一种改进的区间型不确定数据模糊聚类方法

肖满生* 张龙信 张晓丽 胡永祥

(湖南工业大学 计算机学院 株洲 412007)

摘 要：针对区间型不确定数据的特点，该文提出一种改进的模糊C均值聚类算法(IU-IFCM)。首先对区间型数据进行特征变换，由p维特征映射成由$2p$维特征组成的实数据，然后考虑区间中值与区间大小关系，设计一种样本距离计算方法，通过模糊C均值实现对区间型样本聚类。理论分析与对比实验表明，该算法的划分系数(PC)及正确等级(CR)值比其它方法平均提高10%以上，表明有更好的聚类精度，对当前大数据环境下不确定数据的分类提供了一种新的解决方案。

关键词：区间型数据；模糊C均值；影响因子；特征变换

中图分类号：TN911.7; TP391 **文献标识码**：A **文章编号**：1009-5896(2020)08-1968-07

DOI: 10.11999/JEIT190591

An Improved Fuzzy Clustering Method for Interval Uncertain Data

XIAO Mansheng ZHANG Longxin ZHANG Xiaoli HU Yongxiang

(*School of Computer Science*, *Hunan University of Technology*, *Zhuzhou* 412007, *China*)

Abstract: An Improved Fuzzy C-Means clustering algorithm (IU-IFCM) is proposed in this study in accordance with the characteristics of Interval Uncertain data. First, the interval data is transformed into real data composed of $2p$ dimension feature, which is mapped from that of p dimension feature. Second, a method for calculating sample distance, which realizes the interval sample clustering by fuzzy c-mean algorithm, is designed while considering the relationship between interval median value and interval size. Theoretical analysis and comparison experiments show that the presented algorithm surpaes the compared algorithms by more than 10% on average in terms of the Partition Coefficient (PC) and Correct Rank(CR) value. These results indicate that the algorithm presents in this study has better clustering accuracy and provides a new solution for the classification of uncertain data in current big data environments.

Key words: Interval data; Fuzzy C-means; Impact factor; Feature transformation

1 引言

在大数据环境下，存在一类模糊的、不确定的数据，如人们交流中的语言数据、天气预报的气温数据、各种仪器工具测量得到的不精确数据等，这些数据常用区间值形式表示，因此如何分析与处理该类数据是当今数据分析与研究的主要内容之一[1–4]。

在对精确数据的聚类分析中，模糊C均值(Fuzzy C-Means, FCM)聚类方法是应用最广泛的一种方法，FCM算法不但理论完善，聚类效率高，而且能实现无监督聚类分析。然而对于包括区间数在内的不确定数据，传统的FCM算法无法直接操作，只能通过对其改进或其它方法来完成该类数据分析与处理[5–9]。如针对区间数的聚类分析，Gao等人[10]率先提出了一种将区间数从p维空间直接变换到$2p$维空间的“实”数据的方法来实现FCM聚类，但没有给出变换时与区间大小、区间中值直接相关的影响因子β的取值依据与大小；Maciel等人[11]提出了参与学习的模糊聚类，它用Hausdorff距离来计算区间数相似度，但方法中使用了太多的参数α,β,t,λ，且所有参数都人为给定，降低了分类的精度；Bao等人[5]与兰蓉[12]通过设计一种新的区间数距离公式来使用FCM算法聚类，然而在计算过程中，区间数中各维特征的权重设计复杂；金萍等人[13]在近似骨架理论的基础上，提出了一种近似骨架启发式聚类算法(APPGUL)，解决了不确定数据聚类对初始值敏感的问题；魏方圆、黄德才[14]首次提出区间数

收稿日期：2019-08-06；改回日期：2020-02-19；网络出版：2020-03-14
*通信作者： 肖满生 xiaomansheng@tom.com
基金项目：国家自然科学基金(61702178)，湖南省自然科学基金(2018JJ4068)，湖南省教育厅科研项目(18C0499)

Foundation Items: The National Natural Science Foundation of China (61702178), The Natural Science Foundation of Hunan Provierce (2018554068), The Research Project of Hunan Provincial Department of Education (18C0499)

密度可达与相连的概念，综合数据的统计信息来表达不确定性数据；此外还有文献[15–17]都对不确定数据或不确定数据流的分类方法进行了研究，这些方法各有优势，但都存在某些不足。

本文在分析上述文献中有关区间数据聚类的基础上，提出了一种改进的区间型不确定数据聚类的FCM算法(IU-IFCM)，将区间型数据变换成由区间中值与区间大小相关的精确实数据，然后采用FCM算法进行聚类实现，方法思路清新，理论依据充分，且易于实现，实验结果表明对区间型不确定数据有较好的聚类效果。

2　基于FCM的区间数据直接聚类与划分聚类

2.1　区间型数据集直接聚类方法

设一观测样本集$\bar{\boldsymbol{X}}=\{\bar{\boldsymbol{x}}_1, \bar{\boldsymbol{x}}_2, \cdots, \bar{\boldsymbol{x}}_k, \cdots, \bar{\boldsymbol{x}}_n\}$包含$n$个样本，其中每个样本$\bar{\boldsymbol{x}}_k=(\bar{x}_{k1}, \bar{x}_{k2}, \cdots, \bar{x}_{kj}, \cdots \bar{x}_{kp})$为$p$维特征矢量，每个特征$\bar{x}_{kj}=[x_{kj}^-, x_{kj}^+]\in I(R^+)$用一个区间数描述，$x_{kj}^-$与$x_{kj}^+$分别为区间左右端点值，其中$R^+$为全体正实数集合，$I(R^+)$是由正实数组成的区间数集合，则$\bar{X}$为一组由区间数组成的观测样本的特征矢量集，假定这组数据有c个自然结构，采用传统的FCM直接聚类，有2种实现算法。

算法1：基于区间端点的直接聚类

该方法直接将区间数的左右2个端点分开，然后采用传统的FCM算法分别对区间值的左右端点直接聚类，分别得到左右端点的聚类中心与隶属度，具体过程如下：

步骤 1　设定迭代停止阈值ε，初始化聚类中心的区间左右端点$v^{(0)}=[v^{-(0)}, v^{+(0)}]$，设计迭代计数器$b=0$；

步骤 2　分别计算样本$\bar{\boldsymbol{x}}_k$的左右端点(x_k^-, x_k^+)与聚类中心$\bar{v}_i(v_i^-, v_i^+)$的左右端点的Euclid距离，即

$$D_{\mathrm{E}}^2(x_k^-, v_i^-)=\sum_{l=1}^{p}\left(x_{kl}^- - v_{il}^-\right)^2, D_{\mathrm{E}}^2(x_k^+, v_i^+)=\sum_{l=1}^{p}\left(x_{kl}^+ - v_{il}^+\right)^2 \tag{1}$$

步骤 3　依据式(1)，计算区间数左端点的隶属度$u_{ik}^{-(b)}$，即如果$\exists i, r$，使$D_{\mathrm{E}}^b(x_r^-, v_i^-)=0$，则$u_{ir}^{-(b)}=1$，且对$k\neq r$, $u_{ir}^{-(b)}=0$，否则按式(2)计算区间左端点的隶属度

$$u_{ik}^{-(b)}=\left\{\sum_{j=1}^{c}\left[\left(D_{\mathrm{E}}^{(b)}, \left(v_i^-\right)/D_{\mathrm{E}}^{(b)}\left(x_k^-, v_j^-\right)\right)^{2/(m-1)}\right]\right\}^{-1} \tag{2}$$

区间右端点的隶属度计算方法同左端点一样。

步骤 4　用式(3)分别更新聚类中心的区间左右端点$v^{\mp(b+1)}$

$$v_i{}^{\mp(b+1)}=\sum_{k=1}^{n}\left(u_{ik}^{\mp(b)}\right)^m\cdot x_k^{\mp}\Big/\sum_{k=1}^{n}\left(u_{ik}^{\mp(b)}\right)^m \tag{3}$$

步骤 5　如果$D_{\mathrm{E}}(v^{\mp(b)}, v^{\mp(b+1)})<\varepsilon$或达到设定的最大迭代次数，算法终止并输出由区间左右端点组成的聚类中心，否则$b=b+1$，转步骤1:。

算法中$i=1, 2, \cdots, c$, $k=1, 2, \cdots, n$，算法1虽然调用了经典FCM算法，但由于该算法分别通过区间数的左右端点来获得聚类中心的区间值的左右端点，因此聚类中心区间值的左右端点相互独立，只与样本的区间左右端点有关。另外，算法所获得的每个样本的左右端点的隶属度也可能不一样，甚至差别很大，即每个样本的隶属度可能有2个不同的值，因而该聚类方法不够准确，甚至会产生错误的数据分类。

算法2：基于区间中值的FCM直接聚类

算法1通过对测量样本的区间左右2端点值分别进行聚类，由于在聚类过程中割裂了样本的左右区间端点的联系，因而聚类结果不理想，基于此，算法2考虑了区间数据的左右端点间联系，将区间数$\bar{x}_{kj}$转化为区间中值$\dot{x}_{kj}$，即$\dot{x}_{kj}=\left(x_{kj}^+ + x_{kj}^-\right)/2$, $k=1, 2, \cdots, n$, $j=1, 2, \cdots, p$，其中n为样本个数，p为每个样本的特征数，然后直接调用传统的FCM算法对样本中值进行聚类，得到最终的隶属度与聚类中心，如式(4)与式(5)所示，具体过程与算法1类似，此处不再赘述。

$$u_{ik}^{(b)}=\left\{\sum_{j=1}^{c}\left[\left(d^{(b)}\left(\dot{x}_k, \dot{v}_i\right)/d^{(b)}\left(\dot{x}_k, \dot{v}_j\right)\right)^{2/(m-1)}\right]\right\}^{-1} \tag{4}$$

$$\dot{v}_i^{(b+1)}=\sum_{k=1}^{n}\left(u_{ik}^{(b)}\right)^m\cdot\dot{x}_k\Big/\sum_{k=1}^{n}\left(u_{ik}^{(b)}\right)^m \tag{5}$$

式(4)中的$d(\dot{x}_k, \dot{v}_i)$为样本与聚类中心区间中值的欧氏距离，聚类完成后再利用式(6)恢复聚类中心的区间型值

$$v_{ij}^{\mp}=\sum_{k=1}^{n}\left(u_{ik}^*\right)^m\cdot x_{kj}^{\mp}\Big/\sum_{k=1}^{n}\left(u_{ik}^*\right)^m \tag{6}$$

其中u_{ik}^*为算法收敛后的隶属度。

本算法通过把区间值直接转化为区间中值而调用经典FCM算法进行聚类，由于没有考虑区间大小(宽度)对聚类的影响，使得只要是相同中值的区间数就具有相同的隶属度，这种聚类显然存在着缺陷与不足。

2.2 区间型数据的划分聚类算法

该算法算法3使用的测量样本同2.1节一致，其核心是采用划分的方法求出区间数的距离，详见文献[5]，然后基于该距离调用经典的FCM获得隶属度与聚类中心，进而实现区间数聚类。

算法的距离求取方法为：将两区间数$\bar{\boldsymbol{x}}_k=[x_k^-,x_k^+]$与$\bar{\boldsymbol{x}}_j=[x_j^-,x_j^+]$分别划分为$l$份，每1份标记为$[A_{h-1},A_h]$与$[B_{h-1},B_h],h=1,2,\cdots,l$，这里$A_{h-1}=x_k^-+(h-1)(x_k^+-x_k^-)/l, A_h=x_k^-+h(x_k^+-x_k^-)/l$，$B_{h-1}=x_j^-+(h-1)(x_j^+-x_j^-)/l, B_h=x_j^-+h(x_j^+-x_j^-)/l$，参照文献[18]中区间数距离定义方法，两区间数$\bar{x}_k$与$\bar{x}_j$的距离定义为

$$\begin{aligned}&D_{\mathrm{I}}(\bar{\boldsymbol{x}}_k,\bar{\boldsymbol{x}}_j)\\&=\frac{1}{l}\sum_{h=1}^{l}\iint_D\{[(A_{h-1}+A_h)/2+(A_h-A_{h-1})x]\\&\quad-[(B_{h-1}+B_h)/2+(B_h-B_{h-1})y]\}^2\mathrm{d}x\mathrm{d}y\\&=\frac{1}{l}\sum_{h=1}^{l}\Big\{\big[\big(x_k^--x_j^-\big)+(2h-1)\big(\big(x_k^+-x_k^-\big)\\&\quad-\big(x_j^+-x_j^-\big)\big)/(2l)\big]^2\\&\quad+\frac{1}{12l}\left[\big(x_k^+-x_k^-\big)^2+\big(x_j^+-x_j^-\big)^2\right]\Big\}\end{aligned}\tag{7}$$

从式(7)可以看出，两区间数之间的距离只与区间端点值与区间划分份数l有关，而且当$l\to\infty$时，其距离最小，因此，两区间数的最小距离为

$$\begin{aligned}\min D(\bar{x}_k,\bar{x}_j)&=\lim_{l\to\infty}D_{\mathrm{I}}(\bar{\boldsymbol{x}}_k,\bar{\boldsymbol{x}}_j)\\&=\big(x_k^--x_j^-\big)\big(x_k^+-x_j^+\big)\\&\quad+\big[\big(x_k^+-x_k^-\big)-\big(x_j^+-x_j^-\big)\big]^2/3\end{aligned}\tag{8}$$

有了距离的定义，就可以参照2.1节中的算法1进行迭代求出隶属度，然后再分别计算出聚类中心的左右区间端点，如式(9)、式(10)所示。

$$v_{ij}^{-(b+1)}=\frac{3}{2}\sum_{k=1}^{n}(u_{ik})^m\cdot\left(\frac{2}{3}x_{kj}^-+\frac{1}{3}x_{kj}^+-\frac{1}{3}v_{ij}^{+(b)}\right)\Big/\sum_{k=1}^{n}(u_{ik})^m\tag{9}$$

$$v_{ij}^{+(b+1)}=\frac{3}{2}\sum_{k=1}^{n}(u_{ik})^m\cdot\left(\frac{1}{3}x_{kj}^-+\frac{2}{3}x_{kj}^+-\frac{1}{3}v_{ij}^{-(b)}\right)\Big/\sum_{k=1}^{n}(u_{ik})^m\tag{10}$$

从上面可以看出，本算法在计算样本与聚类中心的距离时，不但考虑了区间数的左右端点，而且通过区间划分兼顾了区间大小，因此该算法得到的聚类结果更精准与客观，然而，由于距离计算与划分份数有关，且在每一次聚类中心左右端点值计算过程中都要用到上一次的聚类中心区间左右端点值，因此迭代次数增加，其收敛性也难以证明。

3 改进的区间型不确定数据的FCM聚类方法

上一节阐述了基于区间数的3种聚类算法，这些算法都存在不足，其中算法1对区间数左右端点直接聚类，它分裂了区间左右端点间的联系，算法2只考虑区间中值而未兼顾区间大小对聚类的影响，算法3虽然考虑了区间端点与大小对聚类的影响，但其聚类中心的迭代计算不但与隶属度有关，而且与上一次的计算结果有关，这样反复迭代有可能陷入死循环，其收敛性难以证明。基于此，本文兼顾区间大小与区间中值，设计一种改进的区间数FCM聚类算法来克服上述不足，具体如下。

3.1 区间型特征变换

观测样本集$\bar{\boldsymbol{X}}$如上节所述，其中每个样本$\bar{\boldsymbol{x}}_k$由p维特征组成，每个特征是一个区间数，考虑区间中值与区间大小的关系，将区间数$\bar{\boldsymbol{x}}_k$映射到由区间中值$\dot{x}_k$与区间大小$\hat{x}_k$所张成的特征空间$S(\dot{x}_k,\hat{x}_k)$中，形成特征空间中的一个实值点$\boldsymbol{x}_k$，即

$$\mathrm{M}:\ \bar{\boldsymbol{x}}_k\in I^p(R^+)\to\boldsymbol{x}_k\in R^{2p}\tag{11}$$

其中$\bar{\boldsymbol{x}}_k=(\bar{x}_{k1},\bar{x}_{k2},\cdots\bar{x}_{kp})\in I^p(R^+)$为$p$维区间型特征数据，$\bar{x}_{kj}=[x_{kj}^-,x_{kj}^+]\in I(R^+)$为样本的第$j$维特征，变换后的样本$\boldsymbol{x}_k=\left(\dot{x}_k,\lambda\hat{x}_k\right)=\left(\dot{x}_{k1},\cdots,\dot{x}_{kp},\lambda\hat{x}_{k1},\cdots,\lambda\hat{x}_{kp}\right)$就成为了普通数据，即$2p$维空间的一个点，其中$\dot{x}_k=\left(x_k^-+x_k^+\right)/2$, $\hat{x}_k=\left(x_k^+-x_k^-\right)$, λ为加权因子，这是一个非常重要的参数，用来控制区间大小对聚类的影响，λ的定义：分析区间数可知，在区间中值已确定的情况下，区间越大，即中值离左右区间端点距离越大，则区间大小对该区间数聚类的影响越大，例如，今天的气温在25°左右，则25°是区间中值，其变化范围可以是区间数[24, 26]，也可以是[23, 27]，显然变化范围越大，其对区间数的影响也越大，因此，λ的定义为

$$\lambda=(\dot{x}-x^-)/\dot{x}=(x^+-\dot{x})/\dot{x}\tag{12}$$

将中值$\dot{x}=(x^-+x^+)/2$代入式(12)，可得$\lambda=(x^+-x^-)/(x^++x^-)$。

由式(12)可以看出，$\lambda\in[0,1]$，且区间越大(区间越宽)，则λ值越大，即区间宽度对该区间数的影响越大，当$\lambda=1$，此时区间大小与区间中值同等重要，当区间变窄直至趋于0，即$x^-=x^+$时，$\lambda=0$，这时区间数变成普通的精确数，区间大小对聚类无

影响，因此λ可看着是区间大小对区间数聚类影响的度量。

3.2 改进的区间型不确定数的FCM聚类算法实现

有了区间大小影响因子，改进的FCM算法进行聚类过程如下：

步骤 1　分析待聚类的样本集$\bar{\boldsymbol{X}}$，按照式(11)将其变换成普通确定数的样本集$\boldsymbol{X}$，并依据式(12)，设定影响因子λ的合理取值；

步骤 2　初始化，设置迭代停止阈值ε，初始化聚类中心模式$V^{(0)}$，设置迭代计数器$b=0$；

步骤 3　根据转换后的样本特征，按式(13)定义样本$\bar{\boldsymbol{x}}_k$与聚类中心$\bar{v}_i$间的距离

$$d^2(\bar{\boldsymbol{x}}_k,\bar{\boldsymbol{v}}_i)=\sum_{l=1}^{p}\left[\left(\dot{x}_{kl}-\dot{v}_{il}\right)^2+\lambda^2\left(\hat{\dot{x}}_{kl}-\hat{\dot{v}}_{il}\right)^2\right] \tag{13}$$

其中$\dot{x}_{kl}$为样本$\bar{x}_k$的第l维特征区间中值，$\hat{\dot{x}}_{kl}$为其第l维特征的区间值大小，$\dot{v}_{il}$为聚类中心$\bar{v}_i$的第l维特征区间中值，$\hat{\dot{v}}_{il}$为其第l维区间值大小，重新定义的距离既考虑样本区间中值，又兼顾了区间大小，因而在聚类时能客观表达样本信息。

步骤 4　基于式(13)定义的距离，调用2.1节的FCM算法进行迭代计算，满足终止条件后，获得样本集的最佳划分隶属度u_{ik}^*和聚类中心$\boldsymbol{v}_i^*$，其中$\boldsymbol{v}_i^*=\left(\dot{v}_{i1},\dot{v}_{i2},\cdots,\dot{v}_{ip},\lambda\hat{\dot{v}}_{i1},\lambda\hat{\dot{v}}_{i2},\cdots,\lambda\hat{\dot{v}}_{ip}\right)$；

步骤 5　反变换，将步骤4中获得的聚类中心复原，从而获得聚类中心$\bar{v}_i$的区间端点值，即：$v_{il}^-=\dot{v}_{il}-\hat{\dot{v}}_{il}/2, v_{il}^+=\dot{v}_{il}+\hat{\dot{v}}_{il}/2\,, l=1,2,\cdots,p$。

本算法通过空间映射变换将区间型样本值变换成由区间中值和区间大小组成的特征实向量，设计一影响因子来兼顾区间中值与区间大小对聚类的共同作用，然后调用FCM算法进行聚类，算法终止后通过反变换恢复聚类中心的区间数形式。因此，本算法克服了第2节中所提出的算法弱点，而且设计简单，理论依据充分，具有一定的优越性。

4　实验结果分析

为了验证本文所提出的改进区间型不确定数据FCM聚类的有效性，用文献[10]中给出的Fat-Oil、文献[9]给出的Fish以及文献[11]所述人工合成的数据集进行实验，这3个数据集各有特征，其中Fat-Oil没有给出先验的数据类属关系，Fish数据集虽然有先验的数据类属关系，但数据类之间的差别小，可分性差，而人工合成数据集中各类之间的分离性较好，因此通过这3个数据集可以全面检验算法的实际效果。实验环境采用Matlab2014a软件编程，实验评价指标：一个是划分系数(Partition Coefficient, PC)，其定义为

$$\mathrm{PC}=\frac{1}{n}\sum_{i=1}^{c}\sum_{k=1}^{n}u_{ik}^2 \tag{14}$$

其中n为待分类的样本个数，c为聚类数，$u_{ik}\in[0,1]$为样本$\bar{x}_k$与划分类i的隶属度，PC值在$[1/c,1]$之间，且u_{ik}越大，划分系数PC就越大，聚类效果越好，因此PC可以作为评价聚类效果的指标；另一个是正确等级(Correct Rank, CR)指标[5]，其定义为：设$U=\{u_1,u_2,\cdots,u_r\}$与$V=\{v_1,v_2,\cdots,v_c\}$分别表示已给出的先验类和通过算法得到的实际类，则

$$\mathrm{CR}=\frac{\sum_{i=1}^{r}\sum_{j=1}^{c}C_{n_{ij}}^2-(C_n^2)^{-1}\sum_{i=1}^{r}C_{n_i}^2\sum_{j=1}^{c}C_{n_j}^2)}{\frac{1}{2}\left(\sum_{i=1}^{r}C_{n_i}^2+\sum_{j=1}^{c}C_{n_j}^2\right)-(C_n^2)^{-1}\sum_{i=1}^{r}C_{n_i}^2\sum_{j=1}^{c}C_{n_j}^2} \tag{15}$$

其中组合$C_{n_{ij}}^2=n_{ij}(n_{ij}-1)/2$, n_{ij}为既属于类u_i又属于类v_j中的样本个数，n_i,n_j分别为类u_i与类v_j中样本个数，n为样本总数，很显然，CR值在[–1, 1]范围内，其值越接近1说明算法的划分性能越好，当其值接近于0或为负数时，算法性能很差，因此可以通过CR指标值来检验算法在聚类划分时性能的好坏。

另外将本文所提出的区间型不确定数据改进的模糊C均值算法(IU-IFCM)与第2节所阐述的区间值端点直接聚类算法1(E_FCM)、基于区间中值直接聚类算法2(M_FCM)以及基于区间型数据划分的聚类算法3(D_FCM)进行对比实验，以便更清晰的验证聚类结果。

实验1　Fat-Oil数据集

该数据集组成如表1所示，为一组实际数据，包含8个4维特征矢量，各维特征值均为区间数，分别采用前面所述的E_FCM, M_FCM, D_FCM及本文所提出的区间型不确定数据的改进算法IU_IFCM进行聚类实验，按式(14)计算划分系数，

表 1　Fat_Oil数据集

样本	比重(g/cm^3)	冰点(°C)	io值	sa值
亚麻油	[0.930 0.935]	[–27 –8]	[170 204]	[118 196]
紫苏油	[0.930 0.937]	[–5 –4]	[192 208]	[188 197]
棉籽油	[0.916 0.918]	[–6 –1]	[99 113]	[189 198]
芝麻油	[0.920 0.926]	[–6 –4]	[104 116]	[187 193]
山茶油	[0.916 0.917]	[–21 –15]	[80 82]	[189 193]
橄榄油	[0.914 0.919]	[0 6]	[79 90]	[187 196]
牛油	[0.860 0.870]	[30 38]	[40 48]	[190 199]
猪油	[0.858 0.864]	[22 32]	[53 77]	[190 202]

需要说明的是算法1在聚类分析时，因其对区间左右端点分别聚类得到了2个不同的隶属度u_{ik}^{-}, u_{ik}^{+}，因此最后的划分系数PC取左右端点聚类后的平均值，另外由于Fat-Oil数据集没有给出先验类别划分，因此指标CR值不能在实验中给出。实验中，聚类数设置为$c=3$、模糊加权指数$m=2$，迭代终止阈值(相邻两次迭代的聚类中心之差)$\varepsilon=0.05$，本文IU_IFCM算法中提出的区间大小影响因子λ是本算法区别于其它3种算法的主要特征，其取值大小直接影响聚类结果的质量，是实验中要分析的主要参数，依据式(12)，分别计算表1中样本各维特征值的区间大小影响因子，得到该数据集区间值大小影响因子平均值λ=0.2(因各特征的区间大小影响因子基本相同，均为0.2)，每种算法分别进行10次实验后取平均值，其划分系数对比如图1所示。

可以看出，4种算法中，IU_IFCM的PC值最大，表明该算法的聚类效果最好；D_FCM算法次之，主要是该算法聚类时，由于Fat-Oil数据集没有类别划分先验知识，虽然样本的划分与距离表达完备，但聚类中心区间2端点值本身在迭代计算时要相互引用，左区间端点要引用右区间端点，见式(9)与式(10)，即在聚类过程中即使没有收敛，但满足了终止条件算法就终止；E_FCM算法最差，这是因为E_FCM聚类时，2个区间端点分别独立完成，它们之间没有联系，进而在聚类过程中可能出现2区间端点完全不一致的隶属度；而M_FCM算法比E_FCM虽然有所改进，但在聚类计算过程中忽略了区间大小产生的影响，因而结果也不够理想。

实验2 Fish数据集

该数据集包括12个鱼种样本，每个样本由13个区间特征值与1个先验分类值描述，详见文献[9]，区间值特征用于聚类分析，先验分类值表明该种鱼的事先分类情况，包括肉食性鱼类、草食性鱼类、腐屑食性鱼类和杂食性鱼类4种类别，采用上述4种方法进行实验，实验主要参数设置：聚类数$c=4$，模糊加权指数$m=2$，迭代终止阈值$\varepsilon=0.01$。本文提出的IU_IFCM算法中区间大小影响因子根据式(12)提出的λ因子的计算方法，分别对13个特征的区间值进行计算，得到13个区间大小影响因子$\lambda_i(i=1,2,\cdots,13)$，由于在同一数据集中各$\lambda_i$相差不大，故最后取其平均值即$\lambda=0.48$为该数据集区间大小影响因子，且从这里可以看出，区间大小影响因子的值较大。由于该数据集已有鱼种分类先验知识，因此采用4种算法进行实验时，既可计算划分系数PC，也可计算CR指标值，每种算法各进行10次重复实验取实验结果平均值，4种算法进行聚类后的最后分类结情以及PC, CR值分别见表2和图2，表2中的数字代表鱼种类别。

从表2可以直观看出，本文所提出的IU_IFCM算法的聚类结果与先验分类最接近，其它3种算法聚类结果难以直观分辨，可借助于图2中的实验结果对比分析，可以看出，无论是PC值还是CR值，4种算法中IU_IFCM算法的聚类效果最好，D_FCM次之，E_FCM最差，原因是Fish数据集中区间大小影响很大($\lambda=0.48$)，而本文提出的IU_IFCM算法兼顾了区间大小影响，故聚类效果好，D_FCM算法虽然考虑了区间大小的影响，但其距离计算过程中区间划分份数$l\to\infty$才有最佳效果与收敛性，而实际上$l\to\infty$是不可能的，故其结果次之，其它2种算法M_FCM、E_FCM都没有考虑区间大小，因而在这样一个区间大小影响大的数据集中聚类效果较差。该实验也验证了本文所提出改进区间型不确定数据聚类算法的有效性。

表 2 4种算法对Fish数据集的分类结果

	腐屑性	肉食性	杂食性	草食性
先验分类	1 2 3 4	5 6 7 8	9 10	11 12
E_FCM	1 2 5	4 6 3	7 10	8 9 11 12
M_FCM	1 3 4	6 10 11	2 8	5 7 9 12
D_FCM	1 2 4	5 6 8 9	3 10 11	7 12
IU_IFCM	1 2 3 4	6 7 8	5 9 10	11 12

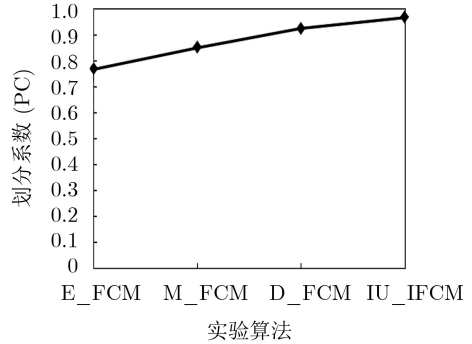

图 1 4种算法的划分系数比较

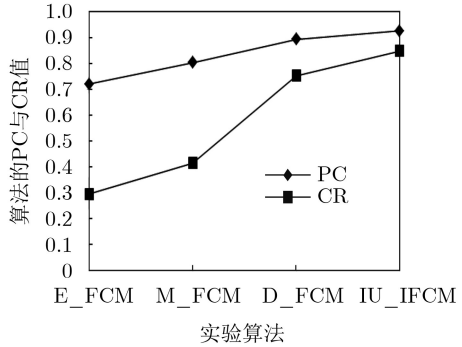

图 2 Fish数据集4种算法的PC, CR比较

实验3　人工合成数据集

该数据集是一个人工合成的2维空间数据集，详见文献[11]，共包含3类350个样本点数据，每一个样本$\boldsymbol{x}=(x_1, x_2)$服从独立的2维正态分布，其主要参数见表3，其中$m_1, m_2$分别表示样本第1维、第2维均值，$\sigma_1^2, \sigma_2^2$分别表示第1,2维协方差，该样本集有较好的分离性，其中一类包含50个圆形分布样本点，另2类各包含150个椭圆形分布的样本点，将该数据集的每个样本点作为“种子”，按照$\boldsymbol{x}=([x_1-\gamma_1/2, x_1+\gamma_1/2], [x_2-\gamma_2/2, x_2+\gamma_2/2])$的方法生成区间型数据，其中$\gamma_1, \gamma_2$分别表示样本点第1, 2维宽度，其值在区间[1, 8]范围内随机产生，从而得到该样本集的区间型数据集，如图3所示。

采用实验1,2中所述的4种方法分别对该数据集进行实验，实验环境与条件同前面，其中聚类数$c=3$，通过计算可得IU_IFCM算法中区间大小影响因子$\lambda=0.1$，终止阈值$\varepsilon=0.01$，共进行10次重复实验取平均值，实验结果如图4所示。

从图4可以看出，本文所提出的IU_IFCM算法

表 3　人工数据集

参数	类1	类2	类3
m_1	28	60	45
m_2	22	30	38
σ_1^2	100	9	9
σ_2^2	9	144	9

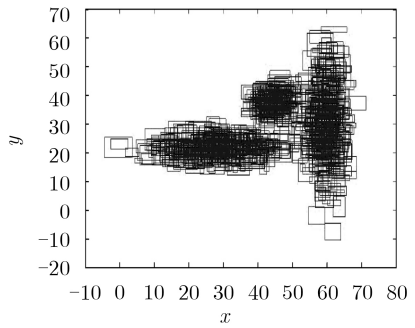

图 3 人工合成区间数据集

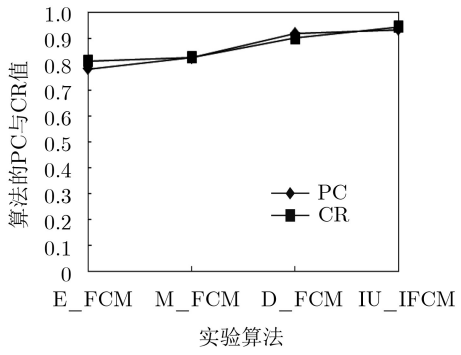

图 4 人工合成数据集4种算法的PC、CR比较

的PC, CR指标值虽然大于其它3种算法，但都相差不大，原因是本数据集中，区间大小影响因子λ值不大，即本数据集中区间大小对聚类的影响不明显，因而4种算法聚类结果相差不大。另外对比实验2还可以看出，针对不同结构的数据集，算法得到的聚类结果也不同，本实验的人工合成数据集因有较好的类属分离性，故PC, CR值大，而实验2中的Fish鱼种分类，因某些鱼种(如腐屑性类、肉食性类)分类不明显，有一定的重叠数据，故其值较小。

5　结束语

区间型不确定数据是目前大数据环境中重要组成部分，本文通过分析比较目前已有的基于区间型数据聚类FCM算法的不足，给出了一种改进的区间型不确定数据集FCM聚类算法，并通过理论分析与实验对比验证其有效性。本文的主要创新点有2个，一是将p维区间值数据变换为$2p$维实数据的过程中兼顾了区间中值与区间大小的关系，使得数据聚类更为客观；二是在聚类计算时，引入区间大小影响因子λ，阐述了其取值的理论依据与推导方法，且在实验中给出了它对实验结果影响的具体分析。因而本文提出的方法在在大数据环境下对区间型数据的分类与描述有一定的应用价值。需要说明的是，在区间数变换以及区间数距离的计算过程中本文提出的算法没有考虑不同特征对聚类结果的影响不同这一情况，即所有特征的影响因子λ取值相同，可能会导致某些区间数据集聚类结果不够准确，这一问题的解决将是下一步的研究目标。

参 考 文 献

[1] JIANG Bin, PEI Jian, TAO Yufei, *et al*. Clustering uncertain data based on probability distribution similarity[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2013, 25(4): 751–763. doi: 10.1109/TKDE.2011.221.

[2] GULLO F and TAGARELLI A. Uncertain centroid based partitional clustering of uncertain data[J]. *Proceedings of the VLDB Endowment*, 2012, 5(7): 610–621. doi: 10.14778/2180912.2180914.

[3] DALLACHIESA M, JACQUES-SILVA G, GEDIK B, *et al*. Sliding windows over uncertain data streams[J]. *Knowledge and Information Systems*, 2015, 45(1): 159–190. doi: 10.1007/s10115-014-0804-5.

[4] 彭宇, 罗清华, 彭喜元. UIDK-means: 多维不确定性测量数据聚类算法[J]. 仪器仪表学报, 2011, 32(6): 1201–1207. doi: 10.19650/j.cnki.cjsi.2011.06.001.

PENG Yu, LUO Qinghua, and PENG Xiyuan. UIDK-

means: A Multi-dimensional uncertain measurement data clustering algorithm[J]. *Chinese Journal of Scientific Instrument*, 2011, 32(6): 1201–1207. doi: 10.19650/j.cnki.cjsi.2011.06.001.

[5] BAO Chaozheng, PENG Hongming, HE Di, *et al*. Adaptive fuzzy c-means clustering algorithm for interval data type based on interval-dividing technique[J]. *Pattern Analysis and Applications*, 2018, 21(3): 803–812. doi: 10.1007/s10044-017-0663-2.

[6] D'URSO P, MASSARI R, DE GIOVANNI L, *et al*. Exponential distance-based fuzzy clustering for interval-valued data[J]. *Fuzzy Optimization and Decision Making*, 2017, 16(1): 51–70. doi: 10.1007/s10700-016-9238-8.

[7] BRITO P, SILVA A P D, and DIAS J G. Probabilistic clustering of interval data[J]. *Intelligent Data Analysis*, 2015, 19(2): 293–313. doi: 10.3233/IDA-150718.

[8] HAMDAN H. Maximum likelihood estimation from interval-valued data. Application to fuzzy clustering[C]. The 13th International Conference on Theory and Application of Fuzzy Systems and Soft Computing -ICAFS-2018. Istanbul, Turkey, 2019: 3–10. doi: 10.1007/978-3-030-04164-9_3.

[9] 谢志伟, 王志明. 一种区间型数据的自适应模糊C均值聚类算法[J]. 计算机工程与应用, 2012, 48(17): 193–198, 237. doi: 10.3778/j.issn.1002-8331.2012.17.038.
XIE Zhiwei and WANG Zhiming. Self-adapting fuzzy c means clustering algorithm for interval data[J]. *Computer Engineering and Applications*, 2012, 48(17): 193–198, 237. doi: 10.3778/j.issn.1002-8331.2012.17.038.

[10] GAO Xinbo, JI Hongbing, and XIE Weixin. A novel FCM clustering algorithm for interval-valued data and fuzzy-valued data[C]. The 5th International Conference on Signal Processing Proceedings. The 16th World Computer Congress 2000, Beijing, China, 2000: 1551–1555. doi: 10.1109/ICOSP.2000.893395.

[11] MACIEL L, BALLINI R, GOMIDE F, *et al*. Participatory learning fuzzy clustering for interval-valued data[C]. The 16th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, Eindhoven, The Netherlands, 2016: 687–698. doi: 10.1007/978-3-319-40596-4_57.

[12] 兰蓉. 模糊信息距离及其若干应用[D]. [博士论文], 西安电子科技大学, 2013: 61–73.
LAN Rong. Fuzzy information distances and their some applications[D]. [Ph.D. dissertation], Xidian University, 2013: 61–73.

[13] 金萍, 宗瑜, 屈世超, 等. 面向不确定数据的近似骨架启发式聚类算法[J]. 南京大学学报: 自然科学, 2015, 51(1): 197–205. doi: 10.13232/j.cnki.jnju.2015.01.027.
JIN Ping, ZONG Yu, QU Shichao, *et al*. Approximate backbone guided heuristic clustering algorithm for uncertain data[J]. *Journal of Nanjing University*: *Natural Sciences*, 2015, 51(1): 197–205. doi: 10.13232/j.cnki.jnju.2015.01.027.

[14] 魏方圆, 黄德才. 基于区间数的多维不确定性数据UID-DBSCAN聚类算法[J]. 计算机科学, 2017, 44(11A): 442–447. doi: 10.11896/j.issn.1002-137X.2017.11A.094.
WEI Fangyuan and HUANG Decai. UID-DBSCAN clustering algorithm of multi-dimensional uncertain data based on interval number[J]. *Computer Science*, 2017, 44(11A): 442–447. doi: 10.11896/j.issn.1002-137X.2017.11A.094.

[15] ZHANG Qin, FANG Zhigeng, LIU Sifeng, *et al*. On variable weight clustering model of generalized interval grey numbers for multiple uncertain data[J]. *Journal of Grey System*, 2019, 31(1): 84–99.

[16] 陆亿红, 任胜亮. 基于区间数的不确定数据流2κ近邻聚类算法[J]. 浙江工业大学学报, 2018, 46(3): 321–326. doi: 10.3969/j.issn.1006-4303.2018.03.015.
LU Yihong and REN Shengliang. The clustering algorithm of uncertain data stream 2κ-near neighbors based on interval number[J]. *Journal of Zhejiang University of Technology*, 2018, 46(3): 321–326. doi: 10.3969/j.issn.1006-4303.2018.03.015.

[17] 张新猛, 蒋盛益. 一种基于相似度概率的不确定分类数据聚类算法[J]. 山东大学学报: 工学版, 2011, 41(3): 12–16.
ZHANG Xinmeng and JIANG Shengyi. An algorithm for clustering uncertain categorical data based on similarity probability[J]. *Journal of Shandong University*: *Engineering Science*, 2011, 41(3): 12–16.

[18] TRAN L and DUCKSTEIN L. Comparison of fuzzy numbers using a fuzzy distance measure[J]. *Fuzzy Sets and Systems*, 2002, 130(3): 331–341. doi: 10.1016/s0165-0114(01)00195-6.

肖满生：男，1968年生，教授，主要研究方向为智能计算和智能信息处理.

张龙信：男，1983年生，博士，讲师，研究方向为大数据与数据安全.

张晓丽：女，1994年生，硕士，研究方向为智能信息处理.

责任编辑：阮 望