控制与决策
Control and Decision

基于边缘峰度度量的特征缩减模糊聚类算法

潘兴广， 王士同

引用本文:
潘兴广, 王士同. 基于边缘峰度度量的特征缩减模糊聚类算法[J]. 控制与决策, 2021, 36(11): 2665−2673.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2020.0220

基于边缘峰度度量的特征缩减模糊聚类算法

潘兴广[1,2†], 王士同[1]

(1. 江南大学 数字媒体学院，江苏 无锡 214122；2. 贵州民族大学 工程实训中心，贵阳 550025)

摘 要: 对含有不重要特征、冗余特征的数据进行聚类,采用特征缩减模糊聚类(feature reduction fuzzy c-means, FRFCM)算法是有效的. 该算法使用特征的均值方差比(mean-to-variance ratio, MVR)度量特征的重要性,删除权重小于阈值的特征,仅保留重要特征进行聚类,以提升算法的性能和速度. 但该算法存在以下不足: 1) 数据归一化后,特征的MVR值会发生改变,重要特征的MVR值可能会变小,不重要特征的MVR值可能会变大; 2) 一些数据的重要特征,其MVR指标未必大; 3) FRFCM算法特征权重分配依赖于初始化,不恰当的初始化会使算法给出错误的权重分配,使得聚类过程中算法会删除重要特征而保留不重要特征,造成FRFCM算法的聚类结果不正确. 对此,首先构造边缘峰度度量(marginal kurtosis measuree, MKM)指标来度量特征的重要性; 然后基于该指标提出一种新的、具有鲁棒的特征缩减模糊聚类算法. 通过在人工数据集和真实数据集上的验证,表明所提出的算法是有效的.

关键词: 模糊聚类；特征缩减；边缘峰度度量；均值方差比

中图分类号: TP391 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2020.0220 **开放科学(资源服务)标识码(OSID):**

引用格式: 潘兴广, 王士同. 基于边缘峰度度量的特征缩减模糊聚类算法[J]. 控制与决策, 2021, 36(11): 2665-2673.

Feature-reduction fuzzy clustering algorithm based on marginal kurtosis measure

PAN Xing-guang[1,2†], *WANG Shi-tong*[1]

(1. Digital Media School，Jiangnan University，Wuxi 214122，China；2. Engineer Training Center，Guizhou Minzu University，Guiyang 550025，China)

Abstract: The feature reduction fuzzy c-means (FRFCM) algorithm has been proven effective for clustering data with redundant features. The FRFCM can automatically compute individual feature weight, and simultaneously reduce these redundant feature component. However, it still has the following disadvantages: 1) the large MVR value of original features may become small if the data is normalized, and vice versa. 2) the MVR value of important features of some datasets is not necessarily large. 3) feature assignment is sensitive to initialization. The FRFCM may produce wrong weights if initialization is improper, which can deteriorate the clustering accurancy. Therefore, we first devise a new index, named marginal kurtosis measure (MKM), to measure the importance of features instead of using MVR index. Then, a novel and robust feature reduction fuzzy c-means clustering algorithm based marginal kurtosis measure is proposed. Experiments on synthetic and real-world dataset demonstrate that the proposed method is effective and efficient.

Keywords: fuzzy clustering； feature reduction； marginal kurtosis measure； mean-to-variance ration

0 引 言

聚类分析是一种数据驱动的数据分析方法,它广泛用于统计、模式识别和机器学习等领域. 给定一个由n个数据点$\boldsymbol{x}_i$(其中$i=1,2,\ldots,n$)组成的数据集,聚类算法旨在将其划分为k个簇,使得同一簇的数据点具有最大的相似性,不同簇之间的数据点相似性最小. 在基于划区的聚类算法[1-7]中, k-means[5]和模糊聚类(FCM)[6]是两个经典的算法. k-means以其简单而称著,而FCM因具有表达隶属于不同簇的数据点的能力而在数据分析领域中广受欢迎. 但k-means和FCM并未考虑特征权重,它们对每个特征一视同仁,也就是说,重要和不重要的特征对聚类的贡献相同. 因此, k-means和FCM会受到不重要特征或冗余特征的影响,导致k-means和FCM在含有不重要特征和冗余特征的数据集上得到不正确的聚类结果[8],算法的性能在某些应用场景中会显著下

收稿日期: 2020-03-02；**修回日期:** 2020-09-10.
基金项目: 国家自然科学基金面上项目(61572236).
责任编委: 薛建儒.
†通讯作者. E-mail: 408206387@qq.com.

降[9]. 为了克服这一缺点, k-means 和FCM 算法引入了特征选择和特征加权技术. 特征选择的方法假定每个被选的特征具有相同的重要性;特征加权方法是特征选择技术的扩展[10],它允许特征权重在[0,1]范围内取值,并且特征越重要,权重值越大. 特征加权是聚类分析中的一项重要技术. 特征加权技术被引入到 k-means 和FCM 算法后,基于这两个算法的一些改进算法被相继提出,例如: WKM[11]、EWKM[12]、MWKM[13]、MWFCM[14]、WFCM[15]、SCAD2[16] 和ESSC[17] 等算法. 这些算法使用特征加权技术后,性能普遍得到提高. 但是,这些算法没有考虑数据中的不重要特征或冗余特征,同时也没有在聚类过程中引入删除不重要特征和冗余特征的机制. 特征缩减模糊聚类(FRFCM)[8] 算法最先把特征缩减技术用于聚类,在进行聚类的同时进行特征缩减,提升了聚类的准确率和算法的运算速度. 在多视角 k-means[18] 算法中也使用了特征缩减技术.

本文的主要贡献有: 1) 提出一个度量特征重要性的MKM 指标; 2) 提出一个特征缩减的模糊聚类新算法; 3) 设计一个具有理论依据的阈值,在特征缩减机制中判断哪些特征需要保留,哪些特征需要删除.

1 相关工作介绍

本部分先对包括FRFCM 算法在内的特征加权聚类算法进行介绍,再对这些算法的特点进行详细讨论.

1.1 特征加权的 k-means 算法

Huang 等[11] 提出了加权 k-means 算法(weighted k-means, WKM), 该算法在 k-means 算法中增加一个附加步骤,以计算聚类过程中每次迭代的特征权重. 由于每个特征的权重与特征的类内方差之和成反比, WKM 算法可以识别数据的不重要特征. 使用特征加权的技术,可以显著降低不重要特征对聚类结果的影响,提升算法的性能. Jing 等[12] 提出了一种熵加权 k-means (entropy weighted k-means, EWKM)算法. EWKM 算法是一种子空间聚类算法,该算法使用特征加权的技术,对高维稀疏数据聚类较为有效. EWKM 算法最小化类内方差,同时最大化特征权重的熵,使其能使用更多的维度对高维稀疏数据进行聚类,避免了仅使用几个维度来对高维数据进行聚类. 但EWKM 算法的计算量大,增加了算法的时间复杂度.

1.2 特征加权的模糊聚类算法

Wang 等[15] 提出了一种特征加权的FCM(WFCM)算法. WFCM 算法对FCM 算法进行改进,赋予特征不同的权重,以提升FCM 算法的性能. Frigui 等[16] 提出了另一种特征加权FCM 算法,称为同时聚类和属性区分(simultaneous clustering and attribute discrimination, SCAD1) 算法. 在SCAD1 算法中,既考虑了特征权重,又考虑了不同簇类的权重分配. 同时, Frigui 等[16] 在特征权重中添加了一个指数 q,提出了另外一个同时聚类和属性区分(simultaneous clustering and attribute discrimination, SCAD2) 算法, SCAD1 和SCAD2 都具有相似的聚类行为和结果. Deng 等[17] 将簇间信息引入到加权软子空间聚类算法中,提出了增强软子空间聚类(enhanced soft subspace clustering, ESSC) 算法. 引入簇间信息和数据集的模糊划分后,虽然ESSC 算法的性能优于大多数的特征加权聚类算法,但ESSC 算法存在不能正确为特征分配权重的问题.

1.3 特征缩减的模糊聚类算法

Yang 等[8] 最先把特征缩减的思想引入到模糊聚类算法中,提出了FRFCM 算法. 他们使用特征的均值方差比(mean-to-variance ration, MVR) 指标来度量特征的重要性,并在目标函数中使用特征的MVR 值控制特征加权方差项和熵项. FRFCM 算法使用迭代公式计算并更新特征权重,通过使用一个阈值,删除权重小于阈值的不重要特征. 在删除数据不重要的特征之后,提高了FRFCM 算法聚类的准确性,同时也提升了聚类速度. FRFCM 算法的目标函数如下:

$$\begin{aligned}&\min_{\boldsymbol{U},\boldsymbol{V},\boldsymbol{w}} J(\boldsymbol{U},\boldsymbol{V},\boldsymbol{w}) = \\ &\sum_{i=1}^{n}\sum_{k=1}^{c}u_{ik}^{m}\sum_{j=1}^{d}\delta_j w_j(x_{ij}-v_{kj})^2+\frac{n}{c}\sum_{j=1}^{d}w_j\log\delta_j w_j;\\ &\text{s.t.}\ \sum_{k=1}^{c}u_{ik}=1,\ \sum_{j=1}^{d}w_j=1,\ \forall j=1,2,\ldots,n.\end{aligned}\tag{1}$$

其中: $\delta_j=\left(\dfrac{\text{mean}(\boldsymbol{X})}{\text{var}(\boldsymbol{X})}\right)_j,\ j=1,2,\cdots,d.$

w_j 的迭代公式如下:

$$w_j=\frac{\dfrac{1}{\delta_j}\exp\left(-\dfrac{c}{n}\sum\limits_{k=1}^{c}\sum\limits_{i=1}^{n}u_{ik}^{m}\delta_j(x_{ij}-v_{kj})^2\right)}{\sum\limits_{t=1}^{d}\dfrac{1}{\delta_t}\exp\left(-\dfrac{c}{n}\sum\limits_{k=1}^{c}\sum\limits_{i=1}^{n}u_{ik}^{m}\delta_t(x_{it}-v_{kt})^2\right)}.\tag{2}$$

虽然FRFCM 算法在一些数据集上得到较好的聚类结果,但FRFCM 算法存在如下不足: 1) FRFCM 算法使用特征的MVR 指标度量特征的重要性,但重要特征的MVR 值不一定大; 2) 对一些数据集归一化后,特征的MVR 值会改变,较大的MVR 值会变

小,较小的MVR值会变大; 3)特征加权的聚类算法不能正确分配权重的问题[18-20],FRFCM算法属于特征加权算法的范畴,亦不能为特征分配合理权重. 因此,在聚类过程中,如果重要特征分配到的权重值较小,则进行特征缩减时,重要特征将被删除,这会严重降低聚类算法的性能. 为了克服FRFCM算法的不足,本文构造一个边缘峰度度量(marginal kurtosis measure, MKM)的特征重要性指标,并基于该指标提出一个新的具有鲁棒性的特征缩减模糊聚类算法.

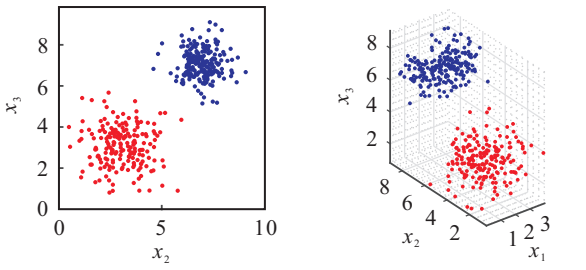
2　基于边缘峰度度量的特征缩减模糊聚类算法

给定数据集$\boldsymbol{X}$,设$\eta = (\boldsymbol{X} - E(\boldsymbol{X}))^2$,本文使用统计量$\eta$的均值与标准差之比(mean-to-standard-deviation-ration, MSDR)代替η的均值与方差之比(mean-to-variance ration, MVR)来度量特征的重要性,并定义δ_j如下:

$$\delta_j = (E\eta\sqrt{\mathrm{var}(\eta)})_j = \left(\frac{E((\boldsymbol{X}-E(\boldsymbol{X}))^2)}{\sqrt{E((\boldsymbol{X}-E(\boldsymbol{X}))^2 - E((\boldsymbol{X}-E(\boldsymbol{X}))^2))^2}}\right)_j = \left(1\Big/\sqrt{\frac{E((\boldsymbol{X}-E(\boldsymbol{X}))^4)}{(E((\boldsymbol{X}-E(\boldsymbol{X}))^2))^2} - 1}\right)_j = (1/\sqrt{\mathrm{kurtosis}(\boldsymbol{X}) - 1})_j. \tag{3}$$

在式(3)中,δ_j是η的均值与标准差之比,本质上它是数据矩阵$\boldsymbol{X}$各维特征的峰度的函数,称之为边缘峰度度量(marginal kurtosis measure, MKM). 与MVR指标相比,MKM指标具有对数据归一化不变的性质. 由式(3)可知,峰度$(\mathrm{kurtosis}(\boldsymbol{X}))_j$越小,$\delta_j$越大,特征$j$越重要. 即数据$\boldsymbol{X}$的特征的MKM值越大,特征越重要. 下面通过例子揭示文献[8]中使用的MVR指标度量特征重要性的不足,并说明本文所构造指标的合理性.

现按文献[8]的方法构造含400个数据点的人工数据集,数据集的第2、3维特征x_2和x_3由高斯混合模型$\sum_{k=1}^{2}\alpha_k N(\mu_k, \Sigma_k)$生成. 其中: 参数$\mu_1 = [5,5]$, $\mu_2 = [7,7]$; $\Sigma_1 = \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$, $\Sigma_1 = \begin{bmatrix}0.5 & 0\\ 0 & 0.5\end{bmatrix}$; $\alpha_1 = \alpha_2 = 0.5$. 见图1(a),第1维特征$x_1$和第4维特征$x_4$分别在区间$[0,4]$和$[0,2]$上服从均匀分布,记该数据集为$D_1$. 观察图1(a)可知,数据集$D_1$中第2、3维是重要特征. 而根据表1中MKM指标大小,亦可判断第2、3维是重要特征. 但表1中第2、3行数据表明,如果D_1未归一化,则根据特征的MVR指标大小,可判断第1、4维是重要特征; 对数据归一化后,第2、3维变成重要特征(见图2). 但对于MKM指标,不论数据是否归一化,指标值均不变. 再考虑iris数据集的特征指标,不论数据归一化与否,第1、2维特征MVR指标值较大,但实际上第3、4维特征才是重要特征. 通过对比,说明本文设计的MKM指标用于度量特征重要性较为合理.



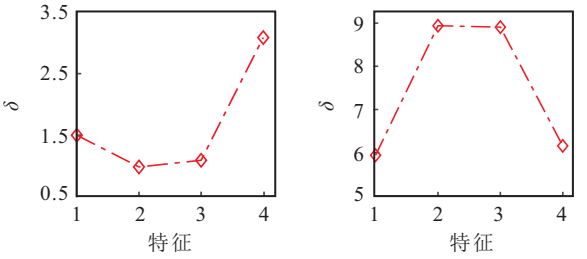
(a) 二维高斯混合样本点　　(b) 样本点扩充特征x_1

图 1　D_1在二维子空间和三维子空间的分布

表 1　D_1数据集特征的MVR指标与MKM指标对比

指标	x_1	x_2	x_3	x_4
MKM	1.072	1.360	1.341	1.143
MVR	1.497	0.981	1.089	3.085
MVR*	5.944	8.940	8.906	6.161

注: MVR*为归一化后的指标.



(a) 未归一化D_1的δ_{MVR}值　　(b) 归一化后D_1的δ_{MVR}值

图 2　归一化和未归一化D_1的δ_{MVR}值

表 2　iris数据集特征的MVR指标与MKM指标对比

指标	x_1	x_2	x_3	x_4
MKM	0.834	0.666	1.282	1.222
MVR	8.522	16.244	1.207	2.054
MVR*	8.103	13.455	5.228	4.527

注: MVR*为归一化后的指标.

2.1　基于边缘峰度度量的特征缩减模糊聚类算法

受Yang等[8]方法的启发,本文构造度量特征重要性的边缘峰度度量(MKM)指标后,基于该指标提出一种新的、具有鲁棒性的特征缩减模糊聚类算法,新算法克服了FRFCM算法的缺点. 新算法使用MKM指标度量特征的重要性,每个特征的权重使用

迭代公式进行更新. 在聚类过程中,权重小于阈值的特征从数据集中删除,而权重大于阈值的特征则被算法保留. 设$\boldsymbol{X}=\{\boldsymbol{x}_i|i=1,2,\dots,n\}$为数据矩阵,特征的权重$\boldsymbol{w}=[w_1,\dots,w_d]^{\mathrm{T}}$,新算法为如下目标函数的最优化问题:

$$\min_{\boldsymbol{U},\boldsymbol{V},\boldsymbol{w}} J(\boldsymbol{U},\boldsymbol{V},\boldsymbol{w})=\sum_{i=1}^{n}\sum_{k=1}^{c}\sum_{j=1}^{d}u_{ik}^{m}w_j(x_{ij}-v_{kj})^2+\gamma\sum_{j=1}^{d}w_j(\log w_j-\log\delta_j);\quad(4)$$

$$\text{s.t}\sum_{k=1}^{c}u_{ik}=1,\ \sum_{j=1}^{d}w_j=1,\ w_j\geqslant 0.\quad(5)$$

其中δ_j由式(3)给出,δ_j为边缘峰度度量,用来控制特征权重w_j. 由于上述聚类问题是一个有约束条件的优化问题,考虑到$\boldsymbol{U}$、$\boldsymbol{V}$和$\boldsymbol{w}$是连续变量,为了求出最优解,使用拉格朗日乘子法进行求解. 现将拉格朗日函数分别对$\boldsymbol{U}$、$\boldsymbol{V}$和$\boldsymbol{w}$求一阶偏导数,并令其为零,可得

$$L=J+\sum_{i=1}^{n}\lambda_i\Big(1-\sum_{k=1}^{c}\mu_{ik}\Big)+\beta\Big(1-\sum_{j=1}^{d}w_j\Big),\quad(6)$$

$$\frac{\partial L}{\partial u_{ik}}=mu_{ik}^{m-1}\sum_{j=1}^{d}w_j(x_{ij}-v_{kj})^2-\lambda_i=0,\quad(7)$$

$$\frac{\partial L}{\partial v_{kj}}=-2w_j\sum_{i=1}^{n}u_{ik}^{m}v_{kj}(x_{ij}-v_{kj})=0,\quad(8)$$

$$\frac{\partial L}{\partial w_j}=\sum_{k=1}^{c}\sum_{i=1}^{n}u_{ik}^{m}(x_{ij}-v_{kj})^2+\gamma(\ln w_j-\ln\delta_j+1)+\beta=0.\quad(9)$$

根据式(7),可得u_{ik}的迭代公式为

$$u_{ik}=\frac{\Big[\sum_{j=1}^{d}w_j(x_{ij}-v_{kj})^2\Big]^{\frac{1}{1-m}}}{\sum_{s=1}^{c}\Big[\sum_{j=1}^{d}w_j(x_{sj}-v_{kj})^2\Big]^{\frac{1}{1-m}}}.\quad(10)$$

由式(8),可得

$$v_{kj}=\sum_{i=1}^{n}u_{ik}^{m}x_{ij}\Big/\sum_{i=1}^{n}u_{ik}^{m}.\quad(11)$$

根据式(9),推导出w_j的迭代公式为

$$w_j=\frac{\delta_j\exp\Big(-\frac{c}{\lambda n}\sum_{k=1}^{c}\sum_{i=1}^{n}u_{ik}^{m}(x_{ij}-v_{kj})^2-1\Big)}{\sum_{j=1}^{d}\delta_j\exp\Big(-\frac{c}{\lambda n}\sum_{k=1}^{c}\sum_{i=1}^{n}u_{ik}^{m}(x_{ij}-v_{kj})^2-1\Big)}.\quad(12)$$

阈值的选择是特征缩减聚类算法的重要步骤. 在FRFCM算法中,Yang等[8]使用$1/\sqrt{nd}$作为阈值. 但本文设法寻找一个合适的阈值来确定哪些特征需要删除,而不是使用$1/\sqrt{nd}$作为阈值. 考虑到δ_j是特征重要性的度量,在式(4)中,如果对δ_j进行归一化,让$\gamma\to+\infty$,则当算法收敛时有$w_j\approx\delta_j$,因此,可使用δ_j提供的信息来构造合适的阈值. 由算术平均数、几何平均数和调和平均数的关系可知

$$d\Big/\sum_{j=1}^{d}\frac{1}{\delta_j}<\sqrt[d]{\prod_{j=1}^{d}\delta_j}<\frac{1}{d}\sum_{j=1}^{d}\delta_j=\frac{1}{d}.$$

因此,可以选择$d\Big/\sum_{j=1}^{d}\frac{1}{\delta_j}$作为新算法的阈值. 为了让算法更具灵活性,可以选择$\alpha\cdot d\Big/\sum_{j=1}^{d}\frac{1}{\delta_j}$为阈值,其中参数$\alpha$的大小视$w_j$取值情况而定.

2.2 新算法迭代过程

初始化:固定$\varepsilon>0$,给定聚类簇类数,随机初始化聚类中心$\boldsymbol{V}^{(0)}$和特征权重$\boldsymbol{w}^{(0)}$,令$t=1$.

step 1: 对数据集$\boldsymbol{X}$,根据式(3)计算δ_j.

step 2: 根据式(10),使用δ_j、$\boldsymbol{V}^{(t-1)}$和$\boldsymbol{w}^{(t-1)}$计算隶属度函数$\boldsymbol{U}^{(t)}$.

step 3: 根据式(11),使用$\boldsymbol{U}^{(t)}$更新聚类中心$\boldsymbol{V}^{(t)}$.

step 4: 使用δ_j、$\boldsymbol{U}^{(t)}$和$\boldsymbol{V}^{(t)}$,根据式(12)更新$\boldsymbol{w}^{(t)}$,对$\boldsymbol{w}^{(t)}$归一化.

step 5: 从数据集中删除$|R|$个特征,其中

$$R=\Big\{j\Big|w_j^{(t)}\leqslant\sqrt{d}\Big/\sum_{j=1}^{d}\frac{1}{\delta_j}\Big\},$$

并令$d^{(\mathrm{new})}=d-|R|$.

step 6: 利用式(12)计算$\boldsymbol{w}^{(t)}$.

step 7: 如果$\|\boldsymbol{U}^{(t)}-\boldsymbol{U}^{(t-1)}\|<\varepsilon$,则结束循环;否则令$t=t+1,d=d^{(\mathrm{new})}$,返回step 2.

3 实验研究

将WKM、EWKM、WFCM、SCAD2、增强软子空间聚类ESSC、FRFCM算法与所提出算法进行对比实验,然后评估所提出算法性能. 实验使用的数据集包含人工数据集和真实数据集. 在大多数情况下,对原始数据运行聚类算法效果不佳,因此,先按下式对每个特征$\boldsymbol{f}_j=[x_{1j},x_{2j},\dots,x_{nj}]^{\mathrm{T}}$进行归一化处理:

$$\boldsymbol{y}_j=\frac{\boldsymbol{f}_j^{\mathrm{T}}-\boldsymbol{f}_j\cdot\mathbf{1}^{\mathrm{T}}\cdot\mathbf{1}}{\max(\boldsymbol{f}_j)-\min(\boldsymbol{f}_j)},\quad(13)$$

其中$\mathbf{1}$是元素均为1的向量.

在所有算法中,分别设置阈值$\varepsilon=10^{-5}$和最大迭代次数$t=500$. 所有属于模糊聚类范畴的算法中,均设置模糊指数$m=2$. 实验的所有代码均用Matlab

编写, 并在配置为 4 GB 内存、CUP 为 Intel Core i5-4590 的 PC 上运行. 实验使用正确率 (AC)[21]、归一化互信息 (NMI)[22] 和 ARI[23] 三个性能指标来评价聚类算法的性能. 使用文献 [24] 的初始化策略对聚类中心进行随机初始化, 同时对特征权重也进行随机初始化, 实验结果为算法运行 30 次的结果取平均.

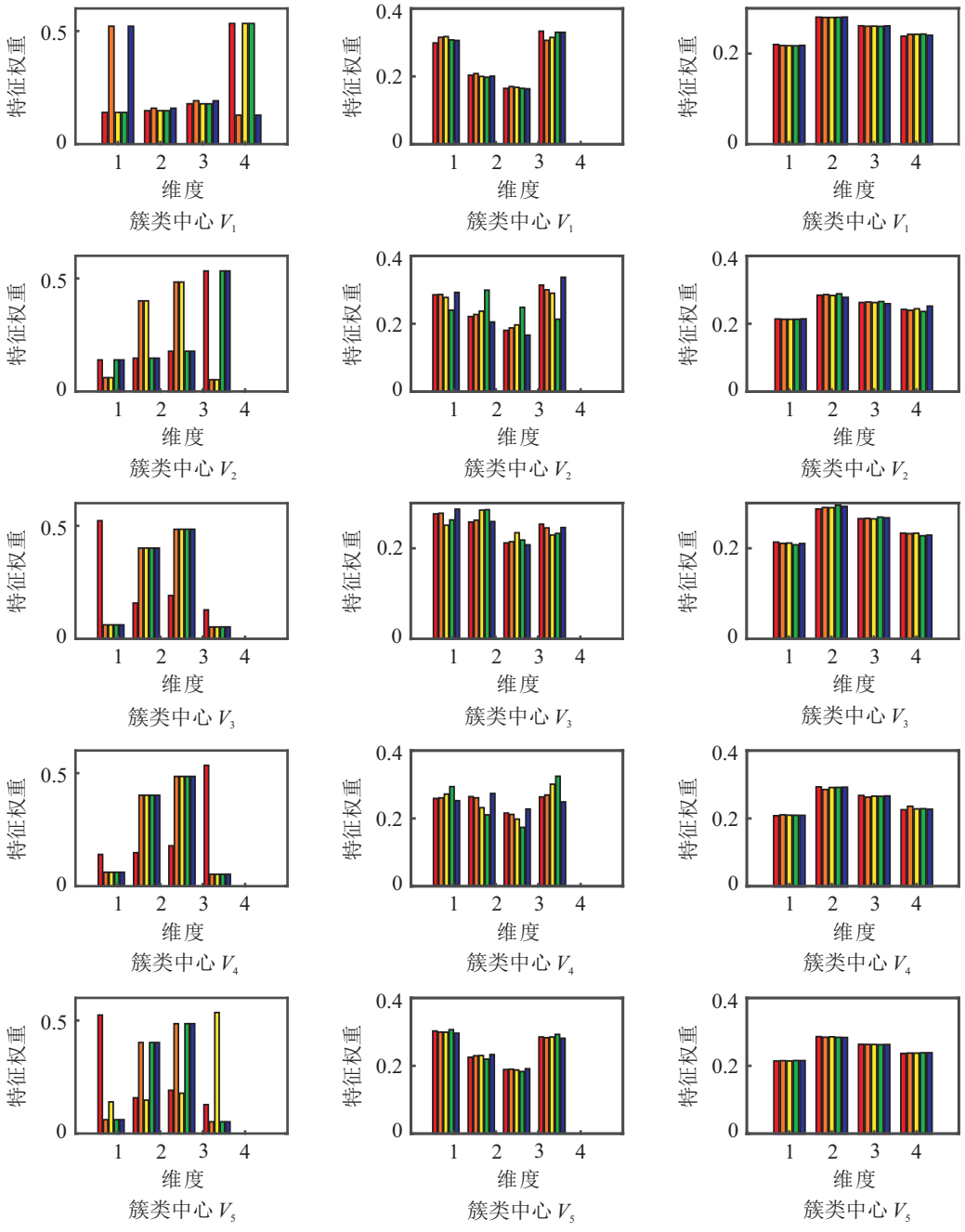
3.1　实验过程和实验分析

3.1.1　实验 1 (本文算法能为特征正确分配权重)

为了便于与 FRFCM 算法作对比, 本实验生成文献 [8] 中例 1 使用的数据集, 记为 D_2. 生成该数据集的方法与第 2 节中生成数据集 D_1 的方法相同, 不同的是, 特征 x_1 和特征 x_4 分别在区间 $[0, 10]$ 和 $[0, 12]$ 服从均匀分布. 现利用式 (13) 对 D_2 进行归一化. 按文献 [22] 的初始化策略从 D_2 中随机选择 5 个样本点 V_1, V_2, V_3, V_4, V_5, 作为 5 组初始的簇类中心 (见表 3); 然后在该数据集执行 WKM、FRFCM 和本文算法各 30 次, 以验证本文算法能否为特征分配正确的权重.

表 3　实验选取的 5 组初始簇类中心

	初始簇类中心
V_1	[0.805, 0.924, 0.699, 0.831; 0.495, 0.699, 0.701, 0.392]
V_2	[0.155, 0.708, 0.761, 0.963; 0.142, 0.436, 0.367, 0.321]
V_3	[0.232, 0.776, 0.761, 0.677; 0.512, 0.190, 0.327, 0.682]
V_4	[0.276, 0.719, 0.690, 0.357; 0.369, 0.231, 0.428, 0.757]
V_5	[0.086, 0.524, 0.412, 0.582; 0.191, 0.335, 0.264, 0.646]



(a) WKM 算法　　(b) FRFCM 算法　　(c) 本文算法

图 3　WKM、FRFCM 和本文算法在 D_2 上的特征权重分配

为了节省空间,实验中仅绘制前5次运行的特征权重的直方图. 对于WKM绘制算法终止时权重的直方图,而对于FRFCM和本文算法,则是完成第1次迭代后的直方图. 如图3所示,对于这5组中心,随机初始化权重,然后运行WKM算法和FRFCM算法,实验表明,这两个算法并没有将权重正确分配给各维特征. 不同的初始权重会产生不同的权重分配结果(见图3),说明两个算法对权重初始化较为敏感. 不恰当的权重初始化,所产生的特征权重不能为特征分配正确的权重. 对于FRFCM算法而言,这一缺点在未归一化的数据集D_2上表现尤为明显. 固定特征的初始权重为$\boldsymbol{w}=\left[\frac{1}{4},\frac{1}{4},\frac{1}{4},\frac{1}{4}\right]^{\mathrm{T}}$时,随机初始化聚类中心,此时WKM和FRFCM算法亦未能正确为特征分配权重(见表4).

表4 各算法在D_2上的聚类结果

评价指标	AC	NMI	ARI
WKM	0.839 ± 0.213	0.621 ± 0.459	0.635 ± 0.467
EWKM	0.864 ± 0.199	0.712 ± 0.431	0.683 ± 0.452
WFCM	0.932 ± 0.165	0.841 ± 0.342	0.852 ± 0.347
SCAD2	0.998 ± 0.000	0.977 ± 0.000	0.990 ± 0.000
ESSC	0.998 ± 0.000	0.977 ± 0.000	0.990 ± 0.000
FRFCM	0.998 ± 0.000	0.977 ± 0.000	0.990 ± 0.000
本文算法	$\mathbf{0.998\pm0.000}$	$\mathbf{0.977\pm0.000}$	$\mathbf{0.990\pm0.000}$

本文所提出的算法,对于固定的初始中心随机初始化权重,对于固定初始权重随机初始化中心,这两种情况下均能给特征正确分配权重,该算法分配给第2维和第3维特征权重较大,分配给第1维和第4维特征权重较小. 此时执行本文算法,算法完成第1轮迭代便删除第1维和第4维特征,仅保留第2维和第3维重要特征,从而提升了算法的运行速度和聚类的正确率.

现按文献[24]的方法初化簇类中心,并随机初始化特征权重,执行WKM、EWKM、WFCM、SCAD2、ESSC、FRFCM和本文算法,然后比较各算法在数据D_2上的聚类结果(见表4). 由于WKM、EWKM不能正确分配特征权重,聚类结果不理想,WKM各聚类指标的标准差较大,说明算法对初始化非常敏感;SCAD2和ESSC能给特征分配正确的权重(见表5),聚类结果较好;FRFCM同样不能产生正确的特征权重,但在这几个数据集上的聚类性能指标较高,其原因是对D_2归一化后,FRFCM算法分配给各维的特征权重差别不大(见图3),但若按$1/\sqrt{nd}$计算则得到的阈值为0.025,特征权重值均大于该阈值,算法保留了全部特征,并没有进行特征缩减.

表5 SCAD2和ESSC算法在归一化D_2上的特征分配

算法	第1维特征	第2维特征	第3维特征	第4维特征
SCAD2	0.246	0.253	0.265	0.235
	0.233	0.264	0.275	0.227
ESSC	0.198	0.321	0.329	0.152
	0.133	0.367	0.380	0.120

3.1.2 实验2(各算法在Dim 032～Dim 1 024六个人工数据集上的实验对比)

实验2所使用的6个数据集均含1 024个样本,共16类,每个类含64个样本,维数从32到1 024. 数据集可从文献[25]的网页下载获取. 这6个数据集的详细信息见表6. 表7为7个不同聚类算法在6个人工数据集上的聚类结果.

表6 高维空间的6个人工数据集

数据集	样本数	特征数	簇类数	每簇类样本数
Dim 032	1 024	32	16	64
Dim 064	1 024	64	16	64
Dim 128	1 024	128	16	64
Dim 256	1 024	256	16	64
Dim 512	1 024	512	16	64
Dim 1 024	1 024	1 024	16	64

本文算法、SCAD2、ESSC和FRFCM在这6个数据集上聚类平均准确率均为0.997,NMI指标的平均值均为1,ARI指标的均值均为1. 与WKM、EWKM和WFCM聚类算法相比,本文算法的聚类效果是最好的. 虽然FRFCM的表现力与本文算法相当,SCAD2、ESSC和FRFCM方法的聚类效果和本文算法相近或相同,但是这3个算法的运行时间相对较高. FRFCM算法能识别出的不重要特征较少,保留的特征较多. 在数据集Dim1 024上,FRFCM算法识别出523个不重要特征,算法迭代终止时保留了501个重要特征;而本文算法识别出700个不重要特征,仅保留了324个重要特征用于聚类. 可见,本文算法把数据中的不重要特征和冗余特征删除后,更有利于算法正确且快速地将数据划分成不同的簇类. 若采取一般的随机初始化方法初始化聚类中心,则本文算法在人工数据集上的聚类效果好于对比算法,在6个人工数据集上的3个指标AC、NMI和ARI的均值分别为0.999、1.000和0.999. ESSC算法的聚类效果次之,3个指标的均值分别为0.879、0.957和0.856. 本文算法的各指标分别比ESSC算法高出了13.65%、4.49%和16.70%. FRFCM算法的聚类性能最低,其原因是FRFCM算法对初始化敏感,导致其不能为特征正确分配权重,于是重要特征被删除,而保留不重要特征用于聚类,因此,FRFCM的聚类效果不理想.

表 7　各算法在 Dim 032 ~ Dim 1 024 数据集上的聚类结果

数据集维数	评价指标	WKM	EWKM	WFCM	SCAD2	ESSC	FRFCM	本文算法
Dim 032	AC	0.682 ± 0.083	0.933 ± 0.043	0.994 ± 0.021	0.995 ± 0.017	0.995 ± 0.019	**0.995 ± 0.017**	0.984 ± 0.036
	NMI	0.905 ± 0.031	0.982 ± 0.013	0.998 ± 0.006	0.999 ± 0.005	0.999 ± 0.005	**0.999 ± 0.005**	0.996 ± 0.010
	ARI	0.681 ± 0.093	0.932 ± 0.046	0.995 ± 0.020	0.995 ± 0.018	0.995 ± 0.019	**0.995 ± 0.018**	0.984 ± 0.035
Dim 064	AC	0.640 ± 0.083	0.938 ± 0.035	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	**1.000 ± 0.000**
	NMI	0.890 ± 0.032	0.983 ± 0.010	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	**1.000 ± 0.000**
	ARI	0.639 ± 0.100	0.938 ± 0.034	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	**1.000 ± 0.000**
Dim 128	AC	0.654 ± 0.082	0.948 ± 0.047	0.302 ± 0.081	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	**1.000 ± 0.000**
	NMI	0.895 ± 0.031	0.986 ± 0.013	0.683 ± 0.077	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	**1.000 ± 0.000**
	ARI	0.649 ± 0.092	0.946 ± 0.050	0.270 ± 0.091	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	**1.000 ± 0.000**
Dim 256	AC	0.606 ± 0.101	0.985 ± 0.028	0.263 ± 0.072	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	**1.000 ± 0.000**
	NMI	0.876 ± 0.041	0.996 ± 0.007	0.657 ± 0.063	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	**1.000 ± 0.000**
	ARI	0.605 ± 0.108	0.985 ± 0.028	0.237 ± 0.070	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	**1.000 ± 0.000**
Dim 512	AC	0.613 ± 0.066	0.977 ± 0.031	0.344 ± 0.072	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	**1.000 ± 0.000**
	NMI	0.883 ± 0.026	0.994 ± 0.008	0.720 ± 0.057	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	**1.000 ± 0.000**
	ARI	0.623 ± 0.066	0.976 ± 0.032	0.305 ± 0.071	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	**1.000 ± 0.000**
Dim 1 024	AC	0.612 ± 0.095	0.989 ± 0.025	0.460 ± 0.076	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	**1.000 ± 0.000**
	NMI	0.875 ± 0.044	0.997 ± 0.007	0.809 ± 0.045	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	**1.000 ± 0.000**
	ARI	0.598 ± 0.119	0.989 ± 0.025	0.457 ± 0.099	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	**1.000 ± 0.000**

3.1.3　实验 3 (各算法在 12 个真实数据集上的实验对比)

表 8 给出了 12 个真实数据集的数据信息.

表 8　12 个真实数据集的数据信息

数据集	样本数	特征数	簇类数
iris	150	4	3
mammographic	830	5	3
bupa	345	6	2
contraceptive	1 473	9	3
wine	178	13	3
waveform	800	21	3
wdbc_all	569	30	2
wpbc	198	32	2
sonar	208	60	2
ORL	400	2 576	40
SMK_CAN_1 987	187	19 993	2
GLI_85	85	22 283	2

表 9 列出了本实验中全部算法的聚类结果. 与其他方法相比, 本文算法的聚类效果优于其他 6 个对比算法, 在 9 个数据集上聚类的评价指标是最高的. 在 waveform 数据集上, 本文算法使用 4 个重要特征聚类, 准确率 AC = 0.511, 而 FRFCM 使用了 7 个特征, AC = 0.522; 在 wpbc 数据集上, 本文算法使用 11 个重要特征聚类, 准确率 AC = 0.576, 而 FRFCM 使用了 31 个特征, AC = 0.581. 在这两个数据集上, 本文算法使用较少的特征便可获得与 FRFCM 接相近的聚类效果, 从这个角度看, 本文算法优于 FRFCM 算法. 在 ORL 数据集上, 本文算法使用 215 个重要特征聚类, AC = 0.503, 而 EWKM 算法的准确率 AC = 0.600 是最高, 但它使用了全部的 2 567 维特征聚类, 算法消耗的时间较多. 虽然在这 3 个数据集上, 本文算法的聚类效果不是最好, 但使用较少的特征就可以得到与对比算法很接近的聚类效果, 这体现了本文算法的优势. 另外, 除了 ORL 数据集, 本文算法在其他数据集上聚类性能指标的标准差均为零, 说明本文算法具有很强的鲁棒性.

本文算法的时间复杂度取决于 3 个更新阶段: 1) 计算隶属度矩阵时间复杂度为 $O(nc^2d)$; 2) 计算簇类中心 V_k 需要 $O(nc)$; 3) 更新权值 w_j 需要 $O(ncd^2)$, 总的时间复杂度为 $O(nc^2d + ncd^2)$.

4　结　论

本文研究了模糊聚类中的特征缩减学习方法. 首先构造边缘峰度度量的特征重要性度量指标, 然后基于该指标提出了一个新的、鲁棒性的特征缩减模糊聚类算法. 经过与 WKM、EWKM、FCM、WFCM、SCAD2 和 FRFCM 算法的聚类性能进行比较, 表明本文算法对初始化具有较强的鲁棒性, 克服了特征加权聚类算法不能正确为特征分配权重的问题. 本文算法使用 MKM 指标来度量特征的重要性, 克服了 MVR 指标的不足; 而且在聚类过程中同时进行特征缩减, 提升了聚类的效果和聚类速度.

表 9 各算法在12个真实数据集上的聚类结果

数据集	评价指标	WKM	EWKM	WFCM	SCAD2	ESSC	FRFCM	本文算法
	AC	0.857±0.178	0.852±0.104	0.937±0.083	0.880±0.000	0.893±0.000	0.900±0.000	**0.967±0.000**
iris	NMI	0.759 ± 0.230	0.689 ± 0.095	0.850 ± 0.088	0.702 ± 0.000	0.729 ± 0.000	0.744 ± 0.000	**0.904 ± 0.000**
	ARI	0.681 ± 0.093	0.932 ± 0.046	0.995 ± 0.020	0.995 ± 0.018	0.995 ± 0.019	0.995 ± 0.018	**0.984 ± 0.035**
	AC	0.693 ± 0.163	0.516 ± 0.000	0.789 ± 0.000	0.795 ± 0.000	0.585 ± 0.059	0.792 ± 0.000	**0.797 ± 0.000**
mammographic	NMI	0.198 ± 0.177	0.016 ± 0.012	0.278 ± 0.000	0.285 ± 0.000	0.106 ± 0.064	0.278 ± 0.000	**0.288 ± 0.000**
	ARI	0.234 ± 0.219	0.000 ± 0.000	0.334 ± 0.000	0.348 ± 0.000	0.039 ± 0.035	0.339 ± 0.001	**0.350 ± 0.000**
	AC	0.548 ± 0.005	0.548 ± 0.003	0.548 ± 0.000	0.507 ± 0.000	0.522 ± 0.000	0.507 ± 0.000	**0.572 ± 0.000**
bupa	NMI	0.503 ± 0.002	0.503 ± 0.000	0.503 ± 0.000	0.499 ± 0.000	0.499 ± 0.000	0.499 ± 0.000	**0.507 ± 0.000**
	ARI	−0.006 ± 0.004	−0.006 ± 0.001	−0.007 ± 0.000	−0.007 ± 0.000	−0.008 ± 0.000	−0.007 ± 0.000	**0.007 ± 0.000**
	AC	0.405 ± 0.004	0.401 ± 0.018	0.425 ± 0.019	0.396 ± 0.000	0.402 ± 0.018	0.409 ± 0.005	**0.452 ± 0.000**
contraceptive	NMI	0.366 ± 0.000	0.015 ± 0.006	0.031 ± 0.013	0.028 ± 0.000	0.017 ± 0.005	0.038 ± 0.003	**0.044 ± 0.000**
	ARI	0.252 ± 0.000	−0.001 ± 0.013	0.017 ± 0.013	0.030 ± 0.000	0.005 ± 0.013	0.031 ± 0.001	**0.035 ± 0.000**
	AC	0.663 ± 0.036	0.685 ± 0.053	0.691 ± 0.000	0.691 ± 0.000	0.691 ± 0.000	0.685 ± 0.000	**0.695 ± 0.000**
wine	NMI	0.268 ± 0.046	0.321 ± 0.053	0.311 ± 0.000	0.322 ± 0.000	0.324 ± 0.000	0.308 ± 0.000	**0.324 ± 0.000**
	ARI	0.263 ± 0.053	0.317 ± 0.060	0.317 ± 0.000	0.323 ± 0.000	0.326 ± 0.000	0.312 ± 0.000	**0.330 ± 0.000**
	AC	**0.522 ± 0.004**	0.516 ± 0.004	0.516 ± 0.000	0.511 ± 0.000	0.514 ± 0.000	0.514 ± 0.000	0.511 ± 0.000
waveform	NMI	**0.477 ± 0.000**	0.627 ± 0.030	0.615 ± 0.000	0.615 ± 0.000	0.627 ± 0.000	0.609 ± 0.000	0.747 ± 0.000
	ARI	**0.252 ± 0.000**	0.250 ± 0.001	0.248 ± 0.000	0.239 ± 0.000	0.246 ± 0.000	0.241 ± 0.000	0.245 ± 0.000
	AC	0.852 ± 0.000	0.926 ± 0.006	0.921 ± 0.000	0.928 ± 0.000	0.930 ± 0.000	0.926 ± 0.000	**0.956 ± 0.000**
wdbc_all	NMI	0.477 ± 0.000	0.627 ± 0.030	0.615 ± 0.000	0.615 ± 0.000	0.627 ± 0.000	0.609 ± 0.000	**0.747 ± 0.000**
	ARI	0.486 ± 0.000	0.722 ± 0.021	0.706 ± 0.000	0.730 ± 0.000	0.736 ± 0.000	0.724 ± 0.000	**0.831 ± 0.000**
	AC	0.522 ± 0.004	0.516 ± 0.004	**0.516 ± 0.000**	0.511 ± 0.000	0.514 ± 0.000	0.514 ± 0.000	0.511 ± 0.000
wpbc	NMI	0.013 ± 0.011	0.024 ± 0.006	**0.027 ± 0.000**	0.015 ± 0.000	0.015 ± 0.000	0.021 ± 0.000	0.013 ± 0.000
	ARI	0.015 ± 0.018	0.027 ± 0.016	**0.035 ± 0.000**	0.020 ± 0.000	0.020 ± 0.000	0.022 ± 0.000	0.018 ± 0.000
	AC	0.582 ± 0.015	0.541 ± 0.025	0.536 ± 0.012	0.548 ± 0.001	0.548 ± 0.007	0.540 ± 0.003	**0.625 ± 0.000**
sonar	NMI	0.059 ± 0.021	0.024 ± 0.023	0.005 ± 0.002	0.007 ± 0.000	0.008 ± 0.002	0.005 ± 0.001	**0.049 ± 0.000**
	ARI	0.023 ± 0.008	0.004 ± 0.008	0.001 ± 0.003	0.004 ± 0.000	0.005 ± 0.003	0.002 ± 0.001	**0.058 ± 0.000**
	AC	0.490 ± 0.006	**0.600 ± 0.046**	0.367 ± 0.016	0.038 ± 0.002	0.112 ± 0.033	0.088 ± 0.623	0.503 ± 0.192
ORL	NMI	0.775 ± 0.030	**0.819 ± 0.018**	0.656 ± 0.010	0.361 ± 0.012	0.426 ± 0.025	0.354 ± 0.041	0.709 ± 0.008
	ARI	0.385 ± 0.056	**0.597 ± 0.044**	0.227 ± 0.013	0.054 ± 0.008	0.048 ± 0.009	0.040 ± 0.010	0.343 ± 0.015
	AC	0.537 ± 0.015	0.549 ± 0.015	0.551 ± 0.000	0.604 ± 0.000	0.604 ± 0.000	0.604 ± 0.000	**0.613 ± 0.000**
SMK_CAN_ 1987	NMI	0.004 ± 0.004	0.007 ± 0.004	0.007 ± 0.000	0.031 ± 0.000	0.031 ± 0.000	0.031 ± 0.000	**0.351 ± 0.000**
	ARI	0.002 ± 0.005	0.006 ± 0.006	0.006 ± 0.000	0.038 ± 0.000	0.038 ± 0.000	0.038 ± 0.000	**0.040 ± 0.000**
	AC	0.017 ± 0.018	0.672 ± 0.052	0.721 ± 0.040	0.719 ± 0.005	0.737 ± 0.006	0.729 ± 0.000	**0.729 ± 0.000**
GLI_85	NMI	0.021 ± 0.011	0.092 ± 0.076	0.203 ± 0.052	0.206 ± 0.018	0.188 ± 0.003	0.213 ± 0.000	**0.213 ± 0.000**
	ARI	0.076 ± 0.084	0.080 ± 0.075	0.193 ± 0.047	0.183 ± 0.009	0.216 ± 0.010	0.202 ± 0.000	**0.202 ± 0.000**

参考文献(References)

[1] 朱林, 王士同, 邓赵红. 改进模糊划分的FCM聚类算法的一般化研究[J]. 计算机研究与发展, 2009, 46(5): 814-822.
(Zhu L, Wang S T, Deng Z H. Research on generalized fuzzy c-means clustering algorithm with improved fuzzy partitions[J]. Computer Research and Development, 2009, 46(5): 814-822.)

[2] 张远鹏, 周洁, 邓赵红, 等. 代表点一致性约束的多视角模糊聚类算法[J]. 软件学报, 2019, 30(2): 282-301.
(Zhang Y P, Zhou J, Deng Z H, et al. Multi-view fuzzy clustering approach based on medoid invariant

constraint[J]. Journal of Software, 2019, 30(2): 282-301.)

[3] Deng Z H, Jiang Y Z, Chung F L, et al. Transfer prototype-based fuzzy clustering[J]. IEEE Transactions on Fuzzy Systems, 2016, 24(5): 1210-1232.

[4] Lloyd S P. Least squares quantization in PCM[J]. IEEE Transactions on Information Theory, 1982, 28: 129-137.

[5] 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.

[6] Bezdek J C. Pattern recognition with fuzzy objective function algorithms[M]. Norwell: Kluwer Academic Publishers, 1981.

[7] Zhang Y P, Chung F L, Wang S T. Fast exemplar-based clustering by gravity enrichment between data objects[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2020, 50(8): 2996-3009.

[8] Yang M S, Nataliani Y. A feature-reduction fuzzy clustering algorithm based on feature-weighted Entropy[J]. IEEE Transactions on Fuzzy Systems, 2018, 26(2): 817-835.

[9] Zhou J, Chen L, Chen C L P, et al. Fuzzy clustering with the entropy of attribute weights[J]. Neurocomputing, 2016, 198: 125-134.

[10] Guyon I, Gunn S, Ben-Hur A, et al. Result analysis of the NIPS 2003 feature aelection challenge[C]. Neural Information Processing Systems (NIPS). 2004: 545-552.

[11] Huang J Z, Ng M K, Rong H, et al. Automated variable weighting in k-means type clustering[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(5): 657-668.

[12] Jing L P, Ng M K, Huang J Z. An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data[J]. IEEE Transactions on Knowledge and Data Engineering, 2007, 19(8): 1026-1041.

[13] Amorim R C D, Mirkin B. Minkowski metric, feature weighting and anomalous cluster initializing in K-means clustering[J]. Pattern Recognition, 2012, 45(3): 1061-1075.

[14] Svetlova L, Mirkin B, Lei H. MFWK-means: Minkowski metric fuzzy weighted K-means for high dimensional data clustering[C]. Proceedings of the 14th IEEE International Conference on Information Reuse and Integration. San Francisco: IEEE, 2013: 692-699.

[15] Wang X Z, Wang Y D, Wang L J. Improving fuzzy c-means clustering based on feature weight learning[J]. Pattern Recognition Letters, 2004, 25(10): 1123-1132.

[16] Frigui H, Nasraoui O. Unsupervised learning of prototypes and attribute weights[J]. Pattern Recognition, 2004, 37(3): 567-581.

[17] Deng Z H, Choi K S, Chung F L, et al. Enhanced soft subspace clustering integrating within-cluster and between-cluster information[J]. Pattern Recognition, 2010, 43(3): 767-781.

[18] Yang M S, Sinaga K P. A feature-reduction multi-view k-means clustering algorithm[J]. IEEE Access, 2019, 7: 114472-114486.

[19] Hashemzadeh M, Golzari O A, Farajzadeh N. New fuzzy C-means clustering method based on feature-weight and cluster-weight learning[J]. Applied Soft Computing, 2019, 78: 324-345.

[20] Xing H J, Wang X Z, Ha M H. A comparative experimental study of feature-weight learning approaches[C]. Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics. Anchorage, 2011: 3500-3505.

[21] Cai D, He X, Han J. Document clustering using locality preserving indexing[J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(12): 1624-1637.

[22] Strehl A, Ghosh J. Cluster ensembles — A knowledge reuse frame-work for combining multiple partitions[J]. Journal of Machine Learning Research, 2003, 3: 583-617.

[23] Hubert L, Arabie P. Comparing partitions[J]. Journal of Classification, 1985, 2(1): 193-218.

[24] Arthur D, Vassilvitskii S. k-means++: The advantages of careful seeding[J]. Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms, 2007: 1027-1035.

[25] Speech and Image Processing Unit, School of Computing University of Eastern Finland. Clustering datasets[DB/OL]. http://cs.joensuu.fi/sipu/datasets/.

作者简介

潘兴广 (1979−), 男, 博士生, 从事机器学习、数据挖掘的研究, E-mail: 408206387@qq.com;

王士同 (1964−), 男, 教授, 博士生导师, 从事模式识别、人工智能等研究, E-mail: wxwangst@jiangnan.edu.cn.

(责任编辑：李君玲)