DOI: 10.13718/j.cnki.xdzk.2024.05.001
何宇豪，陈颖悦，曾高发，等. 一种邻域粒谱聚类方法[J]. 西南大学学报(自然科学版)，2024，46(5)：2-10.

一种邻域粒谱聚类方法

何宇豪[1]，　陈颖悦[1]，　曾高发[2]，　刘培谦[3]

1. 厦门理工学院 计算机与信息工程学院，福建 厦门 361024；2. 厦门市执象智能科技有限公司，福建 厦门 361000；
3. 厦门理工学院 经济与管理学院，福建 厦门 361024

摘要：谱聚类是一种无监督学习的聚类方法，其具有能够收敛至全局最优且适用于任意形状样本空间的优点. 然而，传统方法构造的相似矩阵有时难以准确反映出数据之间的近似关系，从而导致聚类结果不佳. 粒计算技术能够很好地解决这一问题. 通过将数据邻域粒化，从粒子的视角重新衡量数据之间的近似关系，提出了一种基于邻域粒的谱聚类方法. 首先，将样本的单一属性通过邻域粒化的方式形成邻域粒子；然后，将属于同一样本的粒子组合构造成粒子向量；接着，利用定义的 2 种邻域粒距离公式，对构造出的粒向量进行距离度量，并通过径向基函数生成相似矩阵，从而进行谱聚类；最后，使用 UCI 数据集进行验证，将谱聚类算法与邻域粒结合，从邻域参数和邻域粒向量的距离度量方式 2 个方面进行性能测试，并与传统聚类算法进行对比. 实验结果表明，基于邻域粒构造的相似矩阵在谱聚类中是可行且有效的.

关　键　词：粒计算；谱聚类；聚类；邻域；粒向量
中图分类号：TP391　　**文献标志码**：A
文 章 编 号：1673-9868(2024)05-0002-09
开放科学(资源服务)标识码(OSID)：

A Neighborhood Granular Spectral Clustering Method

HE Yuhao[1]，　CHEN Yingyue[1]，　ZENG Gaofa[2]，　LIU Peiqian[3]

1. *College of Computer and Information Engineering*，*Xiamen University of Technology*，*Xiamen Fujian 361024*，*China*；
2. *Zhixiang Intelligent Technology Co. Ltd.*，*Xiamen Fujian 361000*，*China*；
3. *School of Economics and Management*，*Xiamen University of Technology*，*Xiamen Fujian 361024*，*China*

Abstract: Spectral clustering is an unsupervised learning clustering method, which has the advantages of convergence to the global optimum and is applicable to arbitrary shape sample space. However, the similarity matrix constructed by traditional methods can not reflect the approximate relation between data sometimes, so the clustering results are not good. Granular computing can solve this problem well. By granulating the data in the neighborhood and re-measuring the approximate relation between the data from the perspective of granules, a spectral clustering method based on neighborhood granules is proposed. Firstly, the single attribute of the sample was formed into the neighborhood granules by the way of neighborhood granulation, and then the granule vector was constructed by combining the granules belonging to

收稿日期：2023-05-29
基金项目：国家自然科学基金项目(61976183)；厦门市科技计划项目(2022CXY0428)；住房和城乡建设部研究开发项目(2022-K-169).
作者简介：何宇豪，硕士研究生，主要从事粒计算、机器学习与数据挖掘方面的研究.

the same sample. By using two kinds of neighborhood granule distance formula defined, the constructed grain vector was measured by distance, and the radial basis function was used to generate a similar matrix for spectral clustering. Finally, the performance of the spectral clustering algorithm combined with neighborhood granules was tested using the UCI datasets for validation. The algorithm's performance was evaluated in two aspects: neighborhood parameters and distance measurement methods of neighborhood granule vectors. The results were compared with those of traditional clustering algorithms. The experimental results showed that the similarity matrix constructed using neighborhood granulation is feasible and effective for spectral clustering.

Key words: granular computing; spectral clustering; clustering; neighborhood; granule vectors

Zadeh 通过发表的“Fuzzy sets and information granularity”[1]，对信息粒进行了定义，开启了信息粒度化思想的探索. 随后，Lin[2]提出了粒计算(granular computing)的概念. Yao 等[3]则不断深入研究粒计算与粗糙集领域，并将其应用在知识挖掘、机器学习等领域. 21 世纪初，众多研究学者纷纷加入粒计算这一新兴研究领域，并不断取得突破与创新. 苗夺谦等[4]通过对粒计算理论与方法的研究，展示了粒计算在人工智能领域的前景与应用. Ruan 等[5]提出的信息粒化方法有助于减少时间序列分析的时间成本. 此外，朱鹏飞等[6-8]在邻域约简和分类等领域也有显著突破. 粒计算中的粒化概念是人类认识世界的重要组成部分，在特征选择、处理复杂数据等方面发挥着重要作用.

在人类活动和自然研究中，普遍存在聚类问题. 当前提出的聚类方法旨在对大量未知标注的数据进行分类，通过数据中个体与类簇相关的属性，将个体分类到各个类簇中. 在这个过程中，同一类簇中的个体属性更为相似. 人们常用的聚类算法[9]包括划分聚类、密度聚类、基于混合分布的聚类等. 其中，最为人熟知的是 k-means 算法[10-11]，它是一种划分聚类方法. 由于 k-means 算法易于实现且运行时间较短，因此成为一种广泛使用的聚类分析算法，并且在各个领域都取得了不错的效果. 然而，该算法也存在一些缺点，尤其是对于稀疏或局部分布不均匀的数据，其效果往往并不理想. 另一种常见的聚类方法是密度聚类[12]，其原理是只要区域内的点密度高于某一阈值，就将其归入相近的类簇中. 代表性的算法有 DBSCAN[13]以及针对其进行优化的 OPTICS[14]. 为了解决密度分布不均匀导致聚类效果不佳的问题，孙璐等[15]提出了 GFDBSCAN，Rodriguez 等[16]提出了基于密度峰值的聚类方法. 与划分聚类相比，密度聚类依赖于密度而不是距离的特点，因此能够对任意形状的簇进行聚类，但是其空间复杂度相较其他算法要高许多，计算开销更大. 基于混合分布的聚类方法是通过假设簇内符合多元正态分布，判断数据样本符合各个样本分布的概率，不断迭代更新，直至得到数据集的聚类情况. 这种方法主要被人们运用于图像聚类目标识别等领域，并不断加以改进. 此外，聚类还有基于网格聚类[18]、层次聚类[19]、谱聚类等算法.

谱聚类[20]是一种源自图论的聚类方法，它具有处理任意形状的数据集的能力，并且在稀疏数据聚类方面表现十分出色. 因此，谱聚类在数据挖掘、图像分割、模式识别以及遥感等领域中得到广泛应用. 近年来，由于谱聚类独特的优势，引起了学术界的广泛关注. 孔万增等[21]提出了利用矩阵的特征向量和本征间隙来自动确定类别个数的优化方法. Chen 等[22]对划分准则进行了改进，将算法的时间复杂度降至 $O(n^2c)$. 在谱聚类优化中，相似矩阵的构造方法是一个关键的焦点，因为谱聚类对相似矩阵极为敏感. 如果构造出的相似矩阵不能很好地反映数据之间的近似关系，其效果就难以保证. 因此，一个良好的相似矩阵的构造对于谱聚类算法至关重要. 邻域粒[23]能够有效地反映数据之间的近似关系，本文针对相似矩阵的构造，结合邻域粒及其距离度量方式，提出了一种新的相似矩阵构造方法，从而得到了基于邻域粒的谱聚类算法. 该方法首先通过在单一特征上对样本进行邻域粒化的方式形成邻域粒子，然后将属于同一样本的粒子组合构造成粒子向量. 通过定义邻域粒向量的距离度量，用粒向量的距离代替常规径向基核函数中的欧式距离，从而得到经过邻域粒化后的数据相似矩阵. 由于样本粒化是通过全局信息进行的，使得样本具有了全局性，从而能够更好地反映数据之间的近似关系. 最后，通过与传统聚类算法在 UCI 数据集上的比较，实验结果表明邻域粒化的方式在谱聚类中取得了不错的效果.

1 邻域粒向量表示

传统的聚类方法输入的数据是样本数据，本文通过粒子与粒向量的构造方法，将数据在单一特征上进行邻域粒化，获得了粒子并推广到多个特征上，通过数据的多个粒子进行组合，构造出粒向量. 并且在粒子与粒向量上，同样能进行数学运算.

设聚类系统为$C=(R, P)$，其中数据集合为$R=\{r_1, r_2, \cdots, r_n\}$，特征集合为$P=\{p_1, p_2, \cdots, p_m\}$. 对于确定数据$r \in R$，其单一特征$p \in P$，$v(r, p) \in [0, 1]$表示进行归一化后，数据$r$在单一特征$p$上的值.

对于聚类系统$C=(R, P)$，对于不同的数据$r_1, r_2 \in R$，其单一特征$p \in P$，则数据r_1和r_2在单一特征p的曼哈顿距离为

$$D_p(r_1, r_2)=|\ v(r_1, p)-v(r_2, p)\ | \tag{1}$$

其中：$v(r_1, p)$表示数据r_1在单一特征p上的值.

定义1 给定聚类系统$C=(R, P)$，对于不同的数据$r_1, r_2 \in R$，其单一特征$p \in P$，并设定邻域参数为δ，则定义数据与数据是否为邻域的判别函数为

$$\varphi(r_1, r_2)=\begin{cases}0 & D_p>\delta\\ 1 & D_p\leqslant\delta\end{cases} \tag{2}$$

其中：$\varphi(r_1, r_2)=1$表示数据与数据之间在该邻域范围内互为邻域；$\varphi(r_1, r_2)=0$则表示数据与数据在该邻域范围内不相邻.

定义2 给定聚类系统$C=(R, P)$，对于不同的数据$r_1, r_2 \in R$，其单一特征$p \in P$，则数据r在特征p上的邻域粒子定义为

$$g_p(r)=\{n_j\}_{j=1}^{n}=\{l_1, l_2, \cdots, l_n\} \tag{3}$$

上述$l_j=\varphi(r_1, r_2)$，代表二者是否相邻.

定义3 给定聚类系统$C=(R, P)$，对于确定数据$r \in R$，其有特征子集$A \subseteq P$，设$A=\{p_1, p_2, \cdots, p_m\}$，则$r$在特征子集$A$上的邻域粒向量定义为

$$\boldsymbol{G}_{\boldsymbol{A}}(\boldsymbol{r})=(g_1(r), g_2(r), \cdots, g_m(r))^{\mathrm{T}} \tag{4}$$

上述$g_m(r)$代表数据r在特征p_m上构造的邻域粒子.

由此可知，邻域粒子由0和1组成，表示数据在设定的邻域参数范围内的相邻关系. 由于邻域粒向量是由邻域粒子构成的，因此与传统向量的实数元素不同，其元素是由集合组合而成的.

定义4 给定聚类系统$C=(R, P)$，对于确定数据$r \in R$，其单一特征$p \in P$，则数据r在特征p上的邻域粒子$g_p(r)$的大小定义为

$$Size(g_p(r))=|\ g_p(r)\ |=\sum_{j=1}^{n} l_j \tag{5}$$

由邻域粒子的定义可知，其范围为：$1 \leqslant g_1(r) \leqslant n$.

定义5 给定聚类系统$C=(R, P)$，对于确定数据$r \in R$，其有特征子集$A \subseteq P$，设$A=\{p_1, p_2, \cdots, p_m\}$，则数据$r$在特征子集$A$上的邻域粒向量的大小定义为

$$Size(\boldsymbol{G}_{\boldsymbol{A}}(\boldsymbol{r}))=|\ \boldsymbol{G}_{\boldsymbol{A}}(\boldsymbol{r})\ |=\sqrt{\sum_{i=1}^{m}|\ g_i(r)\ |^2} \tag{6}$$

由邻域粒向量的定义可知，其大小范围为：$\sqrt{m} \leqslant|\ \boldsymbol{G}_{\boldsymbol{A}}(\boldsymbol{r})\ |\leqslant n * \sqrt{m}$.

2 邻域粒距离度量

定义6 给定聚类系统$C=(R, P)$，特征集合为$P=\{p_1, p_2, \cdots, p_m\}$. 对于不同数据$r_1, r_2 \in R$，其邻域粒向量在特征集合$P$上，分别表示为

$$\boldsymbol{G}_{\boldsymbol{P}}(\boldsymbol{r})=(g_1(r_1), g_2(r_1), \cdots, g_m(r_1))^{\mathrm{T}}$$

$$\boldsymbol{G}_{\boldsymbol{P}}(\boldsymbol{r})=(g_1(r_2), g_2(r_2), \cdots, g_m(r_2))^{\mathrm{T}}$$

则其加、减、交、并与异或运算定义为

$$\boldsymbol{G_P}(\boldsymbol{r}_1)+\boldsymbol{G_P}(\boldsymbol{r}_2)=(g_1(r_1)+g_1(r_2),\ g_2(r_1)+g_2(r_2),\ \cdots,\ g_m(r_1)+g_m(r_2))^{\mathrm{T}} \tag{7}$$

$$\boldsymbol{G_P}(\boldsymbol{r}_1)-\boldsymbol{G_P}(\boldsymbol{r}_2)=(g_1(r_1)-g_1(r_2),\ g_2(r_1)-g_2(r_2),\ \cdots,\ g_m(r_1)-g_m(r_2))^{\mathrm{T}} \tag{8}$$

$$\boldsymbol{G_P}(\boldsymbol{r}_1)\wedge\boldsymbol{G_P}(\boldsymbol{r}_2)=(g_1(r_1)\wedge g_1(r_2),\ g_2(r_1)\wedge g_2(r_2),\ \cdots,\ g_m(r_1)\wedge g_m(r_2))^{\mathrm{T}} \tag{9}$$

$$\boldsymbol{G_P}(\boldsymbol{r}_1)\vee\boldsymbol{G_P}(\boldsymbol{r}_2)=(g_1(r_1)\vee g_1(r_2),\ g_2(r_1)\vee g_2(r_2),\ \cdots,\ g_m(r_1)\vee g_m(r_2))^{\mathrm{T}} \tag{10}$$

$$\boldsymbol{G_P}(\boldsymbol{r}_1)\oplus\boldsymbol{G_P}(\boldsymbol{r}_2)=(g_1(r_1)\oplus g_1(r_2),\ g_2(r_1)\oplus g_2(r_2),\ \cdots,\ g_m(r_1)\oplus g_m(r_2))^{\mathrm{T}} \tag{11}$$

定义 7　给定聚类系统 $C=(R,\ P)$，特征集合为 $P=\{p_1,\ p_2,\ \cdots,\ p_m\}$. 对于不同数据 $r_1,\ r_2\in R$，其邻域粒向量在特征集合 P 上，分别表示为

$$\boldsymbol{G_P}(\boldsymbol{r})=(g_1(r_1),\ g_2(r_1),\ \cdots,\ g_m(r_1))^{\mathrm{T}}$$

$$\boldsymbol{G_P}(\boldsymbol{r})=(g_1(r_2),\ g_2(r_2),\ \cdots,\ g_m(r_2))^{\mathrm{T}}$$

则 2 个邻域粒向量的相对距离定义为

$$d(\boldsymbol{G_P}(\boldsymbol{r}_1),\ \boldsymbol{G_P}(\boldsymbol{r}_2))=\frac{1}{m}\sum_{i=1}^{m}\frac{|\ g_i(r_1)\oplus g_i(r_2)\ |}{|\ g_i(r_1)\vee g_i(r_2)\ |} \tag{12}$$

其中：$|\ P\ |=m$，由上述定义可知，其相对距离满足：$0\leqslant d(\boldsymbol{G_P}(\boldsymbol{r}_1),\ \boldsymbol{G_P}(\boldsymbol{r}_2))\leqslant 1$.

定义 8　给定聚类系统 $C=(R,\ P)$，特征集合为 $P=\{p_1,\ p_2,\ \cdots,\ p_m\}$. 对于不同数据 $r_1,\ r_2\in R$，其邻域粒向量在特征集合 P 上，分别表示为

$$\boldsymbol{G_P}(\boldsymbol{r})=(g_1(r_1),\ g_2(r_1),\ \cdots,\ g_m(r_1))^{\mathrm{T}}$$

$$\boldsymbol{G_P}(\boldsymbol{r})=(g_1(r_2),\ g_2(r_2),\ \cdots,\ g_m(r_2))^{\mathrm{T}}$$

则 2 个邻域粒向量的绝对距离定义为

$$h(\boldsymbol{G_P}(\boldsymbol{r}_1),\ \boldsymbol{G_P}(\boldsymbol{r}_2))=\frac{1}{m*n}\sum_{i=1}^{m}|\ g_i(r_1)\oplus g_i(r_2)\ | \tag{13}$$

其中：$|\ P\ |=m$，$|\ R\ |=n$. 由上述定义可知，其绝对距离满足：

$$0\leqslant h_p(\boldsymbol{G_P}(\boldsymbol{r}_1),\ \boldsymbol{G_P}(\boldsymbol{r}_2))\leqslant 1$$

定理 1　邻域粒向量的相对距离作为距离度量，要满足以下 3 个性质：

性质 1　非负性，$0\leqslant d(\boldsymbol{G_P}(\boldsymbol{r}_1),\ \boldsymbol{G_P}(\boldsymbol{r}_2))\leqslant 1$；

性质 2　对称性，$d(\boldsymbol{G_P}(\boldsymbol{r}_1),\ \boldsymbol{G_P}(\boldsymbol{r}_2))=d(\boldsymbol{G_P}(\boldsymbol{r}_2),\ \boldsymbol{G_P}(\boldsymbol{r}_1))$；

性质 3　三角不等式，$d(\boldsymbol{G_P}(\boldsymbol{r}_1),\ \boldsymbol{G_P}(\boldsymbol{r}_2))+d(\boldsymbol{G_P}(\boldsymbol{r}_2),\ \boldsymbol{G_P}(\boldsymbol{r}_3))\geqslant d(\boldsymbol{G_P}(\boldsymbol{r}_1),\ \boldsymbol{G_P}(\boldsymbol{r}_3))$.

关于以上性质证明如下：

1）由 $g_i(r_1)\oplus g_i(r_2)=g_i(r_1)\vee g_i(r_2)-g_i(r_1)\wedge g_i(r_2)$ 可知

$$\frac{|\ g_i(r_1)\oplus g_i(r_2)\ |}{|\ g_i(r_1)\vee g_i(r_2)\ |}=\frac{|\ g_i(r_1)\vee g_i(r_2)-g_i(r_1)\wedge g_i(r_2)\ |}{|\ g_i(r_1)\vee g_i(r_2)\ |}$$

则

$$0\leqslant\frac{|\ g_i(r_1)\oplus g_i(r_2)\ |}{|\ g_i(r_1)\vee g_i(r_2)\ |}\leqslant 1$$

由 $P=\{p_1,\ p_2,\ \cdots,\ p_m\}$ 可知 $|\ P\ |=m$，因此

$$0\leqslant\sum_{i=1}^{m}\frac{|\ g_i(r_1)\oplus g_i(r_2)\ |}{|\ g_i(r_1)\vee g_i(r_2)\ |}\leqslant m$$

则

$$0\leqslant\frac{1}{m}\sum_{i=1}^{m}\frac{|\ g_i(r_1)\oplus g_i(r_2)\ |}{|\ g_i(r_1)\vee g_i(r_2)\ |}\leqslant 1$$

所以，$0\leqslant d(\boldsymbol{G_P}(\boldsymbol{r}_1),\ \boldsymbol{G_P}(\boldsymbol{r}_2))\leqslant 1$ 成立.

2）因为 $g_i(r_1)\vee g_i(r_2)=g_i(r_2)\vee g_i(r_1)$，$g_i(r_1)\wedge g_i(r_2)=g_i(r_2)\wedge g_i(r_1)$，可知

$$\frac{1}{m}\sum_{i=1}^{m}\frac{|\ g_i(r_1)\oplus g_i(r_2)\ |}{|\ g_i(r_1)\vee g_i(r_2)\ |}=\frac{1}{m}\sum_{i=1}^{m}\frac{|\ g_i(r_2)\oplus g_i(r_1)\ |}{|\ g_i(r_2)\vee g_i(r_1)\ |}$$

因此，$d(\boldsymbol{G}_{\boldsymbol{P}}(\boldsymbol{r}_1),\boldsymbol{G}_{\boldsymbol{P}}(\boldsymbol{r}_2))=d(\boldsymbol{G}_{\boldsymbol{P}}(\boldsymbol{r}_2),\boldsymbol{G}_{\boldsymbol{P}}(\boldsymbol{r}_1))$ 成立.

3) 因

$$\frac{|g_i(r_1)\oplus g_i(r_2)|}{|g_i(r_1)\vee g_i(r_2)|}+\frac{|g_i(r_2)\oplus g_i(r_3)|}{|g_i(r_2)\vee g_i(r_3)|}\geqslant\frac{|g_i(r_1)\oplus g_i(r_3)|}{|g_i(r_1)\vee g_i(r_3)|}$$

所以

$$\frac{1}{m}\sum_{i=1}^{m}\frac{|g_i(r_1)\oplus g_i(r_2)|}{|g_i(r_1)\vee g_i(r_2)|}+\frac{1}{m}\sum_{i=1}^{m}\frac{|g_i(r_2)\oplus g_i(r_3)|}{|g_i(r_2)\vee g_i(r_3)|}\geqslant\frac{1}{m}\sum_{i=1}^{m}\frac{|g_i(r_1)\oplus g_i(r_3)|}{|g_i(r_1)\vee g_i(r_3)|}$$

成立，所以 $d(\boldsymbol{G}_{\boldsymbol{P}}(\boldsymbol{r}_1),\boldsymbol{G}_{\boldsymbol{P}}(\boldsymbol{r}_2))+d(\boldsymbol{G}_{\boldsymbol{P}}(\boldsymbol{r}_2),\boldsymbol{G}_{\boldsymbol{P}}(\boldsymbol{r}_3))\geqslant d(\boldsymbol{G}_{\boldsymbol{P}}(\boldsymbol{r}_1),\boldsymbol{G}_{\boldsymbol{P}}(\boldsymbol{r}_3))$ 成立.

同理可证，邻域粒的绝对距离也满足非负、对称、三角不等式以上3个性质.

3 基于邻域粒的谱聚类算法

基于邻域粒的谱聚类算法是一种方法，它首先将样本进行邻域粒化，然后通过计算邻域粒向量之间的距离来构造相似矩阵，最后进行正常的谱聚类. 不同于局部处理，邻域粒化是从全局出发对样本进行处理的. 邻域粒向量由邻域粒子构成，因此邻域粒化后的粒子与粒向量具有全局信息. 由于谱聚类对相似矩阵极为敏感，利用该方法构建的相似矩阵能更好地反映数据之间的近似关系，从而使得谱聚类达到更好的效果.

3.1 基于邻域粒的谱聚类原理

基于邻域粒的谱聚类算法的流程如下：首先对样本进行邻域粒化，将每个样本的单一特征邻域粒化为粒子，然后将这些粒子组合起来构成粒向量. 由于邻域粒化是通过全局信息进行的，因此同一簇内的粒向量之间的距离更为紧密，而不同簇之间的粒向量之间的距离相对更大. 基于这样的构造方式，得到的相似矩阵能够更好地反映数据之间的关系. 构建的邻接矩阵方法如下所示：

先将样本数据进行邻域粒化，粒化结果为 $\boldsymbol{GT}=(\boldsymbol{G}_{\boldsymbol{P}}(\boldsymbol{r}_1),\boldsymbol{G}_{\boldsymbol{P}}(\boldsymbol{r}_2),\cdots,\boldsymbol{G}_{\boldsymbol{P}}(\boldsymbol{r}_m))$. 此时，样本被邻域粒化后，可以采用粒向量的距离公式进行距离计算，并以此构建邻接矩阵 $\boldsymbol{W}$. 若采用粒向量的绝对距离，表示为 $h(\boldsymbol{G}_{\boldsymbol{P}}(\boldsymbol{r}_1),\boldsymbol{G}_{\boldsymbol{P}}(\boldsymbol{r}_2))$，此时邻接矩阵 $\boldsymbol{W}$ 定义如下：

$$\boldsymbol{W}_{ij}=S_{ij}=\exp\left(-\frac{h(\boldsymbol{G}_{\boldsymbol{P}}(\boldsymbol{r}_i),\boldsymbol{G}_{\boldsymbol{P}}(\boldsymbol{r}_j))}{2\sigma^2}\right)\tag{14}$$

若采用粒向量的相对距离，表示为 $d(\boldsymbol{G}_{\boldsymbol{P}}(\boldsymbol{r}_1),\boldsymbol{G}_{\boldsymbol{P}}(\boldsymbol{r}_2))$，此时邻接矩阵 $\boldsymbol{W}$ 定义如下：

$$\boldsymbol{W}_{ij}=S_{ij}=\exp\left(-\frac{d(\boldsymbol{G}_{\boldsymbol{P}}(\boldsymbol{r}_i),\boldsymbol{G}_{\boldsymbol{P}}(\boldsymbol{r}_j))}{2\sigma^2}\right)\tag{15}$$

谱聚类的主要思想是将所有的数据放置到图中，并用点表示，点之间的关系用边来表示. 其边的权值由点之间的相似性决定，如果相似性高，则权值高；如果相似性低，则权值低. 然后通过对图进行切割，确保子图内部的权值高，而子图与子图之间的权值低. 邻域粒化后的粒子具有全局性，能很好地满足这一思想，从而让谱聚类的性能更好，更有效地完成聚类任务.

3.2 邻域粒谱聚类算法

具体的算法如算法1所示.

算法1 邻域粒谱聚类算法.

输入：聚类系统 $C=(R,P)$，其中数据集合为 $R=\{r_1,r_2,\cdots,r_n\}$，特征集合为 $P=\{p_1,p_2,\cdots,p_m\}$，降维后的维度 K_1，聚类维数 K，邻域参数 δ.

输出：簇划分 $F=(f_1,f_2,\cdots,f_k)$.

① 样本集 U 邻域粒化为 $\boldsymbol{GT}=(\boldsymbol{G}_{\boldsymbol{P}}(\boldsymbol{r}_1),\boldsymbol{G}_{\boldsymbol{P}}(\boldsymbol{r}_2),\cdots,\boldsymbol{G}_{\boldsymbol{P}}(\boldsymbol{r}_m))$；

② 计算邻域粒向量 $\boldsymbol{G}_{\boldsymbol{P}}(\boldsymbol{r}_i)$ 之间的距离后通过邻域粒距离使用径向基函数构建相似矩阵 $\boldsymbol{W}$；

③ 通过相似矩阵 $\boldsymbol{W}$ 构建度矩阵 $\boldsymbol{D}$；

④ 计算出拉普拉斯矩阵 $\boldsymbol{L}$ 以及标准化拉普拉斯矩阵 $\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{L}\boldsymbol{D}^{-\frac{1}{2}}$；

⑤ 计算标准化的拉普拉斯矩阵最小的 K_1 个特征值所各自对应的特征向量 $\boldsymbol{v}$；

⑥ 将各自对应的特征向量 $\boldsymbol{v}$ 组成的矩阵按行标准化，最终组成 $n \times K_1$ 维的特征矩阵 $\boldsymbol{V}$；

⑦ 将特征矩阵 $\boldsymbol{V}$ 的每一行视为一个 K_1 维的样本，共 n 个样本，用聚类方法将其分为 K 类；

⑧ 输出簇划分 $F=(f_1, f_2, \cdots, f_k)$.

谱聚类算法的运作需要一个聚类维数 K，这通常需要借助于真实的标签信息来选取一个恰当的 K 值，或者使用肘部法等方法获得一个适合的 K 值. 此外，在邻域粒化过程中要对邻域参数 δ 进行设定，其与数据间的距离相关联，需要通过检验来确定一个合适的值.

4　实验分析

实验通过多个 UCI 数据集验证了提出的相似矩阵构造算法的有效性. 表 1 描述了各个 UCI 数据集具体的相关信息. 在实验过程中，对数据进行了归一化处理，以消除算法运行过程中奇异样本带来的不良影响. 归一化后，所有数据会介于[0，1]之间. 归一化公式如下：

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

在数据归一化后，将样本进行邻域粒化，形成粒向量. 在衡量粒向量之间距离时，有相对距离和绝对距离 2 种选择. 本次实验将分别验证相对距离和绝对距离的粒谱聚类算法的效果差别，并与使用谱聚类和其他传统聚类算法进行比较，以验证算法的聚类效果. 实验过程中，使用常用的聚类指标：聚类纯度(Cluster Purity，CP)和调整兰德指数(Adjusted Rand Index，ARI)作为评估指标，能够很好地展现对比实验的结果.

表 1　数据集信息

数据集	样本数	特征数	类别数
Iris	150	4	3
Glass	214	9	6
Wdbc	569	30	2
Pima-indians-diabetes	768	8	2
Heart	297	13	2
Seed	210	7	3

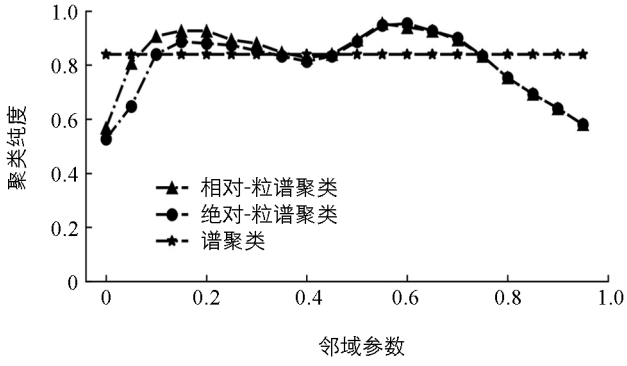
4.1　邻域参数和距离度量方式的影响

选用不同的邻域参数和距离度量方式对于样本之间的距离具有重要影响，可能导致构造出的相似矩阵结果大相径庭，从而直接影响聚类的效果. 为了深入了解在基于邻域粒的谱聚类算法中邻域参数和邻域粒距离度量对聚类效果的影响，我们进行了一系列实验. 选取的数据集具有真实标签，因此选用聚类纯度(CP)作为聚类性能评估指标. 鉴于之前对数据进行过归一化处理，将邻域参数设置为从 0 开始，以 0.05 为间隔，总共 20 组邻域参数，并分别采用邻域粒的相对距离和绝对距离进行实验. 不同数据集在这 20 组邻域参数和 2 种度量距离方式下的实验结果如图 1 所示：

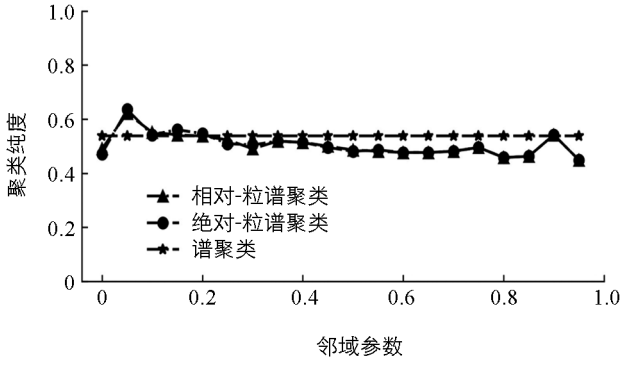
根据图 1(a)，(c)，(f)的实验结果，在 Iris、Wdbc 和 Seed 数据集上观察到以下情况：在 Iris 数据集中，当邻域参数为 0.55 时，2 种不同距离的谱聚类算法均达到了相同的最高值 0.953 3. 在 Wdbc 数据集上，当邻域参数为 0.2 时，基于邻域粒相对距离的谱聚类算法的 CP 值为 0.935，而基于邻域粒绝对距离的谱聚类算法的 CP 值达到了 0.933 2. 在 Seed 数据集上，当邻域参数为 0.3 时，基于邻域粒相对距离的谱聚类算法的 CP 值为 0.881，而基于邻域粒绝对距离的谱聚类算法的 CP 值达到了 0.895 2. 从结果可以看出，基于邻域粒相对距离构造的相似矩阵和基于绝对距离构造的相似矩阵在谱聚类算法中表现出的 CP 值最佳效果基本相同. 另外，CP 值曲线均呈现“凸”型，表明过高或过低的邻域参数会导致构造出的相似矩阵不能很好地反映数据的情况，进而影响聚类算法的性能并导致聚类性能降低.

根据图 1(b)，(d)，(e)的实验结果，在 Glass、Pima-indians-diabetes 和 Heart 数据集上观察到以下情况：在 Glass 数据集上，当邻域参数为 0.05 时，基于邻域粒绝对距离和相对距离的谱聚类算法的 CP 值分

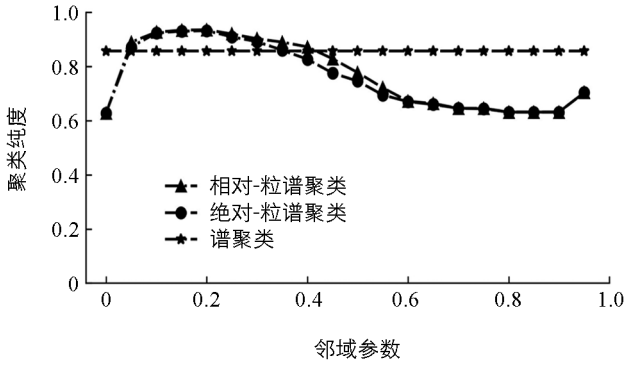
别达到了0.616 8和0.640 2，随后在传统谱聚类算法的CP值0.542 1附近波动. 在Pima-indians-diabetes数据集上，当邻域参数为0.4时，2种算法的CP分别达到最大值0.677 1和0.678 4. 在Heart数据集上，当邻域参数为0.3时，2种算法的CP值同时达到相同最大值0.835. 在这3个数据集中，粒谱聚类算法的聚类纯度CP对邻域参数不是很敏感，大致在传统谱聚类算法的CP值附近波动.



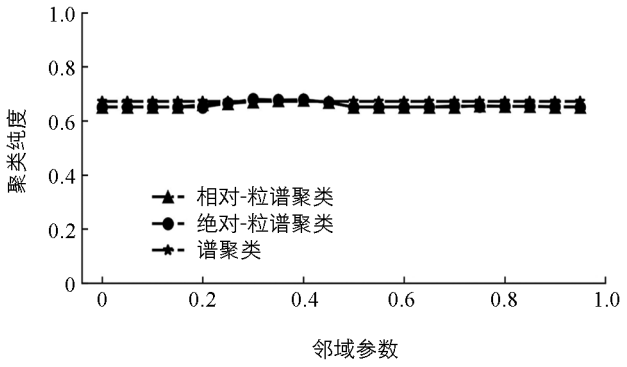
a. Iris数据集的聚类纯度CP



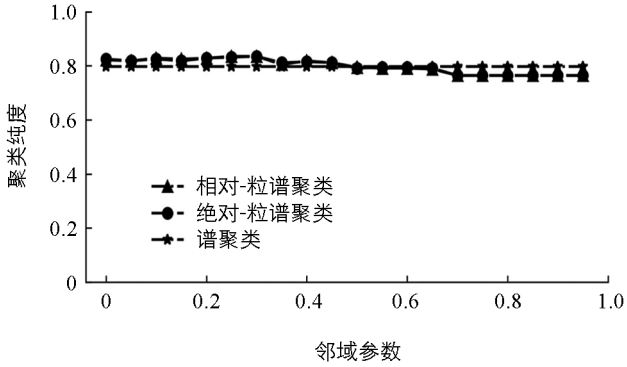
b. Glass数据集的聚类纯度CP



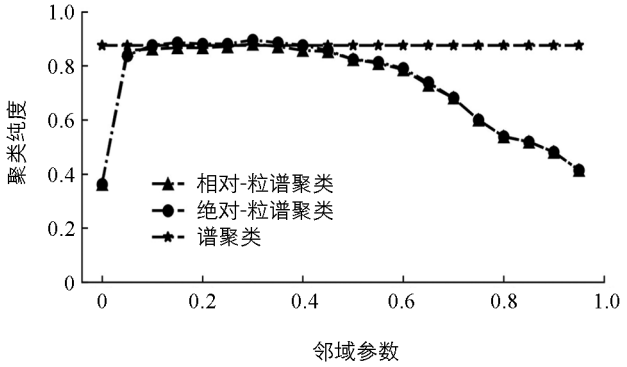
c. Wdbc数据集的聚类纯度CP



d. Pima-indians-diabetess数据集的聚类纯度CP



e. Heart数据集的聚类纯度CP



f. Seed数据集的聚类纯度CP

图1 各个数据集上在不同邻域参数的聚类纯度CP

综上所述，从邻域参数的角度来看，过大或过小的邻域参数都可能对聚类性能产生不良影响. 而从邻域粒距离度量的角度来看，距离度量对算法性能的影响不是很显著. 总体而言，存在适当的参数取值，可以使聚类效果达到最佳，甚至超过传统的谱聚类算法. 因此，与传统的谱聚类算法相比，利用邻域粒构建的相似矩阵能够在大多数数据集上取得更好的聚类性能效果.

4.2 聚类算法比较

本次实验将基于邻域粒相对距离和绝对距离的谱聚类算法与谱聚类算法、基于K近邻构造相似矩阵的谱聚类算法、GMM算法、Agglomerative Clustering算法和K-means算法进行对比. 采用上述UCI数据集，通过CP和ARI这2个聚类性能指标进行评价，其中2个指标的最大值为1，性能指标越高越好. 具体

对比结果如表 2 和表 3 所示.

表 2　选取的算法在不同数据集上的聚类纯度 CP 结果对比

数据集	粒谱聚类（相对距离）	粒谱聚类（绝对距离）	谱聚类	谱聚类（K 近邻）	K-means	GMM	Agglomerative Clustering
Iris	0.953 3	0.953 3	0.840 0	0.900 0	0.886 7	**0.966 7**	0.886 7
Seed	0.881 0	**0.895 2**	0.876 2	0.857 1	0.890 5	0.842 8	0.871 4
Wdbc	0.935 0	0.933 2	0.855 9	**0.947 2**	0.927 9	0.926 2	0.868 2
Glass	0.616 8	**0.640 2**	0.537 4	0.626 1	0.546 7	0.523 3	0.546 7
pima-indians-diabetes	**0.677 1**	**0.678 4**	0.671 8	0.651 0	0.66 8	0.651 0	0.651 0
Heart	**0.835 0**	**0.835 0**	0.798 0	0.649 8	0.804 7	0.754 2	0.703 7

表 3　选取的算法在不同数据集上的调整兰德指数 ARI 结果对比

数据集	粒谱聚类（相对距离）	粒谱聚类（绝对距离）	谱聚类	谱聚类（K 近邻）	K-means	GMM	Agglomerative Clustering
Iris	**0.868 0**	**0.868 2**	0.623 0	0.744 5	0.716 3	0.507 3	0.719 5
Seed	**0.679 0**	**0.712 6**	0.656 8	0.634 1	0.704 8	0.611 5	0.675 2
Wdbc	0.754 7	0.748 6	0.497 1	**0.798 4**	0.730 1	0.672 1	0.538 3
Glass	0.173 7	**0.236 0**	0.112 2	0.204 1	0.161 7	0.146 7	0.158 8
pima-indians-diabetes	**0.114 5**	**0.117 2**	0.105 4	0.009 8	0.102 3	0.069 4	0.072 7
Heart	**0.447 0**	**0.447 0**	0.352 9	0.086 1	0.369 2	0.275 1	0.162 6

由表 2 可知，在通过 CP 进行性能评估时，在 Heart 和 pima-indians-diabetes 数据集中，基于邻域粒距离的 2 种谱聚类算法的性能表现优于其他算法. 在 Glass 和 Seed 数据集中，基于邻域粒绝对距离的谱聚类算法性能最优. 在 Iris 和 Wdbc 数据集中，2 种粒谱聚类算法得分虽然都高于原始谱聚类算法，显示出良好的聚类效果，但性能得分略低于 GMM 算法和基于 K 近邻的谱聚类算法.

由表 3 可知，在通过 ARI 进行性能评估时，在 Iris、Heart、pima-indians-diabetes 和 Seed 4 个数据集中，2 种距离的粒谱聚类算法的性能表现优秀，性能得分高于其他算法. 在 Glass 数据集中，基于邻域粒绝对距离的谱聚类算法性能是最优的. 在 Wdbc 和 Glass 数据集中，基于 K 近邻的谱聚类算法的性能要优于其他算法.

从以上实验结果可以看出，在所列举的数据集中，基于邻域粒距离的谱聚类算法的聚类性能均优于传统的谱聚类算法，并且优于大多数其他算法. 尽管在某些数据集上，该算法的性能指标略逊于 GMM 和基于 K 近邻的谱聚类算法，但通过邻域粒距离优化相似矩阵的构造，其性能指标得分与最优算法的性能指标得分基本相当. 与传统的算法不同，粒谱聚类算法利用邻域粒化技术在结构上取得了突破，使得数据在算法进行之前就得到处理. 这样处理后，簇内的粒向量距离更为紧密，而簇间的粒向量距离相较簇内的粒向量则更大. 因此构建的相似矩阵更能反映数据之间的关系，从而提升了谱聚类的聚类性能. 这使得谱聚类能够在不同形状的数据集上都展现出良好的性能.

5　结论

本文结合谱聚类的相似矩阵构建过程，将粒计算思想与谱聚类相融合，提出了基于邻域粒的谱聚类算法. 该算法将邻域粒化技术引入到谱聚类算法的相似矩阵构造中. 通过全局信息进行邻域粒化，使得生成的粒向量具有全局性. 因此，簇内的粒向量距离更为紧密，而簇间的粒向量距离相对较大. 由此构建的相似矩阵更能够准确地反映数据之间的关系，从而提高了谱聚类的聚类性能. 最后，本文在常见的 UCI 数据集上进行了实验验证，结果表明基于邻域粒的谱聚类算法相比传统方法在聚类效果上表现更佳. 这一实验结果在不同形状的数据集上得到了验证，证明了基于邻域粒的谱聚类算法的可行性和有效性.

参考文献:

[1] ZADEH L A. Toward a Theory of Fuzzy Information Granulation and Its Centrality in Human Reasoning and Fuzzy Logic [J]. Fuzzy Sets and Systems, 1997, 90(2): 111-127.

[2] LIN T Y. Data Mining and Machine Oriented Modeling: a Granular Computing Approach [J]. Applied Intelligence, 2000, 13(2): 113-124.

[3] YAO Y Y, YAO B X. Covering Based Rough Set Approximations [J]. Information Sciences, 2012, 200: 91-107.

[4] 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.

[5] RUAN J H, WANG X P, SHI Y. Developing Fast Predictors for Large-Scale Time Series Using Fuzzy Granular Support Vector Machines [J]. Applied Soft Computing, 2013, 13(9): 3981-4000.

[6] 朱鹏飞, 胡清华, 于达仁. 基于随机化属性选择和邻域覆盖约简的集成学习 [J]. 电子学报, 2012, 40(2): 273-279.

[7] 段洁, 胡清华, 张灵均, 等. 基于邻域粗糙集的多标记分类特征选择算法 [J]. 计算机研究与发展, 2015, 52(1): 56-65.

[8] WANG C Z, HU Q H, WANG X Z, et al. Feature Selection Based on Neighborhood Discrimination Index [J]. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(7): 2986-2999.

[9] 章永来, 周耀鉴. 聚类算法综述 [J]. 计算机应用, 2019, 39(7): 1869-1882.

[10] 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.

[11] SINAGA K P, YANG M S. Unsupervised K-Means Clustering Algorithm [J]. IEEE Access, 2020, 8: 80716-80727.

[12] 陈叶旺, 申莲莲, 钟才明, 等. 密度峰值聚类算法综述 [J]. 计算机研究与发展, 2020, 57(2): 378-394.

[13] 李文杰, 闫世强, 蒋莹, 等. 自适应确定DBSCAN算法参数的算法研究 [J]. 计算机工程与应用, 2019, 55(5): 1-7, 148.

[14] 王红, 葛丽娜, 王苏青, 等. 基于OPTICS聚类的差分隐私保护算法的改进 [J]. 计算机应用, 2018, 38(1): 73-78.

[15] 孙璐, 梁永全. 融合网格划分和DBSCAN的改进聚类算法 [J]. 计算机工程与应用, 2022, 58(14): 73-79.

[16] RODRIGUEZ A, LAIO A. Clustering by Fast Search and Find of Density Peaks [J]. Science, 2014, 344(6191): 1492-1496.

[17] 汤峥, 宋余庆, 刘哲. 基于粒子群优化和EM算法的图像聚类研究 [J]. 小型微型计算机系统, 2015, 36(7): 1602-1606.

[18] 张东月, 倪巍伟, 张森, 等. 一种基于本地化差分隐私的网格聚类方法 [J]. 计算机学报, 2023, 46(2): 422-435.

[19] COHEN-ADDAD V, KANADE V, MALLMANN-TRENN F, et al. Hierarchical Clustering: Objective Functions and Algorithms [M] //Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms. Philadelphia, PA: Society for Industrial and Applied Mathematics, 2018: 378-397.

[20] 白璐, 赵鑫, 孔钰婷, 等. 谱聚类算法研究综述 [J]. 计算机工程与应用, 2021, 57(14): 15-26.

[21] 孔万增, 孙志海, 杨灿, 等. 基于本征间隙与正交特征向量的自动谱聚类 [J]. 电子学报, 2010, 38(8): 1880-1885, 1891.

[22] CHEN X J, HONG W J, NIE F P, et al. Spectral Clustering of Large-Scale Data by Directly Solving Normalized Cut [C] //Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. London, United Kingdom: ACM, 2018: 1206-1215.

[23] 闫静茹, 陈颖悦, 曾高发, 等. 基于邻域粒化的逻辑回归算法 [J]. 山西大学学报(自然科学版), 2024, 47(1): 40-47.

责任编辑 崔玉洁