基于自适应松弛的鲁棒模糊C均值聚类算法

高云龙[①] 王志豪*[①] 潘金艳[②] 罗斯哲[①] 王德鑫[①]

[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：噪声是影响聚类结果的最重要的因素之一，现有的模糊聚类算法主要通过对隶属度约束进行松弛的方式来降低噪声样本的影响。这种方式仍然存在两个基本问题需要解决：第一，如何评估一个样本是噪声的可能性；第二，如何在抑制噪声样本影响力的同时，保留正常样本的作用力。针对这两问题，该文提出了基于自适应松弛的鲁棒模糊C均值聚类算法(AR-RFCM)。新模型基于K最近邻的方式(KNN)来估计样本的可靠性，自适应地调整松弛参数，从而实现在降低噪声样本影响力的同时，保留可靠样本的作用力。此外，AR-RFCM利用了C均值聚类模型中隶属度的稀疏性来提高可靠样本的作用力，从而提高数据簇的内聚程度，进而降低噪声样本的影响。实验表明，AR-RFCM不仅在处理噪声样本时具有良好的鲁棒性，同时在25个UCI 数据集实验中，分类正确率(兰德指数)平均高于FCM算法7.7864%。

关键词：噪声；聚类；模糊C均值；自适应；松弛

中图分类号：TP391; TP273 **文献标识码**：A **文章编号**：1009-5896(2020)07-1774-08

DOI: 10.11999/JEIT190556

Robust Fuzzy C-Means Based on Adaptive Relaxation

GAO Yunlong[①] WANG Zhihao[①] PAN Jinyan[②] LUO Sizhe[①] WANG Dexin[①]

[①](*College of Aeronautics and Astronautics, Xiamen University, Xiamen* 361102, *China*)

[②](*College of Information Engineering, Jimei University, Xiamen* 361021, *China*)

Abstract: Noise is one of the most important influences for clustering. Existing fuzzy clustering methods try to reduce the impact of noise by relaxing the constraint condition of membership. But there are still two basic problems to be solved. The first is how to evaluate the probability that a sample point is a noise. The second is how to retain the effect of normal points while suppressing the impact of noise. To solve these two problems, Robust Fuzzy C-Means based on Adaptive Relaxation (AR-RFCM) is proposed. The new model estimates the reliability of sample points by the method of the K-Nearest Neighbor (KNN). It adjusts adaptively the relaxation parameters to reduce the impact of noise, and keeps the effect of reliable sample points at the same time. In addition, AR-RFCM utilizes the sparsity of membership in K-means to improve the effect of reliable sample points. Therefore, the compactness of clusters is improved and the impact of noise is suppressed. Experiments demonstrate that AR-RFCM has a good robustness for noise, and also achieves higher rand index in all 25 UCI data sets, even averagely higher than FCM 7.7864%.

Key words: Noise; Clustering; Fuzzy C-Means (FCM); Adaptive; Relaxation

1 引言

聚类是模式识别中最重要的无监督学习方法之一，常用于学习并揭示数据的本质结构和潜在关系。在过去的几十年中，成百上千的聚类算法被提出[1]。目前受到最广泛关注的聚类算法主要包含如下4大类：(1)基于划分的聚类算法[2,3]；(2)基于密度的聚类算法[4,5]；(3)基于图论的聚类算法[6,7]；(4)基于其它模型的聚类算法[8,9]。其中基于划分的聚类算法，特别是K均值聚类算法(K-Means)[10]，受到广泛的关注和研究。

为了解决K-Means的聚类中心趋同性问题，Bezdek等人[11]在模糊理论的基础上，提出了模糊C均值(Fuzzy C-Means, FCM)聚类算法。模糊逻

收稿日期：2019-07-24；改回日期：2020-03-13；网络出版：2020-04-09

*通信作者： 王志豪 zhwang@stu.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 Provincial Natural Science Foundation of Fujian Province (2013J05098, 2016J01756)

辑的引入使得每一个样本可以在不同程度上同时隶属于不同的数据簇，使得样本对每个数据簇都存在作用力，引入了簇间作用力的概念。因此FCM能得到比K-Means可分性更好的数据簇结构。然而和绝大多数学习算法相同，FCM也存在受噪声(包括异常值)影响较大的问题。其原因主要在于除簇中心外，任意样本点的隶属度均会被分摊到不同数据簇中：(1)本点无论距离簇中心多远，都能对数据簇中心产生较大的拉扯力，增大了噪声样本点的影响力；(2)具有明确类别特征的样本点的隶属度仍会被分配到其它数据簇中，降低了各数据簇的内聚程度，间接增大了噪声样本点的影响力。

为了解决这一问题，有学者提出了这样的观点：对于噪声样本，应该降低它属于所有数据簇的隶属度[12]。许多有关噪声的模糊聚类算法正是基于该观点所提出的。为了降低噪声样本的隶属度，Dave[13]通过对整个数据集添加一个虚拟数据簇的方式，提出了噪声聚类算法(Noise Clustering, NC)来松弛样本的隶属度约束。Krishnapuram和Keller[14]在可能性模糊C均值聚类算法 (Possibilistic C-Means, PCM)中，通过给每个数据簇增加一个相应的虚拟数据簇的方式，来松弛样本的隶属度约束，从而降低噪声样本的影响。此外，Zarinbal等人[15]把相对熵加到FCM的模型中，提出了基于相对熵的模糊C均值聚类算法(Relative Entropy Fuzzy C-Means, REFCM)。REFCM引入朗伯函数求解模型，松弛了样本的隶属度约束条件，同时利用了相对熵的数学特性，提高了数据簇之间的可分性。尽管这些模糊聚类算法在一定程度上降低了噪声样本的影响力，却存在着一定的局限性，具体表现为：(1)近乎无差别地处理所有样本(包括噪声样本和正常样本)，使得噪声样本的绝对作用力被弱化的同时，正常样本的绝对作用力也被弱化；(2)模糊隶属度的引入使得任意样本对任意数据集都能造成影响。对于具有明确类别特征的样本而言，其作用力仍被分摊到其它非所属数据簇上，使得各数据簇的内聚程度降低，间接地增大了噪声样本的影响。

本文在分析了国内外最新研究的基础上，提出了基于自适应松弛的鲁棒模糊C均值聚类算(Robust Fuzzy C-Means based on Adaptive Relaxation, AR-RFCM)。AR-RFCM在噪声聚类(NC)的基础上，基于K最近邻方法(K-Nearest Neighbor, KNN)对样本可靠性的估计，自适应调整不同样本的松弛参数，有差异地处理可靠样本和噪声样本。因此，AR-RFCM能在降低噪声样本影响力的同时保留可靠样本的作用力，真正地降低了噪声样本的相对作用力。此外，AR-RFCM利用了K-Means模型中对隶属度的稀疏表征，提高了数据簇中具有明确类别特征样本的作用力，进而增大数据簇的类内聚结度，间接地弱化了噪声样本的影响。实验结果表明：(1)AR-RFCM在松弛异常值样本隶属度约束的同时(使异常值样本隶属度和远低于1)，保留了正常样本的作用力；(2)AR-RFCM在处理含人造噪声的异常值数据集时，展现了算法对噪声良好的鲁棒性；(3) AR-RFCM在处理真实数据集实验中取得了明显高于其它比较算法的分类指标。在25个UCI 数据集上，AR-RFCM取得了平均高于FCM算法7.7864%的分类正确率(兰德指数)，体现了算法的实用性。

2　算法模型

2.1　模糊C均值聚类算法

FCM模型如式(1)所示

$$\left.\begin{aligned}&\min_{U,V} J_{\mathrm{FCM}}(\boldsymbol{U},\boldsymbol{V})=\sum_{i=1}^{n}\sum_{j=1}^{c}u_{ij}^{m}d_{ij}^{2}\\&\text{s.t.}\quad 0\le u_{ij}\le 1,\sum_{i=1}^{c}u_{ij}=1,0<\sum_{i=1}^{c}u_{ij}<n\end{aligned}\right\}\tag{1}$$

其中，n为样本个数，c为数据簇个数，m是模糊系数($m>1$), u_{ij}是第i个样本x_i隶属于第j个数据簇的隶属程度，d_{ij}是第i个样本x_i到第j个数据簇的簇中心v_j的欧式距离。

2.2　噪声聚类算法

NC聚类算法通过引入虚拟的第c类的方式，分摊噪声样本的隶属度值。其模型为

$$\left.\begin{aligned}&\min_{U,V} J_{\mathrm{NC}}(\boldsymbol{U},\boldsymbol{V})=\sum_{i=1}^{n}\sum_{j=1}^{c}u_{ij}^{m}d_{ij}^{2}\\&\text{s.t.}\quad 0\le u_{ij}\le 1,\sum_{i=1}^{c}u_{ij}=1,0<\sum_{i=1}^{c}u_{ij}<n\end{aligned}\right\}\tag{2}$$

其中，$d_{ij}^2=\|x_i-v_j\|^2\,(j=1,2,\cdots,c-1)$，而对于虚拟数据簇而言$d_{ic}^2=\delta$，其中$\delta$为给定的松弛参数。其它符号均和FCM模型中的符号相同。

从NC模型中可得，所有样本到虚拟的第c类中心的欧式距离均为给定的常数，并没有考虑到样本之间的差异，在降低噪声样本作用力的同时降低了正常样本的作用力，因而不能很好地解决噪声问题。

2.3　模型及其分析

在无监督学习中，噪声的界定一直是一大难题。在数据集中，当一个样本周围存在若干相近的样本时，该样本为异常值的可能性较低。此时如果把该样本区别于周围的样本而当作噪声样本来处理，显然是不合理的。而当某一样本远离其它样本

时，那么这个样本是噪声的可能性较大。因此，我们可以通过数据集的分布情况，来辨别样本的可靠程度。通过增大可靠样本的作用力，可以在提高各数据簇内聚程度的同时，间接地降低噪声样本的作用力；而通过降低不可靠样本的作用力，可以直接地降低噪声样本的影响。

基于上述观点，本文提出了一种基于评估样本可靠性，实现在降低噪声样本影响力的同时，保证可靠样本作用力的聚类算法，称为基于自适应松弛的鲁棒模糊C均值聚类算法(AR-RFCM)。AR-RFCM的初始模型为

$$\left.\begin{aligned}&\min_{U,V} J_{\mathrm{AR-RFCM}}(\boldsymbol{U},\boldsymbol{V})=\sum_{i=1}^{n}\sum_{j=1}^{c}\left(u_{ij}^{m}+\alpha u_{ij}\right)d_{ij}^{2}\\&\quad+\sum_{i=1}^{n}\beta_i\left[\left(1-\sum_{j=1}^{c}u_{ij}\right)^{m}+\alpha\left(1-\sum_{j=1}^{c}u_{ij}\right)\right]\\&\text{s.t.}\quad 0\le u_{ij}\le 1,\sum_{j=1}^{c}u_{ij}\le 1,0<\sum_{j=1}^{n}u_{ij}<n\end{aligned}\right\}\tag{3}$$

其中，α为比例系数，β_i为自适应松弛系数，其它符号均与NC模型相同。

通常情况下$m=2$，最终模型可以简化为

$$\left.\begin{aligned}&\min_{U,V} J_{\mathrm{AR-RFCM}}(\boldsymbol{U},\boldsymbol{V})=\sum_{i=1}^{n}\sum_{j=0}^{c}\left(u_{ij}^{2}+\alpha u_{ij}\right)d_{ij}^{2}\\&\text{s.t.}\quad 0\le u_{ij}\le 1,\sum_{j=0}^{c}u_{ij}=0,0<\sum_{j=1}^{n}u_{ij}<n\end{aligned}\right\}\tag{4}$$

其中，虚拟类满足$d_{i0}^{2}=\beta_i=k\beta\Big/\sum_{l=1}^{k}\|x_i-N_l(x_i)\|^2$，$N_l(x_i)$为样本$x_i$的第$l$近邻样本。

在AR-RFCM模型中，引入虚拟的第0类后，对于任意样本x_i而言，其隶属度满足$u_{i0}=1-\sum_{j=1}^{c}u_{ij}$。自适应松弛参数$\beta_i$可以看作是任意样本$x_i$到虚拟的第0类簇中心的欧式距离。以KNN的方式求样本的邻域方差$\sum_{l=1}^{k}\|x_i-N_l(x_i)\|^2/k$来估计样本的可靠性，满足$d_{i0}^{2}=\beta_i=k\beta\Big/\sum_{l=1}^{k}\|x_i-N_l(x_i)\|^2$，其中$\beta$为给定参数。当样本的邻域内紧凑性较高时，该样本是噪声的可能性越低，样本的可靠性越高；当样本的邻域内紧凑性较低时，该样本属于该数据集的可能性越低，是噪声的可能性较大。此外，该模型中，在k给定的基础上，每个样本的可靠性可以通过预处理的方式进行计算，不需要随着优化算法迭代，这样可以避免KNN的方法在迭代过程中带来的难以承受的计算复杂度。

为了充分比较FCM, K-Means和AR-RFCM，本文将K-Means, FCM和AR-RFCM 3种算法在1维数据集(包含2类数据)上学得的隶属度分布函数展示在图1中。在FCM 模型中，无论样本离数据簇中心有多远，该样本属于该类的隶属度都不为0。此时簇间作用力不仅由位于数据簇交界处的样本提供，还由距离较远的样本提供。在K-Means模型中，样本属于数据簇的隶属度值非0即1，数据簇间不存在簇间作用力。在AR-RFCM模型中，对位于数据簇交界处的样本而言，保留了样本对非所属数据簇的簇间作用力，相比于K-Means提高了聚类算法的簇间可分性；而对于远离相交区域的样本，隶属于所属数据簇的隶属度为1，相比于FCM提高了数据簇的内聚程度，间接地降低了噪声样本的影响力。在这种隶属度分布结构下，AR-RFCM突出了数据簇交界处样本在数据簇边界划分时起的作用力，进而提升了数据簇间数据簇间的可分性，提高了AR-RFCM学得的数据簇结构的泛化能力。

2.4 模型求解

由AR-RFCM算法模型式(4)，建立拉格朗日辅助函数

$$\begin{aligned}L_1(\boldsymbol{U},\boldsymbol{\Lambda})=&\sum_{i=1}^{n}\sum_{j=0}^{c}\left(u_{ij}^{2}+\alpha u_{ij}\right)d_{ij}^{2}\\&-\sum_{i=1}^{n}\lambda_i\left(\sum_{j=0}^{c}u_{ij}-1\right)\end{aligned}\tag{5}$$

对隶属度$u_{ij}\,(i=1,2,\cdots,n;j=1,2,\cdots,c)$和辅助参数$\lambda_i$分别求偏导数得

$$\frac{\partial L_1(\boldsymbol{U},\boldsymbol{\Lambda})}{\partial u_{ij}}=(2u_{ij}+\alpha)d_{ij}^{2}-\lambda_i=0\tag{6}$$

$$\frac{\partial L_1(\boldsymbol{U},\boldsymbol{\Lambda})}{\partial \lambda_i}=\sum_{j=0}^{c}u_{ij}-1=0\tag{7}$$

化简后联立两式可得

$$u_{ij}=\frac{1+\dfrac{c\alpha}{2}}{\displaystyle\sum_{k=0}^{c}\frac{d_{ij}^{2}}{d_{ik}^{2}}}-\frac{\alpha}{2}\tag{8}$$

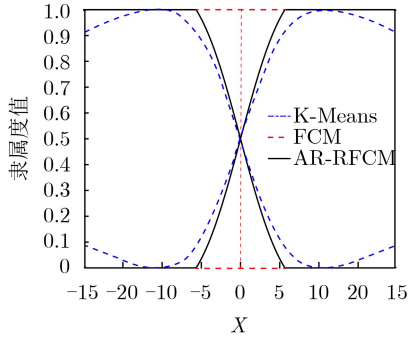


图 1 不同聚类算法的隶属度函数

簇中心的迭代计算式对应的辅助函数为

$$L_2(\boldsymbol{V})=\sum_{i=1}^{n}\sum_{j=1}^{c}\left(u_{ij}^2+\alpha u_{ij}\right)\|x_i-v_j\|^2+\sum_{i=1}^{n}\beta_i\left[\left(1-\sum_{j=1}^{c}u_{ij}\right)^2+\alpha\left(1-\sum_{j=1}^{c}u_{ij}\right)\right] \tag{9}$$

对$v_j\,(j=1,2,\cdots,c)$求导得

$$\frac{\partial L_2(\boldsymbol{V})}{\partial v_j}=-2\sum_{i=1}^{n}\left(u_{ij}^2+\alpha u_{ij}\right)(x_i-v_j)=0 \tag{10}$$

化简得

$$v_j=\frac{\sum_{i=1}^{n}\left(u_{ij}^2+\alpha u_{ij}\right)x_i}{\sum_{i=1}^{n}\left(u_{ij}^2+\alpha u_{ij}\right)} \tag{11}$$

2.5 算法求解步骤

AR-RFCM算法求解步骤如下所示：

步骤 1　确定模糊系数m，数据簇个数c，设定比例系数α，松弛系数β, KNN算法参数k，最大迭代次数$t_{\max}$和目标函数收敛阈值ε，令迭代次数t=0；

步骤 2　随机选取c个数据簇中心，计算每个样本的可靠性估计参数β_i；

步骤 3　通过式(8)更新隶属度矩阵$\boldsymbol{U}$；

步骤 4　通过式(11)更新隶属度矩阵$\boldsymbol{V}$；

步骤 5　令t=t+1，若t<$t_{\max}$并且$\|J(t)-J(t+1)\|>\varepsilon$，则返回到步骤3，反之算法迭代停止。

3　实验结果

为了验证AR-RFCM聚类算法的性能，本文分别在人造数据集(含噪声)和UCI真实数据集上进行了实验，选择FCM, PCM, NC, REFCM 和FDCM_SSR[16]算法作为AR-RFCM的比较算法。实验环境: PC, CPU: 2.60 GHz, RAM: 8 GB, Program: MATLAB R2017a。

3.1 实验评价指标

UCI真实数据集实验中采用兰德指数(Rand Index, RI)来衡量聚类结果的分类准确率。对于人造数据集而言，实验采用上文提及的准确率(accuracy)，精确度(precision)，灵敏度(sensitivity)和特异度(specificity)指标来衡量聚类算法的性能。其中RI的定义为

$$\mathrm{RI}=\frac{f_{00}+f_{11}}{n(n-1)/2} \tag{12}$$

其中，f_{00}是样本中，在聚类结果中属于同一类，并且在数据集标签中也属于同一类的样本对数，f_{11}是样本中，在聚类结果中属于不同类，并且在数据集标签中也属于不同类的样本对数，n是样本个数。RI可以看作是利用样本之间的关系来衡量聚类结果正确率的指标，满足RI∈[0, 1]。RI数值越大，表示聚类结果越好。

准确率、精确度、灵敏度和特异度等指标定义为

$$\left.\begin{aligned}\text{准确率}&=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}\\ \text{精确度}&=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}\\ \text{特异度}&=\frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}\\ \text{灵敏度}&=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{TN}}\end{aligned}\right\} \tag{13}$$

其中，TP是分类正确的正类样本的个数，TN是分类正确的负类样本的个数，FP是分类错误的正类样本的个数，FN是分类错误的负类样本的个数。准确率、精确度、灵敏度和特异度数值都是在0到1之间，且数值越大表示聚类效果越好。

3.2 人造数据集

为了验证AR-RFCM算法对各类噪声的鲁棒性，这里准备了2种不同数据集，分别是包含了异常值和非异常值噪声的人造数据集。

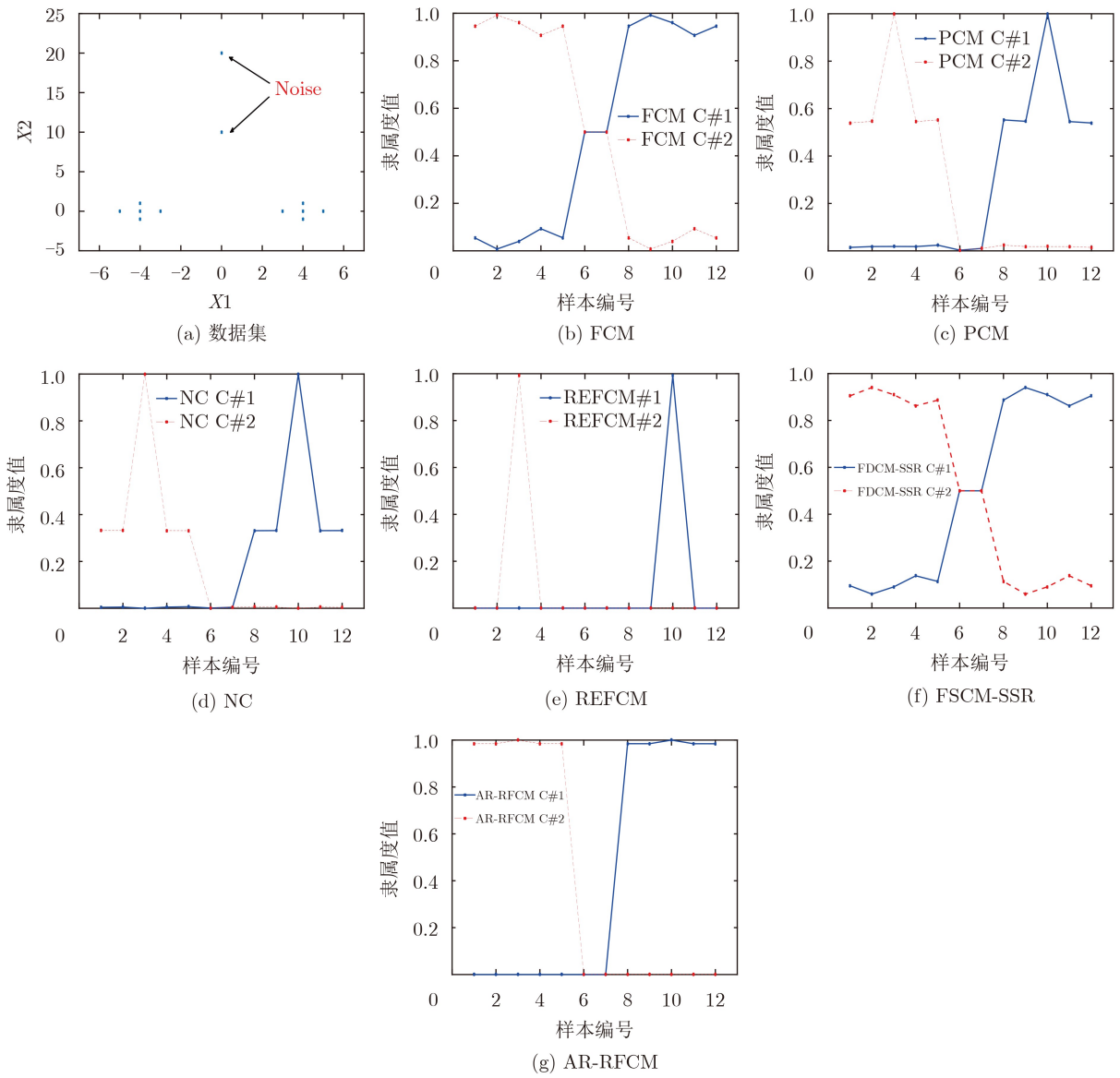
3.2.1 异常值数据集

数据集如图2(a)所示，包含2个可分性较好的数据簇，每个数据簇包含5个样本。此外，数据集中还加入了两个异常值样本。在图2(a)中，样本按照从左到右，从上到下的顺序编号。AR-RFCM及其比较算法的隶属度函数分别如图2(b)—图2(f)所示，具体的隶属度值也被记录在表1中。

如表1、表2和图2所示，不同的模糊聚类算法在处理异常值样本时，表现出了不同的特性。对于FCM和FDCM_SSR算法而言，异常值样本的隶属度满足约束条件，因而两个可分性较好的数据簇受到噪声样本较大的拉扯力，数据簇中心发生了偏移。而对于PCM, NC, REFCM算法而言，尽管异常值样本的隶属度约束条件被松弛了，降低了异常值样本的影响，数据簇中心没有发生偏移，但其他正常样本的作用力同时也被降低了，数据簇的内聚程度被降低。相反地，AR-RFCM 在降低异常值样本作用的同时，很好地保留了正常样本的作用力，增大了各数据簇的内聚程度，提高了算法的鲁棒性。

3.2.2 噪声数据集

第2个实验是在噪声数据集上展开，用来证明



(a) 数据集 (b) FCM (c) PCM

(d) NC (e) REFCM (f) FSCM-SSR

(g) AR-RFCM

图 2 不同聚类算法的隶属度函数

表 1 各算法求得的隶属度值

样本编号	FCM		PCM		NC		REFCM		FDCM_SSR		AR-RFCM	
	C#1	C#2	C#1	C#2	C#1	C#2	C#1	C#2	C#1	C#2	C#1	C#2
1	0.054	0.946	0.015	0.539	0.004	0.332	0	0	0.095	0.905	0	0.984
2	0.008	0.992	0.018	0.546	0.005	0.332	0	0	0.060	0.940	0	0.985
3	0.040	0.960	0.019	0.999	0	1.000	0	0.992	0.090	0.910	0	1.000
4	0.093	0.907	0.018	0.545	0.005	0.332	0	0	0.137	0.863	0	0.985
5	0.055	0.945	0.024	0.552	0.007	0.331	0	0	0.114	0.886	0	0.985
6	0.500	0.500	0.003	0.003	0.001	0.001	0	0	0.500	0.500	0	0
7	0.500	0.500	0.010	0.010	0.004	0.004	0	0	0.500	0.500	0	0
8	0.945	0.055	0.552	0.024	0.331	0.007	0	0	0.886	0.114	0.985	0
9	0.992	0.008	0.546	0.018	0.332	0.005	0	0	0.941	0.059	0.985	0
10	0.960	0.040	0.999	0.019	1.000	0	0.992	0	0.910	0.090	1.000	0
11	0.907	0.093	0.545	0.018	0.332	0.005	0	0	0.863	0.137	0.985	0
12	0.946	0.054	0.539	0.015	0.332	0.004	0	0	0.905	0.095	0.984	0

AR-RFCM聚类算法对噪声数据具有良好的鲁棒性。数据集中包含了3个数据簇，每个数据簇都有100个样本，同时数据集中还包含了100个噪声样本，增大模糊聚类算法对数据簇本质结构的提取难度。3个数据簇的样本和噪声样本分别满足不同的4种高斯分布，其参数如表3所示。完整数据集如图3(a)所示。

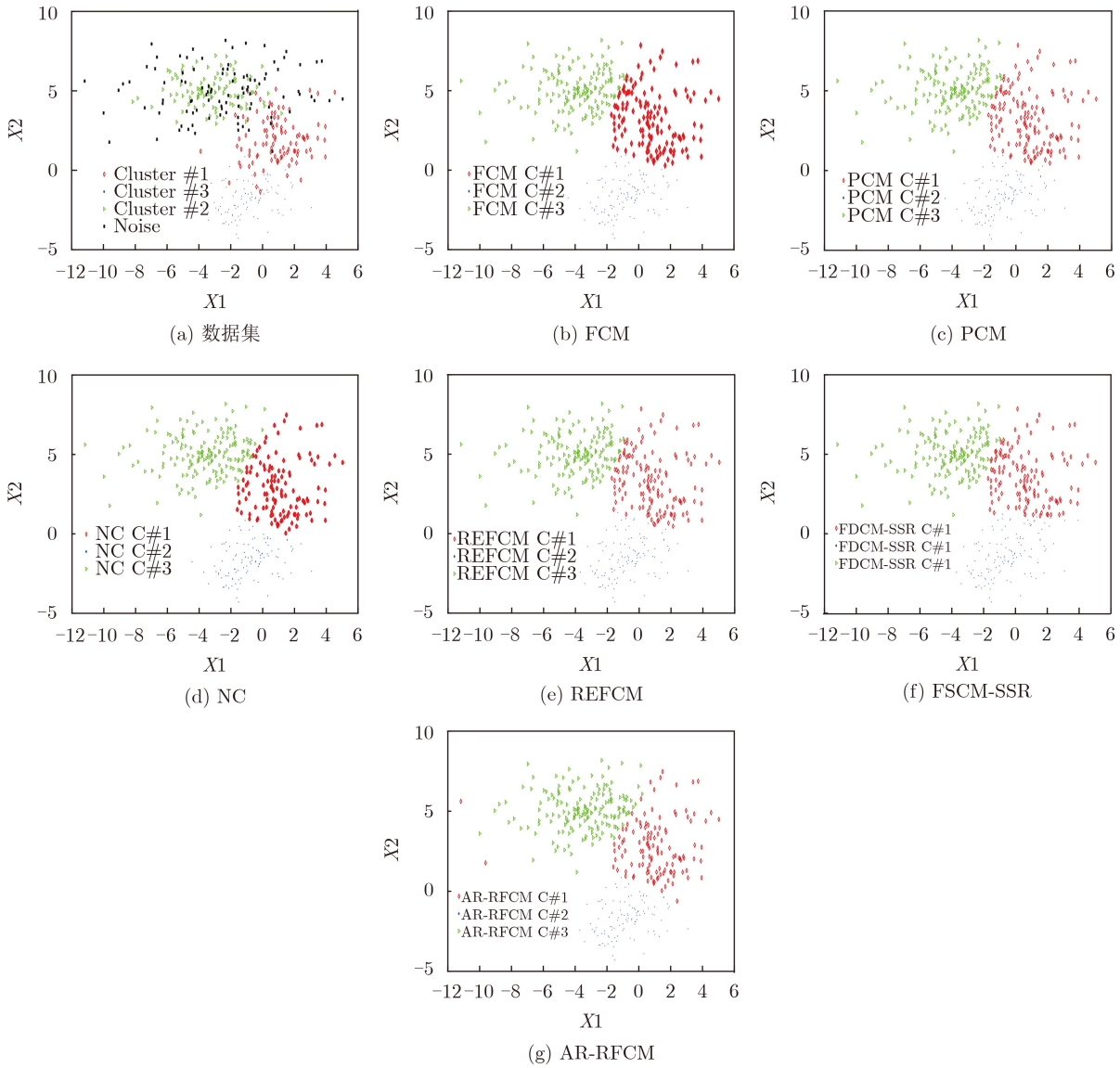
在带有非异常值噪声样本的人造数据集上的实验结果如表4所示，AR-RFCM在准确率，精确

表 2　各算法所得的聚类簇中心

	簇中心1	簇中心2
FCM	$\begin{bmatrix}3.9870\\0.0011\end{bmatrix}$	$\begin{bmatrix}-3.9870\\0.0011\end{bmatrix}$
PCM	$\begin{bmatrix}3.9870\\0.0011\end{bmatrix}$	$\begin{bmatrix}-3.9870\\0.0011\end{bmatrix}$
NC	$\begin{bmatrix}3.9996\\0.0002\end{bmatrix}$	$\begin{bmatrix}-3.9996\\0.0002\end{bmatrix}$
REFCM	$\begin{bmatrix}4.0000\\0.0000\end{bmatrix}$	$\begin{bmatrix}-4.0000\\0.0000\end{bmatrix}$
FDCM_SSR	$\begin{bmatrix}3.4833\\1.6530\end{bmatrix}$	$\begin{bmatrix}-3.4833\\1.6530\end{bmatrix}$
AR-RFCM	$\begin{bmatrix}3.9999\\0.0000\end{bmatrix}$	$\begin{bmatrix}-3.9999\\0.0000\end{bmatrix}$

表 3　噪声数据集信息

	均值	协方差
簇1	[1　2]	$\begin{bmatrix}2 & 0.2\\0.2 & 2\end{bmatrix}$
簇2	[−1　−2]	$\begin{bmatrix}2 & 0\\0 & 1\end{bmatrix}$
簇3	[−3　−5]	$\begin{bmatrix}3 & 0\\0 & 1\end{bmatrix}$
噪声	[−3　5]	$\begin{bmatrix}10 & 0\\0 & 2\end{bmatrix}$



(a) 数据集　(b) FCM　(c) PCM

(d) NC　(e) REFCM　(f) FSCM-SSR

(g) AR-RFCM

图 3 不同聚类算法的隶属度函数

度，灵敏度和特异度4项指标上，均优于其它聚类算法，取得了最好的聚类结果。如图3所示，尽管数据集被污染得较严重，数据簇结构变得难以识别，但AR-RFCM聚类算法学得的数据簇结构更为准确，因而取得的聚类效果更好，体现了算法对噪声良好的鲁棒性。

3.3 UCI数据集

本文选取了25个UCI数据集，来检验AR-RFCM及其比较算法在处理真实数据集时的表现。在该实验中，实验评价指标采用兰德指数作为聚类结果的衡量指标。实验采用随机初始数据簇中心的方式，为降低随机性带来的影响，每个算法在每个数据集上分别运行10次，保留并记录10次实验的兰德指数的均值和标准差，实验结果如表5所示。

在25个UCI数据集中，FCM的平均正确率为70.6536%, PCM的平均正确率为69.2292%, NC的平均正确率为70.114%, REFCM的平均正确率为72.4208%, FDCM_SSR的平均正确率为75.1216，而AR-RFCM的平均正确率为78.44%。如表5所示，AR-RFCM在绝大多数的数据集上均取得了最高的兰德指数，并且在UCI数据集上兰德指数的平均值比FCM平均高出7.7864%，比PCM高出9.2108%，比NC高出8.326%，比REFCM高出6.0192%，比FDCM_SSR高出3.3184%。这表明

表 4 噪声数据集实验结果

	FCM	PCM	NC	REFCM	FDCM_SSR	AR-RFCM
准确率	0.8683	0.8683	0.8767	0.8583	0.8667	**0.8833**
精确度	0.6775	0.6775	0.6900	0.6625	0.6750	**0.7000**
灵敏度	0.9033	0.9033	0.9200	0.8833	0.9000	**0.9333**
特异度	0.8567	0.8567	0.8622	0.8500	0.8556	**0.8667**

表 5 不同聚类算法在UCI数据集的聚类结果的RI(%)

	FCM	PCM	NC	REFCM	FDCM_SSR	AR-RFCM
Ecoli	79.66±0.47	78.60±3.28	78.25±1.11	80.12±1.13	79.83±0.45	**88.40±1.62**
Auto-mpg	76.23±0	75.21±4.02	77.27±0.46	75.64±0	**85.62±0.30**	77.64±2.09
Dermatology	83.49±1.18	81.17±3.26	69.82±6.61	81.33±5.49	84.88±1.59	**92.55±0**
Iris	83.68±0	79.33±7.90	82.27±4.23	85.68±0	**87.37±0**	85.68±0
Zoo	87.91±1.76	87.56±6.15	88.90±5.82	85.76±5.22	89.78±2.33	**96.65**±0
Transfusion	53.10±0	55.47±5.59	56.30±3.75	53.16±0	**63.90±0.09**	63.68±0
Parkinsons	52.19±0	58.86±5.86	64.40±5.31	60.27±7.74	63.12±0.57	**73.83±0**
Banknote	51.52±0	59.55±9.75	61.32±6.89	59.68±12.12	52.44±1.73	**66.32±2.79**
Creadit-approval	67.57±0.37	55.78±5.82	64.89±0	53.96±6.53	67.51±0	**68.06±0**
Breast-cancer	90.53±0	78.42±14.25	86.24±16.97	91.59±0	94.31±0	**94.86±0**
Wine	95.43±0	73.35±4.57	91.85±6.62	90.36±0	95.43±0	**96.40±0.73**
Automobile	71.83±0.44	70.81±1.78	72.08±1.62	71.79±0.38	71.99±0.29	**74.24±0.29**
Messidor Features	50.49±0	50.62±0.35	50.63±0.01	50.86±0	50.65±0.53	**51.80±0.61**
Fertility	50.00±0	65.74±12.93	50.08±0.18	57.99±6.55	**79.13±0.71**	78.85±1.75
Seeds	89.92±0	78.70±5.28	88.53±0.52	87.06±0	89.92±0	**91.26±0.52**
Blance	58.82±4.58	58.49±5.70	62.66±4.69	60.55±5.09	60.56±4.07	**65.97±3.54**
House Votes	77.52±0	71.46±8.61	75.49±6.07	72.41±6.76	77.86±0	**78.90±0**
Vowel	67.15±2.47	82.68±1.61	53.52±4.44	82.91±0.94	66.81±1.81	**85.47±0.20**
Glass	71.31±0.45	69.29±1.19	71.58±0.98	71.07±0.86	71.59±0.39	**72.45±0.77**
Mammographic	67.96±0	64.00±7.05	67.98±0.04	62.35±6.02	**68.55±0**	68.11±0
Pima Indians Diabetes	59.07±0	56.23±3.27	52.41±0.13	58.09±0	59.06±0.03	**59.18±0.29**
Qualitative Bankruptcy	94.53±0	74.58±16.72	**97.62±0**	94.53±0	94.53±0	**97.62±0**
Seismic Bumps	51.88±0	71.15±12.53	56.19±10.29	**87.70±0**	87.69±0.10	**87.70±0**
Phishing Data	68.72±0.29	60.73±5.18	69.13±0.28	62.71±0	68.93±0.39	**69.67±0.09**
Yeast	65.83±1.61	72.95±1.21	63.44±2.90	72.95±2.06	66.58±1.58	**75.71±0.03**

AR-RFCM聚类算法不仅在处理噪声问题上展现出良好的鲁棒性，同时在处理真实数据集上有着优异的表现，体现了算法的实用价值。

4　结束语

噪声一直是模糊聚类算法难以解决的问题。PCM，NC和REFCM等模糊聚类算法通过给噪声样本在所有数据簇均分配较低的隶属度的方式，来降低噪声样本带来的影响。尽管这类算法能在一定程度上降低噪声样本的影响，但是它们无差别地松弛样本的隶属度约束条件，使得聚类算法在降低噪声样本作用力的同时，降低了正常样本的作用力。在这种情况下，噪声样本的相对作用力难以降低，对聚类结果的影响力仍然很大。而FDCM_SSR等算法在解决噪声问题时，同样难以区分正常样本和噪声样本。为了解决这些问题，AR-RFCM引入了KNN的方式来评估样本的可靠程度，给不同可靠程度的样本以不同的松弛参数，从而实现在降低噪声样本作用力的同时，保留可靠样本的作用力。此外，AR-RFCM还把K-Means模型引入到FCM模型中，提升了可靠样本对所属数据簇的作用力，提升了各数据簇的内聚程度，间接地降低了噪声样本的影响，提高了聚类算法对噪声的鲁棒性。当然，AR-RFCM 也存在着不足之处，当模糊系数从2扩展到m的时候，模型求解的解析解难以得到，这也是后续我们将关注并展开进一步研究的方向。

参 考 文 献

[1] JAIN A K. Data clustering: 50 years beyond k-means[J]. *Pattern Recognition Letters*, 2010, 31(8): 651–666. doi: 10.1016/j.patrec.2009.09.011.

[2] DENG Zhaohong, JIANG Yizhang, CHUNG Fulai, *et al*. Transfer prototype-based fuzzy clustering[J]. *IEEE Transactions on Fuzzy Systems*, 2016, 24(5): 1210–1232. doi: 10.1109/TFUZZ.2015.2505330.

[3] 张洁玉, 李佐勇. 基于核空间的加权邻域约束直觉模糊聚类算法[J]. 电子与信息学报, 2017, 39(9): 2162–2168. doi: 10.11999/JEIT161317.
ZHANG Jieyu and LI Zuoyong. Kernel-based algorithm with weighted spatial information intuitionistic fuzzy c-means[J]. *Journal of Electronics & Information Technology*, 2017, 39(9): 2162–2168. doi: 10.11999/JEIT161317.

[4] LV Yinghua, MA Tinghuai, TANG Meili, *et al*. An efficient and scalable density-based clustering algorithm for datasets with complex structures[J]. *Neurocomputing*, 2016, 171: 9–22. doi: 10.1016/j.neucom.2015.05.109.

[5] QIN Xiaoyu, TING Kaiming, ZHU Ye, *et al*. Nearest-neighbour-induced isolation similarity and its impact on density-based clustering[C]. The 33rd AAAI Conference on Artificial Intelligence, Honolulu, USA, 2019: 4755–4762. doi: 10.1609/aaai.v33i01.33014755.

[6] 赵小强, 刘晓丽. 基于公理化模糊子集的改进谱聚类算法[J]. 电子与信息学报, 2018, 40(8): 1904–1910. doi: 10.11999/IEIT170904.
ZHAO Xiaoqiang and LIU Xiaoli. An improved spectral clustering algorithm based on axiomatic fuzzy set[J]. *Journal of Electronics & Information Technology*, 2018, 40(8): 1904–1910. doi: 10.11999/IEIT170904.

[7] YIN Hao, BENSON A R, LESKOVEC J, *et al*. Local higher-order graph clustering[C]. The 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, Canada, 2017: 555–564. doi: 10.1145/3097983.3098069.

[8] MOSLEHI Z, TAHERI M, MIRZAEI A, *et al*. Discriminative fuzzy c-means as a large margin unsupervised metric learning algorithm[J]. *IEEE Transactions on Fuzzy Systems*, 2018, 26(6): 3534–3544. doi: 10.1109/TFUZZ.2018.2836338.

[9] 刘解放, 王士同, 王骏, 等. 一种具有最优保证特性的贝叶斯可能性聚类方法[J]. 电子与信息学报, 2017, 39(7): 1554–1562. doi: 10.11999/JEIT160908.
LIU Jiefang, WANG Shitong, WANG Jun, *et al*. Bayesian possibilistic clustering method with optimality guarantees[J]. *Journal of Electronics & Information Technology*, 2017, 39(7): 1554–1562. doi: 10.11999/JEIT160908.

[10] MACQUEEN J. Some methods for classification and analysis of multivariate observations[C]. The 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, USA, 1965: 281–297.

[11] BEZDEK J C, EHRLICH R, and FULL W. FCM: The fuzzy c-means clustering algorithm[J]. *Computers & Geosciences*, 1984, 10(2/3): 191–203.

[12] DE OLIVEIRA J V and PEDRYCZ W. Advances in Fuzzy Clustering and its Applications[M]. Chichester: John Wiley & Sons, Ltd., 2007. doi: 10.1002/9780470061190.

[13] DAVE R N. Characterization and detection of noise in clustering[J]. *Pattern Recognition*, 1991, 12(11): 657–664. doi: 10.1016/0167-8655(91)90002-4.

[14] KRISHNAPURAM R and KELLER J M. A possibilistic approach to clustering[J]. *IEEE Transactions on Fuzzy Systems*, 1993, 1(2): 98–110. doi: 10.1109/91.227387.

[15] ZARINBAL M, ZARANDI M H F, and TURKSEN I B. Relative entropy fuzzy c-means clustering[J]. *Information Sciences*, 2014, 260: 74–97. doi: 10.1016/j.ins.2013.11.004.

[16] GU Jing, JIAO Licheng, YANG Shuyuan, *et al*. Fuzzy double c-means clustering based on sparse self-representation[J]. *IEEE Transactions on Fuzzy Systems*, 2018, 26(2): 612–626. doi: 10.1109/TFUZZ.2017.2686804.

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

王志豪：男，1993年生，硕士生，研究方向为机器学习和模式识别.

潘金艳：女，1978年生，副教授，主要研究方向为人工智能和机器学习理论与方法.

罗斯哲：男，1995年生，硕士生，研究方向为模式识别和维数约简.

责任编辑：马秀强