一种基于属性空间相似性的模糊聚类算法

施伟锋* 卓金宝 兰 莹
(上海海事大学 上海 201306)

摘 要：模糊C均值(FCM)聚类算法及其相关改进算法基于最大模糊隶属度原则确定聚类结果，没有充分利用迭代后的模糊隶属度矩阵和簇类中心的样本属性特征信息，影响聚类准确度。针对这个问题，该文提出一种新的改进思路：改进FCM算法输出定类原则。给出二元属性拓扑子空间中属性相似度的定义，最终提出一种基于属性空间相似性的改进FCM算法(FCM-SAS)：首先，选择FCM算法聚类后模糊隶属度低于聚类置信度的样本作为存疑样本；然后，计算存疑样本与聚类后聚类中心的属性相似度；最后，基于最大属性相似度原则更新存疑样本的簇类标签。通过UCI数据集实验，证明算法不仅有效，还较一些基于最大模糊隶属度原则定类的改进算法具有更优的聚类评价指标。

关键词：模糊C均值聚类；属性拓扑子空间；拓扑相似度；聚类置信度；最大属性相似度原则
中图分类号：TP391 **文献标识码**：A **文章编号**：1009-5896(2019)11-2722-07
DOI: 10.11999/JEIT180974

A Novel Fuzzy Clustering Algorithm Based on Similarity of Attribute Space

SHI Weifeng ZHUO Jinbao LAN Ying
(*Shanghai Maritime University*, *Shanghai* 201306, *China*)

Abstract: With the attribute feature information of the fuzzy membership matrix and cluster centers after the iteration not fully utilized, the results of Fuzzy C-Means (FCM) Clustering and related modified algorithms are determined based on the principle of maximum fuzzy membership, causing bad influence on the clustering accuracy. To solve this problem, the improvement ideas are proposed: to improve classification principle of FCM. The formula definition of attribute similarity in binary topological subspaces is given. Then, the improved FCM algorithm based on the Similarity of Attribute Space (FCM-SAS) is proposed: First, samples with fuzzy membership degree lower than the clustering reliability are selected as suspicious samples. Next, the attribute similarity between the suspicious samples and the cluster centers after clustering are calculated. Finally, cluster labels of suspicious samples based on the principle of maximum attribute similarity are updated. The validity and superiority of the proposed algorithm is verified by the UCI sample set experiments and comparisons with other modified algorithms based on the principle of maximum fuzzy membership.

Key words: Fuzzy C-Means (FCM) clustering; Attribute topology subspace; Attribute similarity; Clustering reliability; Principle of maximum attribute similarity

1 引言

模糊C均值(FCM)聚类由Bezdek[1]在1981年提出，经过多年的研究，目前已经大量应用于机器学习、图像处理和数据挖掘等领域，是无监督类模式识别研究的重要内容之一。

随着各类应用领域的深入研究，出现大量改进算法。从算法参数结构和输入特征的角度看，FCM算法的改进基本上可以分为两类。第1类是FCM算法参数优化选取。为了自动确定聚类簇数，文献[2]将鲁棒竞争凝聚理论引入FCM中，所提出的RCA算法可以有效解决在聚类簇数个数未知情况下FCM算法聚类问题。文献[3]在此基础上提出一种基于鲁棒学习的RL-FCM算法，聚类效果得到大幅提高。为了生成最优初始聚类中心，文献[4]分别基于网格估计和密度估计生成最优初始聚类中心，继而提出两种改进算法GB-FCM和DG-FCM，相较于FCM更适合大数据聚类。为了选择最优模糊加

收稿日期：2018-10-17；改回日期：2019-02-28；网络出版：2019-04-25
*通信作者： 施伟锋 wfshi@shmtu.edu.cn
基金项目：国家自然科学基金(61503240)，上海海事大学研究生创新基金(2016ycx078)
Foundation Items: The National Natural Science Foundation of China (61503240), Shanghai Maritime University Graduate Student Innovation Fund Project (2016ycx078)

权指数，文献[5]提出一种熵指数约束竞争凝聚聚类算法EICCA，不仅可以自动确定聚类个数，还可以自动选择最佳模糊加权指数。为了选择最适合的距离函数，除了现在广泛采用的欧式距离和马氏距离，文献[6]通过分析磁共振图像的特点，用局部和非局部区域的加权欧式距离替代传统的欧氏距离作为FCM距离函数，图像聚类准确度较传统算法更高。文献[7]用样本数据的几何散度值作为距离函数，改进的FCM算法不仅聚类准确度有所提高，还具有更强的抗躁性能。文献[8]基于样本影响值改进簇类中心计算方法和距离函数，所提出的SCMS-FCM算法对含有大噪声样本具有较好的处理效果。第2类是对算法输入数据进行分析处理。基于样本数据特征的一类改进方法，充分挖掘和利用样本点、各个簇类和样本总体三者间在样本空间的相对分布特征信息。实际上，由于样本特征差异的程度不同，此类改进方法的难点仍然是如何提取最显著的样本特征，然后对不同特征赋权[9]。另外，在所提取的数据特征冗杂的情况下，选取尽量少的特征聚类也是当前关注的热点[10]。在这两类改进思路的实现过程中，常将FCM算法与其他方法的混合或集成，如粒子群[11]、遗传算法[12,13]、神经网络[14]、核函数[15]等。

虽然上述改进算法在不同程度上提高了聚类性能，但是最后一步仍然采用传统的定类准则——最大模糊隶属度准则，简单地处理迭代后模糊隶属度矩阵以确定簇类标签，未对最大隶属度准则进行分析和改进。传统模糊理论认为，最大模糊隶属度准则可以保证最终所确定的样本隶属关系相对而言是最显著的。所以，在样本的各个模糊隶属度取值相差较大的情况下，此原则是简单且有效的。然而，在采用FCM算法进行实际聚类时会出现以下情况：在各个模糊隶属度取值差异较小的情况下，当设置不同的模糊加权指数、目标函数最小误差和迭代次数等基本参数时，继续使用最大模糊隶属度准则进行定类可能造成这些样本被错误分类，继而影响聚类准确度。

近年来，为了解决大数据背景下复杂多属性数据分析问题，除了传统的样本点间数据分析研究，研究人员也从属性降维技术[16,17]、属性约简[18,19]和属性空间的构建与表述[20,21]等方面开展了大量样本属性空间研究。主成分分析(Principal Components Analysis, PCA)是应用最为广泛的属性空间降维技术之一，文献[17]基于PCA和极限学习机提出一种堆叠隐空间FCM改进算法，具有更高效的非线性数据处理能力，克服了传统FCM算法对模糊指数的敏感性问题。文献[19]将张量正则分解用于属性约简并与传统FCM结合，提出一种适用于物联网大数据的FCM改进算法，虽然聚类准确率略有降低，但是聚类效率得到了大幅提升，在物联网数据智能信息挖掘方面具有较大应用前景。但是，属性降维技术和属性约简将样本高维属性空间投影到低维属性空间，不同程度地造成原样本信息的丢失。因此，在降维或维数约简之前需要充分分析样本属性间的数据关联性影响。文献[21]分析成对属性间的关联性影响，构建可以体现属性间相互影响程度的成对关联属性空间，并应用于凝聚层次聚类算法的改进，结果比传统属性空间的结果更好。但是，文献[21]只给出了基于成对属性间余弦夹角的簇类间距离计算方法，证明过程和属性空间中的物理意义并未详细说明。

受到文献[21,22]启发，针对传统FCM中定类准则可能产生误判的问题，本文提出一种FCM算法新的改进思路：改进FCM算法的输出定类方法，即最大模糊隶属度原则。可以构建存疑样本与簇类中心在属性空间中的相似度计算模型，度量模糊隶属度取值无明显差异的样本与各个簇类中心的相似度，将相似度最大的簇类中心所在类作为此存疑样本所隶属的簇类。该思路可以充分利用迭代后簇类中心内涵丰富的不同簇类间的属性特征信息，实现对最大隶属度原则的有效改进。

因此，本文首先在分析FCM算法和最大模糊隶属度原则定类缺点的基础上，给出样本与簇类中心属性相似度的定义，详细说明样本属性相似度计算的空间几何模型的建立和推导过程，然后，提出一种基于属性空间相似性的改进FCM算法(FCM based on Similarity of Attribute Space, FCM-SAS)，可以在不增加算法复杂度的情况下，明显提高聚类准确度。最后，通过UCI数据实验验证改进算法的有效性，聚类指标较FCM算法和其他改进算法更优。

2　FCM算法

假设具有s个属性的样本$\boldsymbol{x}$组成的样本集$\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$，则FCM算法的目标函数$J$如式(1)所示

$$\min\ J(\boldsymbol{U},\boldsymbol{V})=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^{m}\left\|\boldsymbol{x}_j-\boldsymbol{v}_i\right\|^2 \tag{1}$$

其中，$\boldsymbol{V}=\{\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_c\}$是$c$个可能的簇类中心；$\boldsymbol{U}=[u_{ij}]\in[0,1]$是样本的模糊隶属矩阵，$u_{ij}$表示样本点$x_j$属于第$i$个簇类的模糊隶属度，数值越大隶属程度越强；$m$是模糊加权指数，一般取为2。

在满足式(2)所示约束条件的情况下，根据Lagrange乘子法，按照式(3)和式(4)经过多次迭代更新模糊隶属矩阵$\boldsymbol{U}$和簇类中心$\boldsymbol{V}$后，便可得到理论上的最优解。

$$\sum_{i=1}^{c} u_{ij}=1\ ,\ j=1,2,\cdots,n \tag{2}$$

$$u_{ij}=\left[\sum_{k=1}^{c}\left(\frac{\|\boldsymbol{x}_j-\boldsymbol{v}_i\|}{\|\boldsymbol{x}_j-\boldsymbol{v}_k\|}\right)^{\frac{2}{m-1}}\right]^{-1},\ i\neq k \tag{3}$$

$$\boldsymbol{v}_i=\frac{\sum_{j=1}^{n} u_{ij}^{m}\boldsymbol{x}_j}{\sum_{j=1}^{n} u_{ij}^{m}} \tag{4}$$

3 最大模糊隶属度原则分析与改进思路

传统FCM算法及其一些改进型算法基于最大模糊隶属度原则进行定类，在处理上述样本集的聚类问题时存在以下两个共性问题。

(1)只对模糊隶属矩阵$\boldsymbol{U}$进行取极大值运算，存在误判的可能性。例如，在处理样本类别特征不鲜明的样本点$\boldsymbol{x}_1$和$\boldsymbol{x}_2$时，$\boldsymbol{x}_1\boldsymbol{v}_1\approx\boldsymbol{x}_1\boldsymbol{v}_2\approx\boldsymbol{x}_1\boldsymbol{v}_3$和$\boldsymbol{x}_2\boldsymbol{v}_3\gg\boldsymbol{x}_2\boldsymbol{v}_1\approx\boldsymbol{x}_2\boldsymbol{v}_2$，样本点$\boldsymbol{x}_1$和$\boldsymbol{x}_2$的隶属关系就取决于其模糊隶属度微小的数值大小差别。当算法设置不同的初始参数或者选择不同度量距离时，会使这些差别发生改变，可能影响最终的定类结果。

(2)在确定簇类标签时，忽略了迭代后的簇类中心。簇类中心的迭代变化可以看做是一种由大数据量样本集映射到小数据量样本集的压缩变换，蕴含了样本集全部簇类簇内和簇间的特征信息。忽略迭代后的簇类中心会降低这部分特征信息的有效利用，不利于聚类准确度的提高。

为了解决这两个问题，本文提出一种基于最大属性相似度原则定类的改进方法，其改进思路是初步筛选模糊隶属度分布不明显的样本作为存疑样本，再结合聚类中心的属性特征信息，对存疑样本进行2次定类。例如，对于$\boldsymbol{x}_3$, $\boldsymbol{x}_4$和$\boldsymbol{x}_5$等模糊隶属度分布差异明显的样本点，可以根据最大模糊隶属度原则直接定类；然后，筛选出最大模糊隶属度低于已给定数值的样本作为存疑样本，此给定数值可以称为聚类置信度η，像样本点$\boldsymbol{x}_1$和$\boldsymbol{x}_2$这样的存疑样本便需要2次定类；最后，再分别计算存疑样本点与各个簇类中心的相似度，选取相似度最大的相似关系更新类别标签。

4 最大属性相似度

首先定义样本的2元属性拓扑空间，然后在此空间中先后定义拓扑相似度集合和属性相似度。

定义1 设待聚类的样本点$\boldsymbol{x}$具有s个属性$\boldsymbol{x}=\{x_1,x_2,\cdots,x_s\}$, $\mathcal{T}$是由这些属性的某些子集所组成的集类，取$\mathcal{T}=\mathcal{T}_2\cup\varnothing\cup\boldsymbol{x}$，其中$\mathcal{T}_2$如式(5)所定义

$$\mathcal{T}_2=\{\{x_p,x_q\}\,|p,q\in\{1,2,\cdots,s\}\,,p\neq q\} \tag{5}$$

则称$(\boldsymbol{x},\mathcal{T})$为$\boldsymbol{x}$的属性拓扑空间，$\mathcal{T}$为$\boldsymbol{x}$的属性拓扑，$(\boldsymbol{x},\mathcal{T}_2)$为$\boldsymbol{x}$的2元属性拓扑子空间，$\mathcal{T}_2$为$\boldsymbol{x}$的2元属性拓扑。

定义2 样本集$\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$中任意两个样本点$\boldsymbol{x}_1=\{a_1,a_2,\cdots,a_s\}$和$\boldsymbol{x}_2=\{b_1,b_2,\cdots,b_s\}$，对其2元属性拓扑先后取$\rho$运算和$\gamma$运算，得到式(6)—式(8)

$$\begin{aligned}\rho(\boldsymbol{x}_1,\mathcal{T}_2)&=\rho(\boldsymbol{x}_1)\\&=\{\{a_p-a_q\}\,|p,q\in\{1,2,\cdots,s\}\,,p\neq q\}\end{aligned} \tag{6}$$

$$\begin{aligned}\rho(\boldsymbol{x}_2,\mathcal{T}_2)&=\rho(\boldsymbol{x}_2)\\&=\{\{b_p-b_q\}\,|p,q\in\{1,2,\cdots,s\}\,,p\neq q\}\end{aligned} \tag{7}$$

$$\begin{aligned}&\gamma(\rho(\boldsymbol{x}_1),\rho(\boldsymbol{x}_2))=\\&\left\{\left\{\frac{a_p-a_q}{b_p-b_q}\right\}\cup\left\{\frac{a_p}{b_p}\right\}|p,q\in\{1,2,\cdots,s\}\,,p\neq q\right\}\end{aligned} \tag{8}$$

称$\gamma(\rho(\boldsymbol{x}_1),\rho(\boldsymbol{x}_2))$是样本点$\boldsymbol{x}_1$和$\boldsymbol{x}_2$在2元属性拓扑子空间中的拓扑相似度集合，简记为$\gamma(\boldsymbol{x}_1,\boldsymbol{x}_2)$；此集合中元素个数称为拓扑相似度集合长度，简记为$\gamma_{\text{dis}}=C_s^2$。

例如，对于存疑样本点$\boldsymbol{x}_3=\{d_1,d_2,d_3,d_4\}$和簇类中心$\boldsymbol{v}_1=\{e_1,\ e_2,\ e_3,\ e_4\}$，其拓扑相似度集合$\gamma(\boldsymbol{x}_3,\boldsymbol{v}_1)$如式(9)所示

$$\begin{aligned}&\gamma(\boldsymbol{x}_3,\boldsymbol{v}_1)=\\&\left\{\frac{d_1-d_2}{e_1-e_2},\frac{d_1-d_3}{e_1-e_3},\frac{d_1-d_4}{e_1-e_4},\right.\\&\left.\frac{d_2-d_3}{e_2-e_3},\frac{d_2-d_4}{e_2-e_4},\frac{a_3-a_4}{b_3-b_4}\right\}\\&\cup\left\{\frac{d_1}{e_1},\frac{d_2}{e_2},\frac{d_3}{e_3},\frac{d_4}{e_4}\right\}\end{aligned} \tag{9}$$

如果两者属于同一种簇类，那么在2元属性拓扑子空间中体现为$\gamma(\boldsymbol{x}_3,\boldsymbol{v}_1)$中所有或者大部分元素逼近1。

定义3 如果两个样本点$\boldsymbol{x}_1$和$\boldsymbol{x}_2$的拓扑相似度集合为$\gamma(\boldsymbol{x}_1,\boldsymbol{x}_2)$，记$\gamma(\boldsymbol{x}_1,\boldsymbol{x}_2)$中落在以1为中心，以$\delta$为半径的邻域$U^o(1,\delta)$的元素个数为$\lambda$，则称

$$\psi(\boldsymbol{x}_1,\boldsymbol{x}_2)=\frac{\boldsymbol{\lambda}}{s+C_s^2} \tag{10}$$

是样本点$\boldsymbol{x}_1$和$\boldsymbol{x}_2$在二元属性拓扑子空间中的属性相似度。显然，$\psi\in[0,1]$。两个样本越相似，则ψ越接近1；反之，则ψ越接近0。

5　FCM-SAS算法

在上述分析和定义的基础上，提出一种基于属性空间相似性的模糊聚类算法(FCM-SAS)，具体步骤如表1所示。

表 1　FCM-SAS算法具体步骤

输入：	样本集$\boldsymbol{X}$、样本数n，聚类个数c、加权指数m、迭代阈值ε、最大迭代次数T、聚类存疑率ξ、属性占比率κ；
输出：	样本标签集$\boldsymbol{X}_l'$；
1	按表1的传统FCM算法步骤得到迭代后模糊隶属度矩阵$\boldsymbol{U}$、簇类中心$\boldsymbol{V}$和样本标签集$\boldsymbol{X}_l$，令$j=0$；
2	计算所有样本点$\boldsymbol{x}$的模糊隶属度最大值，按递增顺序排序并组成数组，选出此数组中第$[\xi\times n]$个元素的数值作为聚类置信度η；
3	令$j=j+1$，判断第j个样本的模糊隶属度$\max(\{u_{ij}\mid i=1,2,\cdots,c\})$是否不大于聚类置信度$\eta$，若是则此样本为存疑样本，转步骤4；否则转步骤8；
4	按式(8)计算第j个样本与各个簇类中心在2元属性拓扑子空间中的拓扑相似度集合$\gamma(\boldsymbol{x}_j,\boldsymbol{v}_i)$，将所有集合中的元素取绝对值后按递增的顺序排序并组成数组，计算此数组中第$[n\times\gamma_{dis}\times\kappa]$个元素与数值1之间差的绝对值作为邻域半径$\delta$；
5	以δ为邻域半径，按式(10)计算第j个样本与各个簇类中心的属性相似度$\psi(\boldsymbol{x}_j,\boldsymbol{v}_i)$；
6	若最大属性相似度$\max(\psi(\boldsymbol{x}_j,\boldsymbol{v}_i))$只有一个，则选出最大属性相似度时的簇类中心所在的类别作为此样本更新后的标签x_{lj}'，转步骤8；否则，转步骤7；
7	若最大属性相似度不止一个，则选择这些簇类中最大拓扑相似度集合之和$\widehat{S}=\max(\Sigma\gamma(\boldsymbol{x}_j,\boldsymbol{v}_i))$时的簇类中心所在的类别作为$x_{lj}'$；
8	判断$j<n$，若是则转步骤3，否则输出更新后的样本标签集$\boldsymbol{X}_l'$。

在此算法中，如果聚类置信度η取值过大，则会筛选出包含大量无需再次定类的样本作为存疑样本，使得算法计算量增加；如果聚类置信度η取值过小，则初步筛选的存疑样本较少，对最终的聚类准确度无显著影响。对于不同簇类的样本，邻域半径δ的取值是不同的，需要根据每个存疑样本与簇类中心的拓扑相似度集合中的元素分布确定。在此情况下，步骤2中的聚类存疑率$\xi\in(0,1)$的应用不仅可以自动确定聚类置信度η，还可以保证筛选出的存疑样本占总体样本的比重适中。步骤4中属性占比率$\kappa\in(0,1)$的应用可以生成合适的邻域半径δ，保证在充分考虑整体属性相似性的基础上，计算存疑样本与簇类中心的属性相似度。

6　仿真实验

6.1　实验说明

实验平台为一台配置Intel i7 3.6 GHz 中央处理器、8 G内存、安装Windows 7操作系统的计算机，在MATLAB R2017a中编写和运行。实验数据为专门用于测试机器学习和数据挖掘算法的UCI数据集：Iris, Wine, Seeds, Breast, Glass[23]。采用普遍应用于聚类性能测试的聚类准确率(Accuracy Rate, AR)、兰德指数[24](Rand Index, RI)和标准化互信息[25](Normalized Mutual Information, NMI)作为聚类评价指标，指标数值越大，聚类性能越好，其定义分别为

$$\mathrm{AR}=\frac{z}{n} \tag{11}$$

$$\mathrm{RI}=\frac{\mathrm{TP}+\mathrm{TN}}{C_n^2} \tag{12}$$

$$\mathrm{NMI}=\frac{\sum\limits_{i=1}^{s_1}\sum\limits_{j=1}^{s_2}P(i,j)\lg\left(\frac{P(i,j)}{P(i)P'(j)}\right)}{\sqrt{\sum\limits_{i=1}^{s_1}p(i)\lg(P(i))\sum\limits_{j=1}^{s_2}P'(j)\lg(P'(j))}} \tag{13}$$

其中，z为正确分类样本个数，n为样本数，TP为实际标签和聚类后标签都相同的样本对的个数，TN为实际标签和聚类后标签都不相同的样本对的个数，C_n^2为所有样本对的总个数，$P(i)$和$P'(j)$分别为实际样本标签组和聚类后标签组中各簇类标签出现的次数与总样本数的比值，s_1和s_2分别为实际样本标签组和聚类后标签组中簇类的个数。

FCM-SAS算法的输入参数如表2所示。

表 2　算法输入参数设置

参数	数值
加权指数m	2
迭代阈值ε	10^{-3}
最大迭代次数T	100
聚类存疑率ξ	0.3
属性占比率κ	0.5

6.2　UCI数据集聚类分析

选择UCI中Iris, Wine, Seeds, Breast, Glass等数据集，进一步测试FCM-SAS算法对真实数据集的聚类效果，其统计描述如表3所示，其中Breast数据集存在有缺省属性的样本16个，本文已删除。

从表3可以看出，Iris和Seeds为低维、簇类少、各类占比均匀的数据集，Wine为高维、簇类少、各类占比略微不均匀的数据集，Breast为高维、簇类少、各类占比不均匀的数据集，Glass为

表 3 UCI数据集的统计描述

数据集	样本数	维数	簇类	各类占比
Iris	150	4	3	50:50:50
Wine	178	13	3	59:71:48
Seeds	210	7	3	70:70:70
Breast	683	9	2	444:239
Glass	214	9	6	70:17:76:13:9:29

高维、簇类多、各类占比不均匀的数据集。这些数据集可以较好地测试聚类算法对不同数据分布特征数据集的聚类效果。数据集经FCM-SAS算法聚类处理后结果的评价指标与FCM算法、第1类相关改进算法(RCA算法[2]、RL-FCM算法[3]、KFCM算法[26])、第2类相关改进算法(WFCM算法[9]、FRCM算法[10])和混合算法(PSO-FCM算法[11]、GA-FCM算法[11]、ABC-FCM算法[11]、WGFCM算法[12])对比。聚类结果对比如表4和表5所示，这些用来对比的算法都是以最大隶属度原则确定聚类标签的，评价指标数值为算法多次运行后的平均值，在相同数据集和同一指标的情况下加粗更优的指标数值。

表 4 UCI数据集聚类结果评价指标对比(1)

		FCM	RL-FCM	RCA	WFCM	FRCM	FCM-SAS(标准化样本)	FCM-SAS(未标准化样本)
Iris	AR	0.893	0.907	0.967	0.957	0.960	**0.987**	0.953
	RI	0.880	0.892	0.958	0.934	0.952	**0.983**	0.942
	NMI	0.750	–	–	0.831	0.873	**0.949**	0.8498
Seeds	AR	0.895	0.895	0.903	0.895	0.895	**0.919**	0.900
	RI	0.874	0.874	0.884	0.873	0.876	**0.899**	0.877
	NMI	0.695	–	–	0.677	0.697	**0.717**	0.671
Breast	AR	0.937	0.953	0.655	0.938	0.947	**0.965**	0.946
	RI	0.876	0.910	0.548	0.884	0.911	**0.932**	0.897
	NMI	0.730	–	–	0.736	0.755	**0.782**	0.688

表 5 UCI数据集聚类结果评价指标对比(2)

		FCM	PSO-IFCM	GA-IFCM	ABC-IFCM	KFCM	WGFCM	FCM-SAS(标准化样本)	FCM-SAS(未标准化样本)
Iris	AR	0.893	0.807	0.849	0.787	0.895	0.973	**0.987**	0.953
Wine	AR	0.949	0.655	0.652	0.642	0.942	**0.966**	0.955	0.781
Glass	AR	0.421	0.419	0.393	0.467	0.460	**0.733**	0.533	0.472

分析表4和表5，可以发现FCM-SAS算法对这5种不同类型实际数据集的聚类指标都比FCM算法更优，尤其是在对高维、簇类多、各类占比不均匀数据集Glass的聚类准确度提高了0.112，说明基于属性最大相似度原则的改进方法是有效的。是否对样本进行标准化处理对FCM-SAS算法的聚类性能会有影响，样本经过标准化处理的聚类评价指标整体上比未经过标准化处理的更优，这也与前文属性相似度原理分析相符。对于在聚类过程中量纲影响程度较大的数据集，比如Wine数据集，如果未经过标准化处理，FCM-SAS算法的聚类效果甚至远低于传统FCM算法。同时，相比于其他类型的改进算法，除了在处理Wine和Glass数据集时聚类指标比WGFCM算法低之外，FCM-SAS的聚类指标整体上是最优的，这说明FCM-SAS算法比多数现有改进算法更优。

表6是FCM-SAS算法聚类过程中的统计数据，其中，1次定类错误的样本数实际上就是传统FCM聚类错误的样本数，存疑样本数是需要进行2次定类的样本数。

表 6 FCM-SAS算法聚类过程统计数据

样本集	1次定类错误样本数	1次定类正确的存疑样本数	1次定类错误的存疑样本数	存疑样本数	2次定类正确样本数	2次定类错误样本数
Iris	16	29	16	45	43	2
Seeds	21	44	19	63	48	15
Breast	43	173	27	200	191	9
Wine	9	45	8	53	47	6
Glass	126	21	42	63	45	18

从表6可以看出，在聚类存疑率$\xi = 0.3$的情况下，大部分传统FCM算法定类错误的样本(1次定类错误的样本数)被FCM-SAS算法筛选为存疑样本数，经过属性相似度计算与比较，这些样本被2次定类。相较于1次定类错误的存疑样本数，2次定类错误的样本数明显减小；相较于1次定类正确的存疑样本数，2次定类正确的样本数明显增加，说明基于最大属性相似度原则的2次定类是聚类评价指标得到改善的主要原因。

综上：(1)对算法输出定类原则进行改进的新思路是有效的；(2)在处理实际数据时，基于属性最大相似度原则的改进方法显著提高了传统FCM算法的聚类性能；(3)FCM-SAS算法比上述改进算法整体上更优。

7　结束语

本文建立样本属性相似度计算的空间拓扑模型，继而提出一种基于最大属性相似性原则定类的FCM-SAS算法，可以充分利用迭代后的模糊隶属矩阵和簇类中心的属性特征信息，在不显著增加算法复杂度的情况下，各项聚类性能评价指标都有得到提高。UCI数据集的聚类实验证明FCM-SAS算法的有效性，具有更优的聚类性能。在后续的研究中需要进一步考虑如何选取最优算法初始参数的问题。

参 考 文 献

[1] BEZDEK J C. Pattern Recognition with Fuzzy Objective Function Algorithms[M]. Boston: Springer, 1981: 155–201.

[2] FRIGUI H and KRISHNAPURAM R. A robust competitive clustering algorithm with applications in computer vision[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1999, 21(5): 450–465. doi: 10.1109/34.765656.

[3] YANG M and NATALIANI Y. Robust-learning fuzzy C-means clustering algorithm with unknown number of clusters[J]. *Pattern Recognition*, 2017, 71: 45–59. doi: 10.1016/j.patcog.2017.05.017.

[4] SON L H and TIEN N D. Tune up fuzzy C-means for big data: Some novel hybrid clustering algorithms based on initial selection and incremental clustering[J]. *International Journal of Fuzzy Systems*, 2017, 19(5): 1585–1602. doi: 10.1007/s40815-016-0260-3.

[5] HUANG Chengquan, CHUNG F, and WANG Shitong. Generalized competitive agglomeration clustering algorithm[J]. *International Journal of Machine Learning and Cybernetics*, 2017, 8(6): 1945–1969. doi: 10.1007/s13042-016-0572-5.

[6] SINGH C and BALA A. A DCT-based local and non-local fuzzy C-means algorithm for segmentation of brain magnetic resonance images[J]. *Applied Soft Computing*, 2018, 68: 447–457. doi: 10.1016/j.asoc.2018.03.054.

[7] SAHA A and DAS S. Geometric divergence based fuzzy clustering with strong resilience to noise features[J]. *Pattern Recognition Letters*, 2016, 79: 60–67. doi: 10.1016/j.patrec.2016.04.013.

[8] 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.

[9] WANG Xizhao, WANG Yadong, and WANG Lijuan. Improving fuzzy C-means clustering based on feature-weight learning[J]. *Pattern Recognition Letters*, 2004, 25(10): 1123–1132. doi: 10.1016/j.patrec.2004.03.008.

[10] YANG M S and NATALIANI Y. A feature-reduction fuzzy clustering algorithm based on feature-weighted entropy[J]. *IEEE Transactions on Fuzzy Systems*, 2018, 26(2): 817–835. doi: 10.1109/TFUZZ.2017.2692203.

[11] KUO R J, LIN T C, ZULVIA F E, *et al*. A hybrid metaheuristic and kernel intuitionistic fuzzy c-means algorithm for cluster analysis[J]. *Applied Soft Computing*, 2018, 67: 299–308. doi: 10.1016/j.asoc.2018.02.039.

[12] JIANG Zhaohui, LI Tingting, MIN Wengfang, *et al*. Fuzzy c-means clustering based on weights and gene expression programming[J]. *Pattern Recognition Letters*, 2017, 90: 1–7. doi: 10.1016/j.patrec.2017.02.015.

[13] JIE Lilin, LIU Weidong, SUN Zheng, *et al*. Hybrid fuzzy clustering methods based on improved self-adaptive cellular genetic algorithm and optimal-selection-based fuzzy c-means[J]. *Neurocomputing*, 2017, 249: 140–156. doi: 10.1016/j.neucom.2017.03.068.

[14] KIM E H, OH S K, and PEDRYCZ W. Reinforced hybrid interval fuzzy neural networks architecture: Design and analysis[J]. *Neurocomputing*, 2018, 303: 20–36. doi: 10.1016/j.neucom.2018.04.003.

[15] DAGHER I. Fuzzy clustering using multiple Gaussian kernels with optimized-parameters[J]. *Fuzzy Optimization and Decision Making*, 2018, 17(2): 159–176. doi: 10.1007/s10700-017-9268-x.

[16] 王骏, 刘欢, 蒋亦樟, 等. 堆叠隐空间模糊C均值聚类算法[J]. 控制与决策, 2016, 31(9): 1671–1677. doi: 10.13195/j.kzyjc.2015.0768.

WANG Jun, LIU Huan, JIANG Yizhang, *et al*. Cascaded hidden space fuzzy C means clustering algorithm[J]. *Control and Decision*, 2016, 31(9): 1671–1677. doi: 10.13195/j.kzyjc.2015.0768.

[17] LUO Xiong, XU Yang, WANG Weiping, *et al*. Towards

enhancing stacked extreme learning machine with sparse autoencoder by correntropy[J]. *Journal of the Franklin Institute*, 2018, 355(4): 1945–1966. doi: 10.1016/j.jfranklin.2017.08.014.

[18] DAI Jianhua, HU Qinghua, HU Hu, *et al*. Neighbor inconsistent pair selection for attribute reduction by rough set approach[J]. *IEEE Transactions on Fuzzy Systems*, 2018, 26(2): 937–950. doi: 10.1109/TFUZZ.2017.2698420.

[19] BU Fanyu. An efficient fuzzy c-means approach based on canonical polyadic decomposition for clustering big data in IoT[J]. *Future Generation Computer Systems*, 2018, 88: 675–682. doi: 10.1016/j.future.2018.04.045.

[20] MACIEJEWSKI R, JANG Y, WOO I, *et al*. Abstracting attribute space for transfer function exploration and design[J]. *IEEE Transactions on Visualization and Computer Graphics*, 2013, 19(1): 94–107. doi: 10.1109/TVCG.2012.105.

[21] 李保珍, 张亭亭. 成对属性关联分析及其属性空间构建[J]. 情报学报, 2014, 33(11): 1194–1203.
LI Baozhen and ZHANG Tingting. Association analysis of pairwise attributes and construction of attribute space[J]. *Journal of the China Society for Scientific and Technical Information*, 2014, 33(11): 1194–1203.

[22] WEI Cuiping, WANG Pei, and ZHANG Yuzhong. Entropy, similarity measure of interval-valued intuitionistic fuzzy sets and their applications[J]. *Information Sciences*, 2011, 181(19): 4273–4286. doi: 10.1016/j.ins.2011.06.001.

[23] BACHE K and LICHMAN M. UCI irvine machine learning repository[EB/OL]. Irvine, CA: University of California, School of Information and Computer Science, http://archive.ics.uci.edu/ml2013.

[24] RAND W M. Objective criteria for the evaluation of clustering methods[J]. *Journal of the American Statistical Association*, 1971, 66(336): 846–850. doi: 10.1080/01621459.1971.10482356.

[25] COOMBS C H, DAWES R M, and TVERSKY A. Mathematical Psychology: An Elementary Introduction[M]. Englewood Cliffs, USA: Prentice-Hall, 1970: 391–406.

[26] ZHANG Daoqiang and CHEN Songcan. Clustering incomplete data using kernel-based fuzzy C-means algorithm[J]. *Neural Processing Letters*, 2003, 18(3): 155–162. doi: 10.1023/B:NEPL.0000011135.19145.1b.

施伟锋：男，1963年生，博士，教授，主要研究方向为电力系统自动化.

卓金宝：男，1991年生，博士生，研究方向为智能故障诊断与预测.

兰 莹：女，1985年生，博士，讲师，研究方向为多自主体与混杂系统研究.