控制与决策
Control and Decision

一种邻域粒*K*均值聚类方法

陈玉明, 蔡国强, 卢俊文, 曾念峰

引用本文:
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]. 控制与决策, 2023, 38(3): 857−864.

在线阅读 View online: 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,对于非凸、稀疏及模糊的非线性可分数据,其聚类效果不佳. 针对此问题,通过引入粒计算理论,采用邻域粒化技术,提出一种邻域粒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公理化证明. 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比较,结果验证了所提出的邻域粒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):**

引用格式: 陈玉明,蔡国强,卢俊文,等. 一种邻域粒K均值聚类方法[J]. 控制与决策, 2023, 38(3): 857-864.

A neighborhood granular K-means clustering method

CHEN Yu-ming[1†], *CAI Guo-qiang*[1], *LU Jun-wen*[1], *ZENG Nian-feng*[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: For non-convex, sparse and fuzzy nonlinear separable data, the clustering effect of K-means clustering is not good. Therefore, 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 measurements 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的过程,并且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[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[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[4]. 快速、简单的特性使其成为聚类中最常用的方法[5]. K均值聚类的核心思想是使得一个类簇中数据之间的总体差异小于邻近类簇中心的差异[6].

1979年,Zadeh[7]提出信息粒度化的思想,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

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项目(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.

点研究领域[8]. Zadeh一直强调信息粒化的重要性,并且认为需要充分探索信息粒化的计算理论. 随后,粒计算(granular computing)的概念被Lin提出[9],并引起了许多研究学者的兴趣. 其中,苗夺谦教授等[10-12]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. 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均值聚类是一种使用广泛的聚类分析方法,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]利用构建相异性矩阵来优化初始聚类中心的选取;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[18]将粗糙集与K均值聚类算法相结合. 本文通过在算法结构上进行改进,将K均值聚类与粒计算相结合,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. 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数据集进行的实验测试表明,本文算法得到的聚类效果优于K均值聚类算法,为聚类算法探索了一条新的途径.

1 邻域粒向量表示

设数据集IS $=(U,F)$,样本集$U=\{x_1,x_2,\dots,x_n\}$,属性集$F=\{a_1,a_2,\dots,a_m\}$. 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

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

$\varphi(x,y)=1$,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

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

其中$r_j=\varphi(x,x_j)$为样本x和x_j的邻域判别函数,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$P=\{a_1,a_2,\dots,a_m\}$,则x在属性子集P上的邻域粒向量定义为

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

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,邻域粒子是由0或1构成的有序集合,表示样本之间的邻域关系. 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

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

易知邻域粒子的大小满足

$$1\leqslant|(g_a(x)|\leqslant n.$$

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$P=\{a_1,a_2,\dots,a_m\}$,则x的邻域粒向量$G_P(x)$的大小定义为

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

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

$$\sqrt{m}\leqslant|G_P(x)|\leqslant n\times\sqrt{m}.$$

2 邻域粒距离度量

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

$$G_F(x)\bigwedge G_F(y)=(g_1(x)\bigwedge g_1(y),g_2(x)\bigwedge g_2(y),\dots,g_m(x)\bigwedge g_m(y))^{\mathrm{T}}, \tag{7}$$

$$G_F(x)\bigvee G_F(y)=(g_1(x)\bigvee g_1(y),g_2(x)\bigvee g_2(y),\dots,g_m(x)\bigvee g_m(y))^{\mathrm{T}}, \tag{8}$$

$$G_F(x)-G_F(y)=$$
$$(g_1(x)-g_1(y),g_2(x)-g_2(y),\dots,g_m(x)-g_m(y))^{\mathrm{T}},\tag{9}$$

$$G_F(x)\oplus G_F(y)=$$
$$(g_1(x)\oplus g_1(y),g_2(x)\oplus g_2(y),\dots,g_m(x)\oplus g_m(y))^{\mathrm{T}}.\tag{10}$$

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

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

其中$|F|=m$. 易知,邻域粒向量的相对距离满足

$$0\leqslant d(G_F(x),G_F(y))\leqslant 1.$$

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

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

其中:$|F|=m,|U|=n$. 易知,邻域粒向量的绝对距离满足

$$0\leqslant h(G_F(x),G_F(y))\leqslant 1.$$

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3个性质:

1) 非负.

$$0\leqslant d(G_F(x),G_F(y))\leqslant 1.$$

2) 对称.

$$d(G_F(x),G_F(y))=d(G_F(y),G_F(x)).$$

3) 三角不等式.

$$d(G_F(x),G_F(y))+d(G_F(y),G_F(z))\geqslant d(G_F(x),G_F(z)).$$

证明 1) 由$g_i(x)\oplus g_i(y)=g_i(x)\bigvee g_i(y)-g_i(x)\bigwedge g_i(y)$,可知

$$\frac{|g_i(x)\oplus g_i(y)|}{|g_i(x)\bigvee g_i(y)|}=\frac{|g_i(x)\bigvee g_i(y)-g_i(x)\bigwedge g_i(y)|}{|g_i(x)\bigvee g_i(y)|},$$

则

$$0\leqslant\frac{|g_i(x)\oplus g_i(y)|}{|g_i(x)\bigvee g_i(y)|}\leqslant 1.$$

由$F=\{a_1,a_2,\dots,a_m\}$,可知$|F|=m$. 因此,

$$0\leqslant\sum_{i=1}^{m}\frac{|g_i(x)\oplus g_i(y)|}{|g_i(x)\bigvee g_i(y)|}\leqslant m,$$

则

$$0\leqslant\frac{1}{m}\sum_{i=1}^{m}\frac{|g_i(x)\oplus g_i(y)|}{|g_i(x)\bigvee g_i(y)|}\leqslant 1.$$

所以,$0\leqslant d(G_F(x),G_F(y))\leqslant 1$成立.

2) 由$g_i(x)\bigvee g_i(y)=g_i(y)\bigvee g_i(x),g_i(x)\bigwedge g_i(y)=g_i(y)\bigwedge g_i(x)$,可知

$$\frac{1}{m}\sum_{i=1}^{m}\frac{|g_i(x)\oplus g_i(y)|}{|g_i(x)\bigvee g_i(y)|}=\frac{1}{m}\sum_{i=1}^{m}\frac{|g_i(y)\oplus g_i(x)|}{|g_i(y)\bigvee g_i(x)|}.$$

因此,$d(G_F(x),G_F(y))=d(G_F(y),G_F(x))$成立.

3) 由文献[19]中的命题3可知

$$\frac{|g_i(x)\oplus g_i(y)|}{|g_i(x)\bigvee g_i(y)|}+\frac{|g_i(y)\oplus g_i(z)|}{|g_i(y)\bigvee g_i(z)|}\geqslant\frac{|g_i(x)\oplus g_i(z)|}{|g_i(x)\bigvee g_i(z)|},$$

因此

$$\frac{1}{m}\sum_{i=1}^{m}\frac{|g_i(x)\oplus g_i(y)|}{|g_i(x)\bigvee g_i(y)|}+\frac{1}{m}\sum_{i=1}^{m}\frac{|g_i(y)\oplus g_i(z)|}{|g_i(y)\bigvee g_i(z)|}\geqslant\frac{1}{m}\sum_{i=1}^{m}\frac{|g_i(x)\oplus g_i(z)|}{|g_i(x)\bigvee g_i(z)|}$$

成立.

由邻域粒向量的相对距离定义可知,$d(G_F(x),G_F(y))+d(G_F(y),G_F(z))\geqslant d(G_F(x),G_F(z))$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3个性质:

1) 非负.

$$0\leqslant h(G_F(x),G_F(y))\leqslant 1.$$

2) 对称.

$$h(G_F(x),G_F(y))=h(G_F(y),G_F(x)).$$

3) 三角不等式.

$$h(G_F(x),G_F(y))+h(G_F(y),G_F(z))\geqslant h(G_F(x),G_F(z)).$$

证明 1) 由$g_i(x)\oplus g_i(y)=g_i(x)\bigvee g_i(y)-g_i(x)\bigwedge g_i(y),1\leqslant|g_i(x)|\leqslant n$,可知

$$0\leqslant|g_i(x)\oplus g_i(y)|\leqslant n.$$

由$F=\{a_1,a_2,\dots,a_m\}$,可知$|F|=m$. 因此,

$$0\leqslant\sum_{i=1}^{m}|g_i(x)\oplus g_i(y)|\leqslant m\times n,$$

则

$$0\leqslant\frac{1}{m\times n}\sum_{i=1}^{m}|g_i(x)\oplus g_i(y)|\leqslant 1.$$

所以,$0 \leqslant h(G_F(x), G_F(y)) \leqslant 1$成立.

2) 由$g_i(x) \bigvee g_i(y) = g_i(y) \bigvee g_i(x), g_i(x) \bigwedge g_i(y) = g_i(y) \bigwedge g_i(x)$,可知

$$\frac{1}{m \times n} \sum_{i=1}^{m} |g_i(x) \oplus g_i(y)| = \frac{1}{m \times n} \sum_{i=1}^{m} |g_i(y) \oplus g_i(x)|.$$

因此,$h(G_F(x), G_F(y)) = h(G_F(y), G_F(x))$成立.

3) 由文献[19]中的命题6可知

$$|g_i(x) \oplus g_i(y)| + |g_i(y) \oplus g_i(z)| \geqslant |g_i(x) \oplus g_i(z)|,$$

因此

$$\frac{1}{m \times n} \sum_{i=1}^{m} |g_i(x) \oplus g_i(y)| + \frac{1}{m \times n} \sum_{i=1}^{m} |g_i(y) \oplus g_i(z)| \geqslant \frac{1}{m \times n} \sum_{i=1}^{m} |g_i(x) \oplus g_i(z)|$$

成立.

由邻域粒向量的绝对距离定义可知,$h(G_F(x), G_F(y)) + h(G_F(y), G_F(z)) \geqslant h(G_F(x), G_F(z))$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;使簇内的粒向量尽量紧密地连在一起,而簇间的粒向量距离尽量地大.

设样本划分为$(C_1, C_2, \ldots, 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:$h(G_F(x), \mu_i)$为样本x的粒向量与粒质心的绝对距离;μ_i为C_i簇的均值粒向量,也称为粒质心.

损失函数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表示C_i簇中样本的个数,$G_F(x)$表示样本x的粒向量.

粒K均值聚类的目标是使J_e损失函数最小,故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.

算法1 邻域粒K均值聚类算法.

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

输出: 簇划分$C = (C_1, C_2, \ldots, C_K)$.

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

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

3) for $t = 1$ to N

① 将簇划分初始化为$C_j = \varnothing, j = 1, 2, \ldots, K$.

② 对于$i = 1, 2, \ldots, n$,计算邻域粒向量$G_F(x_i)$和各个粒质心向量$\mu_j (j = 1, 2, \ldots, K)$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$\lambda j$,此时更新$C_{\lambda j} = C_{\lambda j} \bigcup x_i$.

③ 对于$j = 1, 2, \ldots, K$,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,则转到下一步.

5) 输出簇划分$C = (C_1, C_2, \ldots, C_K)$.

要注意K值的选择,一般而言,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,且这些质心不能太近. 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共6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.

因为每个数据集中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,这样可以将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

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

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的距离时,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,并将K均值聚类与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,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,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～图4所示.

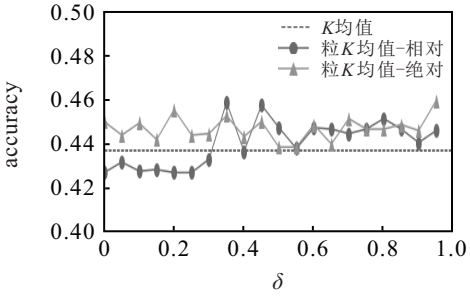


图1　在CMC数据集上不同邻域参数聚类后的accuracy

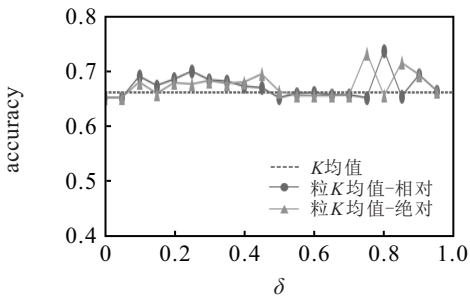
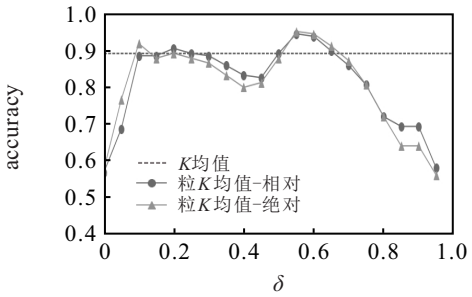
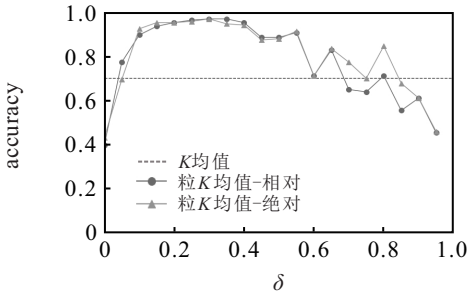


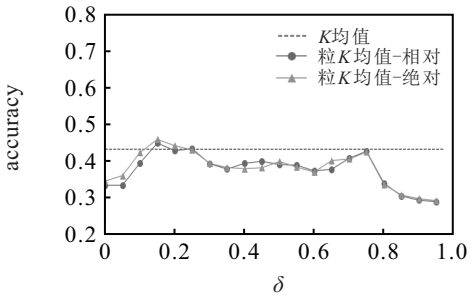
图2　在pima-indians-diabetes数据集上不同邻域参数聚类后的accuracy



(a) 在Iris数据集



(b) 在Wine数据集



(c) 在Yeast数据集

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

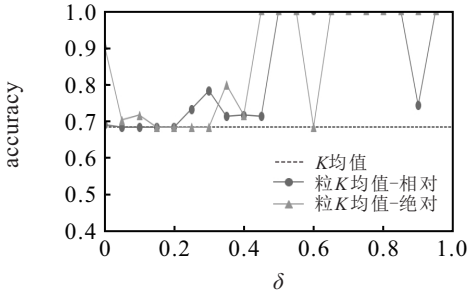


图4　在Heart Disease数据集上不同邻域参数聚类后的accuracy

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0.459 6和0.454 2. 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0.733 1. 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.

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0.946 7、0.953 3. 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0.971 9. 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.

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,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～图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,且超过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得更快,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均值聚类算法在每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	0.454 2	**0.459 6**	0.434 5	0.427 0	0.427 7	0.429 1	0.429 7
Iris	0.953 3	0.946 7	0.886 7	0.793 3	**0.966 7**	0.886 7	0.886 7
Heart Disease	**1.000 0**	**1.000 0**	0.683 2	0.683 2	0.683 2	0.976 9	0.683 2
Wine	0.971 9	0.971 9	0.955 1	0.634 8	0.966 3	**0.977 5**	**0.977 5**
Yeast	**0.544 5**	0.520 9	0.537 0	0.334 9	0.463 6	0.405 0	0.504 7
Pima-indians-diabetes	0.733 1	**0.750 0**	0.668 0	0.652 3	0.651 0	0.690 1	0.651 0

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	**0.544 2**	0.533 4	0.362 9	0.517 2	0.478 0	0.488 3	0.430 3
Iris	0.911 5	0.899 9	0.811 2	0.747 6	**0.935 6**	0.815 9	0.815 9
Heart Disease	**1.000 0**	**1.000 0**	0.553 8	0.570 9	0.553 8	0.960 7	0.571 0
Wine	0.942 5	0.942 5	0.912 6	0.639 9	0.931 0	**0.954 3**	**0.954 3**
Yeast	**0.478 0**	0.474 2	0.307 5	0.474 6	0.285 7	0.364 5	0.298 4
Pima-indians-diabetes	**0.738 3**	**0.738 3**	0.597 2	0.738 0	0.556 0	0.691 8	0.575 6

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5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6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6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5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6种算法的得分; 在Iris数据集中,Gaussian Mixture算法0.935 6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0.954 3.

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 结 论

本文将K均值聚类算法与粒计算相结合,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,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 sets[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, 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]. The Journal of Supercomputing, 2014, 68(2): 574-598.

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

[8] 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]. Caai Transactions on Intelligent Systems, 2007, 2(6): 8-26.)

[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, 10(2): 113-116.
(Miao D Q, Wang J. An information representation of the concepts and operations in rough set theory[J]. Journal of

Software, 1999, 10(2): 113-116.)

[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, 等. 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, et al. 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, 36(12): 3083-3090.
(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, 36(12): 3083-3090.)

[17] 李玥, 穆维松, 褚晓泉, 等. 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, 2022, 37(4): 839-850.
(Li Y, Mu W S, Chu X Q, et al. K-means clustering algorithm based on improved quantum particle swarm optimization and its application[J]. Control and Decision, 2022, 37(4): 839-850.)

[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, 2022, 37(11): 2968-2976.
(Ma F M, Sun J Y, Zhang T F. Rough K-means incremental clustering algorithm consideringneighborhood belonging information of boundary samples[J]. Control and Decision, 2022, 37(11): 2968-2976.)

[19] Chen Y M, Qin N, Li W, et al. 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、机器学习的研究, E-mail: 530279570@qq.com;

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

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

(责任编辑：齐 霁)

下 期 要 目