基于阴影集的截集式可能性C-均值聚类截集门限的选取

雒 僖 范九伦 于海燕 梁 丹
(西安邮电大学通信与信息工程学院 西安 710121)
(电子信息勘验应用技术公安部重点实验室 西安 710121)
(陕西省无线通信与信息处理技术国际合作研究中心 西安 710121)

摘 要 截集式可能性C-均值聚类算法通过引入截集门限,修改典型性值,克服了可能性C-均值聚类算法的最关键问题:一致性聚类。针对算法中截集门限的选取问题,采用阴影集理论,提出了一种新的截集门限的选取方法。该算法利用最优化方法为每一个类确定一个阴影集阈值,并将该阈值作为截集门限;通过分析该选取方法对典型性值和中心偏移量的影响来改进典型性值的修改方式。最后,通过人工数据分析了新的截集门限选取方式对聚类算法性能的影响,利用实际UCI数据分析算法的迭代次数和聚类正确率。实验结果表明,给出的截集门限选取方法能够有效减少迭代次数,提高聚类正确率。

关键词 可能性C-均值聚类,截集式可能性C-均值聚类,聚类核,截集门限,阴影集
中图法分类号 TP311.13 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.08.041

Selection of Cutset Threshold for Cutset-type Possibilistic C-means Clustering Based on Shadowed Set

LUO Xi FAN Jiu-lun YU Hai-yan LIANG Dan
(School of Communication and Information Engineering, Xi'an University of Posts and Telecommunications, Xi'an 710121, China)
(Key Laboratory Electronic Information Application Technology for Scene Investigation, Ministry of Public Security, Xi'an 710121, China)
(Internation Joint Resesrch Center for Wireless Communication and Information Processing, Xi'an 710121, China)

Abstract By introducing the cutset threshold and modifying the typicality, the cutset-type possibilistic C-means clustering algorithm overcomes the most critical problem (consistent clustering) of the possibilistic C-means clustering algorithm. Aiming at the parameter selection problem in the algorithm, this paper proposed a new method based on the shadowed set. This algorithm uses the optimization method to determine the threshold of the shadowed set for each cluster and takes this threshold as the cutset threshold. The modification method of the typicality is improved by analyzing the influence of the selection method on the typicality and the center deviation. Finally, the influence of the new parameter selection method on the performance of the clustering algorithm is analyzed by artificial dataset. The number of iterations and the clustering accuracy of the algorithm are analyzed through the UCI dataset. Experimental results show that the proposed method can effectively reduce the number of iterations and improve the accuracy of clustering.

Keywords Possibilistic C-means clustering, Cutset-type possibilistic C-means clustering, Cluster core, Cutset threshold, Shadowed sets

1 引言

模糊C-均值聚类(Fuzzy C-mans, FCM)算法[1]是一种常用的模糊聚类算法。FCM算法引入模糊集理论来定义模糊隶属度函数。FCM算法要求样本点到各个类的隶属度之和为1。受该约束条件的影响,FCM算法对噪声和奇异点比较敏感。为解决FCM算法对噪声敏感的问题,Krishnapuram等提出了可能性C-均值聚类(Possibilistic C-means, PCM)算法[2],放松了FCM算法中样本点到各个类的隶属度之和为1的约束条件。其中样本点到一个类的隶属度被解释为属于这

到稿日期:2018-07-09 返修日期:2018-11-11 本文受国家自然科学基金项目(61671377,61571361,61601362),西安邮电大学西邮新兴团队(xyt2016-01)资助。

雒 僖(1995-),女,硕士生,主要研究方向为模式识别、聚类分析,E-mail:1450085678@qq.com;范九伦(1964-),男,教授,博士生导师,主要研究方向为模式识别与图像处理、模糊信息处理、图像安全技术研究,E-mail:jiulunf@xupt.edu.cn(通信作者);于海燕(1982-),女,博士,副教授,主要研究方向为模式识别与图像处理;梁 丹(1993-),女,硕士生,主要研究方向为图像处理。

个类的可能性，并且从可能性理论角度构建目标函数。但是PCM算法存在一些问题：对初始值比较敏感，容易产生一致性聚类[3]。

目前，针对PCM算法一致性聚类的缺点，研究人员提出了一些改进的PCM算法。Timm等将各类中心之间距离的反函数引入PCM的目标函数，提出了两种改进的可能性聚类算法[4]。Ferraro等将类间距离引入到目标函数中，提出了一种可能性K均值算法[5]。Pal等将FCM算法与PCM算法相结合，提出了可能性模糊C-均值聚类(Possibilistic Fuzzy C-means，PFCM)算法[6]。Askari等[7]和Sarkar等[8]分别对PFCM算法进行了改进。Xie等提出了一种改进的可能性C-均值(Enhanced Possibilistic C-means，EPCM)算法[9]，该算法通过将数据集划分为两部分(主要部分和辅助部分)来构造具有新约束的目标函数。Hamasuna等通过将L1正则化引入PCM的目标函数，提出了一种Crisp Possibilistic Clustering算法[10]。Xenaki等通过将稀疏理论引入可能性聚类，提出了一种基于稀疏理论的可能性聚类算法[11-12]。

文献[13]提出了截集式可能性C-均值聚类(Cutset-type Possibilistic C-means，C-PCM)算法。该算法将截集的概念引入到PCM算法中，利用截集的概念为每一类产生聚类核，通过对聚类核中样本的隶属度进行修改，将类间关系引入到可能性聚类中，以解决PCM一致性聚类的问题。

在C-PCM算法中，截集门限是一个重要参数，它的选取方法直接影响到聚类的效果。Pedrycz[14-15]提出了阴影集的概念，该概念已被成功地应用于无监督学习[16-18]。因此，本文利用阴影集理论，给出了一种新的截集门限的选择方法。数据实验说明了所提方法的有效性。

2 可能性C-均值聚类与截集式可能性C-均值聚类算法

由于FCM算法对噪声点或奇异点比较敏感，Krishnapuram和Keller提出了PCM算法。PCM算法放松了对隶属度的约束，解决了FCM对噪声的敏感问题。PCM的目标函数定义为：

$$J_{\mathrm{PCM}}(\boldsymbol{T},\boldsymbol{V})=\sum_{i=1}^{c}\sum_{j=1}^{n}t_{ij}^{m}d^{2}(x_j,v_i)+\sum_{i=1}^{c}\eta_i\sum_{j=1}^{n}(1-t_{ij})^{m} \tag{1}$$

约束条件为：

$$t_{ij}\in[0,1],0\leqslant\sum_{j=1}^{n}t_{ij}\leqslant n \tag{2}$$

其中，$\boldsymbol{T}=\{t_{ij}\}_{c\times n}$，$i=1,\cdots,c$，$j=1,\cdots,n$为可能性划分矩阵，$t_{ij}$表示第$j$个样本对第$i$类的典型值。$\eta_i$为惩罚因子，是一个正数，计算公式为：

$$\eta_i=K\frac{\sum_{j=1}^{n}t_{ij}^{m}d^{2}(x_j,v_i)}{\sum_{j=1}^{n}t_{ij}^{m}} \tag{3}$$

其中，K一般取1。典型值和聚类中心的更新公式为：

$$t_{ij}=\left(1+\left(\frac{d^{2}(x_j,v_i)}{\eta_i}\right)^{\frac{1}{m-1}}\right)^{-1} \tag{4}$$

$$v_i=\frac{\sum_{j=1}^{n}t_{ij}^{m}x_j}{\sum_{j=1}^{n}t_{ij}^{m}} \tag{5}$$

PCM放松了隶属度和为1的约束条件，由式(4)可知，样本到某一类的典型值只与该类有关，与其他类无关。这样会使样本点逼近若干个簇，导致一致性聚类的问题[3]。

PCM算法放松了对隶属度的约束条件，对噪声或者奇异点具有强鲁棒性，但是会导致聚类中心重合问题。因此，文献[13]提出了截集式可能性C-均值聚类算法(C-PCM)。该算法将截集的概念引入到可能性聚类中，利用β截集为每一类产生聚类核，对聚类核中样本的典型值进行修改，并将类间关系引入到可能性聚类中，从而克服了PCM一致性聚类的缺点。C-PCM的目标函数定义为：

$$J_{\text{C-PCM}}(\boldsymbol{T},\boldsymbol{V})=\sum_{i=1}^{c}\sum_{j=1}^{n}t_{ij}^{m}d^{2}(x_j,v_i)+\eta\sum_{i=1}^{c}\sum_{j=1}^{n}(1-t_{ij})^{m} \tag{6}$$

约束条件为：

$$\sum_{j=1}^{n}t_{ij}>0,i=1,\cdots,c \tag{7}$$

其中，η为给定的常数。典型值和聚类中心的更新公式与PCM一样。

在C-PCM中，典型值的修改方法如下：首先用PCM方法求出样本点的典型值，然后获得样本点x_j到所有类的最大典型值$t_{qj}=\max\limits_{1\leqslant i\leqslant c}t_{ij}$，若$t_{qj}$满足式(8)：

$$t_{qj}=\frac{1}{1+\left(\frac{d^{2}(x_j,v_q)}{\eta}\right)^{\frac{1}{m-1}}}>\beta \tag{8}$$

则认为样本点x_j属于第q类，不属于其他类。将样本点x_j到其他类的典型值改为0，并使t_{qj}保持不变，即使获胜典型性值保持不变，非获胜典型性值修改为0。具体修改方式为：

如果$t_{qj}>\beta$，那么：

$$\begin{cases}t_{ij}=0, & i=1,\cdots,c\ \&\ i\neq q\\ t_{ij}=t_{ij}, & i=q\end{cases} \tag{9}$$

如果$t_{ij}\leqslant\beta$，那么：

$$t_{ij}=t_{ij},i=1,\cdots,c \tag{10}$$

其中，β为截集门限。当$\beta=1$时，C-PCM算法就退化为PCM算法。截集门限是一个重要参数，可事先给定，也可以随着算法的迭代自适应选取。文献[13]给出了一种截集门限的自适应方法，β_i的计算公式如下：

$$\beta_i=k\cdot\frac{1}{1+\left(\frac{\hat{d}_i^{2}}{\eta}\right)^{\frac{1}{m-1}}} \tag{11}$$

其中，$\hat{d}_i=\dfrac{\sum_{j=1}^{n}t_{ij}^{m}d(x_j,v_i)}{\sum_{j=1}^{n}t_{ij}^{m}}$表示第$i$类的平均距离，$k\in[0,1]$控制聚类核半径的大小。

3 基于阴影集的参数选择

3.1 阴影集概念

阴影集是通过模糊集演化而得到的，它将传统的模糊集

映射到三值逻辑 0,1,[0,1]。阴影集如图 1 所示。给定模糊集 $f(x)$,如果 $f(x)>1-\alpha$,则令 $f(x)=1$;如果 $f(x)<\alpha$,则令 $f(x)=0$。在阴影集中,只有隶属度在区间$[\alpha,1-\alpha]$的元素保留了模糊性,其余的元素都被明确分类。

通过最小化目标函数可以求得最优的隶属度阈值 α。目标函数为:

$$V(\alpha)=|\Omega_1+\Omega_2-\Omega_3|,\alpha\in(0,0.5) \tag{12}$$

其中,$\Omega_1=\int_{x;f(x)\leqslant\alpha} f(x)\,\mathrm{d}x$ 表示隶属度减小的部分,$\Omega_2=\int_{x;f(x)\geqslant 1-\alpha}(1-f(x))\,\mathrm{d}x$ 表示隶属度增大的部分,$\Omega_3=\int_{x;\alpha<f(x)<1-\alpha}\mathrm{d}x$ 表示阴影集。

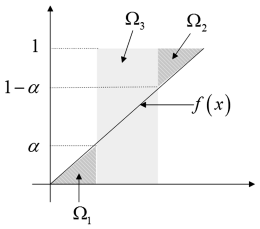


图 1　阴影集

Fig. 1　Shadowed sets

3.2　基于阴影集理论的参数选择方法

假定 $u_1,u_2,\cdots,u_n$ 是不同样本到第 i 类的隶属度值。其中,最小和最大隶属度值为 $u_{i\min}$ 和 $u_{i\max}$,目标函数公式可以修改为:

$$V(\alpha_i)=|\psi_1+\psi_2-\psi_3| \tag{13}$$

$$\alpha_{iopt}=\arg\min_{\alpha_i} V(\alpha_i) \tag{14}$$

其中,$\psi_1=\sum_{u_{ij}\leqslant\alpha_i}u_{ij}$ 表示隶属度减少的部分,$\psi_2=\sum_{u_{ij}\geqslant u_{i\max}-\alpha_i}(u_{i\max}-u_{ij})$表示隶属度增大的部分,$\psi_3=card(\{x_k\mid\alpha_i<u_{ij}<(u_{i\max}-\alpha_i)\})$表示阴影部分。$\alpha_i$ 的取值范围为$[u_{i\min},(u_{i\min}+u_{i\max})/2]$。

本文利用阴影集概念给出了一种新的截集门限选取方式。在阴影集中,通过最小化目标函数 $V(\alpha_i)$求得最优的隶属度阈值 α_i。若样本到这一类的隶属度大于 $1-\alpha_i$,则将隶属度修改为 1。这表明样本点属于这一类。相对应地,在 C-PCM 算法中,记样本点到所有类的最大典型值为 t_{qj},若 $t_{qj}>\beta$,则认为样本点 x_j 属于第 q 类,不属于其他类。我们可以通过阴影集求出每一类所对应的 $u_{i\max}-\alpha_i$,令 $\beta_i=u_{i\max}-\alpha_i$,即可得到每一类所对应的 β_i 值。选取的原理如图 2 所示。β_i 的计算公式如下:

$$\beta_i=u_{i\max}-\alpha_{iopt}=u_{i\max}-\arg\min_{\alpha_i}V(\alpha_i) \tag{15}$$

下面用数据集 $\boldsymbol{X}_{10}$ 来说明参数的选取过程。实验中,参数 $m=2$,$\eta=5$。初始聚类中心 $\boldsymbol{V}_0=\begin{bmatrix}1.294 & 0.292\\ -0.086 & -0.368\end{bmatrix}$。中心偏移量计算公式为 $\Delta v=\|\boldsymbol{v}-\boldsymbol{v}_{\mathrm{real}}\|_F=\sqrt{\sum_{i=1}^{c}\sum_{k=1}^{s}(\boldsymbol{v}(i,k)-\boldsymbol{v}_{\mathrm{real}}(i,k))^2}$,其中 $\boldsymbol{v}$ 表示算法得到的聚类中心,$\boldsymbol{v}_{\mathrm{real}}$表示真实的聚类中心,$\Delta v$ 表示中心的偏移量。实验中,$\boldsymbol{v}_{\mathrm{real}}=\begin{bmatrix}3.34 & 0\\ -3.34 & 0\end{bmatrix}$。图 3 给出了算法在迭代过程中的聚类中心。表 1 列出了算法在迭代过程中的典型值。

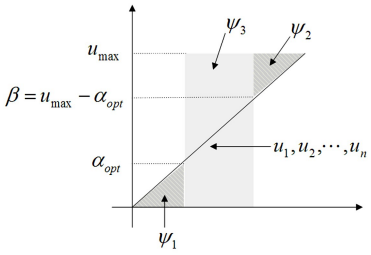
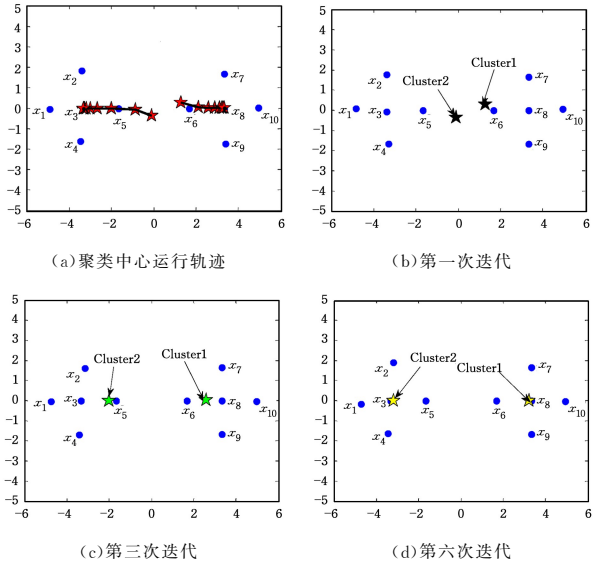


图 2　基于阴影集的参数 β 的选择

Fig. 2　Parameter selection based on shadowed sets



(a)聚类中心运行轨迹　(b)第一次迭代

(c)第三次迭代　(d)第六次迭代

图 3　算法在迭代过程中的聚类中心

Fig. 3　Cluster centers in iterative process of algorithm

1)基于阴影集的参数选取方法对典型值的影响

算法中的典型值代表了样本点到聚类中心的绝对距离。表 1 第一行表示第一次迭代得到的未被修正的典型值。根据式(15),第一次迭代得到的截集门限分别为 $\beta_1=0.5728$,$\beta_2=0.3647$。从表 1 中看出,所有的样本点到第一类的典型值中,只有样本点 $\boldsymbol{x}_6$ 的典型值大于 β_1。相对应地,从图 3(b)中可以看出,样本点 $\boldsymbol{x}_6$ 到第一类的聚类中心的距离比其他样本点到第一类的聚类中心的距离小。所有的样本点到第二类的典型值中,只有样本点 $\boldsymbol{x}_5$ 和 $\boldsymbol{x}_6$ 的典型值大于 β_2。同样地,样本点 $\boldsymbol{x}_5$ 和 $\boldsymbol{x}_6$ 到第二类的聚类中心的距离比其他样本点到第二类的聚类中心的距离小。表 1 第二行表示第一次迭代修改后的典型值。第三行表示第三次迭代得到的典型值。从表 1 中可以看出,样本点 $\boldsymbol{x}_6$,$\boldsymbol{x}_7$,$\boldsymbol{x}_8$,$\boldsymbol{x}_9$ 到第一类的典型值大于 β_1,样本点 $\boldsymbol{x}_3$ 和 $\boldsymbol{x}_5$ 到第二类的典型值大于 β_2。同样地,从图 3(c)中可以看出,$\boldsymbol{x}_6$,$\boldsymbol{x}_7$,$\boldsymbol{x}_8$,$\boldsymbol{x}_9$ 到第一类的聚类中心的距离比其他样本点小,$\boldsymbol{x}_3$ 和 $\boldsymbol{x}_5$ 到第二类的聚类中心的距离比其他样本点小。通过实验发现,阴影集可以反映每一类数据的分布情况。因此,本文可以引用阴影集理论选取每一类对应的截集门限。

表 1 算法在迭代过程中的得到的典型值

Table 1 Typicality values obtained from different iterations

		x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	β_i
第一次	T_1	0.1119	0.1762	0.1882	0.1649	0.3604	**0.9566**	0.4511	0.5393	0.3836	0.2657	0.5728
	T_2	0.1707	0.2532	0.3179	0.2893	**0.6540**	**0.6084**	0.2393	0.2964	0.2713	0.1613	0.3647
修改后	T_1	0.1119	0.1762	0.1882	0.1649	0	**0.9566**	0.4511	0.5393	0.3836	0.2657	—
	T_2	0.1707	0.2532	0.3179	0.2893	**0.6540**	0	0.2393	0.2964	0.2713	0.1613	—
第三次	T_1	0.0802	0.1171	0.1251	0.1168	0.2174	**0.8600**	**0.6010**	**0.8944**	**0.5925**	0.4589	0.5922
	T_2	0.3566	0.5191	**0.7347**	0.5231	**0.9791**	0.2711	0.1377	0.1494	0.1380	0.0927	0.6225
修改后	T_1	0.0802	0.1171	0	0.1168	0	**0.8600**	0.6010	0.8944	0.5925	0.4589	—
	T_2	0.3566	0.5191	**0.7347**	0.5231	**0.9791**	0	0	0	0	0.0927	—
第六次	T_1	0.0688	0.0982	0.1039	0.0982	0.1726	**0.6738**	**0.6412**	**0.9974**	**0.6405**	0.6137	0.6396
	T_2	0.6002	**0.6396**	**0.9946**	**0.6398**	**0.6883**	0.1756	0.0995	0.1054	0.0995	0.0696	0.6390
修改后	T_1	0.0688	0	0	0	0	**0.6738**	**0.6412**	**0.9974**	**0.6405**	0.6137	—
	T_2	0.6002	**0.6396**	**0.9946**	**0.6398**	**0.6883**	0	0	0	0	0.0696	—

2)基于阴影集的参数选取方法对中心偏移量的影响

将本文提出的自适应选取方法与固定值进行实验对比，中心偏移量的结果如表 2 所列。从表 2 可以看出，本文自适应选取方法得到的中心偏移量比 $\beta=0.6$ 时的中心偏移量大。具体原因如下：在第六次迭代时，x_1 到第二类的典型值是 0.6002，到第一类的典型值是 0.0688，两者之间的差值大于 0.5，但 x_1 到第二类的典型值小于第二类所对应的截集门限，因此典型值不进行修改。由式(5)可知，聚类中心是各个样本的加权平均值。因此，在计算第一类的聚类中心时，x_1 也会对其有影响。同理，样本点 x_{10} 到第一类的典型值与到第二类的典型值之间的差值也大于 0.5，且小于第一类所对应的截集门限。同样地，在计算第二类的聚类中心时，x_{10} 也会对其有影响。当固定值取 0.6 时，x_1 到第一类典型值就会被修改为 0，对第一类的聚类中心的计算不会有影响。同理，x_{10} 对第二类的聚类中心的计算不会有影响。通过实验发现，当迭代次数大于 6 时，x_1 到第二类的典型值小于第二类所对应的截集门限，x_{10} 到第一类的典型值小于第一类所对应的截集门限。因此，最终得到的中心偏移量要比取固定值时得到的中心偏移量大。

表 2 中心偏移量的实验结果

Table 2 Experimental results of center deviation

	固定选取方法		本文方法	
	$\beta=0.6$	$\beta=0.7$	旧的隶属度修改方法	新的隶属度修改方法
Δv	0.0064	0.3307	0.0451	0.0064

3)结合阴影集改进典型值的修改方式

根据基于阴影集的参数选取方法对典型值和中心偏移量的影响，本文通过阴影集理论自适应选取截集门限。但是在聚类效果上，本文方法得到的中心偏移量比取固定值时的略大一些，因此为减小中心偏移量，本文对典型值的修改方式进行了修正。修正方式如下：

如果 $t_{qj}-t_{ij}>0.5$，即样本点所对应的最大典型值与其他的典型值之间的差值大于 0.5，则认为样本点属于第 q 类，不属于其他类。因此样本点到其他类的典型值可以修改为 0。

由上述讨论可见，原 C-PCM 算法的典型值修改方法具有局限性，为此本文采用阴影集理论给出了一种新的修正方法。典型值通过以下步骤进行修改。假设 $t_{qj}=\max\limits_{1\leqslant i\leqslant c} t_{ij}$，如果 $t_{qj}>\beta_i$ 或者 $t_{qj}-t_{ij}>0.5$，那么：

$$\begin{cases} t_{ij}=0, & i=1,\cdots,c \ \&\ i\neq q \\ t_{ij}=t_{ij}, & i=q \end{cases} \tag{16}$$

否则：

$$t_{ij}=t_{ij},i=1,\cdots,c \tag{17}$$

本文算法的具体步骤如下。

Step1 给定聚类数目 c，模糊因子 m，惩罚因子 η。设置最大循环次数 $T_{\max}$ 和算法停止的阈值 ε。

Step2 初始化聚类中心 $\boldsymbol{V}^{(0)}$，令迭代次数 $l=0$。

Step3 用式(4)更新典型值 t_{ij}。

Step4 用式(15)计算自适应参数 β_i。

Step5 用以下方式修正典型值 t_{ij}：

令 $t_{qj}=\max\limits_{1\leqslant i\leqslant c} t_{ij}$，如果 $t_{qj}>\beta_i$ 或 $t_{qj}-t_{ij}>0.5$，$i\neq q$，则根据式(16)修正典型值 t_{ij}。

否则，根据式(17)修正典型值 t_{ij}。

Step6 用式(5)更新聚类中心 $\boldsymbol{V}^{(l+1)}$。

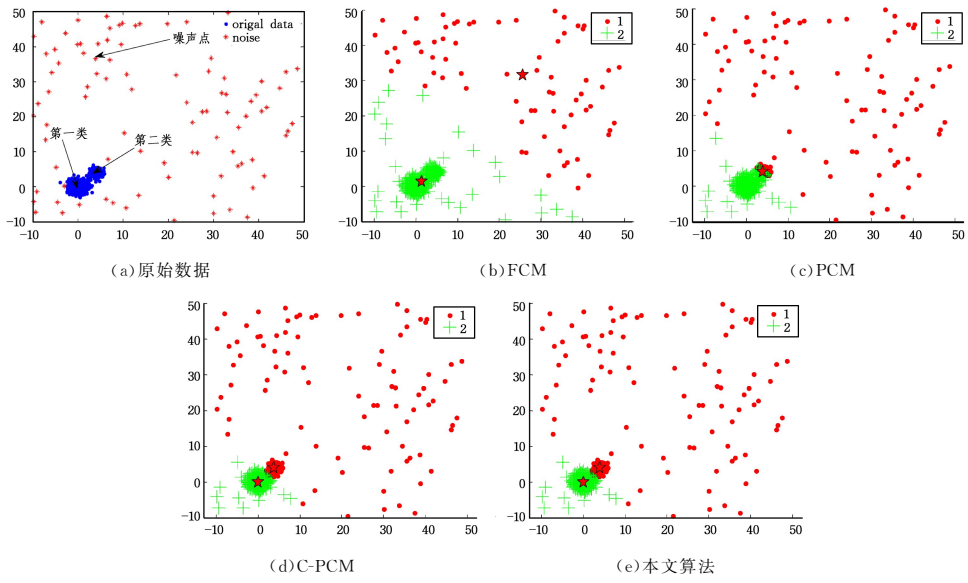
Step7 如果 $\|\boldsymbol{V}^{(l+1)}-\boldsymbol{V}^{(l)}\|<\varepsilon$ 或者迭代次数 $l>T_{\max}$，则算法结束；否则 $l=l+1$，返回到 Step3。

4 实验结果分析

本节将本文算法与 FCM，PCM 和 C-PCM 进行比较。实验数据由两部分组成：人工合成数据和 UCI 数据。在所有实验中，算法的参数均设置为：最大迭代次数 $T_{\max}=100$，算法停止的阈值 $\varepsilon=0.00001$。

4.1 对含噪人工数据集的处理

本节构造了一个球形数据集来分析本文算法的性能。数据集为高斯分布，共有两类：第一类由 1000 个数据点组成，第二类由 500 个数据点组成。其均值分别为 $\boldsymbol{\mu}_1=[0,0]^{\mathrm{T}}$，$\boldsymbol{\mu}_2=[4,4]^{\mathrm{T}}$；协方差矩阵分别为 $\boldsymbol{\Sigma}_1=\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$，$\boldsymbol{\Sigma}_2=\begin{bmatrix}0.5 & 0\\ 0 & 0.5\end{bmatrix}$。加入噪声点，噪声的分布为 $[0,50]\times[0,50]$。实验中参数设置为：$m=2$，$\eta=20$。图 4 给出了不同聚类算法对人造数据的聚类结果，图中☆表示各算法得到的聚类中心。当加入噪声时，FCM 算法受到噪声的影响很明显，PCM 算法容易导致聚类中心重叠，C-PCM 算法和所提出的算法对噪声具有较强的鲁棒性。



(a)原始数据　(b)FCM　(c)PCM

(d)C-PCM　(e)本文算法

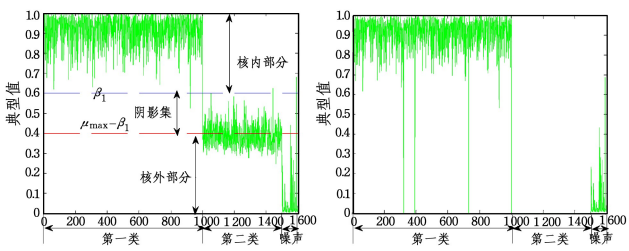
注:1 表示利用聚类算法得到的第一类数据;2 表示利用聚类算法得到的第二类数据

图 4　不同聚类算法对人造数据的聚类结果

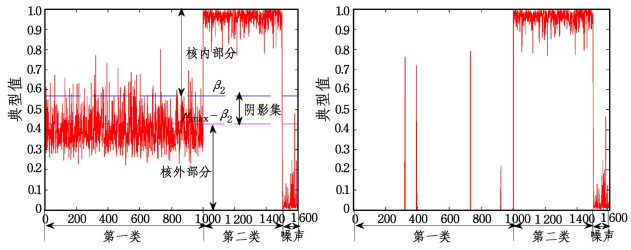
Fig. 4　Clustering results of artificial data by different clustering algorithms

4.2　参数 β 的性能分析

图 5(a)和图 5(c)分别表示最后一次迭代修改之前所有数据点到两类聚类中心的典型值。利用阴影集理论可以求出其对应的阈值 β_i 和 $u_{i\max}-\beta_i$。通过这两个阈值将隶属度分为 3 个部分:大于阈值 β_i 的部分表示属于该类;小于阈值 $u_{i\max}-\beta_i$ 的部分表示不属于这一类;中间部分表示阴影部分。若一个样本点对应的最大典型值大于其对应类的截集门限 β_i,则该样本点属于这一类。图 5(b)表示修改之后所有数据点到第一类聚类中心的典型值;图 5(d)表示修改之后所有数据点到第二类聚类中心的典型值。



(a)修改前数据点到第一类聚类中心的典型值　(b)修改后数据点到第一类聚类中心的典型值



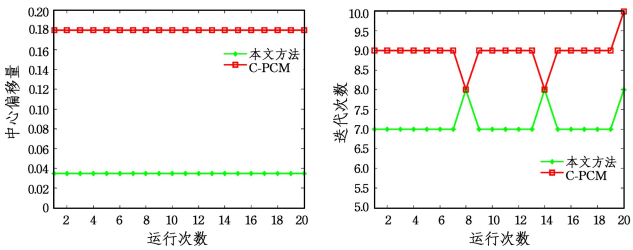
(c)修改前数据点到第二类聚类中心的典型值　(d)修改后数据点到第二类聚类中心的典型值

图 5　最后一次迭代得到的典型值

Fig. 5　Typicality values from last iteration

选取不同的聚类中心,进行 20 次实验。本文使用以下准则:通过中心偏移量、迭代次数来对 C-PCM 和本文算法的性能进行对比。

图 6 表示两种算法得到的中心偏移量值和迭代次数。从图 6 中可以看出,两种算法对初始值不敏感,并且本文所提出的方法得到的中心偏移量小于 C-PCM 算法中得到的中心偏移量,迭代次数也比 C-PCM 算法少一些。



(a)中心偏移量　(b)迭代次数

图 6　C-PCM 和本文算法的性能比较

Fig. 6　Comparison performance between C-PCM and proposed algorithm

4.3　UCI 数据集分析

本节将对 3 个 UCI 数据集进行测试,3 个 UCI 数据集分别为 Iris,Wine 和 Wdbc。Iris 数据集由 150 个样本组成,共 3 类;Wine 数据集由 178 个样本组成,共 3 类;Wdbc 数据集由 569 个样本组成,共 2 类。为了充分利用这些数据集的所有信息,对它们进行归一化处理。

$$\bar{x}_{ip}=\frac{x_{ip}-x_p}{\sqrt{\frac{1}{n-1}\sum_{r=1}^{n}(x_{rp}-x_p)^2}},\ x_p=\frac{1}{n}\sum_{r=1}^{n}x_{rp} \tag{18}$$

其中,$i=1,\cdots,n$,n 为样本的个数;$p=1,\cdots,l$,l 为样本的维数。

FCM,PCM,C-PCM 和本文算法对这些数据集的聚类结

果如表 3 所列。从表 3 可以看出,对于这些数据集,本文算法具有较小的迭代次数。对于 IRIS 数据集,C-PCM 算法和本文算法的聚类正确率是相同的。对于 Wine 数据集和 Wdbc 数据集,本文算法的正确率略优于 C-PCM 算法。

表 3 FCM,PCM,C-PCM 和本文算法对 UCI 数据集的聚类结果

Table 3 Clustering results of FCM,PCM,C-PCM and proposed algorithm on UCI datasets

数据集	FCM		PCM		C-PCM		本文算法	
	迭代次数	聚类正确率%	迭代次数	聚类正确率%	迭代次数	聚类正确率%	迭代次数	聚类正确率%
Iris($m=2,\eta=5$)	23	84.00	36	65.33	21	**84.67**	**14**	**84.67**
Wine($m=2,\eta=30$)	26	96.63	36	64.16	12	96.07	**9**	**97.19**
Wdbc($m=2,\eta=30$)	23	91.56	18	65.55	18	91.74	**15**	**92.09**

结束语 本文利用阴影集的优化理论,提出了一种新的截集门限的选取方式,并给出了一种新的典型性值的修正方法。通过对人工数据和 UCI 数据的实验分析,可以得出相对于 C-PCM 算法,本文算法表现出了更好的性能。但本文算法仍有一些需要改进的地方,惩罚因子的自适应选取和聚类数目的自适应确定是下一步的研究方向。

参 考 文 献

[1] BEZDEK J C. Pattern Recognition with fuzzy objective function algorithms[M]. New York:Plenum Press,1981.

[2] KRISHNAPURAM R,KELLER J M. A possibilistic approach to clustering[J]. IEEE Transactions on Fuzzy Systems,1993,1(2):98-110.

[3] KRISHNAPURAM R,KELLER J M. The possibilistic C-means algorithm:insights and recommendations[J]. IEEE Transactions on Fuzzy Systems,1996,4(3):385-393.

[4] TIMM H,BORGELT C,DÖRING C,et al. An extension to possibilistic fuzzy cluster analysis[J]. Fuzzy Sets and Systems,2004,147(1):3-16.

[5] FERRARO M B,GIORDANI P. On possibilistic clustering with repulsion constraints for imprecise data[J]. Information Sciences,2013,245:63-75.

[6] 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.

[7] ASKARI S,MONTAZERIN N,FAZEL Z M H,et al. Generalized entropy based possibilistic fuzzy C-Means for clustering noisy data and its convergence proof[J]. Neurocomputing,2017,219:186-202.

[8] SARKAR J P,SAHA I,MAULIK U. Rough Possibilistic Type-2 Fuzzy C-Means clustering for MR brain image segmentation [J]. Applied Soft Computing,2016,46:527-536.

[9] XIE Z P,WANG S T,CHUNG F L. An enhanced possibilistic c-means clustering algorithm EPCM[J]. Soft Computing,2008,12(6):593-611.

[10] HAMASUNA Y,ENDO Y. Sequential Extraction by Using Two Types of Crisp Possibilistic Clustering[C]// Proceedings ofIEEE International Conference on Systems,Man,and Cybernetics. New York:IEEE Press,2013:3505-3510.

[11] XENAKI S D,KOUTROUMBAS K D,RONTOGIANNIS A A. Sparsity-Aware Possibilistic Clustering Algorithms[J]. IEEE Transactions on Fuzzy Systems,2016,24(4):1611-1626.

[12] KOUTROUMBAS K D,XENAKI S D,RONTOGIANNIS A A. On the Convergence of the Sparse Possibilistic C-Means Algorithm[J]. IEEE Transactions on Fuzzy Systems,2018,26(1):324-337.

[13] YU H Y,FAN J L. Cutset-type possibilistic c-means clustering algorithm[J]. Applied Soft Computing,2018,64:401-422.

[14] PEDRYCZ W. Shadowed sets:representing and processing fuzzy sets[J]. IEEE Transactions on Systems,Man,and Cybernetics,Part B (Cybernetics),1998,28(1):103-109.

[15] PEDRYCZ W. From fuzzy sets to shadowed sets:interpretation and computing[J]. International Journal of Intelligent Systems,2009,24(1):48-61.

[16] ZHOU J,PEDRYCZ W,MIAO D Q. Shadowed sets in the characterization of rough-fuzzy clustering[J]. Pattern Recognition,2011,14(8):1738-1749.

[17] ZHANG K,KONG W R,LIU P P,et al. Partition region-based suppressed fuzzy C-means algorithm[J]. Journal of Systems Engineering and Electronics,2017,28(5):996-1008.

[18] WANG H L,SHE K,ZHOU M T. Shadowed Sets-based Rough Fuzzy Possibilistic C-means Clustering[J]. Computer Science,2013,40(1):191-194. (in Chinese)
汪海良,佘堃,周明天. 基于阴影集的粗糙模糊可能性 C 均值聚类算法[J]. 计算机科学,2013,40(1):191-194.