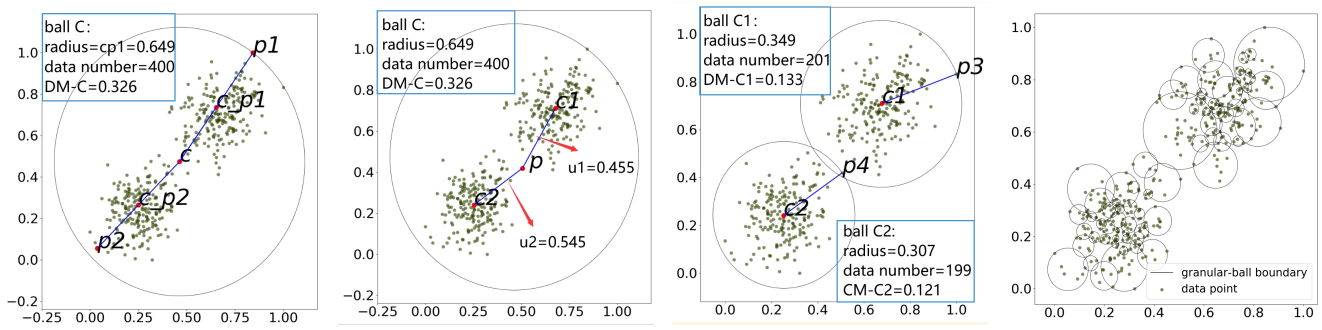
$$\begin{aligned} &\min \lambda_1 * n/\textstyle\sum_{j=1}^{m}\left(|GB_j|\right)/n + \lambda_2 * m, \\ &\text{s.t. quality } (GB_j) \geq T, \end{aligned} \tag{10}$$

其中 λ_1 和 λ_2 是相应的权重系数，m 为粒球的数量。在其他因素保持不变时，覆盖度越高，样本信息损失越少；粒球数量越多，表征就越准确。因此，生成粒球时，应考虑以最少的粒球数量获得最大的覆盖度。通过调整参数 Ai 和 λ_2 ，可以获得最优的粒球生成结果，使整个方程的值最小化。在大多数情况下，目标函数中的两项互不影响，无需权衡，因此λ_1 和 λ_2 默认设为 1。粒球计算可以拟合任意分布的数据 [32],[34]。

III. 所提出的方法

A. 研究动机

现实世界的数据中存在大量模糊边界，硬划分会导致聚类结果具有不确定性，也不合理。如何对具有模糊边界的数据进行软处理，是一个具有重要研究价值的问题。作为模糊聚类中最经典的FCM算法，许多学者已对其改进进行了大量研究。然而，这些FCM算法在噪声环境中存在较大问题，对于样本数量差异较大的簇，其迭代计算结果也不够准确。



(a) (b) (c) (d)

图 1. 模糊粒球的生成。a. p1 是距离中心 c 最远的点，p2 是距离 p1 最远的点；这里我们选择从 c 到 p1、p2 的中点 cp1、cp2 作为初始簇中心 $V=[c\ p1, c\ p2]$。b. 根据初始 V 计算隶属度 U，然后根据条件迭代更新 V 和 U，最终得到最终簇中心 $V=[c1, c2]$；如果 u2>u1，则 p 属于簇 c2。c. 如果 DM-C1<DM-C 且 DM-C2<DM-C，则根据 b 中的 U 和 V 将 C1 拆分为 C1 和 C2。d. 循环执行 a 至 c 的过程，直到生成的模糊粒球数量保持不变，从而得到未归一化模糊粒球。.

现有方法大多从全局视角考虑隶属度迭代。本文从大规模优先策略出发，利用粒球对数据进行模糊迭代。数据的隶属度仅考虑两个粒球，从而提高了迭代效率。所形成的模糊粒子集可针对不同的数据场景采用更多处理方法，增强了模糊聚类算法的实用性。

B. 模糊粒球的生成过程

粒计算模型将粒球作为“粒子”，取代点输入，并采用一种新的粒分类器框架。通过将数据集划分为粒，可以在基本保持数据原始分布的同时，获得清晰的决策边界。粒计算高效、鲁棒且可扩展。在现有的粒计算过程中，整个数据集被视为最粗粒度，然后利用粒球的特征将其从粗粒度细化到细粒度，从而实现可扩展且鲁棒的计算过程。基于此，我们给出以下定义：

定义 1（模糊粒球 FGB，模糊粒球集合 FGB_s）。基于上述粒计算模型，结合现有的经典模糊聚类方法，本文从传统的数据对象隶属度迭代出发，对其进行优化和改进，使用粒球对数据进行模糊迭代；数据的隶属度仅考虑其所在的两个粒球，如此循环，最终通过迭代生成一系列模糊粒球集合。我们的方法基于一个基本假设：分布相似的数据形成一个模糊粒球 FGB，如图 1a 所示；相邻的模糊粒球 FGB_s 形成一个簇。

定义 2（中心 c 和半径 r）。给定数据集 $D \in R_d$，对于每个 FGB，c_i 是 FGB 中所有数据点的重心，r_i 是所有点 p_i 到 c 的最大距离。中心 c 和半径 r 定义如下：

$$c_i = \frac{1}{n}\sum_{i=1}^{n} p_i \tag{11}$$

$$r_i = \max\left(\|p_i - c_i\|\right) \tag{12}$$

此处 $\|.\|$ 表示二范数，n 是 FGB 中数据点的数量。

定义 3（分布式度量 DM）。DM_i 通过计算 FGB 中数据点 n_i 与半径 s_i 的比值来度量。其定义如下：

$$s_i = \sum_{i=1}^{n_i} \|p_i - ci\| \tag{13}$$

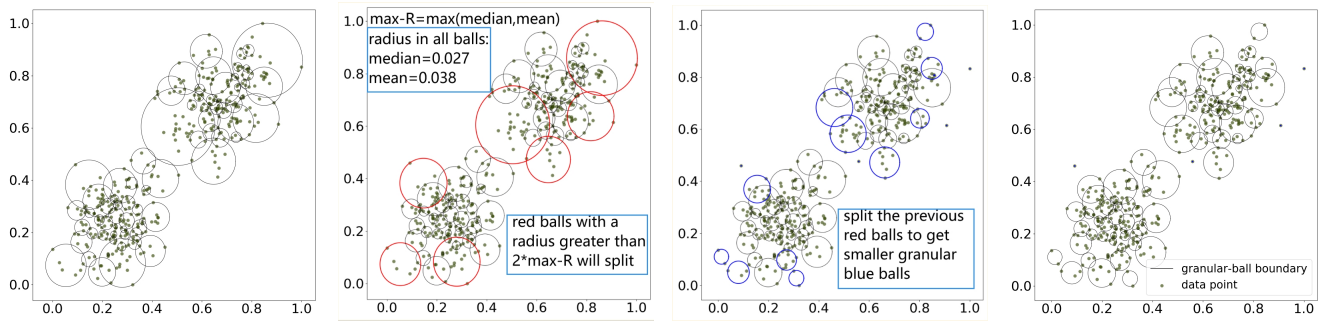
$$DM_i = \frac{s_i}{n_i} \tag{14}$$

定义 4（初始聚类中心 V）。如图 1a 所示：根据给定模糊粒球的中心 c，找到距离 c 最远的数据点 p1，然后找到距离 p1 最远的数据点 p2。在此基础上，选取 c 与 p1 的中点 c_p1，以及 c 与 p2 的中点 c_p2 作为初始聚类中心，即：$V=[c\ p1, c\ p2]$。

$$c_p_1 = \frac{1}{2}(c + p1) \tag{15}$$

$$c_p_1 = \frac{1}{2}(c + p2) \tag{16}$$

定义 5（加权 DM 值 DM_{weight}）。模糊粒球根据 DM 进行划分。如图 1a 所示：我们将整个数据集视为一个模糊粒球 C；确定初始簇中心 $V=[c\ p1, c\ p2]$ 后，使用 FCM算法迭代计算隶属度矩阵 U，并更新图 1b 中的簇中心 $V=[c\ p1, c\ p2]$。随后，将模糊粒球分成两个子模糊粒球 C1 和 C2；在整个过程中，我们计算 C、C1 和 C2 的 DM-C、DM-C1 和 DM-C2 值；最后，通过比较 DM-C、DM-C1 和 DM-C2，确定是否完成图 1c 中的粒球划分。最终，我们得到图 1d 所示的一系列未归一化模糊粒球。在 [x], 中，DM-C1 和 DM-C2 必须小于 DM-C，才能将模糊粒球分开。然而，当噪声很多时，



(a) 未归一化模糊粒球 (b) 归一化前 (c) 归一化后 (d) 最终模糊粒球集

图 2. 模糊粒球的归一化过程。

这一划分规则会导致大量模糊粒球无法继续划分。因此，本文采用加权 DM 值进行比较，以更好地适应噪声情形。为使公式更加简洁明了，我们将 DM-C 记为 DM_C，将 DM-C1 记为 DM_{C1}，将 DM-C2 记为DM_{C2}。因此，DM_{weight} 定义如下：

$$DM_{\text{weight}} = \frac{n_{C1}}{n_C} DM_{C1} + \frac{n_{C2}}{n_C} DM_{C2}. \tag{17}$$

这里，n_C、n_{C1} 和 n_{C2} 分别表示相应模糊粒球中的数据数量。如果 DM_{weight}小于 DM_C，模糊粒球 C 将分裂。图 2a 展示了最终的分裂结果。然而，在图 2a 中，一些半径过大的模糊粒球仍可能受到边界点或噪声点的影响，因此需要进行分割；如果 $r_k \geq 2 \times max(mean(r)$ $median(r))$，则需要分裂 FGB_K。Mean (r) 和 median (r) 分别表示所有模糊粒球半径的均值和中位数。移除半径过大的模糊粒球后，分裂过程即告完成，结果如图 2b 所示。根据以上描述，设计了一种模糊粒球生成算法，如算法 1 所示。

算法 1模糊粒球的生成

输入：D：数据集 输出：FGB_s $sets$：模糊粒球集 初始化：$n = 2, m = 2, iter = 0, max\ count = 100$1：当 true 时执行 2：初始化 $FGB_s = \emptyset$3：根据公式 14、公式 15 初始化 $V = [c\ p1, c\ p2]$；4：根据公式 8 计算 U；5：当 $iter < max\ count$ 时执行 6：根据公式 9 计算 V；7：根据公式 8 更新 U；8：iter=iter+1；9：结束循环 10：根据 U 将 D 预划分为 FGB_k；11：根据公式 16 计算 D、DM_{weight} 的 DM；12：如果 DM $\leq DM_{weight}$，则 13：保留 FGB_k 并将其添加到集合 FGB_s；14：结束 if 15：如果 FGB_k 的数量没有变化，则 16：跳出循环；17：结束 if 18：取 C_k 中的数据 D_k 作为下一次输入，$D = D_k$ 19：结束循环 20：对于每个 $FGB_k \in FGB_s$，执行 21：计算 mean(r)、median(r)，22：如果 $r_k \geq 2 \times max(mean(r)\ median(r))$，则 23：拆分 FGB_k；24：结束 if 25：如果 FGB_k 的数量没有变化，则 26：跳出循环；27：结束 if 28：结束 for 29：返回 FGB_s

C. 连接模糊粒球以形成聚类结果

定义 6（通过 K-均值 FGB K-means 连接模糊粒球集）。根据算法 1，我们获得了最终的模糊粒球集。现在，我们需要连接这些模糊粒球，以形成最终的聚类结果。因此，我们提供了两种连接方式：一种是将这些模糊粒球作为 K均值算法的输入，以完成聚类，该方法定义为 FGB K-means。因此，该算法的设计如算法 2 所示。

定义 7（通过重叠关系连接模糊粒球集，即 FGB 重叠）。另一种是我们提出的方法，定义为 FGB 重叠。它基于一种连通关系，该关系由相邻规则构成。相邻规则描述如下：

$$\|c_i - c_j\| - (r_i + r_j) < \tau_{ij} \tag{18}$$

$$\tau_{ij} = \frac{\min(r_i, r_j)}{1 + \min(o_i, o_j)} \tag{19}$$

其中，o_i 表示 FHB_i 与其相邻模糊粒球之间的重叠次数累积值，o_j 表示 FHB_j 与其相邻模糊粒球之间的重叠次数累积值；τ_{ij} 表示调整系数。该公式表明，如果 FHB_i 与 FHB_j 之间的间隙小于τ_{ij}，则它们满足相邻规则。τ_{ij} 可以动态调整满足相邻规则的判据。如图3所示，为防止对粒

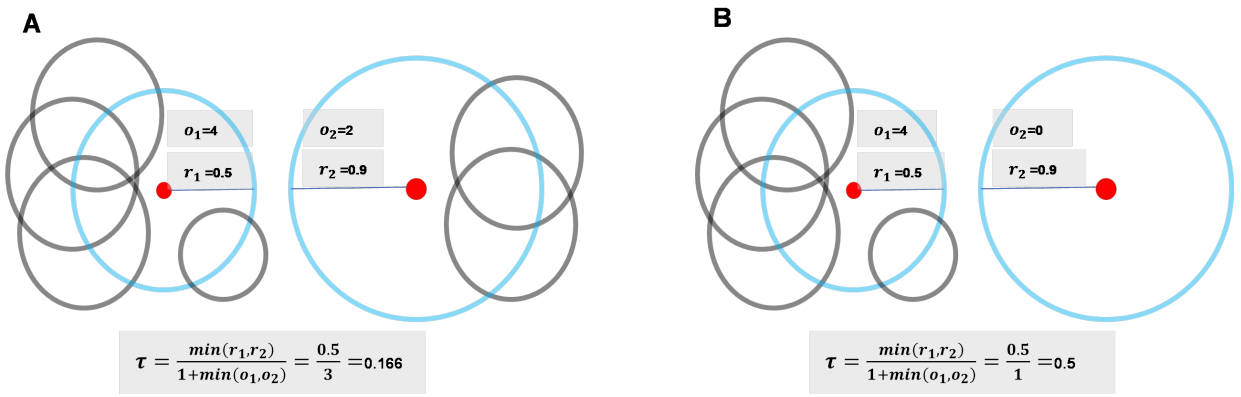


图3. o_1 和 o_2 是邻域模糊粒球重叠次数的累积值，作为?的调整系数，可动态调整满足相邻规则的判据。为防止错误地区分模糊粒球，累积值越大，满足相邻规则的标准越严格。A中的 τ 小于B中的 τ 。

算法 2FGB K-means

输入： FGB_s $sets$：模糊粒球集，簇的数量

簇：K

输出：聚类结果

1：初始化$center\ list = \emptyset$;

2: **对**每个 $FGB_i \in D$ **执行**

3: 计算 FGB 的中心 $_i$,

4: 将 FGB_i 的中心添加到中心列表

5: **结束循环**

6: 将中心列表视为数据集 D，并将数据集 D 和 K 作为 K均值的输入，以将 FGB_s 连接到最终结果聚类结果

7：返回聚类结果

算法3FGB 重叠 _

输入： FGB_s $sets$：模糊粒球集

输出： 聚类结果

1: **对于** 每个 $FGB_i \in FGB_s$ **执行**

2: **对于**每个 $FGB_j (i \neq j) \in FGB_s$ **执行**

3: 计算 o_i 的 FGB_i 与的重叠次数 FGB_j

4: **结束循环**

5: **结束循环**

6: **对于** 每个 $FGB_i \in FGB_s$ **执行**

7: **对于** 每个 $FGB_j (i \neq j) \in FGB_s$ **执行**

8: **如果** FGB_i 和 FGB_j 符合根据式17、E1.18 **则**

9: 将它们连接起来，并设置为相同的标签;

10: **结束 if**

11: **结束 for**

12: **结束 for**

13: **对于** 每个 $FGB_i \in FGB_s$ **执行**

14: **对于** 每个 $FGB_j (i \neq j) \in FGB_s$ **执行**

15: **如果** $o_i = 0$，**那么**

16: 连接 FGB_i 或 FGB_j 与最近的 FGB 以及具有相同标签的集合

17: **结束 if**

18: **结束 for**

19: </style}结束循环

20：返回聚类结果

球的错误区分，累积值越大，满足相邻规则的标准就越严格。如果两个粒球满足相邻规则，则它们属于同一个簇。基于上述描述，设计了一种模糊粒球连接算法，如算法3所示。

IV. 实验

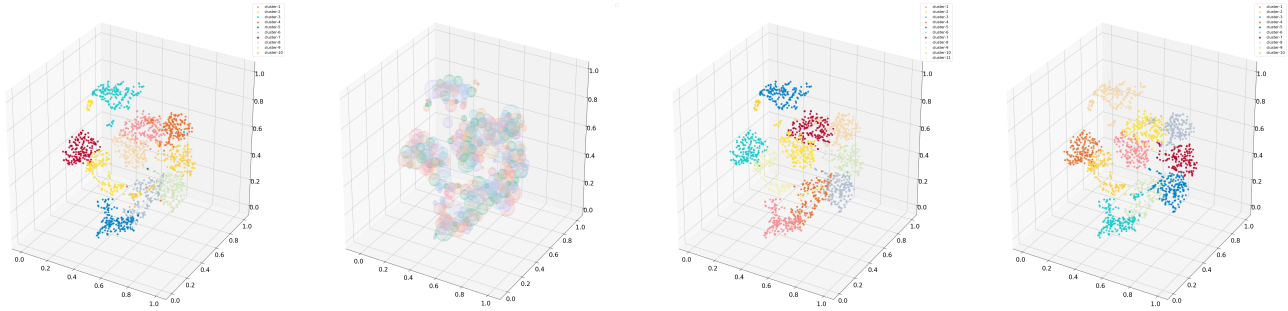
本节在合成数据集和真实数据集上评估所有相关方法。由于我们的方法基于对模糊C均值的应用与改进，因此我们对模糊C均值、ECM-NSGA-II、ECM- MOEA/D、MFS-FCM、DI-FCM、RI-FCM 和 FDPC 进行了对比实验，并将 K-均值算法也作为参照进行比较。实验硬件的详细配置如下：Intel Core i7-10700 CPU @ 2.90GHz，16G 内存。编程环境为 PyCharm。

我们使用三个指标评估带标签数据集上的聚类性能：准确率 [35] 、归一化互信息 [36],[37], 和调整兰德指数 [38]。计算准确率时，需要为每个数据分配一个标签。标签分配过程具有随机性，这会导致分配结果与真实标签不匹配，从而使准确率计算出错。因此，需要构造一个映射函数，将聚类得到的标签与真实标签逐一匹配。基于此映射函数，准确率的计算公式为：

$$ACC = \frac{1}{n}\sum_{i=1}^{n} \delta\left(l_i, \mathrm{map}\left(c_i\right)\right) \tag{20}$$

其中， l_i 是真实标签， c_i 是聚类后得到的标签，而 $\delta(x,\ y) = \begin{cases} 1 & x = y \\ 0 & x \neq y \end{cases}$ 是一个判别函数。

另一个性能评估指标是归一化互信息（NMI）。与准确率不同，归一化互信息不受簇标签排列方式的影响。其计算公式为



(a) 真实标注 (b) 生成的模糊粒球 (c) FGB_重叠 (d) FGB K-means

图4. 我们的方法在 MNIST 数据集上的聚类结果。

归一化互信息为:

$$NMI(L,C) = 2\frac{MI(L,C)}{H(L)+H(C)} \tag{21}$$

其中，L 为真实标签序列，C 为聚类标签序列，M I(L,C) 表示 L 和 C 之间的互信息，H() 表示信息熵，NMI 的计算公式也可表示为:

$$NMI(L,C) = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} n_{i,j}\log\left(n\cdot n_{i,j}\right)}{\sqrt{\left(\sum_i n_i \log\frac{n_i}{n}\right)\left(\sum_j n_j \log\frac{n_j}{n}\right)}} \tag{22}$$

最后一个评价指标调整兰德指数（ARI）定义如下:

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{A_i}{2}\sum_j \binom{B_j}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{A_i}{2}+\sum_j \binom{B_j}{2}\right] - \left[\sum_i \binom{A_i}{2}\sum_j \binom{B_j}{2}\right]/\binom{n}{2}} \tag{23}$$

表I 真实数据集的详细信息

数据集	实例数	维度	簇
cell	8681	22	21
MNIST	1797	3	10
Iris	150	4	3
大豆	47	35	4
皮肤病学	358	34	6
钞票	1372	4	2
ukm	258	5	4
分割	2310	19	7

A. 真实数据集实验

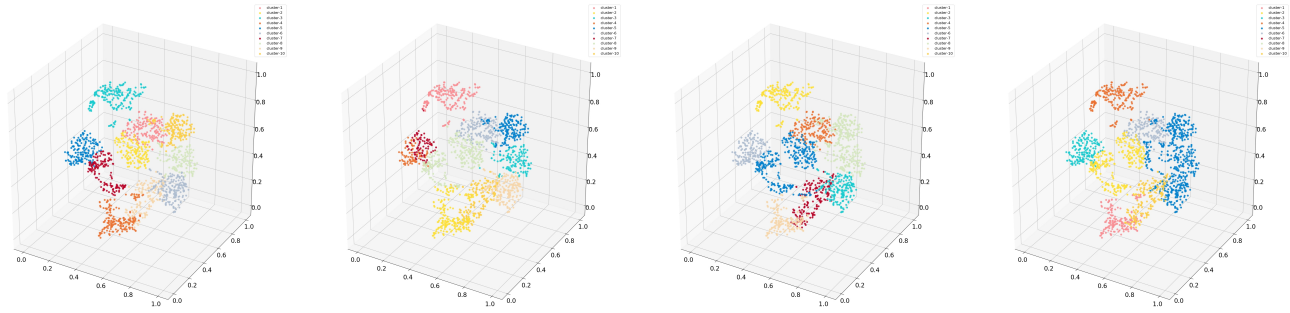
对于真实数据集上的实验，首先，我们在三维真实 MNIST 数据集上进行了实验。图4(a)为原始 MNIST 数据集的真实标注结果，图4(b)为我们的方法生成的模糊粒球结果。如图4(c)所示，这是我们提出的 FGB 重叠算法生成的簇结果；图4(d)则是我们提出的 FGB K-means 算法生成的簇结果。在图5(a)-5(h)中，我们可以看到 K均值、模糊C均值、ECM-NSGA-II、ECM-MOEA/D、MFS-FCM、DI-FCM、RI-FCM 和 FDPC 这8种不同方法在 MNIST 数据集上的聚类结果。可以看出，我们提出的基于模糊粒球的 FGB 重叠算法和_FGB K-means 算法，其聚类效果明显优于其他对比算法，结果也更符合真实标注。同时，一些聚类评估指标的结果也能说明这一点。根据表II可知，在 MNIST 数据集上，我们的 FGB 重叠方法在调整兰德指数和归一化互信息指标上的表现优于其他8种算法；而 FGB K-means 在 MNIST 数据集上的准确率指标也最高。最后，从表4中还可以看出，在程序运行时间方面，除去速度最快的 K均值算法后，我们的方法在 MNIST 数据集上的运行时间也位居前三。

第二个选取的真实数据集是时空生物医学数据集 [39], [40]中共聚焦显微镜下免疫细胞的生物医学图像分割数据集。该数据集尤其适用于免疫细胞图像，因为免疫细胞具有高度可塑性。从表 II 可以看出，在 mnist 数据集上，我们的方法 FGB 重叠的调整兰德指数和归一化互信息指标优于其他 8 种算法；在 mnist 数据集上，FGB K-means 的准确率指标也最高。最后，从表 4 还可以看出，就程序运行时间而言，除去速度最快的 K均值算法后，我们的方法在 mnist 数据集上的运行时间也位居前三。

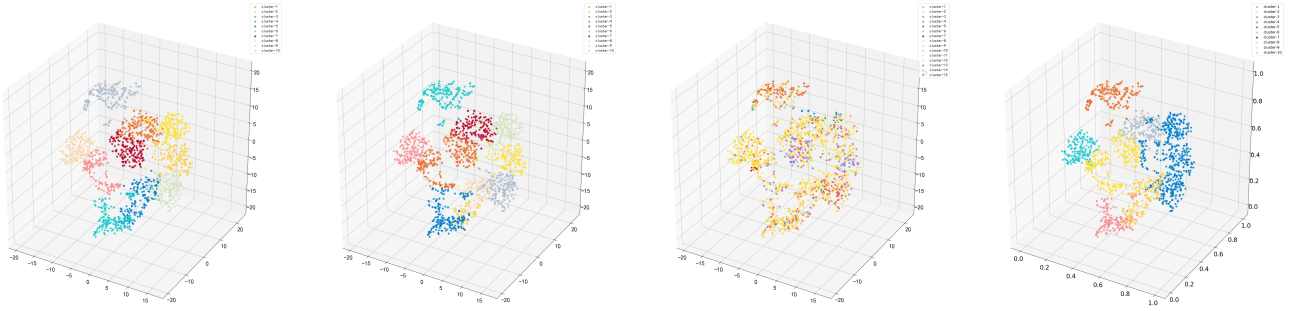
此外，我们还使用了 UCI 机器学习库 [41]中的若干真实数据集。这些数据集的相关信息如表I所示。准确率、归一化互信息和调整兰德指数的比较结果见表 II，这些算法的运行时间见表 III。从表 II 可以看出，与其他 8 种对比算法相比，我们提出的 FGB 重叠和 FGB K-means 在 6 个 UCI 真实数据集上的评价指标值（调整兰德指数、归一化互信息、准确率）大多位居前两名，表明了我们方法的有效性和准确性。此外，根据表 III 还可以看出，我们提出的 FGB 重叠和 FGB K-means 在 6 个 UCI 真实数据集上的运行时间也位居前三，这同样证明了我们方法的效率和可行性。

B. 模拟脑 *MRI* 数据集上的实验

最后，我们还从 BrainWeb 网站上的模拟脑数据库（SBD）中选取了真实数据集进行实验。目前，SBD 包含模拟脑部 MRI



(a) K均值 (b) 模糊C均值 (c) ECM-NSGA-II (d) ECM-MOEA/D



(e) MFS-FCM (f) DI-FCM (g) RI-FCM (h) FDPC

图5. 其他不同方法在 MNIST 数据集上的聚类结果。

表II 不同算法在真实数据集上的准确率、归一化互信息和调整兰德指数比较

数据集	指标	FGB 重叠	FGB K-means	K均值	模糊C均值	ECM-NSGA-II	ECM-MOEA/D	MFS-FCM	DI-FCM	RL-FCM	FDPC
MNIST	ARI	**0.915**	0.905	0.902	0.879	0.614	0.447	0.902	0.786	0.301	0.666
	NMI	**0.934**	0.924	0.921	0.906	0.805	0.727	0.921	0.864	0.568	0.8
	ACC	0.950	**0.954**	0.953	0.934	0.694	0.571	0.953	0.822	0.505	0.737
cell	ARI	**0.773**	0.715	0.718	0.604	0.297	0.317	0.653	0.625	0.651	0.644
	NMI	**0.927**	0.894	0.895	0.845	0.573	0.596	0.873	0.857	0.536	0.612
	ACC	0.831	**0.837**	0.836	0.699	0.398	0.390	0.751	0.728	0.511	0.661
Iris	ARI	0.568	**0.592**	0.583	0.551	0.567	0.564	0.583	0.583	0.453	0.568
	NMI	**0.734**	0.641	0.635	0.612	0.695	0.699	0.635	0.635	0.569	0.734
	ACC	0.667	**0.813**	0.807	0.764	0.726	0.722	0.807	0.807	0.767	0.667
大豆	ARI	0.531	**0.605**	0.563	0.571	0.473	0.504	0.605	0.605	0.455	0.349
	NMI	0.732	**0.740**	0.723	0.730	0.608	0.647	0.74	0.74	0.571	0.635
	ACC	0.617	**0.809**	0.766	0.726	0.686	0.638	0.809	0.809	0.574	0.532
皮肤病学	ARI	0.211	**0.469**	0.445	0.433	0.261	0.233	0.356	0.357	0.319	0.257
	NMI	0.457	**0.597**	0.562	0.509	0.481	0.460	0.564	0.562	0.521	0.582
	ACC	0.508	**0.514**	0.511	0.503	0.465	0.476	0.439	0.444	0.419	0.411
钞票	ARI	**0.391**	0.034	0.049	0.061	0.067	0.038	0.046	0.047	0.049	0.034
	NMI	**0.469**	0.020	0.030	0.040	0.067	0.093	0.029	0.029	0.291	0.098
	ACC	**0.634**	0.596	0.612	0.619	0.595	0.606	0.609	0.610	0.106	0.603
ukm	ARI	0.206	**0.223**	0.185	0.180	0.155	0.203	0.180	0.169	0.114	0.222
	NMI	0.292	0.296	0.269	0.263	0.190	0.249	0.264	0.263	0.193	**0.300**
	ACC	0.485	**0.562**	0.500	0.475	0.505	0.534	0.484	0.481	0.426	0.504
分割	ARI	0.094	0.191	0.186	**0.197**	0.113	0.123	0.186	0.189	0.065	0.134
	NMI	0.210	**0.349**	0.336	0.340	0.299	0.295	0.336	0.339	0.329	0.242
	ACC	0.273	**0.372**	0.367	0.368	0.288	0.307	0.368	0.369	0.135	0.331

表 III 真实数据集上各算法的运行时间比较（单位：秒）

数据集	FGB_重叠	FGB K-means	K均值	模糊C均值	ECM-NSGA-II	ECM-MOEA/D	MFS-FCM	DI-FCM	RL-FCM	FDPC
MNIST	3.624	**2.374**	**0.067**	8.309	33.416	27.019	41.443	19.190	135.895	32.076
cell	57.039	**19.654**	**0.290**	112.430	474.911	507.599	1107.395	629.280	2697.948	830.548
Iris	0.106	**0.119**	**0.022**	**0.053**	7.330	6.093	0.246	0.594	0.578	0.234
大豆	0.028	0.032	**0.016**	**0.014**	8.693	10.721	0.203	0.126	0.094	0.047
皮肤病学	0.375	**0.328**	**0.063**	0.674	12.252	16.744	2.469	3.547	5.689	1.625
钞票	3.644	1.784	**0.037**	**0.484**	14.506	28.931	3.374	3.233	178.996	18.988
ukm	**0.214**	0.236	**0.038**	0.216	11.646	11.273	1.987	2.804	2.249	0.678
分割	6.599	**4.825**	**0.089**	6.533	34.941	34.676	9.544	56.599	390.204	57.082

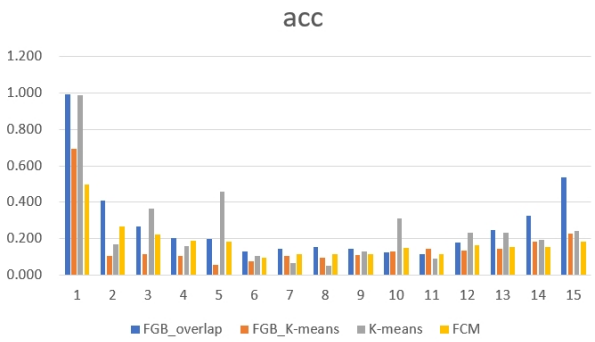


图 6. 模拟脑部 MR 数据集上各算法准确率的比较

基于两种解剖模型的数据：正常模型和多发性硬化症（MS）模型。对于这两种模型，均使用三种序列（T1、T2 和质子密度（PD）加权）模拟完整的三维数据体，并采用不同的层厚、噪声水平和强度不均匀程度。在这两种模型中，我们选择了第二种模型，即多发性硬化症（MS），并通过适当的预处理获得其真实标注。具体操作是先将每个选项设为 T1 模式。在 ICMB协议下，层厚为 1 mm，噪声水平为 0，灰度级不均匀度为 0。然后选取 MS 病理脑图像的前 150 层，每层仅选取 0、1、2、3 和 10 类数据进行实验。分割后，各类别以如下像素值进行渲染。最后将末尾 10 层归为一组。最终获得 15 组真实标注和标签，用于实验和验证。

我们首先直接使用 K均值算法和原始 FCM算法，对这15个数据集进行实验，并统计实验结果。然后，我们使用算法 1为这15个数据集生成模糊超球体。在这些模糊超球体上，我们使用 K均值算法和原始 FCM算法进行聚类，并统计实验结果。最后，我们使用自主提出的重叠方法连接生成的模糊超球体，以生成聚类结果。如图3所示，我们对实验结果进行了统计分析。可以看出，我们提出的通过重叠连接生成模糊超球体的方法具有较为稳定且良好的聚类效果。同时，大多数数据集的聚类准确率也很出色。其准确率大多也高于 K均值和 FCM聚类方法。此外，如图4所示，与

表IV 模拟脑部 MR 数据集上算法运行时间对比（单位：秒）

组	FGB_重叠	FGB K-means	K均值	FCM
1	0.101	0.118	0.014	0.152
2	1.110	1.958	0.021	1.793
3	2.962	5.013	0.027	3.070
4	3.837	7.080	0.031	4.017
5	5.138	12.168	0.037	5.190
6	4.709	14.302	0.046	7.273
7	4.846	13.883	0.042	7.118
8	5.624	14.594	0.043	7.021
9	4.533	12.781	0.043	6.661
10	3.703	11.048	0.038	5.823
11	3.481	9.668	0.038	5.136
12	4.260	8.292	0.034	4.757
13	4.315	6.697	0.030	4.031
14	6.045	7.474	0.024	6.244
15	0.938	1.161	0.017	1.010

运行时间相比，我们可以看到，K均值的运行时间最短，而我们提出的重叠方法也具有较高的效率，仍然优于原始 FCM 聚类方法。

V. 结论

本文通过生成模糊粒球来表征数据的模糊属性，从而表示数据集的适当粒度。所得模糊粒球可通过多种方式用于不同数据集场景下的最终聚类。例如，采用重叠连接可自适应地获得复杂形状的聚类结果；采用K均值则能更快地获得聚类结果。在合成数据集和真实数据集上的实验表明，FGB 重叠和FGB K-means能够识别复杂数据集中的模式，且效果显著。未来的研究仍有许多值得探索的方向。今后，我们将探索如何改进原始FCM迭代方法，并拓展模糊粒球的更多可能性，以进一步提升算法的效果和效率。

VI. 致谢

本研究部分得到了国家自然科学基金委员会项目（批准号：62176033和61936001）、重庆市自然科学基金项目（批准号：cstc2019jcyj-msxmX0485和cstc2019jcyj-cxttX0002）以及美国国家科学基金会资助的NICE：综合计算昆虫学研究生培训项目（项目编号1631776）的支持。

参考文献

[1] Xu W. Xu R., Wunsch D[《R期刊》]. 聚类算法综述，《IEEE 神经网络汇刊》，2005，16(3)：645-678。[2] Jain A K, Murty M N, Flynn P J. 数据聚类：综述[《R期刊》]. ACM 计算调查（CSUR），1999年，31(3)：264-323。[3] Kumar P, Sirohi D. 鸢尾花数据集上的模糊C均值与硬C均值算法比较分析

[ ] 2010, 《R期刊》. 5 2 33 37

国际计算机应用杂志，（）：-。[4] Bharathi S, Venkatesan P. 模糊 C 均值聚类技术在图像分割中的综述[C]//2016年第二届应用与理论计算及通信技术国际会议（iCATccT）。IEEE，2016：604-608。[5] Naouali S, Ben Salem S, Chtourou Z. 分类数据聚类：综述[《R期刊》]. 《国际信息技术与决策制定杂志》，2020年，19(01)：49-96。[6] Zhu L, Chung F L, Wang S. 改进模糊划分的广义模糊C均值聚类算法[《R期刊》]. IEEE《系统、人类与控制论汇刊》B辑（控制论），2009年，39(3)：578-591。[7] Meng, Wang, Dui Hongyan, Zhou Shiyuan, Dong Zhankui and Wu Zige. 核粗糙K均值算法。（2020）。[8] Bezdek J C, Ehrlich R, Full W. 模糊C均值：模糊C均值聚类算法[《R期刊》]. 《计算机与地球科学》，1984年，10(2-3)：191-203。[9] Zhu L F, Wang J S, Wang H Y. 一种新颖的模糊C均值聚类算法有效性函数[《R期刊》]. IEEE Access，2019年，7：152289-152315。[10] Prasad B N, Rathore M, Gupta G等. 硬C均值、模糊C均值及替代C均值算法的性能度量[《R期刊》]. 国际计算机科学与信息技术杂志，2016年，7(2)：878-883。[11] Ruspini E H, Bezdek J C, Keller J M. 模糊聚类：历史回顾

[ ] 2019, 《R期刊》. 14 1 45 55

IEEE《计算智能杂志》，（）：-。[12] Senapati T, Yager R R. 费马模糊集[《R期刊》]. 《环境智能与人性化计算杂志》，2020年，11(2)：663-674。[13] Ibrahim H Z. 模糊集的新扩展及其在粗糙拓扑和医学诊断中的应用[《R期刊》]. 《软计算》，2022年：1-15。[14] Karaaslan F, zl. 对偶二型犹豫模糊集的相关系数及其在聚类分析中的应用[《R期刊》]. 国际智能系统杂志，2020年，35(7)：1200-1229。[15] Liu X, Fan J, Chen Z. 基于密度峰值的改进模糊C均值算法[《R期刊》]. 《国际机器学习与控制论杂志》，2020年，11(3)：545-552。[16] Yang M S. 模糊聚类综述[《R期刊》]. 《数学与计算机建模》，1993年，18(11)：1-16。[17] Feng Q, Chen L, Chen C L P等. 深度模糊聚类：一种表征学习方法

[ ] 2020, 《R期刊》. 28 7 1420 1433

IEEE《模糊系统汇刊》，（）：-。[18] Mau T N, Inoguchi Y, Huynh V N. 一种基于局部敏感哈希的分类数据模糊聚类簇预测方法[《R期刊》]. IEEE Access，2022年，10：34196-34206。[19] Ferraro M B, Giordani P, Serafini A. fclust：用于模糊聚类的R软件包[《R期刊》]. 《R期刊》，2019年，11(1)：198。[20] Jiao L, Yang H, Liu Z等. 利用无监督模糊决策树实现可解释模糊聚类[《R期刊》]. Information Sciences，2022年，611：540-563。[21] Mittal H, Saraswat M. 一种适用于超椭球形或超球形紧密簇（质心相距较远）的新型模糊簇有效性指数[《R期刊》]. IEEE《模糊系统汇刊》，2020年，29(11)：3249-3258。[22] Beigmohamadi R. 分数阶熵及其在模糊C均值聚类中的应用[C]//2022年第九届伊朗模糊与智能系统联合大会（CFIS）。IEEE，2022：1-4。[23] Beigmohamadi R. 分数阶熵及其在模糊C均值聚类中的应用[C]//2022年第九届伊朗模糊与智能系统联合大会（CFIS）。IEEE，2022：1-4。[24] Gao Y, Wang Z, Li H等. 高斯协同模糊C均值聚类[《R期刊》]. 《国际模糊系统杂志》，2021年，23(7)：2218-2234。[25] Zhao K, Dai Y, Jia Z等. 通用模糊C均值聚类策略：利用目标函数控制聚类结果的模糊度[《R期刊》]. IEEE《模糊系统汇刊》，2021年。[26] Tu M, Xu Y. 基于模糊C均值的聚类算法优化与分析[C]//2021年IEEE第三届欧亚物联网、通信与工程会议（ECICE）。IEEE，2021：306-308。[27] Askari S. 针对簇大小不等且含有噪声和离群值的数据的模糊C均值聚类算法：综述与发展[J]。Expert Systems with Applications，2021年，165：113856。[28] Kwon S H，Kim J，Son S H。改进的模糊聚类有效性指标[J]。Electronics Letters，2021年，57(21)：792-794。[29] Chen L。视觉感知中的拓扑结构

[ ] 1982, J。218 4573 699 700

Science，（）：-。[30] Wang G。DGCC：数据驱动的粒状认知计算。Granular Comput 2 (4)： 343355[J]。2017年。[31] Xia S，Peng D，Meng D，等。一种无界快速自适应k均值算法[J]。IEEE Transactions on Pattern Analysis and Machine Intelli [32] Xia S，Wang G，Gao X，等。GBSVM：粒球支持向量机[J]。arXiv预印本 arXiv:2210.03120，2022年。[33] Xia S，Dai X，Wang G，等。分类问题中的高效自适应粒球生成方法[J]。IEEE神经网络与学习系统汇刊，2022年。[34] Xia S，Liu Y，Ding X，等。用于高效、可扩展和稳健学习的粒球计算分类器

[ ] 2019, J。483 136 152

Information Sciences，：-。[35] Wu M，Schlkopf B。用于聚类的局部学习方法[J]。神经信息处理系统进展，2006年，19。[36] Strehl A，Ghosh J。聚类集成——用于组合多个划分的知识复用框架[J]。机器学习研究杂志，2002年，3(12月)：583-617。[37] Cover T M。信息论要素[M]。John Wiley & Sons，1999年。[38] Qian P，Chung F L，Wang S，等。使用最小包围球的大型数据集快速基于图的松弛聚类[J]。IEEE《系统、人类与控制论汇刊》B辑（控制论），2012年，42(3)：672-687。[39] Ulman V，Maka M，Magnusson K E G，等。细胞追踪算法的客观比较[J]。Nature Methods，2017年，14(12)：1141-1152。[40] Maka M，Ulman V，Svoboda D，等。细胞追踪算法比较基准[J]。生物信息学，2014年，30(11)：1609-1617。[41] Asuncion A，Newman D。UCI机器学习库[J]。2007年。