文章编号　1004-924X(2024)07-1045-14

基于自适应近邻信息的模糊C均值聚类算法

高云龙[1]，李建鹏[2]，郑兴莘[1]，邵桂芳[1]，祝青园[1]，曹　超[3*]
(1. 厦门大学 萨本栋微米纳米科学技术研究院，福建 厦门 361102；
2. 厦门大学 自动化系，福建 厦门 361102；
3. 自然资源部 第三海洋研究所，福建 厦门 361005)

摘要：传统的模糊C均值算法直接基于原始数据进行聚类，数据的内在结构可能会被噪声、异常值或其他因素破坏，因此聚类性能会受到影响。为提升FCM算法的鲁棒性，提出了一种基于自适应近邻信息的模糊C均值聚类算法。近邻信息指的是一种基于数据点之间相似度的度量，每个数据点都可以看作其他数据点的近邻，但是不同数据点之间的相似度是不同的。将样本点的近邻信息G_X和类中心点的近邻信息G_V融入基础FCM模型中，为聚类过程提供更多的数据结构信息，用于指导聚类算法中的簇划分过程，以提升算法的稳定性，并提出了3个迭代算法求解本文提出的聚类模型。与其他先进聚类算法对比，在部分基准数据集上聚类性能有10%以上的提升，同时还从参数敏感性、收敛性、消融实验等方面对算法进行评价。实验结果可以充分显示本文提出的聚类算法的可行性与有效性。

关　键　词：模糊C均值聚类；自适应近邻；算法鲁棒性；迭代算法
中图分类号：TP394.1；TH691.9　　**文献标识码**：A　　**doi**：10.37188/OPE.20243207.1045

Fuzzy C-means clustering algorithm based on adaptive neighbors information

GAO Yunlong[1], LI Jianpeng[2], ZHENG Xingshen[1], SHAO Guifang[1], ZHU Qingyuan[1], CAO Chao[3*]

(1. *Pen-Tung Sah Institute of Micro-Nano Science and Technology, Xiamen University, Xiamen* 361102, *China*;
2. *Department of Automation, Xiamen University, Xiamen* 361102, *China*;
3. *Third Institute of Oceanography, Ministry of Natural Resources, Xiamen* 361005, *China*)
* *Corresponding author, E-mail: caochao@tio.org.cn*

Abstract: Traditional FCM algorithms cluster based on raw data, risking distortion from noise, outliers, or other disruptions, which can degrade clustering outcomes. To bolster FCM's resilience, this study introduces a fuzzy C-means clustering algorithm that leverages adaptive neighbor information. This concept hinges on the similarity between data points, treating each point as a potential neighbor to others, albeit with varying degrees of similarity. By integrating the neighbor information of sample points, labeled G_X, and that of cluster centers, labeled G_V, into the standard FCM framework, the algorithm gains additional insights into data structure. This aids in steering the clustering process and enhances the algorithm's robust-

收稿日期：2023-08-28；**修订日期**：2023-10-11.
基金项目：国家自然科学基金资助项目(No. 42076058,No. 52075461)；福建省自然科学基金资助项目(No. 2020J01713, No. 2022J01061)

ness. Three iterative methods are presented to implement this enhanced clustering model. When compared to leading clustering techniques, our approach demonstrates over a 10% improvement in clustering efficacy on select benchmark datasets. It undergoes thorough evaluation across different dimensions, including parameter sensitivity, convergence rate, and through ablation studies, confirming its practicality and efficiency.

Key words: fuzzy C-means clustering; adaptive neighbors; algorithm robustness; iterative algorithm

1 引　言

作为一种无监督方法,聚类的基本任务是将数据点划分为不相交的簇,使得同一簇内数据点之间的相似度最大化,而不同簇之间数据点的相似度最小化。在文本分析方面,聚类算法可以将市场细分为不同的消费者群体,帮助企业了解不同群体的需求和偏好,有助于市场营销策略的制定和产品定价。在计算机视觉领域,聚类算法可以将图像分割为不同的区域或对象,突出不同区域之间色块的差异和相同区域色块的相似,从而实现图像分析和目标检测等任务。聚类算法可以用于分析基因表达数据,帮助研究人员识别基因表达模式并发现疾病相关基因。数据当中的异常点或离群点可以通过聚类算法进行检测,可应用于故障诊断和网络安全等领域。总的来说,聚类算法在模式识别、图像处理和数据挖掘等领域有着十分广泛的应用,可以帮助人们分析数据,理解数据的本质结构特征,从数据中获取有用信息。

当标签信息不可用时,将数据分区成不同的块是很困难的。为了解决这个问题,聚类算法被提出,用以探索样本之间的内在相关性和差异。在过去的几十年里,许多类型的聚类算法被提出,具有代表性的有K-Means聚类[1]、模糊C均值聚类[2-4]和谱聚类[5]等。其中,由于算法理论的简单高效,K-Means聚类和模糊C均值聚类引起了很多关注。K-Means聚类也被称为硬聚类,其中每个样本被分配到距离最近的聚类原型。然而,随着信息技术的高速发展,数据的维度和规模也在快速增长,维数灾难问题出现,高维空间中的样本分布复杂,各个类别之间的边界模糊不清。因此,K-Means聚类的性能会受到严重影响。为了解决这个问题,模糊C均值聚类(Fuzzy C-Means Clustering, FCM)算法被提出。对于FCM聚类,根据隶属度将样本与每个类别相关联,并使用模糊指数来控制隶属度的稀疏性。Yu等分析了选择适当的模糊指数的规则,并在多个数据集上进行实验。结果表明,在大多数情况下,推荐使用[1.5,2.5]的范围[6-7]。

FCM聚类面临的常见问题是对噪声和异常值的敏感性。为了解决该问题,研究人员采用稀疏规范化来减少异常值的干扰,通常将FCM算法中距离的度量方式从平方范数替换为一种稀疏范数,通过这种方式,异常值对目标函数的贡献将被抑制。Xu等提出了一种稳定的FCM算法,使用$l_{2,1}$范数和截断的l_1范数替换原有的平方范数,分别构建了两个模糊聚类模型并提出了两种迭代加权算法来求解[8]。Chang等通过使用稀疏规范化范数(l_p范数)重新构造FCM目标函数,提出了一个非凸优化模型,通过这种方式评估每个特征对目标函数的贡献[9]。Zhang等修改谱聚类,并提出了一种模糊聚类和谱聚类结合的算法,引入σ-norm,以自适应地提高FCM对微小或较大异常值的稳定性[10-11]。另一种方法是引入局部空间信息,为了提高图像分割的性能,Chuang等提出了一个两步过程的FCM算法[12]。在第一步中,通过常规FCM算法获得隶属度矩阵。之后,通过空间信息更新该矩阵的元素,其中每个像素都落在一个小窗口中,其属于某个类别的概率由窗口中各像素属于该类别的概率的加权平均值确定。Cai等将局部空间关系和局部灰度关系都纳入模糊聚类模型中,以保证图像的抗噪性和保留细节的能力[13]。Nie等基于距离较小的数据点应该具有更大的概率成为邻居这一前提假设,提出了

一种新的视角来解决聚类问题，为每个数据点分配自适应最优近邻，基于局部连通性学习数据相似性矩阵；并对学习到的相似性矩阵的拉普拉斯矩阵施加秩约束，以实现理想的邻居分配，从而使数据中的连通分量恰好等于聚类数，并且每个连通分量对应一个簇，以达到优异的聚类结果[14]。上述算法主要依赖于数据的原始分布结构进行聚类。然而，现实世界中的数据往往包含噪声，数据中的噪声可能破坏其结构并影响聚类结果。受收缩模式[15-17]的思想启发，研究人员通过在灵活的流形上进行聚类，而不是在原始数据空间中，可以避免噪声对数据结构的影响。为了获得原始数据合适的流形结构，进行收缩模式的学习。收缩模式可以视为一种映射，它将数据映射到具有相同维数，但不是更低维数的灵活流形上。流形空间比原始数据空间具有更好的抵抗噪声，能增强聚类的稳定性[18]。

受局部结构信息在提升聚类性能的多个成功算法应用的启发，本文提出了一种基于自适应近邻信息的模糊C均值聚类算法（Adaptive Neighour Fuzzy C-Means，ANFCM）。具体来说，对于每个样本点，根据其余样本点与其欧氏距离度量，基于距离较近的样本点成为近邻的可能性更大这一先验假设，挖掘数据的局部结构信息指导聚类过程，从而达到减弱噪声、离群点影响的作用。首先，通过近邻信息学习样本点的相似性以及簇中心和样本点之间的相似性，挖掘簇中心和样本点局部结构信息，指导聚类过程；其次，将上述两种相似性引入传统FCM框架，补偿FCM单一欧式距离平方的度量方式，使得算法在考虑全局聚类结构的同时，也能关注局部邻域信息，提升算法的稳定性，降低对噪声和异常值的敏感性。

2　相关工作

2.1　模糊C均值聚类和K均值聚类

FCM是最早提出的处理重叠聚类的算法之一。FCM的核心是将每个数据点根据隶属度分配到多个聚类原型中。形式上，给定一个数据集$X=[x_1,x_2,\cdots,x_n]\in\mathbf{R}^{d\times n}$，其中$d$是维度，$n$是数据集中样本点的个数。$x_i\in\mathbf{R}^d$是第$i$个数据点。假设这些数据点来自$c$个类簇。在标签信息不可用的情况下将$n$个数据点分到$c$个类中，FCM算法的目标函数及约束条件如下：

$$\min_{U,M}\sum_{i=1}^{n}\sum_{k=1}^{c}u_{ik}^{h}\left\|x_i-m_k\right\|_2^2$$
$$s.t.\ \sum_{k=1}^{c}u_{ik}=1,\quad 0\leqslant u_{ik}\leqslant 1, \tag{1}$$

其中h是模糊指数，用于调整模糊程度，通常为大于1的实数。u_{ik}是矩阵$\boldsymbol{U}\in\mathbf{R}^{n\times c}$的第$(i,k)$个元素，它反映第$i$个数据点属于第$k$个聚类的程度。$m_k$是第$k$个聚类的聚类原型，$M=[m_1,\cdots,m_c]$。根据以下步骤分别更新$U$和$M$的元素，则可以实现对式(2)的求解：

$$u_{ik}=\frac{1}{\sum_{k=1}^{c}\left(\frac{\left\|x_i-m_k\right\|_2}{\left\|x_i-m_k\right\|_2}\right)^{\frac{2}{h-1}}}, \tag{2}$$

$$m_k=\frac{\sum_{i=1}^{n}u_{ik}^{h}x_i}{\sum_{i=1}^{n}u_{ik}^{h}}. \tag{3}$$

当$h=1$时，式(1)等价于：

$$\min_{U,M}\sum_{i=1}^{n}\sum_{k=1}^{c}u_{ik}\left\|x_i-m_k\right\|_2^2,$$
$$s.t.\ \sum_{k=1}^{c}u_{ik}=1,\quad u_{ik}\in\{0,1\} \tag{4}$$

式(4)就是K-Means算法的优化目标函数及约束条件。通常情况下，如果事先给定初始的聚类原型，隶属度矩阵U中各个元素可以根据下式进行计算：

$$u_{ik}=\begin{cases}1, & \left\|x_i-m_k\right\|_2^2=\min\limits_{1\leqslant j\leqslant c}\left\|x_i-m_j\right\|_2^2\\ 0, & \text{otherwise}\end{cases}. \tag{5}$$

2.2　鲁棒稀疏模糊K均值聚类

基于FCM的聚类算法对异常值敏感，为了增强FCM对异常值的稳定性，徐等提出用$l_{2,1}$范数和截断的l_1范数替换原有的平方范数，减小异常值对聚类结果的影响。将$l_{2,1}$范数的鲁棒稀疏模糊K均值聚类（Robust Sparse Fuzzy K-Mean，

RSFKM)模型定义为：

$$\min_{U,M}\sum_{i=1}^{n}\sum_{k=1}^{c}u_{ik}\|x_i-m_k\|_2+\gamma\|U\|_F^2$$

$$s.t.\ \sum_{k=1}^{c}u_{ik}=1,\quad 0\leqslant u_{ik}\leqslant 1. \tag{6}$$

截断的l_1范数的RSFKM模型定义为：

$$\min_{U,M}\sum_{i=1}^{n}\sum_{k=1}^{c}u_{ik}\min\left(\|x_i-m_k\|_2,\varepsilon\right)+\gamma\|U\|_F^2$$

$$s.t.\ \sum_{k=1}^{c}u_{ik}=1,\quad 0\leqslant u_{ik}\leqslant 1, \tag{7}$$

其中：参数γ是用于调整隶属度矩阵$\boldsymbol{U}$的稀疏程度的正则化参数。如果$\gamma=0$，则每个样本点的隶属度向量是稀疏的(只有一个元素是非零的，其他元素都是零)。当$\gamma>0$时，隶属度向量比$\gamma=0$时更稠密。通过调整γ，隶属度向量的稀疏性是一个渐进的变化，随着γ的逐渐增加，隶属度向量中包含越来越多的非零元素。当γ达到一个较大的值时，隶属度向量的所有元素都是非零的，此时隶属度向量是非稀疏的。通过参数寻优可以找到合理的隶属度向量的稀疏性，以获得更准确的聚类结果。参数ε可以视为一个阈值，当样本点与聚类中心点的距离大于给定阈值ε后，距离取值为阈值ε，这样可以显著减少异常值对目标函数的影响。式(6)可以通过交替迭代算法来求解。

2.3 模式收缩模糊K均值聚类

传统FCM算法在原始数据上进行聚类，易受到噪声和异常值的影响。因此，本文提出了模式收缩模糊K均值聚类算法(FKPS)，对原始数据的结构进行放缩，得到理想的数据流形结构，称为收缩模式。收缩模式可看作原始数据结构的近似，近似程度可由参数β控制，进而在学习得到的收缩模式上展开模糊聚类，并提出了迭代算法将缩小模式的学习和模糊聚类集成到一个统一的框架中。由于收缩模式具有所需的理想流形结构，直接进行聚类可以提高聚类性能和模糊聚类的稳定性。

假设数据集$X=[x_1,x_2,\cdots,x_n]\in\mathbf{R}^{d\times n}$，$x_i\in\mathbf{R}^d$是数据集中第$i$个样本点，$x_i$在收缩模式中的对应点定义为$z_i$，且$Z=[z_1,z_2,\cdots,z_n]\in\mathbf{R}^{d\times n}$。FKPS算法的目标函数定义为：

$$\min_{U,M,Z}\sum_{i=1}^{n}\sum_{k=1}^{c}u_{ik}\|z_i-m_k\|_2^2+\gamma\|U\|_F^2+\beta\sum_{i=1}^{n}\|x_i-z_i\|_2^2,\ s.t.\ \sum_{k=1}^{c}u_{ik}=1,\quad 0\leqslant u_{ik}\leqslant 1. \tag{8}$$

3 自适应近邻信息模糊C均值聚类

3.1 模型设计

在聚类领域中，自适应近邻信息指的是一种基于数据点之间相似度的度量，用于指导聚类算法中的簇划分过程。每个数据点都可以被看作是其他数据点的近邻，但是不同数据点之间的相似度是不同的，基于距离较小的数据点应该具有更大的概率成为邻居这一前提假设，可以认为距离较近的样本点同属一个类别的概率较大。因此，自适应近邻信息会根据每个数据点与其他数据点的相似度，自适应地选择最相关的近邻点进行类别划分，从而提高聚类的准确性和稳定性。

本文采用欧氏距离作为距离的度量方式。给定一个数据集$X=[x_1,x_2,\cdots,x_n]\in\mathbf{R}^{d\times n}$，其中$d$是维度，$n$是数据集中样本点的个数。$x_j\in\mathbf{R}^d$是第$j$个样本点，其余样本点与$x_j$成为邻居的概率设为$s_{jk}$，越小的距离$\left\|x_j-x_k\right\|_2^2$对应着更大的近邻概率$s_{jk}$。近邻信息以相似度矩阵$S$的形式体现，上述数据集$X$中各样本点间的相似度矩阵中各元素$s_{jk}$，即为$n$个样本点两两之间的近邻概率。数据集$X$的相似度矩阵$S\in\mathbf{R}^{n\times n}$可用下式求解：

$$\min_{s_i^{\mathrm{T}}1=1,0\leqslant s_i\leqslant 1}\sum_{k=1}^{n}\|x_j-x_k\|_2^2s_{jk}+\lambda s_{jk}^2. \tag{9}$$

本文将样本点x_j的近邻信息定义为G_{x_j}，G_{x_j}为一数值，数据集X中所有样本点的近邻信息构成向量$G_X\in\mathbf{R}^{1\times n}$，$G_{x_j}$为其第$j$个元素，正则化参数$\lambda$的作用是调节相似性矩阵$S$的稀疏性。$G_{x_j}$的定义如下：

$$G_{x_j}=\min_{S}\sum_{k=1}^{n}\|x_j-x_k\|_2^2s_{jk}+\lambda s_{jk}^2$$

$$s.t.\quad s_{jk}\geqslant 0,\ \sum_{k=1}^{n}s_{jk}=1. \tag{10}$$

假设数据集 X 中的 n 个样本点可分为 c 个类别，则类别中心矩阵 $V\in\mathbf{R}^{d\times c}$ 中第 i 列向量 v_i 即为第 i 个类别的中心点。同理，将类中心点 v_i 与 n 个样本点的近邻信息定义为 G_{v_i}，G_{v_i} 为一数值，类别中心矩阵 V 中所有类中心点的近邻信息构成向量 $G_V\in\mathbf{R}^{1\times c}$，$G_{v_i}$ 为第 i 个元素，如图 1 所示 。G_{v_i} 的定义如下：

$$G_{v_i}=\min_{S'}\sum_{k=1}^{n}\|v_i-x_k\|_2^2 s'_{jk}+\lambda s'^{\,2}_{jk}$$
$$s.t.\quad s'_{jk}\geqslant 0,\ \sum_{k=1}^{n}s'_{jk}=1. \tag{11}$$

为利用近邻信息提高 FCM 算法聚类的准确性和稳定性，在得到样本点的近邻信息 G_X 和类中心点的近邻信息 G_V 后，本文将它们融入基础 FCM 模型中得到引入自适应近邻信息的模糊 C 均值聚类算法模型（Adaptive Neighbors Fuzzy C-Means Algorithm，ANFCM）。ANFCM 的模型定义如下：

$$\min_{U,V}J=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^{m}\left(\|x_j-v_i\|_2^2+\alpha\left(G_{x_j}-G_{v_i}\right)^2\right). \tag{12}$$

自适应体现在求解 G_{x_j} 与 G_{v_i} 的过程中，将参数 λ 的选择转换为近邻个数的选择，故模型参数的寻优转化为了自适应调节近邻个数，具体证明见模型优化部分。参数 α 起到调整原聚类结构信息与自适应近邻信息在聚类过程中的重要性，增大参数 α 的值，自适应近邻信息对聚类过程的影响增大，反之减小。

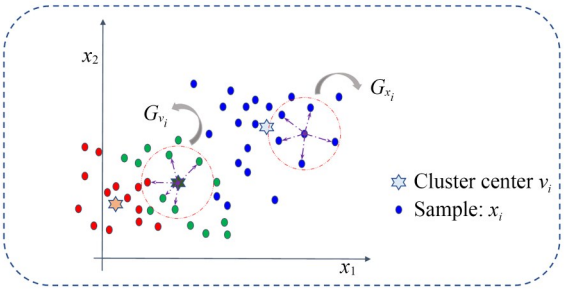


图 1　近邻信息示意图

Fig. 1　Neighborhood information

3.2　模型优化

本文提出的 ANFCM 模型具体定义如下：

$$\begin{cases}\min\limits_{U,V}J=\sum\limits_{i=1}^{c}\sum\limits_{j=1}^{N}u_{ij}^{m}\left(\|x_j-v_i\|_2^2+\alpha\left(G_{x_j}-G_{v_i}\right)^2\right)\\ G_{x_j}=\min\limits_{S}\sum\limits_{k=1}^{N}\|x_j-x_k\|_2^2 s_{jk}+\lambda s_{jk}^2\quad s.t.\quad s_{jk}\geqslant 0,\ \sum\limits_{k=1}^{N}s_{jk}=1\\ G_{v_i}=\min\limits_{S'}\sum\limits_{k=1}^{N}\|v_i-x_k\|_2^2 s'_{jk}+\lambda s'^{\,2}_{jk}\quad s.t.\quad s'_{jk}\geqslant 0,\ \sum\limits_{k=1}^{N}s'_{jk}=1\end{cases} \tag{13}$$

该模型中包含 3 个最小值优化问题。因为 G_{x_j} 的值只与数据 X 有关，所以在一开始就可进行求解。先定义：

$$d_{jk}=\|x_j-x_k\|_2^2, \tag{14}$$

由于：

$$G_X=\min\sum_{j=1}^{N}\sum_{k=1}^{N}s_{jk}d_{jk}+\lambda\|S\|_F^2=\min\sum_{j=1}^{N}S_j^{\mathrm{T}}d_j+\lambda S_j^{\mathrm{T}}S_j\ \lambda\min\sum_{j=1}^{N}\left\|S_j+\frac{d_j}{2\lambda}\right\|_2^2-\frac{1}{4\lambda^2}d_j^{\mathrm{T}}d_j\Leftrightarrow\min\sum_{j=1}^{N}\left\|S_j+\frac{d_j}{2\lambda}\right\|_2^2. \tag{15}$$

针对这个优化问题，可根据拉格朗日法和 KKT 条件求解：

$$L(S_j,\eta,\beta_j)=\frac{1}{2}\left\|S_j+\frac{d_j}{2\lambda_j}\right\|_2^2-\eta(S_j^{\mathrm{T}}-1)-\beta_j^{\mathrm{T}}S_j. \tag{16}$$

根据 Nie 等提出的方法[14]，解决步骤为：

$$\lambda_j=\frac{k}{2}d_{j,k+1}-\frac{1}{2}\sum_{i=1}^{k}d_{ji}, \tag{17}$$

$$\lambda=\frac{1}{N}\sum_{j=1}^{N}\left(\frac{k}{2}d_{j,k+1}-\frac{1}{2}\sum_{i=1}^{k}d_{ji}\right), \tag{18}$$

$$\eta=\frac{1}{k}+\frac{1}{2k\lambda_j}\sum_{i=1}^{k}d_{ji}, \tag{19}$$

$$s_{ji}=\left(-\frac{d_{ji}}{2\lambda_j}+\eta\right)_+. \tag{20}$$

将以上推导整理成求解 G_{x_j} 问题的算法 1。

算法 1：样本点近邻信息 G_X 的求解

输入：数据矩阵 $X\in\mathbf{R}^{d\times n}$，类簇数量 c，自适应近邻个数 k

输入：自适应近邻信息向量 $G_X\in\mathbf{R}^{1\times n}$

1：**开始**

2：**计算** 距离矩阵 $D\in\mathbf{R}^{n\times n}$，矩阵第 j 行第 k 列元素按式(14)定义；

3：　根据给定自适应近邻个数 k，通过式(17)和(18)计算了参数 λ；

4：　拉格朗日乘子 η 根据式(19)计算；

5：　相似度矩阵 $S\in\mathbf{R}^{n\times n}$，矩阵第 j 行第 k 列元素按式(20)定义；

6：　根据 ANFCM 模块(3.5)中 G_{x_j} 的定义式计算样本点 x_j 的近邻信息；

7：**输出** 近邻信息向量 G_X

算法开始时，先随机初始化隶属度矩阵 $\boldsymbol{U}$ 和类中心矩阵 $\boldsymbol{V}$。G_{v_i} 的优化与数据 X 和类中心矩阵 $\boldsymbol{V}$ 有关，每次根据更新后的类中心矩阵 $\boldsymbol{V}$，$\boldsymbol{G}_{v_i}$ 的求解过程与 $\boldsymbol{G}_{x_j}$ 的求解过程类似，使用算法 2 计算其最小值。

算法 2：类中心点近邻信息 G_V 的求解

输入：数据矩阵 $X\in\mathbf{R}^{d\times n}$，聚类原型矩阵 $V\in\mathbf{R}^{d\times c}$，类簇数量 c，自适应近邻个数 k

输入：类中心点自适应近邻信息向量 $G_V\in\mathbf{R}^{1\times c}$

1：**开始**

2：**计算** 距离矩阵 $D\in\mathbf{R}^{c\times n}$，矩阵第 i 行第 k 列元素由 $d_{ik}=\|v_i-x_k\|$ 定义；

3：　根据给定自适应近邻个数 k，通过式(17)和(18)计算参数 λ；

4：　拉格朗日乘子 η 根据式(19)计算；

5：　相似度矩阵 $S\in\mathbf{R}^{c\times n}$，矩阵第 j 行第 k 列元素按式(20)定义；

6：　根据 ANFCM 模块(3.5)中 G_{v_j} 的定义式计算类中心点 v_i 的近邻信息；

7：**输出** 近邻信息向量 G_V

优化目标函数 J 时，G_{x_j} 和 G_{v_i} 当作常数，根据拉格朗日乘数法进行求解。

构造拉格朗日函数：

$$L=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^{m}\left(\|x_j-v_i\|^2+\alpha\left(G_{x_j}-Gv_i\right)^2\right)+\sum_{j=1}^{n}\lambda_j\left(\sum_{i=1}^{c}u_{ij}-1\right). \tag{21}$$

令函数 L 对 u_{ij} 求偏导数，并令其为零：

$$\frac{\partial L}{\partial u_{ij}}=m\left(\|x_j-v_i\|^2+\alpha\left(G_{x_j}-G_{v_i}\right)^2\right)u_{ij}^{m-1}+\lambda_j=0, \tag{22}$$

得到：

$$u_{ij}^{m-1}=-\frac{\lambda_j}{m\left(\|x_j-v_i\|^2+\alpha\left(G_{x_j}-G_{v_i}\right)^2\right)}, \tag{23}$$

$$u_{ij}=\left(-\frac{\lambda_j}{m}\right)^{\frac{1}{m-1}}\frac{1}{\left(\|x_j-v_i\|^2+\alpha\left(G_{x_j}-G_{v_i}\right)^2\right)^{\frac{1}{m-1}}}. \tag{24}$$

结合约束条件 $\sum_{i=1}^{c}u_{ij}=1$，可以得到：

$$u_{ij}=\frac{1}{\sum_{k=1}^{c}\left(\frac{\|x_j-v_i\|^2+\alpha\left(G_{x_j}-G_{v_i}\right)^2}{\|x_j-v_k\|^2+\alpha\left(G_{x_j}-G_{v_i}\right)^2}\right)^{\left(\frac{1}{m-1}\right)}}. \tag{25}$$

令函数 L 对 v_i 求偏导数，并令其为零：

$$\frac{\partial L}{\partial v_i}=\sum_{j=1}^{n}\left(-2u_{ij}^{m}\left(x_j-v_i\right)\right)=0, \tag{26}$$

得到：

$$-2\sum_{j=1}^{n}u_{ij}^{m}x_j+2\sum_{j=1}^{n}u_{ij}^{m}v_i=0, \tag{27}$$

$$v_i=\frac{\sum_{k=1}^{n}u_{ik}^{m}x_k}{\sum_{k=1}^{n}u_{ik}}. \tag{28}$$

ANFCM 算法流程如图 2 所示，整体求解步骤归纳为算法 3。

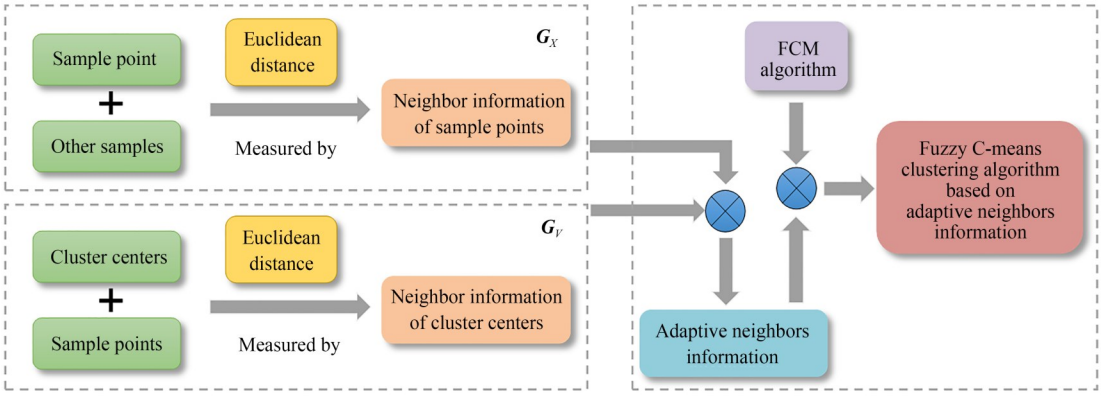


图 2　ANFCM 算法流程

Fig. 2　Flowchart of ANFCM algorithm

算法 3：ANFCM 迭代求解法

输入：数据矩阵 $\boldsymbol{X}\in\mathrm{R}^{d\times n}$，类簇数量 c，样本点和类中心点近邻个数 k_x 和 k_v

输入：隶属度矩阵 $\boldsymbol{U}\in\mathrm{R}^{c\times n}$

1：**开始**

2：**初始化** 随机初始化隶属度矩阵 $\boldsymbol{U}$，使满足 $0\leqslant u_{ij}\leqslant 1$ 且 $\Sigma_i u_{ij}=1$；

3：　　根据式(28)初始化聚类原型矩阵 $\boldsymbol{V}\in\mathrm{R}^{d\times c}$；

4：**计算** 根据**算法** 1 计算样本点的近邻信息 G_X；

5：While $\boldsymbol{U}$ not converge do：

6：　　根据**算法** 2 更新类中心点的近邻信息 G_V；

7：　　根据式(25)更新隶属度矩阵 $\boldsymbol{U}$；

8：　　根据式(28)更新聚类原型矩阵 $\boldsymbol{V}$；

9：End while

10：**输出** 近邻信息向量 G_V

4　实　验

4.1　数据集概况

UCI 数据集是机器学习领域中最常用的验证算法性能的数据集之一。对比实验中，使用 8 个真实基准数据集，如表 1 所示。

表 1　基准数据集情况描述

Tab. 1　Description of benchmark datasets

数据集	样本个数	特征数	类别数
Ionosphere	351	34	2
Jain	373	2	2
WBC	683	9	2
Air	359	64	3
Appendicitis	106	7	2
Mammographic	748	4	2
Pima	768	8	2
WDBC	569	30	2

4.2　对比算法

本文采用 KM，FCM，模糊紧密性与分离性聚类算法(FCS)，ATKM，RSFKM 和 KFS 6 种先进的相关聚类方法作为对比算法：

(1) K-Means 是最为知名的聚类算法之一，它将一组数据点分成若干簇，使得每个数据点被分配到一个簇中，且每个数据点只属于一个簇；

(2) FCM 算法是 K-Means 算法的扩展，每个簇视为一个模糊集合，而隶属度函数测量每个数据点属于簇的可能性，每个聚类原型由所有样本的加权平均值形成；

(3) FCS[19] 为每个聚类分配一个硬核边界，以便硬隶属度和模糊隶属度可以共存于聚类结果中。因此，FCS 可以被看作是介于 K-Means 聚类和模糊 C 均值聚类之间的一种新型聚类算法；

(4) 与 FCM 相比，聚合模糊 K 均值聚类算法(AFKM)[20] 采用正则化参数来调整模糊隶属度，并引入最大熵信息以优化聚类分区；

(5) RSFKM 采用稀疏结构范数来减小异常值对目标函数的影响，并提出了一种重新加权的算法来有效求解模型；

(6) FKPS 直接在得到的收缩模式上执行模糊聚类，收缩模式可以视为没有噪声干扰的干净数据，因此拥有理想的流形结构。

4.3　参数设置

在 FCM 类型的算法中，大量实验表明，隶属度权重指数设置为 2 时能得到较好的结果。根据经验，该部分中所有算法的隶属度权重指数均设置为 2，其余需要调整的参数采用网格搜索策略进行选择。ANFCM 算法有 3 个需要调整的参

数，第一个是正则化参数α，用于调节全局信息与样本点近邻信息对聚类的影响程度，依次取值为[0.001，0.01，0.1，1，10，100，1000，10000]；第二个参数是最近邻样本点个数k_x，第三个参数是最近邻聚类原型个数k_v，两个参数的设置相同，均为[2，3，5，7，9，15，19，25]。AFKM算法的正则化参数λ设置为[0.001，0.01，0.1，1，2，5，10，100]。FKPS算法需要调整的两个参数分别被命名为γ和β，第一个参数γ用于调整隶属度矩阵的模糊程度，取值为[0.1，0.5，1，5，10，50，100，500，1 000]。而第二个参数β则用于调整原始数据和学习到的收缩模式之间的差异，取值为[0.005，0.01，0.05，0.1，0.5，1，5]。在RSFKM算法中，有两个重要的参数，即正则化参数γ和阈值ϵ。正则化参数γ对数据点和聚类中心之间的最小距离设定了限制，并防止隶属度具有极端值，即0和1，其取值为$[10^{-1}, 10]$，步长为0.5。阈值ϵ主要控制离群值的数量，并与表示的残差相关，设置为[0，0.5，1，1.5，2，2.5，3]。

4.4 评价指标

ACC，NMI和Purity是评估聚类结果质量的常用指标。其中，ACC（Accuracy）是聚类的准确率，它度量聚类结果与真实标签之间的匹配程度；NMI（Normalized Mutual Information）是标准化互信息，它测量聚类结果与真实标签之间的相似程度，将互信息归一化，并考虑到聚类结果和真实标签的熵，取值为0到1，值越高表示聚类结果越与真实标签相似；Purity是纯度，它度量聚类结果中同一类别的数据点所占比例。它计算每个聚类中出现次数最多的真实标签，将这些标签的出现次数相加并除以总数据点数得到聚类结果的纯度，取值为0到1，值越高表示聚类结果中同一类别的数据点越多。

4.5 聚类性能评估

由于聚类结果受随机初始化的影响，所有实验结果均为在同等条件下，随机初始化聚类中心的10次聚类结果取平均值。各聚类算法在8个基准数据集上进行聚类，评价指标分别如表2～表4所示，最优算法结果加粗标出。从表4可以

表2 各算法在8个数据集上的Accuracy值

Tab. 2 Accuracy values for each algorithm on 8 datasets

Method	Ionosphere	Jain	WBC	Air	Appendicitis	Mammographic	Pima	WDBC
KM	0.705 1	0.781 0	0.960 6	0.404 2	0.800 9	0.731 6	0.660 2	0.854 1
FCM	0.711 1	0.589 8	0.956 1	0.376 3	0.792 5	0.707 2	0.658 9	0.854 1
FCS	0.709 4	0.595 2	0.956 1	0.381 6	0.792 5	0.707 2	0.658 9	0.852 4
AFKM	0.712 3	0.780 2	0.972 2	0.404 5	0.778 3	0.762 0	0.651 4	0.688 6
RSFKM $l_{2,1}$	0.695 2	0.766 8	0.964 9	0.436 8	0.826 4	0.675 1	0.608 1	0.868 2
RSFKM $capped_{2,1}$	0.641 0	0.740 0	0.774 4	0.442 3	0.837 7	0.762 0	0.651 0	0.627 4
FKPS	0.710 5	0.808 3	0.973 8	0.415 9	0.792 5	0.673 3	0.645 4	0.911 6
ANFCM	**0.863 0**	**0.963 5**	**0.975 1**	**0.507 2**	**0.851 9**	**0.771 4**	**0.739 3**	**0.924 4**

表3 各算法在8个数据集上的NMI值

Tab. 3 NMI values for each algorithm on 8 datasets

Method	Ionosphere	Jain	WBC	Air	Appendicitis	Mammographic	Pima	WDBC
KM	0.118 3	0.332 1	0.743 6	0.010 2	0.174 0	0.014 6	0.026 7	0.422 3
FCM	0.129 3	0.283 7	0.722 3	0.017 7	0.162 1	0.014 8	0.031 7	0.422 3
FCS	0.126 4	0.283 7	0.722 3	0.017 9	0.162 1	0.014 8	0.031 3	0.417 9
AFKM	0.131 2	0.331 1	0.806 8	0.022 1	0.158 3	0.008 5	0.019 7	0.114 0
RSFKM $l_{2,1}$	0.114 4	0.315 9	0.765 1	0.028 9	0.210 2	0.023 4	0.015 1	0.458 7
RSFKM $capped_{2,1}$	0.167 8	0.076 9	0.360 9	0.034 1	**0.243 0**	0.000 4	0.000 0	0.000 0
FKPS	0.129 5	0.381 1	0.816 7	0.024 8	0.162 1	**0.026 1**	0.037 2	0.554 3
ANFCM	**0.413 1**	**0.761 4**	**0.825 5**	**0.064 8**	0.196 3	0.025 9	**0.135 1**	**0.597 3**

表 4　各算法在 8 个数据集上的 Purity 值

Tab. 4　Purity values for each algorithm on 8 datasets

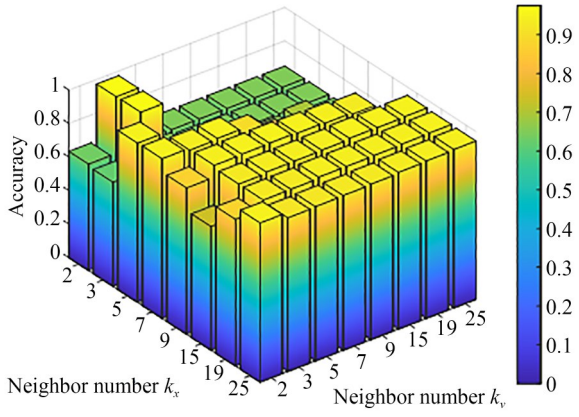
Method	Ionosphere	Jain	WBC	Air	Appendicitis	Mammographic	Pima	WDBC
KM	0.705 1	0.781 0	0.960 6	0.420 6	0.807 5	0.762 0	0.660 2	0.854 1
FCM	0.711 1	0.895 4	0.956 1	0.427 0	0.801 9	0.762 0	0.658 9	0.854 1
FCS	0.709 4	0.895 4	0.956 1	0.424 2	0.801 9	0.762 0	0.658 9	0.852 4
AFKM	0.712 3	0.793 8	0.972 2	0.424 0	0.816 0	0.763 1	0.657 7	0.688 6
RSFKM $l_{2,1}$	0.695 2	0.766 8	0.964 9	0.439 6	0.826 4	0.762 0	0.651 0	0.868 2
RSFKM $capped_{2,1}$	0.657 6	0.759 5	0.774 4	0.445 4	0.837 7	0.762 0	0.651 0	0.627 4
FKPS	0.710 5	0.811 5	0.973 8	0.438 7	0.801 9	0.762 0	0.664 1	0.911 6
ANFCM	**0.863 0**	**0.963 5**	**0.975 1**	**0.507 2**	**0.851 9**	**0.771 4**	**0.739 3**	**0.924 4**

看出，ANFCM 算法在 Ionosphere 和 Jain 数据集上的聚类性能，比对比算法的指标高出 10% 以上；总体上看，与传统 FCM 相比，ANFCM 算法在实验中的所有数据集上的平均表现，即 Accuracy，NMI 和 Purity 3 个指标上分别有 11.872 5%，15.442 5%，6.616 3% 的提升。说明引入局部近邻信息有效指导了隶属度矩阵的学习，进而提高了聚类算法的精度。

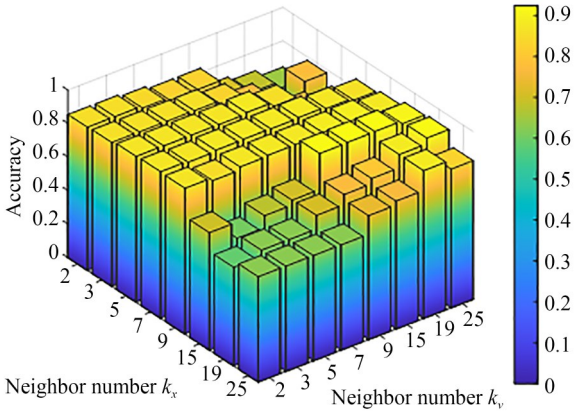
4.6　参数敏感性分析

参数敏感性实验是指在模型开发和优化过程中，对不同参数取值进行实验，以评估模型对参数变化的敏感性，从而确定最佳参数组合的方法。参数敏感性实验的重要性在于可以了解模型在不同参数设置下的性能，从而在确定模型最佳参数组合时提供指导，以最大限度地提高模型的性能，同时也可以评估模型对参数变化的稳定性，即在参数变化的情况下模型的表现是否稳定。ANFCM 算法模型具有近邻样本点个数 k_x，近邻聚类中心点个数 k_v 和正则化参数 α 3 个参数。对于给定某个数据集，首先根据 4.3 节参数的设置范围进行遍历寻优，找出该数据集最佳的一组参数。确定最优参数组合后，本文采用“定一议二”的策略可视化各参数变化对聚类精度的影响。具体方法如下：得到最佳参数组合后，每次固定一个参数为最优值，依据原先设定的取值范围调节其余两个参数，并绘制出聚类准确率随这两个参数的变化曲线，如图 3～图 5 所示。

在 4 个基准数据集上进行实验，聚类精度对参数 k_x 和 k_v 的参数敏感性实验结果如图 3 所示，聚类精度对参数 k_v 和 α 的参数敏感性实验结果如图 4 所示，聚类精度对参数 k_x 和 α 的参数敏感性实验结果如图 5 所示。若聚类性能受参数变化的影响起伏较大，说明引入的这项参数对聚类结果能产生重要影响。由图可以看出，ANFCM 算法的聚类精度对 3 个参数都较为敏感，随参数的变化较大，同时较优的结果集中在较小范围内。



(a) WBC



(b) WDBC

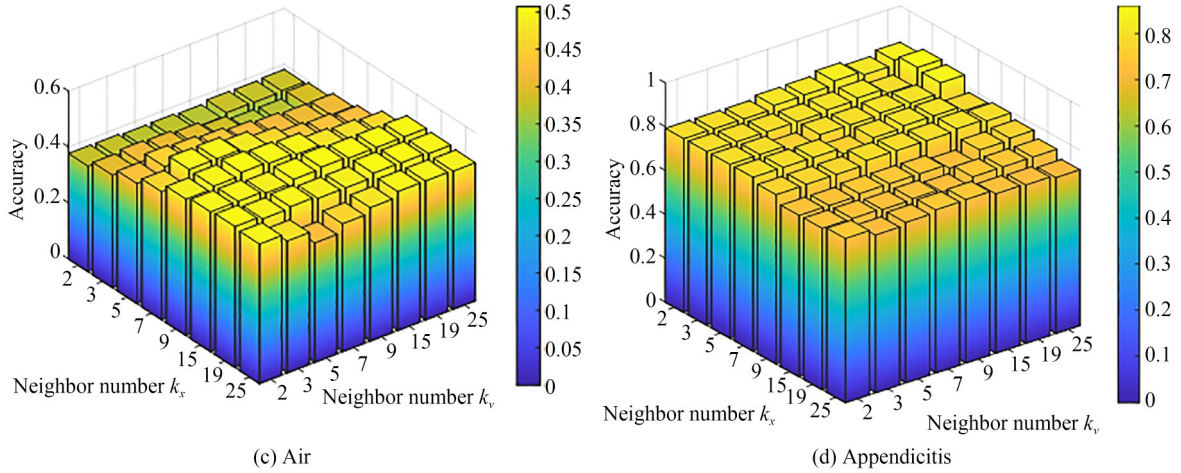


图 3　固定 α 条件下聚类精度对参数 k_x 和 k_v 的参数敏感性实验结果

Fig. 3　Experimental results of parameter sensitivity of clustering accuracy to parameters k_x and k_v under fixed α condition

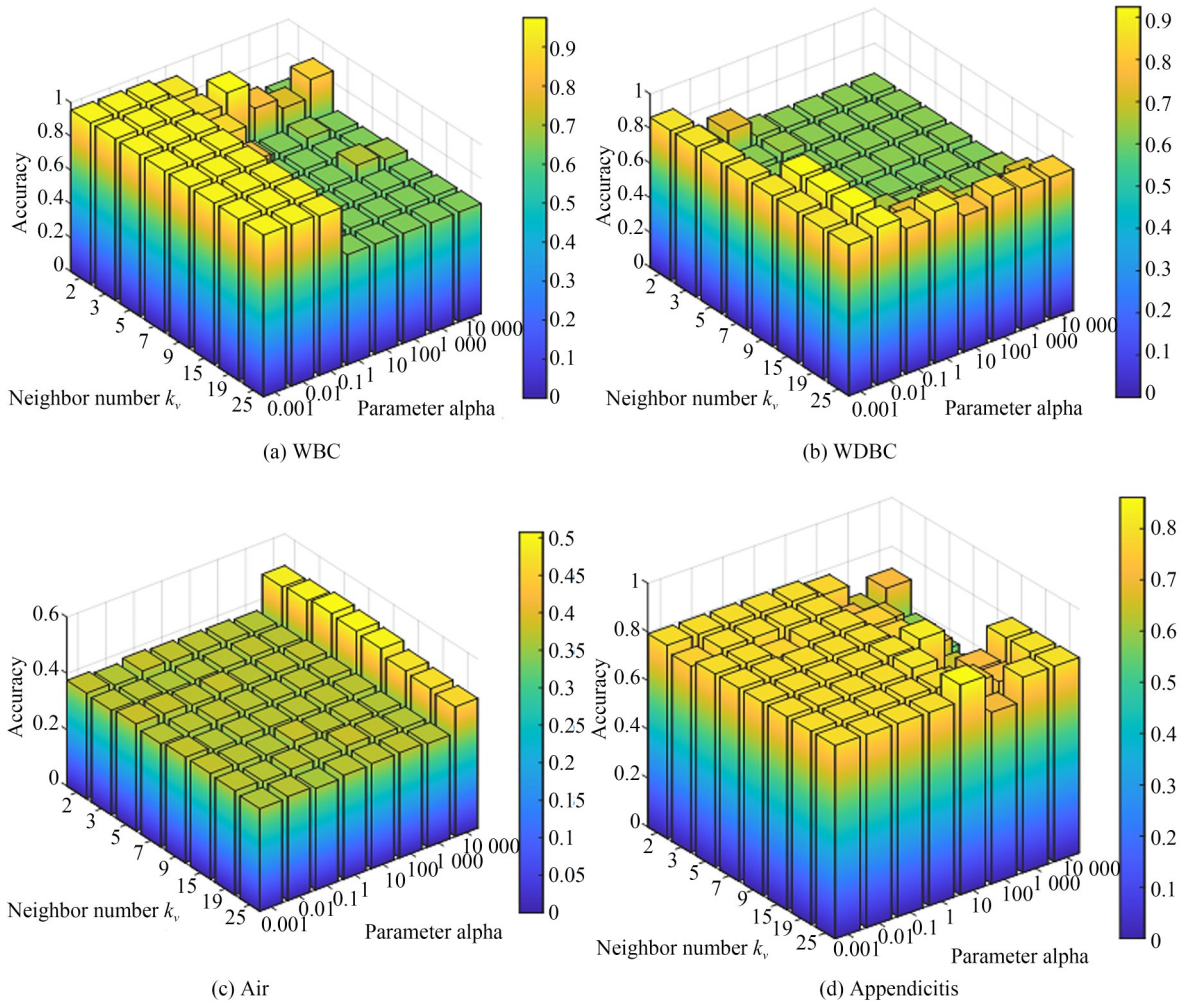


图 4　固定 k_x 条件下聚类精度对参数 k_v 和 α 的参数敏感性实验结果

Fig. 4　Experimental results of parameter sensitivity of clustering accuracy to parameters k_v and α under fixed k_x condition

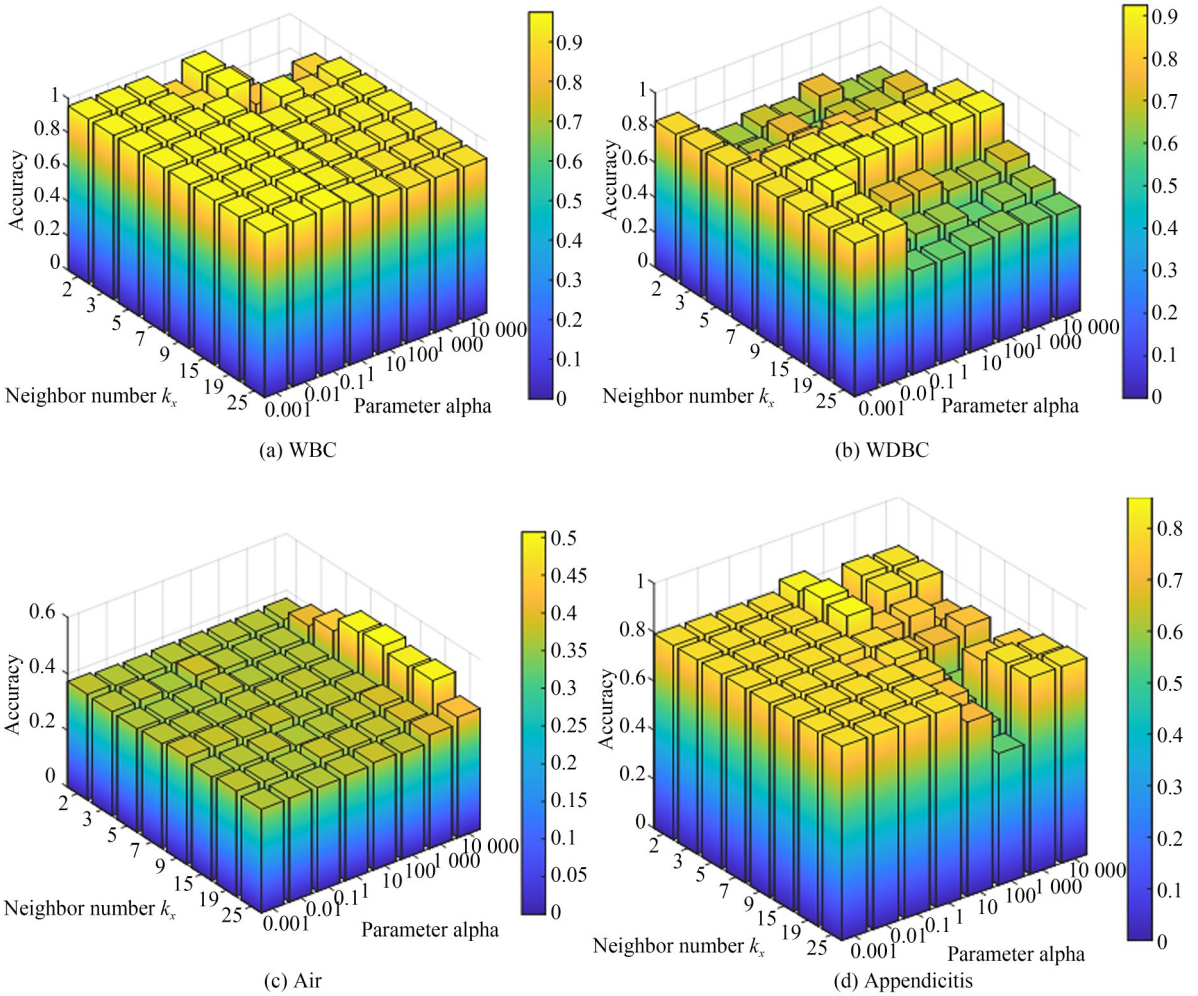


图 5　固定 k_v 条件下聚类精度对参数 k_x 和 α 的参数敏感性实验结果

Fig. 5　Experimental results of parameter sensitivity of clustering accuracy to parameters k_x and α under fixed k_v condition

4.7　收敛性分析

在模糊 C 均值聚类算法中，关键步骤是计算每个数据点与每个簇的隶属度，然后根据这些隶属度来更新每个簇的中心点。算法迭代直到满足收敛条件为止，例如中心点的变化量小于某个阈值。因此，收敛性分析是模糊 C 均值算法的重要组成部分。如果算法无法收敛，无法得到正确的簇划分结果，影响算法的应用效果。模糊 C 均值算法的收敛性分析通常包括以下几个方面：

(1)收敛性证明：证明算法能够在有限的迭代次数内收敛到一个稳定的状态，即每个数据点的隶属度和簇的中心点不再发生明显的变化。

(2)收敛速度分析：分析算法的收敛速度，即算法需要多少次迭代才能达到一个满意的精度，这对算法的实际应用具有重要的指导意义。

(3)收敛性检测方法：设计一些有效的方法来检测算法是否已经收敛，例如通过计算中心点的变化量、隶属度的变化量或目标函数值的变化量来判断算法是否还需要继续迭代。

本文采用目标函数值的下降情况来研究算法的收敛性，ANFCM 算法在 6 个基准数据集上的收敛性实验结果如图 6 所示。可以明显地看出，优化模型的目标函数值在快速下降直至收敛，还可以进一步观察到，所提出的算法通常可以在 10 次迭代内收敛。RSFKM 算法可以在较少迭代次数下达到收敛，在各数据集通常在 50

次迭代以内收敛[8]。此外，FKPS算法在各数据集可以在15次迭代以内达到收敛[18]。对比可以得出，ANFCM模型也具有良好的迭代收敛性能。

为了更加直观地展示聚类模型的性能，在算法迭代过程中，随着迭代次数的增加，聚类性能的变化情况如图6所示。可以看出，随着迭代次数逐渐增加，聚类性能逐渐改善。综上，本文提出的ANFCM模型具有优异的收敛特性。

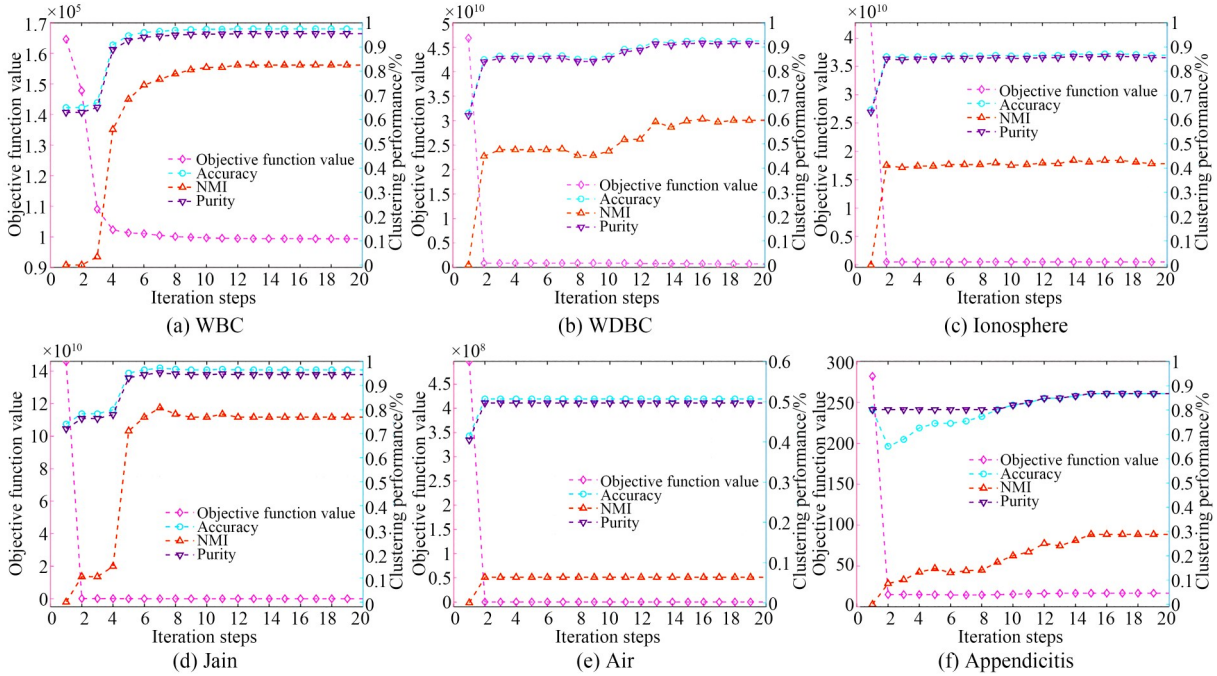


图6　在6个数据集上目标函数值与聚类表现随迭代步数的变化情况

Fig. 6　Changes in objective function values and clustering performance with iteration steps on 6 datasets

4.8　消融实验

消融实验是一种用于评估机器学习模型中各个组成部分对模型性能影响的实验方法。在消融实验中，对模型的某些组成部分进行"切除"，然后观察模型的性能变化，从而确定这些组成部分对模型性能的贡献。

与基础FCM模型相比，本文提出的模型引入局部自适应近邻信息 G_x 与 G_v，即：

$$\min_{U,V} J=\sum_{i=1}^{c}\sum_{j=1}^{N}u_{ij}^{m}\left(\left\| x_j-v_i\right\|_2^2+\alpha\left(G_{x_i}-G_{v_i}\right)^2\right). \tag{29}$$

为体现引入的局部自适应近邻信息对聚类效果的提升，对模型进行如下变动：

（1）ANFCM0：将参数 α 设置为0，同时去除局部自适应近邻信息 G_x 与 G_v，模型退化为基础FCM模型：

$$\min_{U,V} J=\sum_{i=1}^{c}\sum_{j=1}^{N}u_{ij}^{m}\left\| x_j-v_i\right\|_2^2. \tag{30}$$

（2）ANFCM1：去除样本点与近邻样本点之间的位置信息 G_x：

$$\min_{U,V} J=\sum_{i=1}^{c}\sum_{j=1}^{N}u_{ij}^{m}\left(\left\| x_j-v_i\right\|_2^2+\alpha G_{v_i}{}^2\right). \tag{31}$$

（3）ANFCM2：去除样本点与近邻聚类中心点的位置信息 G_v：

$$\min_{U,V} J=\sum_{i=1}^{c}\sum_{j=1}^{N}u_{ij}^{m}\left(\left\| x_j-v_i\right\|_2^2+\alpha G_{x_i}{}^2\right). \tag{32}$$

4个算法在4个基准数据集上进行实验，所有实验结果均在同等条件下，随机初始化聚类中心的10次聚类结果取平均值。对比实验结果如图7所示，可以看出在4个数据集上，ANFCM算法在Accuracy，NMI，Purity 3个评价指标上均取

得最好的效果，在 Ionosphere 数据集和 Jain 数据集上有很大的提升。由此可以得出，加入局部自适应近邻信息 G_x 与 G_v 对提升聚类结果的表现有很大的作用。

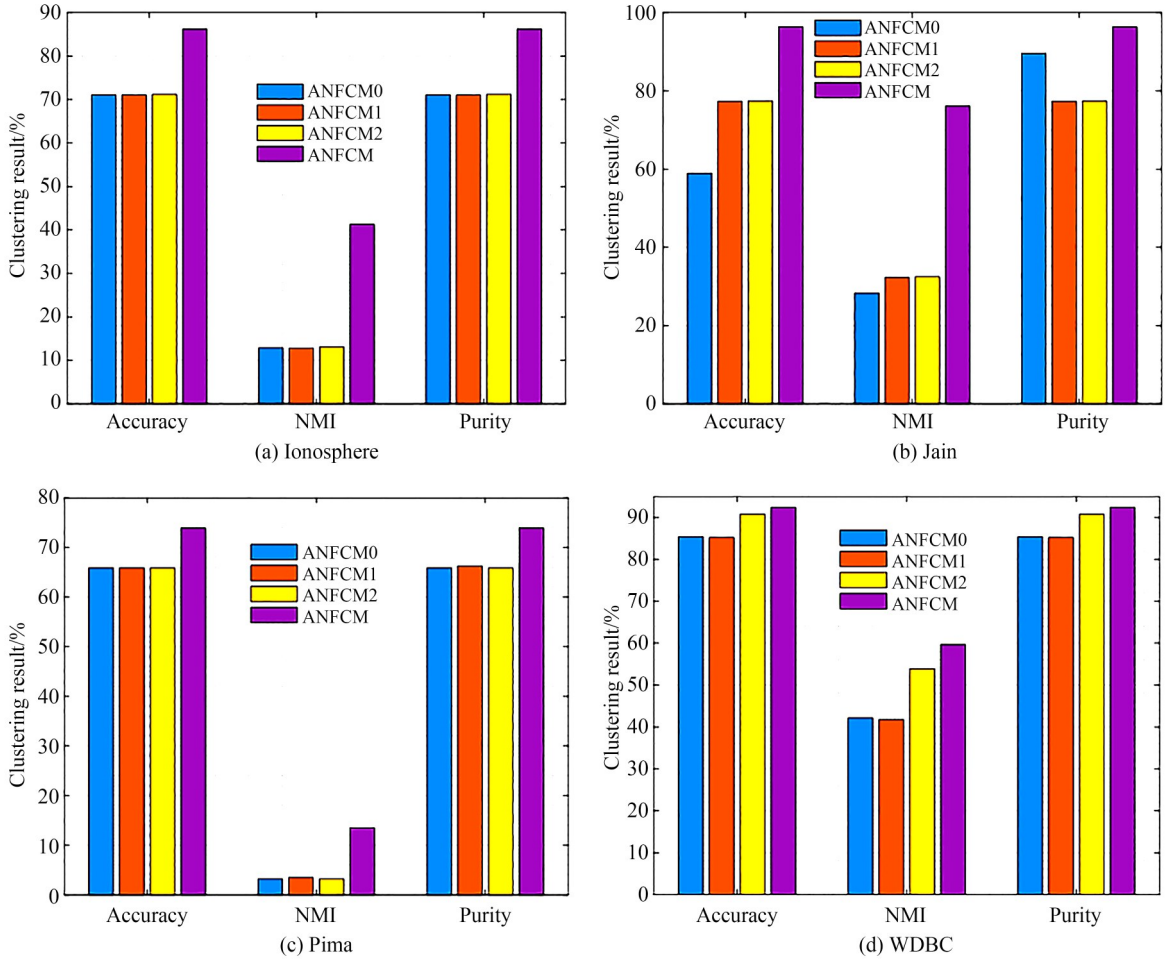


图 7　在 4 个数据集上的消融实验结果

Fig. 7　Ablation experimental results on 4 datasets

5　结　论

本文提出了一种模糊 C 均值聚类算法，同时挖掘簇中心和样本点局部结构信息，指导聚类过程，在保证模糊 C 均值算法优点的同时，减弱噪声、离群点的影响，提升算法的稳定性。通过实验、定量验证以及定性分析了算法的有效性和可行性。但该算法对初始值较为敏感，这是由于 FCM 本质上是非凸优化问题，而算法的实现采用迭代更新的策略，这使得初始值的选取会影响算法的进程。这个问题将在未来做进一步的研究。

参考文献：

[1] MACQUEEN J. Some methods for classification and analysis of multivariate observations[J]. *Proc. Symp. Math. Statist. and Probability*, 5th, 1967, 1.

[2] DUNN J C. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters[J]. *Journal of Cybernetics*, 1973, 3(3): 32-57.

[3] BEZDEK J C, EHRLICH R, FULL W. FCM:

the fuzzy c-means clustering algorithm[J]. *Computers & Geosciences*, 1984, 10(2/3): 191-203.

[4] PAL N R, BEZDEK J C. On cluster validity for the fuzzy c-means model [J]. *IEEE Transactions on Fuzzy Systems*, 1995, 3(3): 370-379.

[5] NG A Y, JORDAN M I, WEISS Y. On Spectral Clustering: Analysis and an algorithm [J]. *proc nips*, 2002.

[6] YU J, YANG M S. Optimality test for generalized FCM and its application to parameter selection[J]. *IEEE Transactions on Fuzzy Systems*, 2005, 13(1): 164-176.

[7] YU J, YANG M S. A generalized fuzzy clustering regularization model with optimality tests and model complexity analysis [J]. *IEEE Transactions on Fuzzy Systems*, 2007, 15(5): 904-915.

[8] XU J L, HAN J W, XIONG K, *et al*. Robust and sparse fuzzy K-means clustering[C]. *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence. July* 9-15, 2016, *New York*, *New York*, *USA*. ACM, 2016: 2224-2230.

[9] CHANG X Y, WANG Q N, LIU Y W, *et al*. Sparse regularization in fuzzy c-means for high-dimensional data clustering[J]. *IEEE Transactions on Cybernetics*, 2017, 47(9): 2616-2627.

[10] ZHANG R, NIE F P, GUO M H, *et al*. Joint learning of fuzzy *k*-means and nonnegative spectral clustering with side information[J]. *IEEE Transactions on Image Processing*, 2019, 28(5): 2152-2162.

[11] Zhang R, Tong H, Xia Y, *et al*. Robust Embedded Deep K-means Clustering[C]. T*he* 28*th ACM International Conference*. ACM, 2019.

[12] CHUANG K S, TZENG H L, CHEN S, *et al*. Fuzzy c-means clustering with spatial information for image segmentation[J]. *Computerized Medical Imaging and Graphics*, 2006, 30(1): 9-15.

[13] CAI W L, CHEN S C, ZHANG D Q. Fast and robust fuzzy C-means clustering algorithms incorporating local information for image segmentation [J]. *Pattern Recognition*, 2007, 40(3): 825-838.

[14] NIE F P, WANG X Q, HUANG H. Clustering and projected clustering with adaptive neighbors [C]. *Proceedings of the* 20*th ACM SIGKDD international conference on Knowledge discovery and data mining. New York New York USA*. ACM, 2014.

[15] ZHOU D Y, BOUSQUET O, LAL T N, *et al*. Learning with local and global consistency [C]. *Proceedings of the* 16*th International Conference on Neural Information Processing Systems. December* 9-11, 2003, *Whistler*, *British Columbia*, *Canada*. ACM, 2003: 321-328.

[16] HOU C P, NIE F P, JIAO Y Y, *et al*. Learning a subspace for clustering via pattern shrinking[J]. *Information Processing and Management: an International Journal*, 2013, 49(4): 871-883.

[17] CHANG X J, NIE F P, MA Z G, *et al*. A convex formulation for spectral shrunk clustering[J]. *ArXiv e-Prints*, 2014: arXiv: 1411.6308.

[18] ZHAO X W, NIE F P, WANG R, *et al*. Robust fuzzy K-means clustering with shrunk patterns learning [J]. *IEEE Transactions on Knowledge and Data Engineering*, 2021: 1.

[19] WU K L, YU J, YANG M S. A novel fuzzy clustering algorithm based on a fuzzy scatter matrix with optimality tests[J]. *Pattern Recognition Letters*, 2005, 26(5): 639-652.

[20] LI M J, NG M K, CHEUNG Y M, *et al*. Agglomerative fuzzy K-means clustering algorithm with selection of number of clusters [J]. *IEEE Transactions on Knowledge and Data Engineering*, 2008, 20(11): 1519-1534.

作者简介：

高云龙(1979—),男,福建厦门人,博士,副教授,2011年于西安交通大学获得博士学位,主要从事统计模式识别、机器学习的研究。E-mail: gaoyl@xmu.edu.cn

通讯作者：

曹　超(1982—),男,博士,硕士生导师,主要从事机器学习与摄影测量在海岸带领域的研究。E-mail: caochao@tio.org.cn