控制与决策
Control and Decision
ISSN 1001-0920,CN 21-1124/TP

《控制与决策》网络首发论文

题目：一种邻域粒 K 均值聚类方法
作者：陈玉明，蔡国强，卢俊文，曾念峰
DOI：10.13195/j.kzyjc.2021.1553
收稿日期：2021-09-04
网络首发日期：2022-02-07
引用格式：陈玉明，蔡国强，卢俊文，曾念峰．一种邻域粒 K 均值聚类方法[J/OL]．控制与决策. https://doi.org/10.13195/j.kzyjc.2021.1553

一种邻域粒 K 均值聚类方法

陈玉明[1†]，蔡国强[1]，卢俊文[1]，曾念峰[2]

(1. 厦门理工学院 计算机与信息工程学院, 福建 厦门 361024；
2. 易成功（厦门）信息科技有限公司, 福建 厦门 361024)

摘　要: K 均值聚类属于无监督学习, 具有简单、易用的特点, 是一种广泛使用的聚类分析方法. 然而, 对于非凸、稀疏及模糊的非线性可分数据, 其聚类效果不佳. 通过引入粒计算理论, 采用邻域粒化技术, 提出了一种邻域粒 K 均值聚类方法. 样本在单特征上使用邻域粒化技术构造邻域粒子, 在多特征上使用邻域粒化技术形成邻域粒向量. 通过定义邻域粒与邻域粒向量的大小、度量和运算规则, 提出两种邻域粒距离度量, 并对所提出的邻域粒距离度量进行了公理化证明. 最后, 采用多个 UCI 数据集进行实验, 将 K 均值聚类算法分别结合两种邻域粒距离度量, 在邻域参数和距离度量两个方面与经典聚类算法进行了比较, 其结果表明了所提出的邻域粒 K 均值聚类方法的可行性和有效性.

关键词: 粒计算；邻域粒；K 均值聚类；聚类；无监督学习；粒向量

中图分类号: TP181　　**文献标志码:** A

DOI: 10.13195/j.kzyjc.2021.1553　　**开放科学 (资源服务) 标识码 (OSID):**

A neighborhood granular K-means clustering method

Yuming Chen[1†], *Guoqiang Cai*[1], *Junwen Lu*[1], *Nianfeng Zeng*[2]

(1.College of Computer and Information Engineering, Xiamen University of Technology, Xiamen 361024, China; 2. E-success (Xiamen) Information Technology Co., Ltd., Xiamen 361024, China)

Abstract: K-means clustering belongs to unsupervised learning and is simple and easy to use. It is a widely used clustering analysis method. However, for non-convex, sparse and fuzzy nonlinear separable data, the clustering effect is not good. By introducing granule computing theory and using neighborhood granulation technology, a neighborhood granule K-means clustering method is proposed. The sample uses neighborhood granulation technology to construct neighborhood granules on a single feature, and to form neighborhood granule vectors on multiple features. By defining the size, measurement and operation rules of neighborhood granules and neighborhood granule vectors, two kinds of neighborhood granule distance measurement are proposed, and the axiomatic proof of the proposed neighborhood granule distance measurement is carried out. Finally, several UCI data sets are used to carry out experiments, the K-means clustering algorithm is combined with two neighborhood granule distance measurements respectively. It is compared with the classical clustering algorithm in two aspects of neighborhood parameters and distance measurement. The results show that the proposed neighborhood granular K-means clustering method is feasible and effective.

Keywords: Granular computing; neighborhood; K-means clustering; clustering; unsupervised learning; granule vectors

0 引　言

聚类是一个根据数据的某些属性将数据划分到相应类簇, 并且在同一个类簇内数据的相似性要尽可能大, 不同类簇间数据的相似性要尽可能小的过程[1]. 聚类过程中使用某种差异度量区分开数据对象, 常见的差异度量有距离度量等[2]. 数据之间存在一种称为“抱团”的性质, 聚类的目的就是为了找到这个性质[3].K 均值 (K-means) 聚类是聚类分析中的一个基础方法, 它是通过随机选取 K 个聚类中心, 计算数据与聚类中心的距离后划分类簇, 对每个类簇的数据求平均值获得新的聚类中心, 不断迭代直到聚类中心不变来完成聚类的 [4]. 由于它快速及简单的特性, 以至于它成为聚类中最常用的方法[5].K-均值聚类的核心思想就是使得一个类簇中数据之间的总体差异要小于邻近类簇中心的差异[6].

1979 年,Zadeh 发表了论文“Fuzzy sets and

收稿日期: 2021-09-04；**录用日期:** 2021-12-30.
基金项目: 国家自然科学基金项目 (61976183, 61871464), 福建省自然科学基金 (2020J01266), 福建省教育厅中青年科研项目 (JAT190679).
†通讯作者. E-mail: ymchen@xmut.edu.cn.

information granularity”, 引入信息粒度化的思想, 同时 Zadeh 还认为信息粒的概念在很多领域中都存在, 只是信息粒的表现形式在不同的领域中是不同的, 随后信息粒成为热点研究领域[7-8]. 从 1994 年开始,Zadeh 一直强调信息粒化的重要性, 还需要充分探索信息粒化的计算理论.1996 年, 粒计算 (granular computing) 的概念被 Lin 提出[9].2000 年之后, 粒计算在国内的热度开始增大, 引起许多研究学者对它的兴趣. 其中苗夺谦教授等在知识粗糙性与信息熵之间的关系、粗糙集理论中的概念和运算等方向展开粒计算的信息论研究[10-12]. 近些年, 邻域粗糙集被引入到特征选择[13-14] 等多个方法中. 当前 AI 研究领域中, 粒计算属于一种新的概念, 在人类较高层次认知机理研究的范畴中就包括了粒计算的信息处理模式、问题求解方法、多粒度表示等[15].

K 均值聚类是一种广泛使用的聚类分析方法, 但对于非凸、稀疏及模糊的非线性可分数据, 其聚类效果并不佳. 近年来, 许多学者更多地是针对 K 均值聚类算法的初始化聚类中心进行研究, 如文献 [16] 是利用构建相异性矩阵来优化初始聚类中心的选取和文献 [17] 通过量子粒子群优化算法降低 K 均值聚类算法对初始聚类中心的依赖, 从而提高聚类的性能. 也有一些学者将粗糙集与 K 均值聚类算法相结合, 如文献 [18]. 本文通过在算法结构上进行改进, 将 K 均值聚类和粒计算相结合, 提出一种邻域粒 K 均值聚类方法. 该方法基于邻域粒化技术, 通过单特征进行邻域粒化形成邻域粒子, 进而在多特征上将多个邻域粒子组成邻域粒向量. 同时利用邻域粒化形式的度量和运算关系定义出邻域粒的距离度量. 将 K 均值聚类的思想和邻域粒化技术相结合得出邻域粒 K 均值聚类算法, 使非线性可分数据的聚类效果得到提升, 同时该方法可以使得每个邻域粒向量都有全局性, 提高聚类的收敛速度. 使用 UCI 数据集进行实验测试表明, 本文算法得到的聚类效果优于 K 均值聚类算法, 为聚类算法探索一条新的途径.

1 邻域粒向量表示

设数据集为 $IS=(U,F)$, 其中样本集为 $U=\{x_1,x_2,...,x_n\}$, 属性集为 $F=\{a_1,a_2,...,a_m\}$. 给定样本 $x\in U$, 对于任一属性 $a\in F$,$v(x,a)\in[0,1]$ 表示样本 x 在属性 a 上归一化后的值.

给定数据集 IS, 对于样本 $x,y\in U$, 单属性 $a\in F$, 则 x 与 y 在单属性 a 上的曼哈顿距离为:

$$s_a(x,y)=|v(x,a)-v(y,a)|. \tag{1}$$

定义 1 给定数据集 IS, 对于样本 $x,y\in U$, 单属性 $a\in F$, 给定邻域参数 δ, 则样本 x,y 的邻域判别函数定义为:

$$\varphi(x,y)=\begin{cases}0, s_a(x,y)>\delta\\ 1, s_a(x,y)\le\delta\end{cases}. \tag{2}$$

当 $\varphi(x,y)=1$, 表示 x,y 互为邻域; $\varphi(x,y)=0$, 则表示 x,y 不相邻.

定义 2 给定数据集 IS, 对于任一样本 $x\in U$ 和任一属性 $a\in F$, 则 x 在属性 a 上进行邻域粒化, 形成的邻域粒子定义为:

$$g_a(x)=\{r_j\}_{j=1}^n=\{r_1,r_2,...,r_n\}, \tag{3}$$

其中 $r_j=\varphi(x,x_j)$ 为样本 x,x_j 邻域判别函数, 表示两者是否相邻.

定义 3 给定数据集 IS, 对于任一样本 $x\in U$, 任一属性子集 $P\subseteq F$, 设 $P=\{a_1,a_2,...,a_m\}$, 则 x 在属性子集 P 上的邻域粒向量定义为:

$$G_P(x)=(g_1(x),g_2(x),...,g_m(x))^T, \tag{4}$$

其中 $g_m(x)$ 是样本 x 在属性 a_m 上的邻域粒子.

邻域粒向量由邻域粒子组成, 邻域粒子由 0 或 1 构成, 表示了样本之间的邻域关系. 邻域粒子是 0 或 1 构成的有序集合. 因此, 邻域粒向量的元素是有序集合, 与传统向量不一样, 传统向量的元素是一个实数.

定义 4 给定数据集 IS, 对于任一样本 $x\in U$, 任一属性 $a\in F$, 邻域粒子 $g_a(x)$ 的大小定义为:

$$Size(g_a(x))=|g_a(x)|=\textstyle\sum_{j=1}^n r_j, \tag{5}$$

易知邻域粒子的大小满足: $1\le|(g_a(x)|\le n$.

定义 5 给定数据集 IS, 对于任一样本 $x\in U$, 任一属性子集 $P\subseteq F$, 设 $P=\{a_1,a_2,...,a_m\}$, 则 x 的邻域粒向量 $G_P(x)$ 的大小定义为:

$$Size(G_P(x))=|G_P(x)|=\sqrt{\textstyle\sum_{i=1}^m|g_i(x)|^2}. \tag{6}$$

邻域粒向量 $G_P(x)$ 的大小也称为邻域粒向量的模, 易知其大小满足: $\sqrt{m}\le|G_P(x)|\le n*\sqrt{m}$.

2 邻域粒距离度量

定义 6 给定数据集 IS, 其中属性集为 $F=\{a_1,a_2,...,a_m\}$. 对于 $\forall x,y\in U$, 存在 F 上的两个邻域粒向量为 $G_F(x)=(g_1(x),g_2(x),...,g_m(x))^T$,$G_F(y)=(g_1(y),g_2(y),...,g_m(y))^T$, 则两个邻域粒向量的交、并、减与异或运算定义为:

$$G_F(x)\wedge G_F(y)=(g_1(x)\wedge g_1(y),g_2(x)\wedge g_2(y),...,g_m(x)\wedge g_m(y))^T; \tag{7}$$

$$G_F(x)\vee G_F(y)=(g_1(x)\vee g_1(y),g_2(x)\vee g_2(y),...,g_m(x)\vee g_m(y))^T; \tag{8}$$

$$G_F(x)-G_F(y)=(g_1(x)-g_1(y),g_2(x)-g_2(y),...,g_m(x)-g_m(y))^T; \quad (9)$$

$$G_F(x)\oplus G_F(y)=(g_1(x)\oplus g_1(y),g_2(x)\oplus g_2(y),...,g_m(x)\oplus g_m(y))^T. \quad (10)$$

定义 7 给定数据集 IS, 其中属性集为 $F=\{a_1,a_2,...,a_m\}$. 对于 $\forall x,y\in U$, 存在 F 上的两个邻域粒向量为 $G_F(x)=(g_1(x),g_2(x),...,g_m(x))^T$,$G_F(y)=(g_1(y),g_2(y),...,g_m(y))^T$, 则两个邻域粒向量的相对距离定义为:

$$d(G_F(x),G_F(y))=\frac{1}{m}\sum_{i=1}^{m}\frac{|g_i(x)\oplus g_i(y)|}{|g_i(x)\vee g_i(y)|}, \quad (11)$$

其中,$|F|=m$. 易知, 邻域粒向量的相对距离满足: $0\le d(G_F(x),G_F(y))\le 1$.

定义 8 给定数据集 IS, 其中属性集为 $F=\{a_1,a_2,...,a_m\}$. 对于 $\forall x,y\in U$, 存在 F 上的两个邻域粒向量为 $G_F(x)=(g_1(x),g_2(x),...,g_m(x))^T$,$G_F(y)=(g_1(y),g_2(y),...,g_m(y))^T$, 则两个邻域粒向量的绝对距离定义为:

$$h(G_F(x),G_F(y))=\frac{1}{m*n}\sum_{i=1}^{m}|g_i(x)\oplus g_i(y)|, \quad (12)$$

其中,$|F|=m$,$|U|=n$. 易知, 邻域粒向量的绝对距离满足: $0\le h(G_F(x),G_F(y))\le 1$.

定理 1 两个邻域粒向量的相对距离是一种距离度量, 满足以下三个性质:

(1) 非负,$0\le d(G_F(x),G_F(y))\le 1$;

(2) 对称,$d(G_F(x),G_F(y))=d(G_F(y),G_F(x))$;

(3) 三角不等式,$d(G_F(x),G_F(y))+d(G_F(y),G_F(z))\ge d(G_F(x),G_F(z))$.

证明. (1) 由 $g_i(x)\oplus g_i(y)=g_i(x)\vee g_i(y)-g_i(x)\wedge g_i(y)$, 可知 $\frac{|g_i(x)\oplus g_i(y)|}{|g_i(x)\vee g_i(y)|}=\frac{|g_i(x)\vee g_i(y)-g_i(x)\wedge g_i(y)|}{|g_i(x)\vee g_i(y)|}$, 则 $0\le\frac{|g_i(x)\oplus g_i(y)|}{|g_i(x)\vee g_i(y)|}\le 1$. 由 $F=\{a_1,a_2,...,a_m\}$, 可知 $|F|=m$. 因此,$0\le\sum_{i=1}^{m}\frac{|g_i(x)\oplus g_i(y)|}{|g_i(x)\vee g_i(y)|}\le m$, 则 $0\le\frac{1}{m}\sum_{i=1}^{m}\frac{|g_i(x)\oplus g_i(y)|}{|g_i(x)\vee g_i(y)|}\le 1$. 所以,$0\le d(G_F(x),G_F(y))\le 1$ 成立.

(2) 因 $g_i(x)\vee g_i(y)=g_i(y)\vee g_i(x)$,$g_i(x)\wedge g_i(y)=g_i(y)\wedge g_i(x)$, 可知 $\frac{1}{m}\sum_{i=1}^{m}\frac{|g_i(x)\oplus g_i(y)|}{|g_i(x)\vee g_i(y)|}=\frac{1}{m}\sum_{i=1}^{m}\frac{|g_i(y)\oplus g_i(x)|}{|g_i(y)\vee g_i(x)|}$. 因此,$d(G_F(x),G_F(y))=d(G_F(y),G_F(x))$ 成立.

(3) 从文献 [19] 中命题 3 可知,$\frac{|g_i(x)\oplus g_i(y)|}{|g_i(x)\vee g_i(y)|}+\frac{|g_i(y)\oplus g_i(z)|}{|g_i(y)\vee g_i(z)|}\ge\frac{|g_i(x)\oplus g_i(z)|}{|g_i(x)\vee g_i(z)|}$. 因此,$\frac{1}{m}\sum_{i=1}^{m}\frac{|g_i(x)\oplus g_i(y)|}{|g_i(x)\vee g_i(y)|}+\frac{1}{m}\sum_{i=1}^{m}\frac{|g_i(y)\oplus g_i(z)|}{|g_i(y)\vee g_i(z)|}\ge\frac{1}{m}\sum_{i=1}^{m}\frac{|g_i(x)\oplus g_i(z)|}{|g_i(x)\vee g_i(z)|}$ 成立. 由邻域粒向量的相对距离定义可知,$d(G_F(x),G_F(y))+d(G_F(y),G_F(z))\ge d(G_F(x),G_F(z))$ 成立.

定理 2 两个邻域粒向量的绝对距离是一种距离度量, 满足以下三个性质:

(1) 非负,$0\le h(G_F(x),G_F(y))\le 1$;

(2) 对称,$h(G_F(x),G_F(y))=h(G_F(y),G_F(x))$;

(3) 三角不等式,$h(G_F(x),G_F(y))+h(G_F(y),G_F(z))\ge h(G_F(x),G_F(z))$.

证明. (1) 由 $g_i(x)\oplus g_i(y)=g_i(x)\vee g_i(y)-g_i(x)\wedge g_i(y)$,$1\le|g_i(x)|\le n$, 可知 $0\le|g_i(x)\oplus g_i(y)|\le n$. 由 $F=\{a_1,a_2,...,a_m\}$, 可知 $|F|=m$. 因此,$0\le\sum_{i=1}^{m}|g_i(x)\oplus g_i(y)|\le m*n$, 则 $0\le\frac{1}{m*n}\sum_{i=1}^{m}|g_i(x)\oplus g_i(y)|\le 1$. 所以,$0\le h(G_F(x),G_F(y))\le 1$ 成立.

(2) 因 $g_i(x)\vee g_i(y)=g_i(y)\vee g_i(x)$,$g_i(x)\wedge g_i(y)=g_i(y)\wedge g_i(x)$, 可知 $\frac{1}{m*n}\sum_{i=1}^{m}|g_i(x)\oplus g_i(y)|=\frac{1}{m*n}\sum_{i=1}^{m}|g_i(y)\oplus g_i(x)|$. 因此,$h(G_F(x),G_F(y))=h(G_F(y),G_F(x))$ 成立.

(3) 从文献 [19] 中命题 6 可知,$|g_i(x)\oplus g_i(y)|+|g_i(y)\oplus g_i(z)|\ge|g_i(x)\oplus g_i(z)|$. 因此,$\frac{1}{m*n}\sum_{i=1}^{m}|g_i(x)\oplus g_i(y)|+\frac{1}{m*n}\sum_{i=1}^{m}|g_i(y)\oplus g_i(z)|\ge\frac{1}{m*n}\sum_{i=1}^{m}|g_i(x)\oplus g_i(z)|$ 成立. 由邻域粒向量的绝对距离定义可知,$h(G_F(x),G_F(y))+h(G_F(y),G_F(z))\ge h(G_F(x),G_F(z))$ 成立.

3 邻域粒 K 均值聚类算法

粒 K 均值聚类算法是无监督的聚类算法, 它以粒向量为单位进行聚类, 粒向量是由粒子构成的, 而粒子是在全局样本空间中进行粒化而形成的, 因此粒子含有全局的信息, 其迭代收敛会加快. 为了设计粒 K 均值聚类算法, 先定义粒聚类的中心点, 阐述邻域粒 K 均值聚类的原理.

3.1 邻域粒 K 均值聚类原理

邻域粒 K 均值聚类算法的思想很简单, 首先对样本进行邻域粒化, 每个样本粒化为一个粒向量, 对于给定的样本集, 按照粒向量之间的距离大小, 将样本集划分为 K 个簇. 让簇内的粒向量尽量紧密的连

在一起, 而让簇间的粒向量距离尽量的大.

设样本划分为 $(C_1, C_2, ..., C_K)$ K 个簇, 则粒 K 均值聚类的损失函数为:

$$J_e = \sum_{i=1}^{K} \sum_{x \in c_i} h(G_F(x), \mu_i), \tag{13}$$

其中 μ_i 为 C_i 簇的均值粒向量, 也称为粒质心,$h(G_F(x), \mu_i)$ 表示样本 x 的粒向量与粒质心的绝对距离. 也可以采用粒向量的相对距离表示, 为:

$$J_e = \sum_{i=1}^{K} \sum_{x \in c_i} d(G_F(x), \mu_i). \tag{14}$$

粒质心公式表示为:

$$\mu_i = \frac{1}{n_i} \sum_{x \in c_i} G_F(x), \tag{15}$$

其中 n_i 为 C_i 簇中样本的个数,$G_F(x)$ 表示样本 x 的粒向量.

粒 K 均值聚类目标则是让 J_e 损失函数最小, 则采用启发式迭代的方法设计邻域粒 K 均值聚类算法.

3.2 邻域粒 K 均值聚类算法

上一小节阐述了粒 K 均值聚类的原理, 本小节将详细阐述邻域粒 K 均值聚类算法.

邻域粒 K 均值聚类算法:

输入: 数据集 $IS = (U, F)$, 其中样本集为 $U = \{x_1, x_2, ..., x_n\}$, 属性集为 $F = \{a_1, a_2, ..., a_m\}$; 类簇参数 K, 邻域参数 δ, 最大迭代次数 N;

输出: 簇划分 $C = (C_1, C_2, ..., C_K)$.

(1) 样本集 U 邻域粒化为 $GT = \{G_F(x_1), G_F(x_2), ..., G_F(x_n)\}$;

(2) 从 GT 中随机选 K 个邻域粒向量作为初始粒质心 $(\mu_1, \mu_2, ..., \mu_K)$;

(3) $For\ t = 1\ to\ N$

(3.1) 将簇划分 C 初始化为 $C_j = \varnothing (j = 1, 2, ..., K)$;

(3.2) 对于 $i = 1, 2, ..., n$, 计算邻域粒向量 $G_F(x_i)$ 和各个粒质心向量 $\mu_j (j = 1, 2, ..., K)$ 的粒距离: $d_{ij} = d(G_F(x_i), \mu_j)$ 或 $d_{ij} = h(G_F(x_i), \mu_j)$;将 x_i 标记为最小的 d_{ij} 所对应的类别 λj; 此时更新 $C_{\lambda j} = C_{\lambda j} \cup x_i$;

(3.3) 对于 $j = 1, 2, ..., K$, 将 C_j 中所有的样本点重新计算新的粒质心 $\mu_j = \dfrac{1}{|C_j|} \sum\limits_{x \in C_j} G_F(x)$;

(4) 如果所有的 K 个粒质心向量都没有发生变化, 则转到步骤 (5);

(5) 输出簇划分 $C = (C_1, C_2, ..., C_K)$.

要注意 K 值的选择, 一般来说, 根据数据的先验知识选择一个合适的 K 值, 若没有先验知识, 则可以通过交叉验证选择一个合适的 K 值. 在确定 K 值后, 需要选择 K 个初始化的质心, 或者就是随机质心.K 个初始化质心的位置选择对最后的聚类结果和运行时间都有很大的影响. 因此, 需要选择合适的 K 个质心, 最好这些质心不能太近. 邻域粒化过程中也有一个超参数: 邻域参数 δ, 这个参数与数据相关, 一般选择较小的邻域参数.

4 实验分析

实验采用 CMC、Iris、Heart Disease、Wine、Yeast、Pima-indians-diabetes 六个 UCI 数据来验证本文算法的有效性, 其中包含线性可分和非线性可分的数据集, 非凸、稀疏及模糊的数据是非线性可分的.

由于每个数据集中的每个特征的值域是不同的, 所以数据预处理采用最大最小值归一化, 这样可以把每个特征的值域转变为在 [0,1] 内, 最大最小值归一化的公式为:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}. \tag{16}$$

在数据预处理之后, 根据邻域参数对数据进行粒化, 形成粒向量. 在计算两个粒向量之间距离的时候, 可以使用相对距离公式, 也可以使用绝对距离公式, 本次实验分别测试基于相对距离的粒 K 均值聚类算法和基于绝对距离的粒 K 均值聚类算法, 并 K 均值聚类及其他聚类算法进行比较来验证算法的聚类效果.

在本次实验中, 使用 Accuracy 和 FMI(Fowlkes Mallows Index) 两种常用的聚类性能评估指标作为本实验对比的精准度.

4.1 邻域参数的影响

不同的邻域参数粒化过程构造了不同的粒向量, 进而影响最后的聚类结果. 为了进一步了解粒 K 均值聚类算法中邻域参数的影响, 因此在每个数据集上以不同的邻域参数进行实验. 由于原始数据集都是有标签的数据, 因此以 Accuracy 作为聚类性能评估指标, 将聚类后的结果与实际的标签进行比较. 实验采取 0 到 1 间隔 0.05 的邻域参数来进行实验, 每个 UCI 数据集在不同邻域参数下的实验结果如图 1-4 所示.

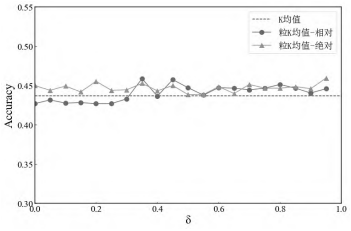

图 1 在 CMC 数据集上不同邻域参数聚类后的 Accuracy

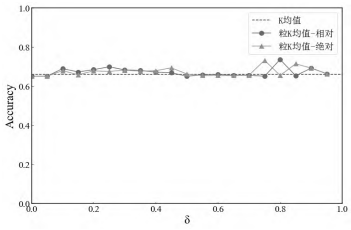

图 2　在 pima-indians-diabetes **数据集上不同邻域参数聚类后的** Accuracy

从图 1 和图 2 可以看出, 在 CMC 和 Pima-indians-diabetes 数据集的实验中, 粒 K 均值聚类算法的聚类性能得分 Accuracy 对邻域参数不是很敏感, 不同邻域参数对应的 Accuracy 差距较小. 对于 CMC 数据集, 基于相对距离的粒 K 均值聚类算法和基于绝对距离的粒 K 均值聚类算法的最高 Accuracy 分别为 0.4596 和 0.4542. 对于 Pima-indians-diabetes 数据集, 当邻域参数取到 0.8 时, 基于相对距离的粒 K 均值聚类算法的 Accuracy 达到了 0.75. 当邻域参数取到 0.75 时, 基于绝对距离的粒 K 均值聚类算法的 Accuracy 达到了 0.7331. 可以得出粒 K 均值聚类算法在这两个数据集上的 Accuracy 变化不大, 均在 K 均值聚类算法的 Accuracy 值上波动.

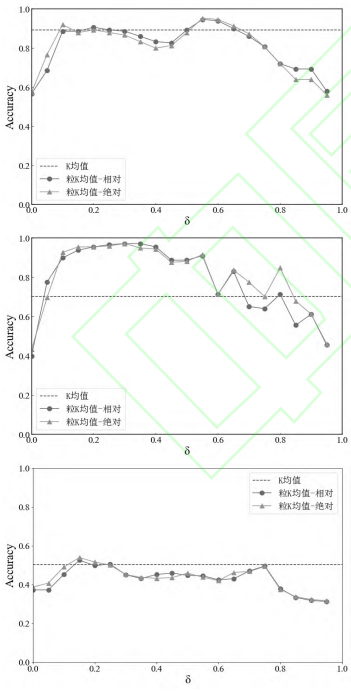

图 3　在 Iris、Wine **和** Yeast **数据集上不同邻域参数聚类后的** Accuracy

从图 3 可以看出, 在 Iris、Wine 和 Yeast 数据集的实验中, 基于相对距离的粒 K 均值聚类算法和基于绝对距离的粒 K 均值聚类算法的 Accuracy 基本保持一致且 Accuracy 曲线均呈现“凸”型, 可知过高或过低的邻域参数会使得粒 K 均值聚类算法的聚类性能降低. 对于 Iris 数据集, 当邻域参数取到 0.55 时, 基于两种距离的 K 均值聚类算法的 Accuracy 达到最高, 分别为 0.9467、0.9533. 对于 Wine 数据集, 基于相对距离的粒 K 均值聚类算法和基于绝对距离的粒 K 均值聚类算法的 Accuracy 均在邻域参数等于 0.3 时得到了相同的最高值 0.9719. 对于 Yeast 数据集, 虽然粒 K 均值聚类算法在邻域参数取到 0.15 时聚类性能要略优于 K 均值聚类算法, 但粒 K 均值聚类算法的聚类性能在总体上是不如 K 均值聚类算法的.

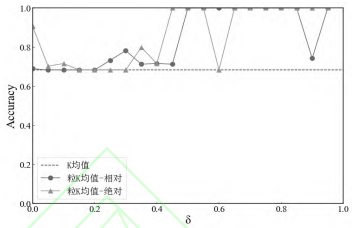

图 4　在 Heart Disease **数据集上不同邻域参数聚类后的** Accuracy

从图 4 可以看出, 在 Heart Disease 数据集的实验中, 不同的邻域参数导致 Accuracy 变化过大, 当邻域参数大于 0.5 时的粒 K 均值的 Accuracy 在整体上要明显高于邻域参数小于 0.5 的 Accuracy. 特别地, 当邻域参数大于 0.5 时粒 K 均值可以达到 Accuracy 为 1 的最好结果.

由图 1-4 可知: 从邻域参数的角度上看, 对于不同数据分布的数据集、不同的邻域参数都会对最终聚类性能造成影响. 从总体上看, 除了 Yeast 数据集, 粒 K 均值聚类算法的 Accuracy 均高于 K 均值聚类算法的 Accuracy, 均能找到合适的邻域参数使得 Accuracy 达到最高值, 超过 K 均值聚类算法. 与 K 均值聚类算法相比, 粒 K 均值聚类算法在算法进行之前就预先对数据进行粒化, 利用邻域粒向量使得算法无论是对于线性数据集还是非线性数据集都可以收敛的更快, 聚类性能更高.

4.2　聚类算法比较

本次实验将基于相对距离的粒 K 均值聚类算法和基于绝对距离的粒 K 均值聚类算法与 K 均值聚类算法、MeanShift 算法、Gaussian Mixture 算法、Birch 算法和 Agglomerative Clustering 算法在前述 6 个数据集上通过 Accuracy 和 FMI 的得分进行对比, 两种性能评估指标均表示为越接近 1 聚类性能越好. 由于 K 均值聚类算法和粒 K 均值聚类算法对初始化的聚类中心比较敏感, 容易导致聚类结果不稳定, 故对 K 均值聚类算法和粒 K 均值聚类算法在一个数据集上都运行 5 次, 选取最高的评估得分作为最后进行对比的得分, 如表 1 和表 2 所示.

表 1 各算法在不同数据集上以 Accuracy 作为性能评估指标的结果对比

数据集	粒 K 均值 (绝对距离)	粒 K 均值 (相对距离)	K 均值	MeanShift	Gaussian Mixture	Birch	Agglomerative Clustering
CMC	0.4542	**0.4596**	0.4345	0.4270	0.4277	0.4291	0.4297
Iris	0.9533	0.9467	0.8867	0.7933	**0.9667**	0.8867	0.8867
Heart Disease	**1.0000**	**1.0000**	0.6832	0.6832	0.6832	0.9769	0.6832
Wine	0.9719	0.9719	0.9551	0.6348	0.9663	**0.9775**	**0.9775**
Yeast	**0.5445**	0.5209	0.5370	0.3349	0.4636	0.4050	0.5047
Pima-indians-diabetes	0.7331	**0.7500**	0.6680	0.6523	0.6510	0.6901	0.6510

表 2 各算法在不同数据集上使用 FMI 作为性能评估指标的结果对比

数据集	粒 K 均值 (绝对距离)	粒 K 均值 (相对距离)	K 均值	MeanShift	Gaussian Mixture	Birch	Agglomerative Clustering
CMC	**0.5442**	0.5334	0.3629	0.5172	0.4780	0.4883	0.4303
Iris	0.9115	0.8999	0.8112	0.7476	**0.9356**	0.8159	0.8159
Heart Disease	**1.0000**	**1.0000**	0.5538	0.5709	0.5538	0.9607	0.5710
Wine	0.9425	0.9425	0.9126	0.6399	0.9310	**0.9543**	**0.9543**
Yeast	**0.4780**	0.4742	0.3075	0.4746	0.2857	0.3645	0.2984
Pima-indians-diabetes	**0.7383**	**0.7383**	0.5972	0.7380	0.5560	0.6918	0.5756

由表 1 可知, 当使用 Accuracy 作为性能评估指标时, 在 Heart Disease 和 Pima-indians-diabetes 数据集中粒 K 均值聚类算法的得分要高于其他五种算法的得分. 在 Yeast 数据集中, 基于绝对距离的粒 K 均值聚类算法的得分都要优于其他六种算法的得分. 在 CMC 数据集中, 基于相对距离的粒 K 均值聚类算法的得分都要优于其他六种算法的得分. 在 Iris 和 Wine 数据集中基于绝对距离的粒 K 均值聚类算法得分虽然都大于 K 均值聚类算法, 但是基于两种距离的粒 K 均值聚类算法得分都分别低于 Gaussian Mixture 算法和 Birch 算法、Agglomerative Clustering 算法.

由表 2 可知, 当使用 FMI 作为性能评估指标时, 在 Heart Disease 和 Pima-indians-diabetes 数据集中粒 K 均值聚类算法的得分都要高于其他五种算法的得分. 在 CMC 和 Yeast 数据集中, 基于绝对距离的粒 K 均值聚类算法的得分都要高于其他六种算法的得分. 在 Iris 数据集中 Gaussian Mixture 算法以 0.9356 的得分高于其他算法的得分. 在 Wine 数据集中 Birch 算法和 Agglomerative Clustering 算法的得分为最高, 分值为 0.9543.

从以上实验可知, 在大部分特征数和类别数较小的数据集上粒 K 均值聚类算法的聚类性能均优于 K 均值聚类算法, 优于大部分的其他算法. 在特征数和类别数比较大的数据集上 (比如 Wine 数据集) 聚类性能要劣于 Birch 算法和 Agglomerative Clustering 算法, 但从性能评估指标得分上来看, 粒 K 均值聚类算法与这两种算法也相差不大. 与传统的算法不同, 粒 K 均值聚类算法利用邻域粒化技术在结构上作出突破, 让数据在算法进行之前就提前得到处理, 提升了算法的收敛速度和聚类性能, 使得算法对于不同类型的数据集都有一个不错的效果.

5 总结

本文将 K 均值聚类算法和粒计算相结合, 通过邻域粒化技术构造基于单特征粒化的邻域粒子、基于多特征粒化的邻域粒向量, 并定义了邻域粒子与邻域粒向量的大小、度量和运算规则, 提出两种邻域粒距离度量. 进一步将邻域粒向量及其运算方式引入到 K 均值聚类算法中, 设计邻域粒 K 均值聚类算法. 由于粒化是在整个样本空间范围内进行, 使得粒向量具有全局的特性, 提高了聚类的精准度. 最后, 利用 UCI 数据集实验验证了邻域粒 K 均值聚类算法的有效性和正确性, 与 K 均值聚类算法相比, 邻域粒 K 均值聚类算法可以得到更好的聚类结果, 更高的聚类效率.

参考文献 (References)

[1] 安秋生, 沈钧毅, 王国胤. 基于信息粒度与 Rough 集的聚类方法研究 [J]. 模式识别与人工智能, 2003, 16(4): 412-417.
(An Q S, Shen J Y, Wang G Y. A clustering method based on information granularity and rough seis[J]. Pattern Recognition and Artificial Intelligence, 2003, 16(4): 412-417.)

[2] 张腾飞, 陈龙, 李云. 基于簇内不平衡度量的粗糙 K-means 聚类算法 [J]. 控制与决策, 2013, 28(10): 1479-1484.
(Zhang T F, Chen L, Li Y. Rough K-means clustering based on unbalanced degree of cluster[J]. Control and Decision, 2013, 28(10): 1479-1484.)

[3] 卜东波, 白硕, 李国杰. 聚类/分类中的粒度原理 [J]. 计算机学报, 2002, 25(8): 810-816.
(Bu D B, Bai S, Li G J. Principle of granularity

in clustering and classification[J]. Chinese Journal of Computers, 2002, 25(8): 810-816.)

[4] 陶莹, 杨锋, 刘洋, 等. K 均值聚类算法的研究与优化 [J]. 计算机技术与发展, 2018, 28(6): 90-92.
(Tao Y, Yang F ,Liu Y, Dai B, et al. Research and optimization of K-means clustering algorithm[J]. Computer Technology and Development, 2018, 28(6): 90-92.)

[5] Hung C H, Chiou H M, Yang W N. Candidate groups search for K-harmonic means data clustering[J]. Applied Mathematical Modelling, 2013, 37(24): 10123-10128.

[6] Abdeyazdan M. Data clustering based on hybrid K-harmonic means and modifier imperialist competitive algorithm[J]. J Supercomput, 2014, 68(2): 574-598.

[7] 王国胤, 张清华, 胡军. 粒计算研究综述 [J]. 智能系统学报, 2007, 2(6): 8-26.
(Wang G Y, Zhang Q H, Hu J. An overview of granular computing[J]. Transactions on Intelligent Systems, 2007, 2(6): 8-26.)

[8] Zadeh L A. Fuzzy sets and information granularity[J]. Fuzzy Sets, Fuzzy Logic, and Fuzzy Systems,1996, 8: 433-448.

[9] Lin T Y. Granular computing on binary relations I: data mining and neighborhood systems[J]. Rough Sets in Knowledge Discovery, 1998, 2: 165-166.

[10] 苗夺谦. Rough Set 理论及其在机器学习中的应用研究 [D]. 北京: 中国科学院自动化研究所, 1997.
(Miao D Q. Research on rough set theory and its application in machine learning[D]. Beijing: Institute of Automation,Chinese Academy of Sciences, 1997.)

[11] 苗夺谦, 王珏. 粗糙集理论中知识粗糙性与信息熵关系的讨论 [J]. 模式识别与人工智能, 1998, 11(1): 34-40.
(Miao D Q, Wang J. On the relationships between information entropy and roughness of knowledge in rough set theory[J]. Pattern Recognition and Artificial Intelligence, 1998, 11(1): 34-40.)

[12] 苗夺谦, 王珏. 粗糙集理论中概念与运算的信息表示 [J]. 软件学报, 1999(2): 2-5.
(Miao D Q, Wang J. An information representation of the concepts and operations in rough set theory[J]. Journal of Software, 1999, 2: 2-5.)

[13] 陈祥焰, 林耀进, 王晨曦. 基于邻域粗糙集的高维类不平衡数据在线流特征选择 [J]. 模式识别与人工智能, 2019, 32(8): 726-735.
(Chen X Y, Lin Y J, Wang C X. Online streaming feature selection for high-dimensional and class-imbalanced data based on neighborhood rough set[J]. Pattern Recognition and Artificial Intelligence, 2019, 32(8): 726-735.)

[14] 白盛兴, 林耀进, 王晨曦, 陈晟煜. 基于邻域粗糙集的大规模层次分类在线流特征选择 [J]. 模式识别与人工智能, 2019, 32(9): 811-820.
(Bai S X, Lin Y J, Wang C X, Chen S Y. Large-Scale hierarchical classification online streaming feature selection based on neighborhood rough set[J]. Pattern Recognition and Artificial Intelligence, 2019, 32(9): 811-820.)

[15] 苗夺谦, 张清华, 钱宇华, 等. 从人类智能到机器实现模型–粒计算理论与方法 [J]. 智能系统学报, 2016, 11(6): 743-757.
(Miao D Q, Zhang Q H, Qian Y H, et al. From human intelligence to machine implementation model: theories and applications based on granular computing[J]. CAAI Transactions on Intelligent Systems, 2016, 11(6): 743-757.)

[16] 廖纪勇, 吴晟, 刘爱莲. 基于相异性度量选取初始聚类中心改进的 K-means 聚类算法 [J]. 控制与决策, 2021, 1-8.
(Liao J Y, Wu S, Liu A L. Improved K-means clustering algorithm for selecting initial clustering centers based on dissimilarity measure[J]. Control and Decision, 2021, 1-8.)

[17] 李玥, 穆维松, 褚晓泉, 傅泽田. 基于改进量子粒子群的 K-means 聚类算法及其应用 [J]. 控制与决策, 2021, 1-10.
(Li Y, Mu W S,Chu X Q,Fu Z T. K-means clustering algorithm based on improved quantum particle swarm optimization and its application[J]. Control and Decision, 2021, 1-10.)

[18] 马福民, 孙静勇, 张腾飞. 考虑边界样本邻域归属信息的粗糙 K-Means 增量聚类算法 [J]. 控制与决策, 2021, 1-9.
(Ma F M, Sun J Y, Zhang T F. Rough K-Means incremental clustering algorithm considering neighborhood belonging information of boundary samples[J]. Control and Decision, 2021, 1-9.)

[19] Chen Y M, Qin N,Li W, Xu F F. Granule structures, distances and measures in neighborhood systems[J]. Knowledge-Based Systems, 2019, 165: 268-281.

作者简介

陈玉明 (1977–), 男, 教授, 博士生导师, 从事粒计算、机器学习等研究, E-mail: ymchen@xmut.edu.cn;

蔡国强 (1997–), 男, 硕士生, 从事粒计算、机器学习等研究, E-mail: 530279570@qq.com;

卢俊文 (1981–), 男, 高级实验师, 硕士生导师, 从事软件评测、机器学习等研究, E-mail: jwlu@xmut.edu.cn;

曾念峰 (1986–), 男, 硕士生导师, 从事人脸识别、系统架构等研究, E-mail: 395373664@qq.com.