簇间可分的鲁棒模糊C均值聚类算法

高云龙[①] 杨程宇[①] 王志豪[①] 罗斯哲[①] 潘金艳*[②]
[①](厦门大学航空航天学院 厦门 361102)
[②](集美大学信息工程学院 厦门 361021)

摘 要：与经典的K均值聚类算法相比，模糊C均值(FCM)聚类算法通过引入模糊因子，考虑不同聚类数据簇之间的相互关系，得到可分性更好的聚类结果。但是模糊因子的引入，使得任意一个样本点都存在模糊性，造成FCM极易受到噪声和离群点的影响，聚类结果泛化性能较差。因此，该文提出一种簇间可分的鲁棒FCM算法(RBI-FCM)。RBI-FCM利用K均值算法对模糊隶属度的稀疏特征，降低不同数据簇之间的相互作用，突出不同数据簇相邻区域的可分性；另外，RBI-FCM在极小化数据簇内部散布度的条件下，考虑不同数据簇之间的可分性，可提高聚类模型的泛化性能。该文设计了有效的模型求解迭代算法。实验结果表明，RBI-FCM算法提高了FCM的鲁棒性，有效降低FCM对数据簇分布差异性和抽样不均衡的敏感性，得到理想的聚类结果。

关键词：聚类；模糊C均值；样本分布；簇间信息
中图分类号：TP311.13 **文献标识码**：A **文章编号**：1009-5896(2019)05-1114-08
DOI: 10.11999/JEIT180604

Robust Fuzzy C-means Clustering Algorithm Integrating Between-cluster Information

GAO Yunlong[①] YANG Chengyu[①] WANG Zhihao[①] LUO Sizhe[①] PAN Jinyan[②]
[①](*School of Aerospace Engineering*, *Xiamen University*, *Xiamen* 361102, *China*)
[②](*Information Engineering College*, *Jimei University*, *Xiamen* 361021, *China*)

Abstract: Comparing with K-means, Fuzzy logic is introduced in Fuzzy C-Means to handle the information between clusters. It can obtain better cluster results. However, fuzzy logic makes observations could belong to more than just one cluster, which results FCM is especially sensitivity to the noisy and outlier and has poor generalization performance. So a Rrobust Fuzzy C-Means clustering integrated Between-cluster Information algorithm (RBI-FCM) is proposed. Taking advantage of the sparsity of K-means, RBI-FCM helps to reduce the interactions among different clusters and improve the separability of sample points which locate in the adjacent domains of different clusters. Beside minimizing the inner-cluster scattering condition, RBI-FCM considers the between-cluster information. The generalization performance of RBI-FCM can be improved. An effective iterative algorithm for solving the model is designed in this paper. The experimental results show that the RBI-FCM improves the robustness of FCM and reduce effectively its sensitivity to size-imbalance and differences on the distribution of clusters of FCM. The great clustering result is obtained.

Key words: Clustering; Fuzzy C-Means (FCM); Sample distribution; Between-cluster information

1 引言

作为模式识别、数据挖掘等方向重要研究内容之一，聚类分析在识别数据内在结构方面发挥着重要作用，被广泛应用于生物、经济、医学、计算机等众多领域[1−6]。根据聚类过程中数据的集聚规则，聚类算法可以被分为4类：基于划分、基于层次、基于密度和基于网格的聚类算法[7]。其中基于划分的聚类算法因其算法流程简单，运算复杂度低而得到广泛研究与运用，K均值(K-means)算法和模糊C均值(Fuzzy C-Means, FCM)算法是该类型聚类算法中最著名的两个算法。与K-means算法相比，FCM算法引入模糊信息使得其对数据样本的

收稿日期：2018-06-20；改回日期：2018-12-24；网络出版：2018-12-28
*通信作者： 潘金艳 gaoyl@xmu.edu.cn
基金项目：国家自然科学基金(61203176)，福建省自然科学基金(2013J05098, 2016J01756)
Foundation Items: The National Natural Science Foundation of China (61203176), The Natural Science Foundation of Fujian Province (2013J05098, 2016J01756)

划分较为柔性，从而得到更加广泛地关注[8]。

在过去很多年，以FCM算法为基础，很多学者从各个不同方面，提出了众多改进的FCM，取得了一系列新的研究成果。例如：为了加强FCM算法的灵活性，文献[7,8]从自动确定数据集聚类个数入手，提出了改进的模糊聚类算法；为了降低FCM算法对噪声的敏感性，文献[9]通过放松FCM算法的隶属度约束，提出了可能性模糊C均值聚类算法(Possibilistic Fuzzy C-Means, PFCM)；针对传统FCM算法对噪声敏感及对边界样本聚类不准确问题，肖满生等人[10]通过将各个样本间的空间距离引入到聚类影响值中，提出了一种空间相关性与隶属度平滑的FCM改进算法；为了能降低FCM算法对样本容量差异的敏感性，Liu等人[11]通过向模糊隶属度迭代式引入类大小参数，提出了一个新的改进的模糊聚类算法；史慧峰等人[12]通过结合有效性函数，提出了自适应的模糊C均值聚类算法，保证了传统的FCM算法在不依赖初始条件情况下，可以得到正确有效的聚类结果，并可自动判定类别数；文献[13]通过改进FCM算法目标函数，提出了可分性更强的广义模糊C聚类算法(Generalized Fuzzy C-Means clustering algorithm with Improved Fuzzy Partitions, GIFP-FCM)，有效提高了FCM算法聚类性能。

这些文献在一定程度上对传统FCM算法进行了成功改进并取得了较好的研究成果，但也存在一定局限性，具体表现在：(1)任意一个样本点都存在模糊性，造成这些改进的方法极易受到噪声和离群点的影响，算法的鲁棒性较差，泛化性能不强；(2)容易对数据簇形成等势划分，受抽样不均衡和不同数据簇分布特征差异性影响极大。本文在分析国内外最新研究进展的基础上，提出一种簇间可分的鲁棒模糊C均值聚类算法(Robust Fuzzy C-Means clustering integrating Between-cluster Information, RBI-FCM)。RBI-FCM通过利用K-means对模糊隶属度的稀疏性特征，有效降低可靠样本点的模糊性，提高对边缘离散样本点聚类鲁棒性；通过引入整体散布度，在极小化数据簇内部散布度的条件下，考虑不同数据簇之间的可分性，有效提高聚类算法的泛化性能。由于RBI-FCM模型存在众多超参数，因此，本文设计了有效的模型求解迭代算法，将对超参数β_j的设置转化为对单一参数β'的设置，极大降低了参数初始值设置的复杂度，简化了算法对于初始值的选取。实验结果表明本文所提算法有效降低了对抽样不均衡和不同数据簇分布特征差异性的敏感性，具有较强的聚类稳定性。

2　簇间可分的鲁棒模糊C均值聚类

2.1　RBI-FCM算法构想

1969年，著名学者Ruspini首次提出模糊划分概念并将模糊集理论引入聚类分析，1974年，Dunn[14]首次提出模糊C均值聚类算法，并在之后由Bezdek发展起来[15]。FCM算法通过优化目标函数，得到聚类中心与每个样本点对各个类中心的隶属度，并以此对样本点进行类的划分。FCM建模为

$$J(u_{ij},v_i)=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^{m}\parallel x_j-v_i\parallel^2,\quad \sum_{i=1}^{c}u_{ij}=1,j=1,2,\cdots,n \tag{1}$$

其中，$m>1$是模糊系数，c为聚类个数，v_i为第i个聚类中心，x_j为第j个样本点，$u_{ij}\in[0,1]$为隶属度值。

FCM算法对所有样本点都以计算其对类中心的欧式距离作为类划分的标准，使得其受噪声和离群点影响较大，聚类稳定性较低。另外FCM算法主要适用于样本容量均匀、分布规律的情况，对簇分布特征差异性敏感。因此，为了增强聚类稳定性，提高聚类效果，需要对FCM算法目标函数模型进行改进。

2.2　RBI-FCM算法模型

簇间可分的鲁棒模糊C均值聚类算法对FCM目标函数进行改进，基于K-means稀疏化特征对隶属度进行自适应加权以提高算法鲁棒性，增强离群点的聚类有效性，并考虑类间可分性，降低对簇分布差异敏感性，其目标函数如式(2)

$$J(u_{ij},v_i)=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^{m}\parallel x_j-v_i\parallel^2+\sum_{i=1}^{c}\sum_{j=1}^{n}\beta_j u_{ij}(1-u_{ij}^{m-1})\parallel x_j-v_i\parallel^2 \tag{2}$$

其中，β_j是一组超参数。对式(2)目标函数进行简单数学变换可得

$$\min_{U,V}J(u_{ij},v_i)=\sum_{i=1}^{c}\sum_{j=1}^{n}(1-\beta_j)\,u_{ij}^{m}\parallel x_j-v_i\parallel^2+\sum_{i=1}^{c}\sum_{j=1}^{n}\beta_j u_{ij}\parallel x_j-v_i\parallel^2 \tag{3}$$

当m=2有

$$\min_{U,V} J(u_{ij},v_i)=\sum_{i=1}^{c}\sum_{j=1}^{n}(1-\beta_j)\left(u_{ij}+\frac{\beta_j}{2(1-\beta_j)}\right)^2\cdot\parallel x_j-v_i\parallel^2-\frac{\beta_j^2}{4(1-\beta_j)}\cdot\sum_{i=1}^{c}\sum_{j=1}^{n}\parallel x_j-v_i\parallel^2 \tag{4a}$$

对于式(4a)中的第2部分有

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n}\parallel x_j-v_i\parallel^2\\&=\frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n}\parallel(x_j-v)-(v_i-v)\parallel^2\\&=\frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n}\left(\parallel x_j-v\parallel^2-2(x_j-v)\cdot(v_i-v)+\parallel v_i-v\parallel^2\right)\\&=\frac{c}{n}\sum_{j=1}^{n}\parallel x_j-v\parallel^2+\sum_{i=1}^{c}\parallel v_i-v\parallel^2\end{aligned} \tag{4b}$$

其中，$v=\dfrac{1}{n}\sum_{j=1}^{n}x_j$, $c\sum_{j=1}^{n}\parallel x_j-v\parallel^2$为固定常数，优化过程中忽略不计。基于式(4b)，式(4a)变为

$$\min_{U,V} J'(u_{ij},v_i)=\sum_{i=1}^{c}\sum_{j=1}^{n}(1-\beta_j)\left(u_{ij}+\frac{\beta_j}{2(1-\beta_j)}\right)^2\cdot\parallel x_j-v_i\parallel^2-\frac{n\beta_j^2}{4(1-\beta_j)}\sum_{i=1}^{c}\sum_{j=1}^{n}\cdot\parallel v_i-v\parallel^2 \tag{4c}$$

式(4c)的目标函数由两部分确定。其中第1个部分确定每个样本点的模糊隶属度，对于任意一个样本点，$(1-\beta_j)$仅影响多目标组合的权重，而不影响模糊隶属度的计算，因此在计算模糊隶属度的过程中可忽略不计，即在β_j与v_i给定的条件下，模糊隶属度由式(4d)优化问题确定

$$\left.\begin{aligned}&\min_{U}\sum_{i=1}^{c}\sum_{j=1}^{n}\left(u_{ij}+\frac{\beta_j}{2(1-\beta_j)}\right)^2\parallel x_j-v_i\parallel^2\\&\text{s.t.}\ \sum_{i=1}^{c}u_{ij}=1\end{aligned}\right\} \tag{4d}$$

通过与式(1)比较可见，模型式(4d)在约束条件下，通过对模糊隶属度进行平移，提高了样本点属于不同数据簇模糊隶属度的对比度，从而降低了样本点的模糊程度。基于文献[16]和文献[17]可知，这一特性提高了模糊聚类算法对边缘样本点的鲁棒性。

式(4c)的目标函数第2个部分考虑了不同数据簇之间的可分性，在极小化数据簇内散布度的同时，极大化不同聚类数据簇之间的可分性。基于文献[18]可知，这一部分有效降低了FCM算法对抽样不均衡和数据簇分布特征差异性的敏感性。基于这两个属性把式(2)称作簇间可分的鲁棒模糊C均值聚类算法。

2.3 RBI-FCM算法公式推导

由式(2)，建立拉格朗日辅助函数

$$\begin{aligned}L(u_{ij},v_i)=&\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^{m}\parallel x_j-v_i\parallel^2\\&+\sum_{i=1}^{c}\sum_{j=1}^{n}\beta_j u_{ij}\left(1-u_{ij}^{m-1}\right)\parallel x_j-v_i\parallel^2\\&-\sum_{j=1}^{n}\lambda_j\left(\sum_{i=1}^{c}u_{ij}-1\right)\end{aligned} \tag{5}$$

令$d_{ij}^2=\parallel x_j-v_i\parallel^2$，将式(5)对$u_{ij}$求偏导，并令结果等于0，可得

$$\begin{aligned}&\frac{\partial L(u_{ij},v_i)}{\partial u_{ij}}=0\Rightarrow mu_{ij}^{m-1}d_{ij}^2(1-\beta_j)+\beta_j d_{ij}^2-\lambda_j\\&=0\Rightarrow u_{ij}=\left(\frac{\lambda_j-(\beta_j d_{ij}^2)}{md_{ij}^2(1-\beta_j)^{1/(m-1)}}\right)\end{aligned} \tag{6}$$

考虑到$\beta=(\beta_1,\beta_2,\cdots,\beta_n)$为一组超参数，为方便求解，简化参数设置，令

$$\beta'=\frac{\beta_j}{\lambda_j},j=1,2,\cdots,n \tag{7}$$

将式(7)代入式(6)得

$$u_{ij}=\left[\lambda_j(1-\beta' d_{ij}^2)\Big/\left(md_{ij}^2(1-\beta_j)\right)\right]^{1/(m-1)} \tag{8}$$

基于约束$\sum_{i=1}^{c}u_{ij}=1$，由式(8)可得

$$\begin{aligned}&\sum_{i=1}^{c}u_{ij}\\&=\sum_{i=1}^{c}\left[\lambda_j(1-\beta' d_{ij}^2)/\left(md_{ij}^2(1-\beta_j)\right)\right]^{1/(m-1)}=1\\&\Rightarrow\lambda_j^{\frac{1}{m-1}}=\frac{1}{\sum_{i=1}^{c}\left[\left(1-\beta' d_{ij}^2\right)/\left(md_{ij}^2(1-\beta_j)\right)\right]^{1/(m-1)}}\end{aligned} \tag{9}$$

将式(9)代入式(8)得

$$\begin{aligned}u_{ij}&=\frac{1}{\sum_{k=1}^{c}\left(\dfrac{1-\beta' d_{kj}^2}{md_{kj}^2(1-\beta_j)}\right)^{1/(m-1)}}\cdot\left(\frac{1-\beta' d_{ij}^2}{md_{ij}^2(1-\beta_j)}\right)^{1/(m-1)}\\&=\frac{\left(1/d_{ij}^2-\beta'\right)^{1/(m-1)}}{\sum_{k=1}^{c}\left(1/d_{kj}^2-\beta'\right)^{1/(m-1)}}\end{aligned} \tag{10}$$

式(10)即为RBI-FCM算法隶属度迭代更新公式，有效地将对超参数的设置转化为对单一参数β'的设置，成功地简化了算法的参数设置。对聚类中心v_i求偏导，并令结果等于0，可得

$$\begin{aligned}\partial L(u_{ij},v_i)/\partial v_i=&-2\sum_{j=1}^{n}u_{ij}^{m}(x_j-v_i)\\&-2\sum_{j=1}^{n}\beta_j u_{ij}(1-u_{ij}^{m-1})(x_j-v_i)\\=&\,0\end{aligned}\tag{11}$$

简化式(11)得

$$v_i=\frac{\sum_{j=1}^{n}u_{ij}^{m}x_j+\sum_{j=1}^{n}\beta_j u_{ij}(1-u_{ij}^{m-1})x_j}{\sum_{j=1}^{n}u_{ij}^{m}+\sum_{j=1}^{n}\beta_j u_{ij}(1-u_{ij}^{m-1})}\tag{12}$$

由给定条件$\beta'=\dfrac{\beta_j}{\lambda_j}\,(j=1,2,\cdots,n)$，并代入式(9)，建立$\beta_j$的迭代更新式如式(13)

$$\begin{aligned}\beta_j^{(\mathrm{new})}&=\lambda_j\beta'\\&=\left(\frac{1}{\sum_{i=1}^{c}\left(\dfrac{1-\beta' d_{ij}^2}{m d_{ij}^2(1-\beta_j^{(\mathrm{old})})}\right)^{1/(m-1)}}\right)^{m-1}\cdot\beta'\end{aligned}\tag{13}$$

在v_i的迭代更新式(12)中含有超参数β_j，所以在更新之前先需对β_j进行迭代更新计算。

2.4 RBI-FCM算法迭代求解流程

RBI-FCM通过下列步骤确定聚类中心V和样本对各个类的隶属度值U：

步骤 1　设定模糊系数m和聚类个数c，随机挑选c个样本点初始化聚类中心$V^{(0)}$，设定最大迭代次数$t_{\max}$和目标函数误差收敛阈值ε，设定非负参数β'，初始化$\beta_j=0$，令迭代次数$t=0$；

步骤 2　通过式(10)更新隶属度矩阵$\boldsymbol{U}^{(t)}$；

步骤 3　通过式(12)、式(13)更新聚类中心$\boldsymbol{V}^{(t+1)}$。

步骤 4　令$t=t+1$；

步骤 5　若$t<t_{\max}$并且$\left\|J^{(t)}-J^{(t-1)}\right\|>\varepsilon$，重复步骤2, 3, 4，否则算法停止。

2.5 RBI-FCM有效性分析

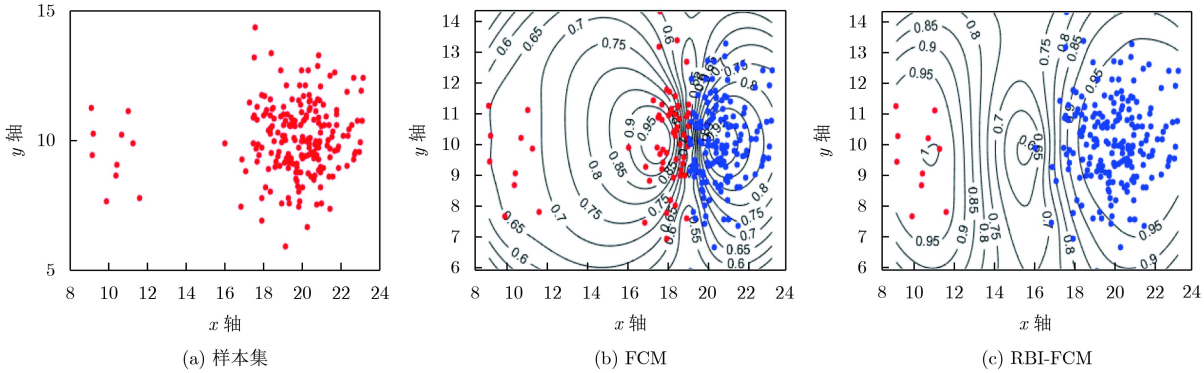
RBI-FCM算法是对FCM算法关于边缘样本有效界定的改进，在极小化类内散布度的条件下，考虑簇间的可分性，减小对数据簇的抽样不均衡性和分布特征差异性的敏感性，增强算法的聚类鲁棒性。如图1所示，当对一个类样本容量有差异的数据集进行聚类时，从聚类结果最大隶属度曲线分布情况不难发现，FCM受数据簇容量差异的影响，小类样本中心发生明显偏移导致样本的误判。RBI-FCM算法基于K-means稀疏化特征，考虑簇与簇之间的可分性，对边缘样本隶属度进行加权处理，减弱了样本数据分布的影响，得到了有效的聚类结果。

3　实验结果分析

为验证本算法的有效性，分别对人造样本数据集和UCI数据集进行了聚类实验，并将聚类结果与FCM, PFCM和GIFP-FCM算法进行比较。其中，为验证RBI-FCM算法的强鲁棒性，对3类人造样本数据集聚类情况进行实验分析，包括各类样本疏密程度有差异的数据集、各类样本容量有差异数据集和各类样本非球形分布数据集。实验环境：Lenovo, CPU: 2 GHz, RAM; 4 GB, Program: MATLAB R2014b。每次实验最大迭代次数$t_{\max}$和收敛阈值ε分别设为100和0.00000001，模糊系数m设为2。

3.1 评价标准

为对实验聚类结果正确率评判，本文采用标准化互信息评价标准(Normalized Mutual Information, NMI)和兰德指数(Rand Index, RI)，标准化互信息以联合熵和个体熵之间的关系来评价聚类结果与标准正确聚类结果的相似度，是常用在聚类中



(a) 样本集　(b) FCM　(c) RBI-FCM

图 1 聚类结果最大隶属度值曲线分布情况

度量聚类结果的一种评价标准。NMI大小为0～1，值越大，意味着聚类结果正确率越高，与正确聚类结果越相近。NMI定义式如式(14)

$$\mathrm{NMI}=\frac{\sum_{i=1}^{c}\sum_{j=1}^{c}P(i,j)\lg\dfrac{P(i,j)}{P(i)P(j)}}{\sqrt{\sum_{i=1}^{c}P(i)\lg P(i)\sum_{j=1}^{c}P(j)\lg P(j)}} \tag{14}$$

其中，i为通过实验所得的聚类结果簇，j为进行对比的正确聚类结果簇。$P(i)$为随机挑选一个样本属于实验聚类结果类i的概率，$P(j)$为随机挑选一个样本属于正确聚类结果类j的概率，$P(i,j)$为随机挑选一个样本点，既属于类i又属于类j的概率。

兰德指数大小为0～1，值越大，意味着聚类结果正确率越高，与正确聚类越相近。RI定义式为

$$\mathrm{RI}=\frac{f_{00}+f_{11}}{n(n-1)/2} \tag{15}$$

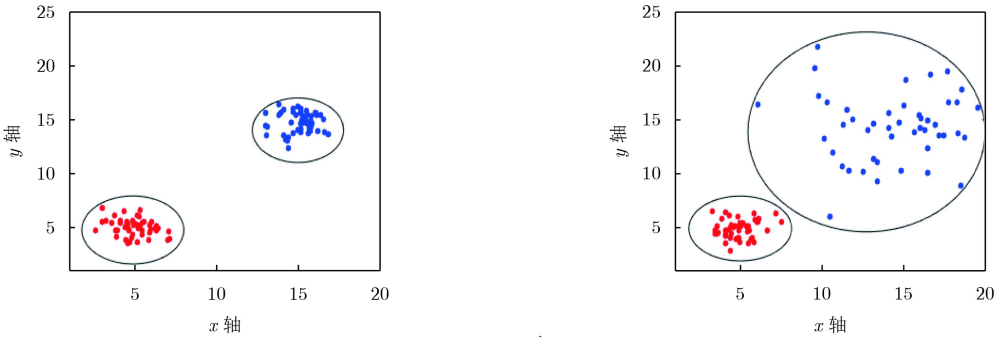
其中，f_{00}是属于不同标签的样本点个数，f_{11}是属于同一标签样本点的个数，n是所有样本点的个数。

3.2 人造数据集聚类实验

实验 1 离散不一的样本数据集聚类实验 如表1所示的一系列人造样本数据集，从第1个样本集到第10个样本集都由两个2维高斯随机分布样本的类别组成，各类样本中心分别为(5, 5)和(15, 15)，类样本数都为50，协方差矩阵其中一类保持[1 0; 0 1]不变，另一类从[1 0; 0 1]变为[10 0; 0 10]。通过协方差矩阵的变化，样本数据集分布变得越来越稀疏，其变化情况如图2所示。在图2中，红色点为第1类数据簇，蓝色点为第2类数据簇，从人造样本数据集1(即图2(a))到人造样本数据集10(即图2(b))，第1类数据簇的疏密分布情况保持不变，第2类数据簇的分布变得越来越稀疏。用FCM和RBI-FCM分别对各样本数据集进行聚类实验，每个样本数据集进行100次聚类试验，计算各样本数据集聚类的NMI平均正确率如图3所示。图3横坐标表示样本数据集序号，纵坐标为对该样本数据集聚类的平均正确率。

表 1 实验1：人造样本数据集主要参数

样本集	类中心	协方差矩阵	各类样本数
1	(5, 5), (15, 15)	[1 0; 0 1], [1 0; 0 1]	50, 50
2	(5, 5), (15, 15)	[1 0; 0 1], [2 0; 0 2]	50, 50
⋮	⋮	⋮	⋮
10	(5, 5), (15, 15)	[1 0; 0 1], [10 0; 0 10]	50, 50



(a) 样本集1 (b) 样本集10

图 2 人造样本疏密分布数据集

从图3聚类结果正确率曲线可见，RBI-FCM算法与FCM算法对类样本松散程度分布都保持一定的敏感度，但在每次试验里，RBI-FCM算法聚类效果略优于传统的FCM算法。

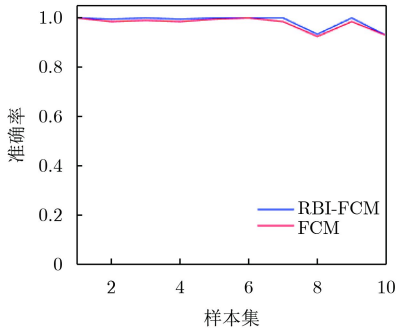


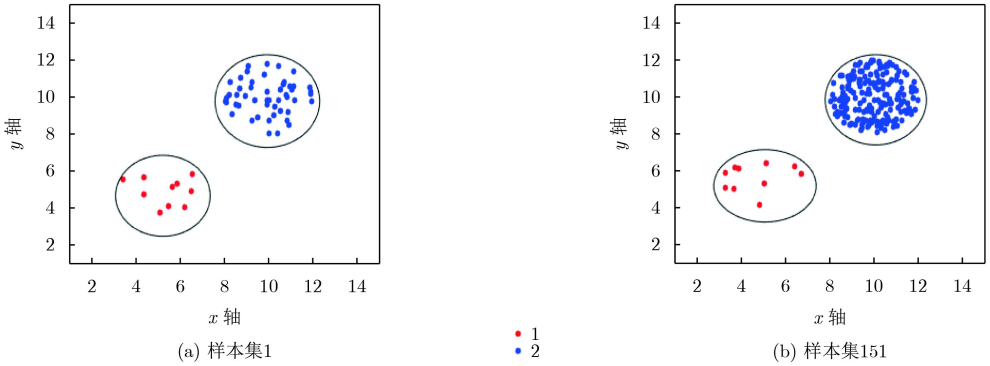
图 3 聚类结果正确率曲线

实验 2 类样本容量差异的数据集聚类实验 如表2所示的一系列人造样本数据集，从第1个样本集到第151个样本集都由两个随机分布样本的类别组成，各类样本点随机分布在一个半径为2的圆内，圆心分别为(5,5)和(10,10)，其中一类样本数保持50不变，另一类样本数从50增加到200。通过类样本数的变化，类样本容量差异性不断增强，其分布变化情况如图4所示。在图4中，红色点为第1类数据簇，蓝色点为第2类数据簇，从人造样本数据集1(即图4(a))到人造样本数据集151(即图4(b))，第

表 2　实验2：人造样本数据集主要参数

样本集	样本随机分布的圆心	各类样本数
1	(5, 5), (15, 15)	50, 50
2	(5, 5), (15, 15)	50, 51
⋮	⋮　⋮	⋮
151	(5, 5), (15, 15)	50, 200

1类数据簇的样本容量保持50不变，第2类数据簇的样本容量从50增加到200。用FCM和RBI-FCM分别对各样本数据集进行聚类实验，每个样本数据集进行100次聚类试验，计算各数据集聚类的NMI平均正确率如图5所示。图5横坐标表示变化类样本容量数，纵坐标为对该样本数据集聚类的平均正确率。



(a) 样本集1　　(b) 样本集151

图 4 人造样本容量分布不均数据集

从图5聚类结果准确率曲线图可见，当变化类样本容量达到120左右时，FCM算法开始变得敏感，聚类结果出现较大波动，而RBI-FCM算法受其变化影响较小，依然保持接近百分之百的正确率。当变化类样本容量达到160时，传统的FCM算法基本上完全崩溃，无法正确聚类，RBI-FCM算法开始变得敏感，但在一定情况下也可得较优的聚类结果，受参数β'影响较大。根据实验结果，RBI-FCM算法相比传统的FCM算法，对类样本容量差异性有更强的包容性。

实验 3　类样本非球形分布数据集聚类实验

如图6(a)、图6(b)所示的非球形分布人造样本数据集，其中图6(a)样本集包含两类，左边样本类呈圆形分布，右边样本类呈长方形分布；图6(b)样本集包含两类，两类皆呈长方形分布。用FCM和RBI-FCM分别对各样本数据集进行聚类实验，聚类结果如图6(c)—图6(f)所示。其中红色点为聚类结果第1类数据簇，蓝色点为聚类结果第2类数据簇。

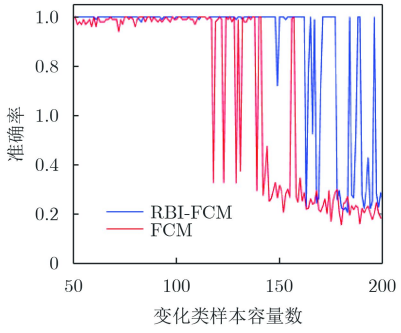


图 5 聚类结果正确率曲线

从图6聚类效果可见，RBI-FCM算法相比FCM算法对样本数据集分布有更好的包容性，聚类结果上保持更好的聚类稳定性。

综合上述3个人造样本数据集实验，本文所提RBI-FCM算法能较好对边缘样本进行有效聚类，成功减弱了FCM算法对样本分布不均衡和数据簇分布差异性数据集聚类的敏感度，有较强的鲁棒性。

3.3　UCI数据集聚类实验

本文选取10个UCI数据集进行聚类实验，每个UCI数据集试验100次，并计算RI与NMI平均正确率，RBI-FCM算法与FCM, PFCM和GIFP-FCM算法对比实验结果如表3所示。由聚类结果可发现，RBI-FCM算法相比FCM, PFCM和GIFP-FCM算法，对UCI数据集聚类表现出更为稳健和有效的聚类效果。

4　结束语

传统的FCM算法没有考虑样本的整体分布，受噪声与离群样本的影响较大，无法对边缘样本进行有效界定，体现出较差的聚类鲁棒性。本文通过引入K-means稀疏化特征，降低可靠样本点的模糊性，引入整体散布度，在极小化内部散布度的条件下，考虑不同数据簇之间的可分性，从而提高FCM的泛化性能。由此改进FCM最优化准则，提出RBI-FCM算法。RBI-FCM算法提高了FCM算法的聚类稳定性，加强了FCM算法对数据簇分布差异和抽样不均衡数据集聚类的鲁棒性。同时，在模型求解、算法公式推导上将对超参数β_j的设置转化为对单一参数β'的设置，简化了参数设置的复杂度，

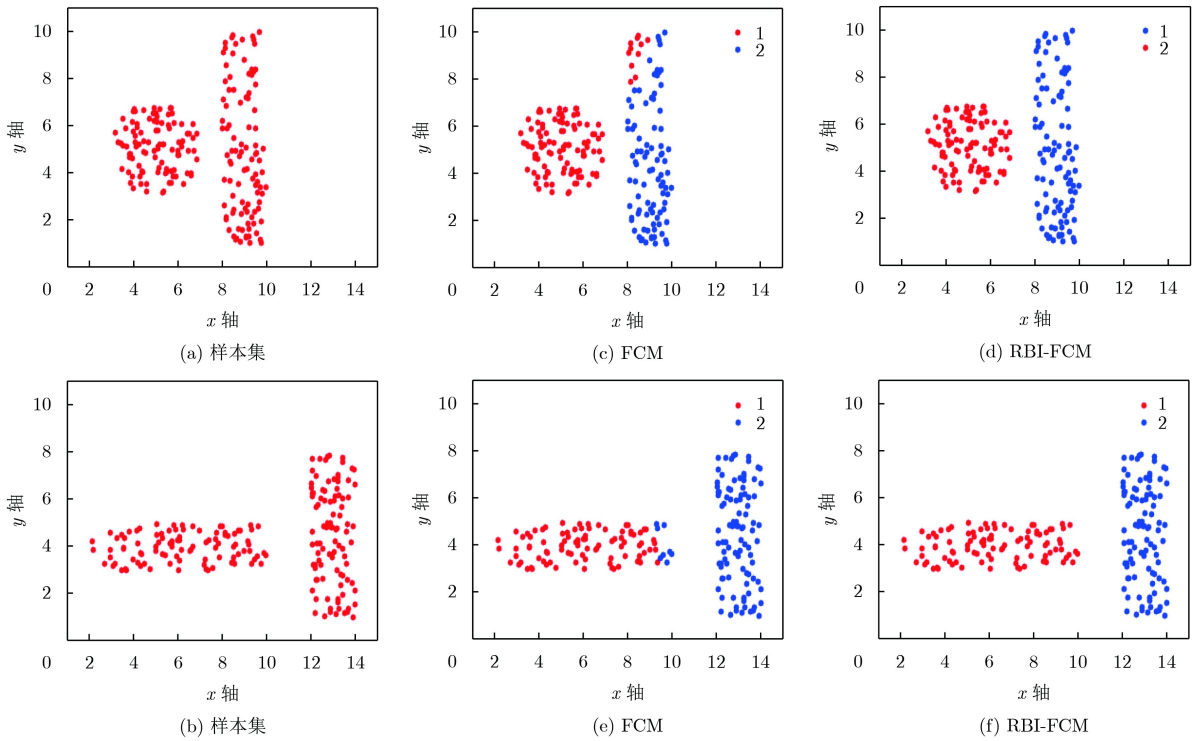


图 6 人造非球形样本数据集及聚类结果

表 3 UCI数据集聚类实验的NMI正确率和RI正确率

UCI数据集	FCM	PFCM	GIFP-FCM	RBI-FCM	UCI数据集	FCM	PFCM	GIFP-FCM	RBI-FCM
Auto-mgp	0.5190	0.5167	0.5008	**0.5443**	Wine	0.4169	0.4168	0.3946	**0.4911**
	0.7534	0.7537	0.7505	**0.7895**		0.7104	0.7105	0.6700	**0.7287**
Zoo	0.6760	0.6824	0.6284	**0.6873**	Balance Scale	0.1223	0.1232	0.1293	**0.1326**
	0.8381	0.8400	0.8236	**0.8464**		0.5887	0.5900	0.5806	**0.5947**
Parkinsons	0.0926	0.0936	0.0526	**0.1071**	House Votes	0.4743	0.4743	0.2917	**0.4948**
	0.5934	0.5934	0.5693	**0.6266**		0.7752	0.7752	0.6688	**0.7890**
Credit Approval	0.0304	0.0304	0.0365	**0.1020**	Vowel	0.3019	0.3127	0.3357	**0.3737**
	0.5048	0.5048	0.5207	**0.5448**		0.7755	0.7988	0.8275	**0.8153**
Banknote Authentication	0.0292	0.0292	0.1145	**0.5249**	Mammographic Masses	0.1054	0.1065	0.1020	**0.1130**
	0.5236	0.5236	0.5555	**0.8053**		0.5676	0.5683	0.5524	**0.5746**

注：每个数据集实验结果的第1行为NMI正确率，第2行为RI正确率

并以上一步所得的β_j求解下一步的β_j，提供了算法设计的新思路。但是，RBI-FCM算法也存在一定的局限性，对于参数β'的设置依赖性太强，一个好的聚类结果依赖一个有效的参数β'的设置，在实际应用中，还需要对参数β'的选取进行理论化的研究。

参 考 文 献

[1] 陈新泉, 周灵晶, 刘耀中. 聚类算法研究综述[J]. 集成技术, 2017, 6(3): 41–49. doi: 10.3969/j.issn.2095-3135.2017.03.004.
CHEN Xinquan, ZHOU Lingjing, and LIU Yaozhong. Review on clustering algorithms[J]. *Journal of Integrati on Technology*, 2017, 6(3): 41–49. doi: 10.3969/j.issn.2095-3135.2017.03.004.

[2] 张传锦, 李璐璐. 基于模糊C均值聚类的无线传感器网络节点定位算法[J]. 电子设计工程, 2016, 24(8): 58–60. doi: 10.14022/j.cnki.dzsjgc.2016.08.017.
ZHANG Chuanjin and LI Lulu. Improving multilateration algorithm based on fuzzy C-means cluster in WSN[J]. *Electronic Design Engineering*, 2016, 24(8): 58–60. doi: 10.14022/j.cnki.dzsjgc.2016.08.017.

[3] 池桂英, 王忠华. 基于分层的直觉模糊C均值聚类图像分割算法[J]. 计算机工程与设计, 2017(12): 3368–3373. doi: 10.16208/j.issn1000-7024.2017.12.031.
CHI Guiying and WANG Zhonghua. Intuitionistic fuzzy C-means clustering algorithm based on hierarchy for image segmentation[J]. *Computer Engineering and Design*,

2017(12): 3368–3373. doi: 10.16208/j.issn1000-7024.2017.12.031.

[4] 黄艳国, 罗云鹏. 基于改进模糊C均值聚类算法的城市道路状态判别方法[J]. 科学技术与工程, 2018, 18(9): 335–342. doi: 10.3969/j.issn.1671-1815.2018.09.052.
HUANG Yanguo and LUO Yunpeng. Identification method of urban road condition based on improved fuzzy C-means method clustering algorithm[J]. *Science Technology and Engineering*, 2018, 18(9): 335–342. doi: 10.3969/j.issn.1671-1815.2018.09.052.

[5] 赵泉华, 刘晓燕, 赵雪梅, 等. 基于可变类FCM算法的多光谱遥感影像分割[J]. 电子与信息学报, 2018, 40(1): 157–165. doi: 10.11999/JEIT170397.
ZHAO Quanhua, LIU Xiaoyan, ZHAO Xuemei, *et al*. Multispectral remote sensing image segmentation based on FCM algorithm with unknown number of clusters[J]. *Journal of Electronics & Information Technology*, 2018, 40(1): 157–165. doi: 10.11999/JEIT170397.

[6] XU Rui and WUNSCH D. Survey of clustering algorithms[J]. *IEEE Transactions on Neural Networks*, 2005, 16(3): 645–678. doi: 10.1109/tnn.2005.845141.

[7] 陈海鹏, 申铉京, 龙建武, 等. 自动确定聚类个数的模糊聚类算法[J]. 电子学报, 2017, 45(3): 687–694. doi: 10.3969/j.issn.0372-2112.2017.03.028.
CHEN Haipeng, SHEN Xuanjing, LONG Jianwu, *et al*. Fuzzy clustering algorithm for automatic identification of clusters[J]. *Acta Electronica Sinica*, 2017, 45(3): 687–694. doi: 10.3969/j.issn.0372-2112.2017.03.028.

[8] YANG MiinShen and NATALIANI Y. Robust-learning fuzzy c-means clustering algorithm with unknown number of clusters[J]. *Pattern Recognition*, 2017, 71: 45–59. doi: 10.1109/nafips.2010.5548175.

[9] PAL N R, PAL K, KELLER J M, *et al*. A possibilistic fuzzy C-means clustering algorithm[J]. *IEEE Transactions on Fuzzy Systems*, 2005, 13(4): 517–530. doi: 10.1109/tfuzz.2004.840099.

[10] 肖满生, 肖哲, 文志诚, 等. 一种空间相关性与隶属度平滑的FCM改进算法[J]. 电子与信息学报, 2017, 39(5): 1123–1129. doi: 10.11999/JEIT160710.
XIAO Mansheng, XIAO Zhe, WEN Zhicheng, *et al*. Improved FCM clustering algorithm based on spatial correlation and membership smoothing[J]. *Journal of Electronics & Information Technology*, 2017, 39(5): 1123–1129. doi: 10.11999/JEIT160710.

[11] LIU Yun, HOU Tao, and LIU Fu. Improving fuzzy c-means method for unbalanced dataset[J]. *Electronics Letters*, 2015, 51(23): 1880–1882. doi: 10.1049/el.2015.1541.

[12] 史慧峰, 马晓宁. 一种自适应的模糊C均值聚类算法[J]. 无线通信技术, 2016, 25(3): 40–45. doi: 10.3969/j.issn.1003-8329.2016.03.009.
SHI Huifeng and MA Xiaoning. An adaptive fuzzy C-means clustering algorithm[J]. *Wireless Communication Technology*, 2016, 25(3): 40–45. doi: 10.3969/j.issn.1003-8329.2016.03.009.

[13] 曲福恒. 模糊聚类算法及应用[M]. 北京: 国防工业出版社, 2011.
QU Fuheng. Fuzzy clustering algorithm and its application[M]. Beijing, National Defense Industry Press, 2011.

[14] DUNN J C. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters[J]. *Journal of Cybernetics*, 1974, 3(3): 32–57. doi: 10.1080/01969727308546046.

[15] BEZDEK J C. Pattern Recognition with Fuzzy Objective Function Algorithms[J]. *Springer US*, 1981. doi: 10.1007/978-1-4757-0450-1.

[16] ZHU Lin, CHUNG FuLai, and WANG Shitong. Generalized fuzzy C-means clustering algorithm with improved fuzzy partitions[J]. *IEEE Transactions on Systems Man & Cybernetics Part B Cybernetics A*, 2009, 39(3): 578–591. doi: 10.3724/sp.j.1087.2013.02355.

[17] HÖPPNER F and KLAWONN F. Improved fuzzy partitions for fuzzy regression models[J]. *International Journal of Approximate Reasoning*, 2003, 32(2): 85–102. doi: 10.1016/s0888-613x(02)00078-6.

[18] DENG Zhaohong, CHOI K S, CHUNG Fulai, *et al*. Enhanced soft subspace clustering integrating within-cluster and between-cluster information[J]. *Pattern Recognition*, 2010, 43(3): 767–781. doi: 10.1016/j.patcog.2009.09.010.

高云龙：男，1979年生，副教授，研究方向为机器学习、时间序列分析和生产制造系统优化与调度.

杨程宇：男，1996年生，本科生，研究方向为机器学习.

王志豪：男，1993年生，硕士生，研究方向为模式识别和机器学习.

罗斯哲：男，1995年生，硕士生，研究方向为维数约简、模式识别和机器学习.

潘金艳：女，1978年生，副教授，研究方向为人工智能和机器学习理论与方法.