距离决策下的模糊聚类集成模型

费博雯*① 邱云飞② 刘万军② 刘大千③
①(辽宁工程技术大学工商管理学院 葫芦岛 125105)
②(辽宁工程技术大学软件学院 葫芦岛 125105)
③(辽宁工程技术大学电子与信息工程学院 葫芦岛 125105)

摘 要：模糊聚类是近年来使用的一类性能较为优越的聚类算法，但该类算法对初始聚类中心敏感且对边界样本的聚类结果不够准确。为了提高聚类准确性、稳定性，该文通过联合多个模糊聚类结果，提出一种距离决策下的模糊聚类集成模型。首先，利用模糊C均值(FCM)算法对数据样本进行多次聚类，得到相应的隶属度矩阵。然后，提出一种新的距离决策方法，充分利用得到的隶属度关系构建一个累积距离矩阵。最后，将距离矩阵引入密度峰值(DP)算法中，利用改进的DP算法进行聚类集成以获取最终聚类结果。在UCI机器学习库中选择9个数据集进行测试，实验结果表明，相比经典的聚类集成模型，该文提出的聚类集成模型效果更佳。

关键词：模糊聚类；集成模型；距离决策；隶属度矩阵；密度峰值算法
中图分类号：TP391 **文献标识码**：A **文章编号**：1009-5896(2018)08-1895-09
DOI: 10.11999/JEIT171065

Fuzzy Clustering Ensemble Model Based on Distance Decision

FEI Bowen① QIU Yunfei② LIU Wanjun② LIU Daqian③
①(*School of Business Administration*, *Liaoning Technical University*, *Huludao* 125105, *China*)
②(*School of Software*, *Liaoning Technical University*, *Huludao* 125105, *China*)
③(*School of Electronic and Information Engineering*, *Liaoning Technical University*, *Huludao* 125105, *China*)

Abstract: Fuzzy clustering is a kind of clustering algorithm which shows superior performance in recent years, however, the algorithm is sensitive to the initial cluster center and can not obtain accurate results of clustering for the boundary samples. In order to improve the accuracy and stability of clustering, this paper proposes a novel approach of fuzzy clustering ensemble model based on distance decision by combining multiple fuzzy clustering results. First of all, performing several times clustering for data samples by using FCM (Fuzzy C-Means), and corresponding membership matrices are obtained. Then, a new method of distance decision is proposed, a cumulative distance matrix is constructed by the membership matrices. Finally, the distance matrix is introduced into the Density Peaks (DP) algorithm, and the final results of clustering are obtained by using the improved DP algorithm for clustering ensemble. The results of the experiment show that the clustering ensemble model proposed in this paper is more effective than other classical clustering ensemble model on the 9 data sets in UCI machine learning database.

Key words: Fuzzy clustering; Ensemble model; Distance decision; Membership matrices; Density Peaks (DP) algorithm

1 引言

聚类可以简单概述为将数据集中的样本对象分成若干类簇集合，使得在同一类簇中的样本具有较高的相似性，不同类簇中的样本差别较大。数据聚类是数据挖掘、智能数据处理、机器学习、模式识别等方面的研究热点之一[1–4]。传统意义的聚类把数据集中各个数据样本严格地划分到某个类簇中，例如K-means聚类等均属于硬划分方法，数据的判别隶属度“非0即1”。为解决聚类中的的硬划分问题，Ruspini首次引入模糊数学理论[5]，主要分为两种思路完成模糊化，第1种思路是在K-Means算法中加入基于模糊权重指数的隶属度函数来实现，第2种则是在K-Means算法中引入信息熵。基于上述

收稿日期：2017-11-15；改回日期：2018-05-09；网络出版：2018-06-07
*通信作者： 费博雯 feibowen2098@163.com
基金项目：国家自然科学基金青年科学基金(61401185)
Foundation Item: The Young Scientists Fund of the National Natural Science Foundation of China (61401185)

两种思路，相继出现了许多优秀的模糊聚类算法，其中最为成功的当属由Dunn提出并由Bezdek推广的模糊C均值(Fuzzy C-means, FCM)算法[6]，由于其计算简单、几何意义较为清晰，被广泛使用在数据聚类中，本文也以FCM算法为基础，构造模糊聚类集成模型。

由于不同的数据集结构差异性较大，单一的聚类方法只能在某些特定的数据集上取得理想的聚类结果，存在一定的局限性，因此本文引入聚类集成学习模型。聚类集成学习是指通过计算多次基聚类之间的结构相关性，集成这些基聚类结果获得一个联合的鲁棒聚类结果，提高基聚类算法的聚类准确率。现有的聚类集成模型有很多种，如：Strehl等人[7]将聚类集成问题视为基于共享信息的组合优化问题，并提出3种聚类集成模型，分别为：基于簇间相似性关系的CSP (Cluster-based Similarity Partitioning)、基于图划分的HGP (Hyper Graph Partitioning)和基于组合簇类的MCL(Meta-CLustering)。Goswami等人[8]提出基于遗传架构的聚类集成方法，用染色体表示基聚类结果，将相似度较大的染色体之间进行交叉迭代，获得最终的聚类集成最优解。Banerjee等人[9]针对无监督迭代期望最大化(EM)算法的初始化问题，提出了一种聚类集成方法，该算法在一定统计模型的基础上对数据进行拟合，较好地解决遥感多光谱卫星图像的无监督土地覆盖分类问题。Hao等人[10]提出了一种基于链接的相似性度量方法来优化数据聚类关联矩阵，根据所有基本聚类结果之间的聚类的相似度，获得精化的数据聚类关联矩阵。Zhong等人[11]构建基于数据点与基聚类的协关联矩阵，然后将其转化为基于路径的相似矩阵，并利用谱聚类进行聚类集成。褚睿鸿等人[12]提出一种新的聚类集成模型，利用密度峰值(Density Peaks, DP)算法计算基聚类之间的相关性，有效地提高了聚类结果的准确性。

在现有的集成学习模型研究的基础上，本文提出一种距离决策下的模糊聚类集成模型，该模型利用FCM算法对数据集进行m次聚类。然后提出一种距离决策方法，利用得到的m个隶属度矩阵之间的从属关系构建累积距离矩阵，并将距离矩阵引入密度峰值算法中，利用改进的DP算法进行集成以获取最终聚类结果。

本文的聚类集成模型的创新之处：(1)对数据集进行多次模糊聚类，并随机选取初始聚类中心，保证集成模型的多样性、稳定性。(2)利用模糊聚类隶属度关系构建累积距离矩阵，采用软划分的方式计算距离，防止边界样本的丢失，保证集成模型的准确性、全面性。(3)在集成聚类中引入改进的密度峰值算法，自动选取密度最大的数据点为聚类中心，保证集成模型的聚类精度。

2 先验知识

2.1 FCM算法

设数据集为$\boldsymbol{X}=\{x_i|i=1,2,\cdots,n\}$，若将各个数据样本划分到$C$个类簇内，簇中心表示为$c_k(k=1,2,\cdots,C)$，则$x_i$属于第$k$个类簇中的模糊隶属度为$\boldsymbol{U}=\{u_{ik}|i=1,2,\cdots,n;k=1,2,\cdots,C\}$，由此可定义FCM的目标函数为

$$\begin{aligned} &J_m(u,c)=\sum_{i=1}^{n}\sum_{k=1}^{C}(u_{ik})^m\left\|x_i-c_k\right\|^2,\\ &\text{s.t.}\quad \sum_{k=1}^{C}u_{ik}=1,i=1,2,\cdots,n \end{aligned} \tag{1}$$

其中，$m(1<m<+\infty)$为隶属度因子。为得到$\min J_m(u,c)$，簇中心$c_k(k=1,2,\cdots,C)$和模糊隶属度u_{ik}需满足式(2)和式(3)的更新条件：

$$c_k=\frac{\sum_{i=1}^{n}u_{ik}^m x_i}{\sum_{i=1}^{n}u_{ik}^m},\ k=1,2,\cdots,C \tag{2}$$

$$u_{ik}=\frac{1}{\sum_{j=1}^{C}\left(\frac{d_{ik}}{d_{ij}}\right)^{1/(m-1)}},\ k=1,2,\cdots,C,i=1,2,\cdots,n \tag{3}$$

其中，$d_{ik}=\left\|x_i-c_k\right\|^2$，$d_{ij}=\left\|x_i-c_j\right\|^2$。由上述两个必要的更新条件可知FCM算法是一个简单的迭代过程，算法随机选取范围在(0,1)之间的任意数对模糊隶属度$\boldsymbol{U}$进行初始化，通过式(2)计算C个类簇的中心，然后利用式(1)计算FCM的目标函数，若$J_m(u,c)$小于设定阈值，或相对上次计算的目标函数的变化量ΔJ_m小于阈值，则迭代终止。

2.2 密度峰值算法

密度峰值(Density Peaks, DP)算法是Rodriguez等人[13]2014年在Science上提出的，其结构清晰、计算简单、聚类效果好，为聚类算法提供了一个全新的设计理念。DP的核心主要为聚类中心的设计，具有如下特点：(1)聚类中心被密度均小于它的相邻样本包围；(2)与其他密度更大的样本点之间的距离更大。

设数据集为$\boldsymbol{X}=\{x_i|i=1,2,\cdots,n\}$，对于各个样本点$x_i$均需要计算它的局部密度$\rho_i$和与更高局部密度的点的距离$\delta_i$，其中，$\rho_i$, δ_i均由任意x_i与x_j之间

的距离d_{ij}决定的，因此如何确定由d_{ij}构成的矩阵对于DP算法十分重要，本文将在下一节中对提出的集成模型构造的距离矩阵给予详细介绍。

假设数据样本x_i与x_j之间的距离d_{ij}已知，则局部密度ρ_i为

$$\rho_i=\sum_{j=1,j\neq i}^{n}\exp\left(-\frac{d_{ij}^2}{d_c^2}\right) \tag{4}$$

其中，$d_c(d_c>0)$表示截断距离，需用户事先指定。与更高局部密度的点的距离δ_i为

$$\delta_i=\begin{cases}\min\limits_{j\in I_{\boldsymbol{X}}^i}\{d_{ij}\}, & I_{\boldsymbol{X}}^i\neq\varnothing\\ \max\limits_{j\in I_{\boldsymbol{X}}^i}\{d_{ij}\}, & I_{\boldsymbol{X}}^i=\varnothing\end{cases} \tag{5}$$

其中，$I_{\boldsymbol{X}}^i$为数据集的指标集，具体表示为

$$I_{\boldsymbol{X}}^i=\{k\in I_{\boldsymbol{X}}:\rho_k>\rho_i\} \tag{6}$$

由此，对于数据集$\boldsymbol{X}$中的每一个数据样本x_i都可以得到相应的ρ_i和δ_i，当x_i同时拥有相对较高的ρ_i值和δ_i值时，则将该数据样本视为类中心点，然后将剩余的各个数据样本划分到有高密度的最近类簇中。

3　基于距离决策下的FCM集成模型

3.1　集成模型的组成

本文设计的集成模型由3部分组成，分别为：基聚类结果的获取、累积距离矩阵的构建以及基于DP的聚类集成。

3.1.1 基聚类结果的获取　本文提出的集成模型将单个FCM算法视为一个基聚类成员。由于初始聚类中心选择的不同，使得FCM每次运行都会产生差异的聚类结果，因此构造有效的集成模型可以较好地利用这一缺点，保证集成的多样性。随机选取初始聚类中心，通过m次重复运行FCM算法，得到相应的m个隶属度矩阵$\boldsymbol{U}=[\boldsymbol{U}_1,\boldsymbol{U}_2,\cdots,\boldsymbol{U}_m]$，$\boldsymbol{U}$为$n\times(m\times k)$维矩阵，$n$为数据样本个数，$k$为聚类个数。矩阵每一行表示每个数据样本在$m$次FCM聚类中的隶属度，矩阵的每$k$列表示一次基聚类的计算结果。以wine数据集为例进行隶属度矩阵$\boldsymbol{U}$的计算(其中n=178, k=3)，具体表现形式如式(7)，式中仅给出第1行、第2行以及第178行的具体值，其余省略。

$$\boldsymbol{U}=\left\{\begin{bmatrix}0.06 & 0.91 & 0.03\\ 0.72 & 0.09 & 0.19\\ \vdots & \vdots & \vdots\\ 0.55 & 0.07 & 0.38\end{bmatrix}_{n\times k}\cdots\begin{bmatrix}0.12 & 0.75 & 0.13\\ 0.69 & 0.07 & 0.24\\ \vdots & \vdots & \vdots\\ 0.63 & 0.06 & 0.31\end{bmatrix}_{n\times k}\right\}_{n\times(m\times k)},$$

$$\text{s.t.}\quad\sum_{j=1}^{k}\boldsymbol{U}_{ij}=1,\quad i=1,2,\cdots,n \tag{7}$$

3.1.2 累积距离矩阵的构建　在构建距离矩阵过程中，充分利用获取的隶属度信息，本文提出一种距离决策方法，采用软划分方式，使得构建的距离矩阵更为合理，并且在后续的密度集成聚类过程中有效利用样本间的距离进行密度分析、聚类。

在完成基聚类后，利用每次得到的隶属度矩阵$\boldsymbol{U}_j$, j=1,2,$\cdots$,m计算对应的最大隶属类信息矩阵$\boldsymbol{L}_j$，即

$$\boldsymbol{L}_j=[l_1^j,l_2^j,\cdots,l_n^j],\quad\text{s.t.}\quad l_i^j=\arg\max_{1\le k\le C}\{u_{ik}^j\} \tag{8}$$

其中，$\arg\max\{u_{ik}^j\}$表示最大隶属度所在的类编号，C为聚类数，对应隶属度矩阵$\boldsymbol{U}$可得到信息矩阵$\boldsymbol{L}=[\boldsymbol{L}_1,\boldsymbol{L}_2,\cdots,\boldsymbol{L}_m]$。下面以单个隶属度矩阵$\boldsymbol{U}_j$与信息矩阵$\boldsymbol{L}_j$为例进行距离矩阵的构建，任取数据集$\boldsymbol{X}=\{x_i|i=1,2,\cdots,n\}$的两个数据样本$x_a$和$x_b$，相应的隶属度和类信息分别为$(u_{ak}^j,l_a^j)k=1,2,\cdots,C$和$(u_{bk}^j,l_b^j)\quad k=1,2,\cdots,C$。由于数据样本$x_a$和$x_b$被分到某一类中是相互独立的，则可将其视为概率分布问题，设数据样本x_a概率表示为$p_{x_a}=u_{ak}^j$，同理，x_b的概率为$p_{x_b}=u_{bk}^j$。本文利用类信息l_a^j和l_b^j进行判断，若$l_a^j=l_b^j$，则x_a和x_b被归为同一类簇中的联合概率为$p_{(x_a,x_b)}=\sum_{k=1}^{C}u_{ak}^j\times u_{bk}^j$；反之，判定$x_a$与$x_b$属于不同类，则其联合概率为$p_{(x_a,x_b)}=\sum_{k=1}^{C}(1-u_{ak}^j)\times(1-u_{bk}^j)$。由此构造出隶属相似矩阵$\boldsymbol{U}_j'$，即

$$\boldsymbol{U}_j'(x_a,x_b)=\begin{cases}\sum\limits_{k=1}^{C}u_{ak}^j\times u_{bk}^j, & l_a^j=l_b^j\\ \sum\limits_{k=1}^{C}(1-u_{ak}^j)\times(1-u_{bk}^j), & l_a^j\neq l_b^j\end{cases} \tag{9}$$

按照上述方式集成全部隶属度矩阵，得到累积隶属相似矩阵$\boldsymbol{U}'$，即$\boldsymbol{U}'=\sum_{j=1}^{m}\boldsymbol{U}_j'$。在计算累积距离时，遵照数据样本间的相似度越大, 则表示距离越近, 因此利用如式(10)的转换关系：

$$\boldsymbol{D}=m\times\boldsymbol{I}-\boldsymbol{U}' \tag{10}$$

其中，$\boldsymbol{D}$表示累积距离矩阵，$\boldsymbol{I}$为全1矩阵。本文利

用m次FCM聚类隶属度关系构建累积距离矩阵，既保证了集成的全面性，又通过联合概率分布的方式计算距离，保证模型的准确性。

3.1.3 基于DP的聚类集成 本文在原有的DP算法上引入距离矩阵进行改进，经过改进之后，算法对基聚类的隶属度进行密度峰值检测，确定前C个具有高密度的数据样本点作为集成聚类中心。具体过程为：利用改进的DP算法进行聚类集成，将累积距离矩阵$\boldsymbol{D}$作为输入数据，计算出局部密度ρ_i与该数据样本到更高密度样本点的距离δ_i，利用式(11)中的γ_i选择前K个密度峰值点作为集成聚类中心，将其余的数据样本划分到相应类簇中，计算聚类的边界区域，排除光晕点的干扰，从而确定最终的聚类集成结果。

$$\gamma_i=\rho_i\times\delta_i,\qquad i=1,2,\cdots,n \tag{11}$$

需要指出的是截断距离d_c的确定，其决定着DP聚类算法的成败，若d_c取得过大，使得每个数据样本x_i的ρ_i加大以致于类簇之间的区别度不高，难以划分出准确的类别；若d_c取得过小，在同一类簇中的数据样本可能被“过划分”成多个类簇。文献[13]中，作者通过比例系数t确定截断距离d_c，将t的范围限定在1%～2%区间内，遵照原作者思想，本文设定t为2%。

3.2 集成模型的实现

本文提出的距离决策下的模糊集成聚类模型分为获取基聚类FCM的聚类结果、构建累积距离矩阵以及基于DP的聚类集成3部分，该模型的具体实现过程见表1，整体流程框架如图1所示。

4 实验结果与分析

4.1 数据集与评价标准

在UCI(University of California Irvine)库中选择9个数据集进行测试，基本信息见表2。目前已有许多指标用于评价集成模型的优劣，本文选用RI(Rand Index)[14]和NMI(Normalized Mutual Information)[15]两个标准评价各个聚类集成模型。

4.2 模型性能评估

计算机环境：CPU为Intel Core i7 3.40 GHz，内存为16 GB。运行系统：Windows10 64位。实验平台：Matlab 2016a。在实验测试过程中，本文利用FCM算法作为基聚类，通过10次随机选取聚类中心的方式运行FCM算法，获得选取相应的隶属度矩阵$\boldsymbol{U}=[\boldsymbol{U}_1,\boldsymbol{U}_2,\cdots,\boldsymbol{U}_{10}]$，并分别采用RI和NMI进行聚类评价。

4.2.1 m值的选取 为了说明基聚类运行次数m对本文集成模型的影响，分别取m=0, 5, 10, 15, 20，通过实验验证m值的合理性，在9组数据集上的平均RI和NMI值的比较结果如图2所示。

由图中可以看出，当m=0时，即采用单一的FCM算法得到的聚类结果较差，平均RI和NMI值为0.655和0.508，且算法的标准差较大，表明算法

表 1 算法实现流程

输入：实验数据集$\boldsymbol{X}=\{x_i|i=1,2,\cdots,n\}$，基聚类器运算次数$m$，实验数据集的总类簇数$C$

输出：数据集$\boldsymbol{X}=\{x_i|i=1,2,\cdots,n\}$的聚类集成标签Q

步骤 1 获取基聚类FCM的聚类结果：

(1) 判断基聚类器运算次数是否小于m；

(2) 利用范围在(0,1)之间的随机数初始化模糊隶属度$\boldsymbol{U}_j$, j=1,2,$\cdots$,m，满足式(1)的约束条件；

(3) 通过式(2)计算C个类簇的中心$c_k^j(k=1,2,\cdots,C)$；

(4) 通过式(1)计算FCM的目标函数，若$J_{hj}(u,c)$小于设定阈值，或相对上次计算的目标函数的变化量ΔJ_{hj}小于阈值，则迭代终止；

(5) 利用式(3)重新计算新的模糊隶属度$\boldsymbol{U}_j$，返回(3)；

(6) 保存模糊隶属度$\boldsymbol{U}_j$, $j=j+1$，返回(1)；

步骤 2 构建累积距离矩阵：

(1) 利用式(8)计算每次得到的隶属度矩阵$\boldsymbol{U}_j$, j=1,2,$\cdots$,m对应的最大隶属类信息矩阵$\boldsymbol{L}_j$；

(2) 利用式(9)计算单个隶属度矩阵$\boldsymbol{U}_j$与信息矩阵$\boldsymbol{L}_j$构造出的隶属相似矩阵$\boldsymbol{U}_j'$为例进行距离矩阵的构建；

(3) 重复执行m次(2)，得到累积隶属相似矩阵$\boldsymbol{U}'$；

(4) 利用式(10)构建累积距离矩阵$\boldsymbol{D}$；

步骤 3 基于DP的聚类集成：

(1) 利用步骤2得到的累积距离矩阵$\boldsymbol{D}$计算数据样本间的两两距离d_{ij}，并确定截断距离d_c；

(2) 按照式(4)和式(5)分别计算数据样本x_i的局部密度ρ_i和与更高局部密度的点的距离δ_i；

(3) 利用式(11)中的γ_i选择前K个密度峰值点作为集成聚类中心$\{c_k,k=1,2,\cdots,C\}$，对非数据中心的数据样本进行归类；

(4) 计算聚类的边界区域，排除光晕点的干扰。

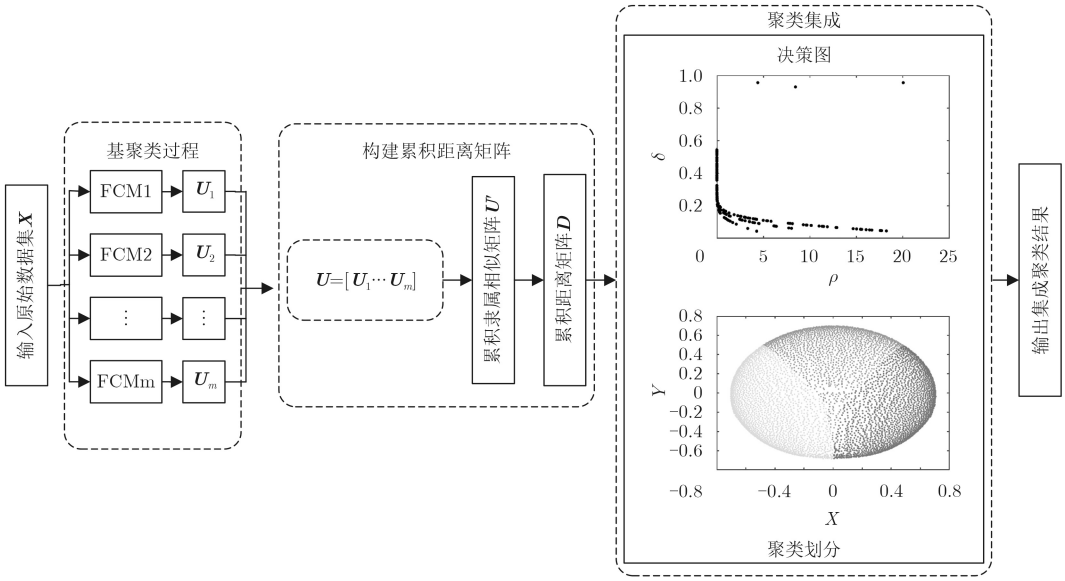


图 1 聚类集成模型整体框架图

表 2　实验数据集信息

序号	1	2	3	4	5	6	7	8	9
数据集	parkinsons	wdbc	ionosphere	german	iris	wine	waveform	vehicle	x8d5k
数据样本数	195	569	351	1000	150	178	5000	846	1000
属性	22	30	34	24	4	13	21	18	8
类别数	2	2	2	2	3	3	3	4	5

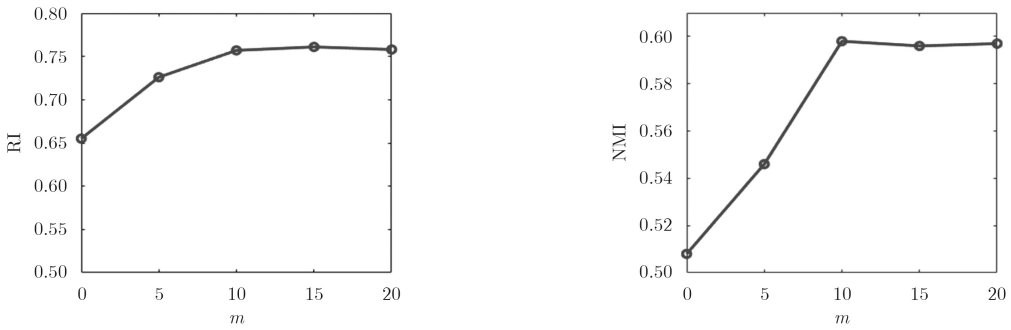
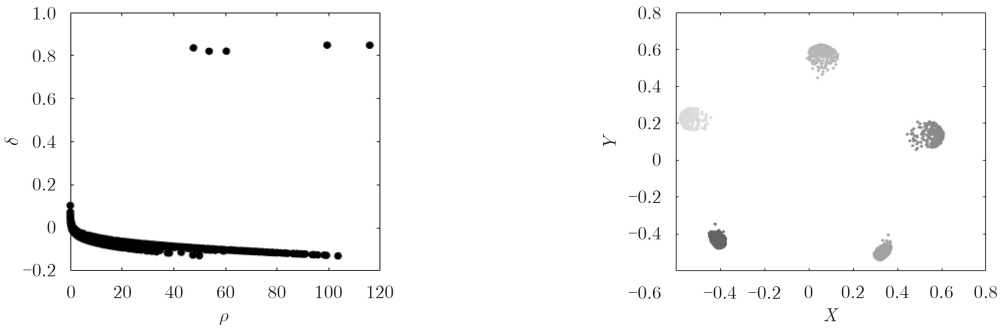


图 2 不同m值下的平均RI、NMI值的比较结果

的稳定性较差。当m=10时，该模型的聚类准确性有了较大的提高，平均RI和NMI值为0.757和0.598。需要指出的是，当$m>10$时，模型的聚类准确率(RI和NMI值)趋于稳定，但在构建距离矩阵$\boldsymbol{D}$更为复杂，增加了模型的计算复杂度。因此，本文为了兼顾模型的聚类准确性和计算复杂度，保证实验的统一性，将m值取为10更为合理。

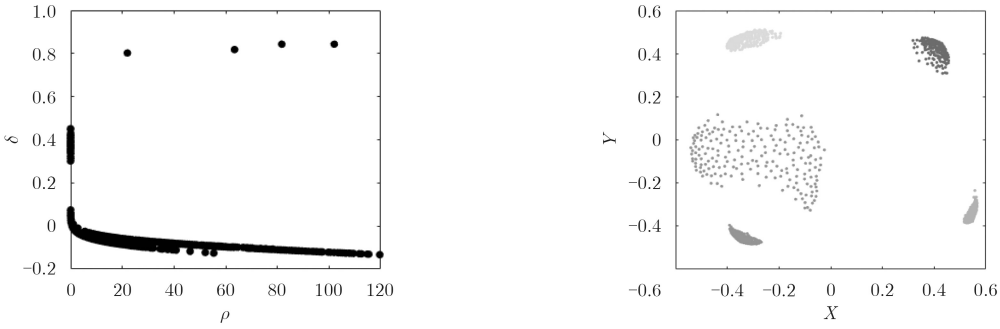
4.2.2 D的有效性验证　为了验证累积距离矩阵$\boldsymbol{D}$的有效性，本文主要对式(9)隶属相似矩阵$\boldsymbol{U}_j'$进行讨论，将$\boldsymbol{U}_j'$分为以下两种情况：一种为加入距离决策，按照式(9)的计算方式；另一种为只加入距离度量。以数据集x8d5k为例，基于不同距离度量的聚类模型聚类效果如图3所示。图3(a1)和图3(b1)表示基于距离的DP集成决策图，图3(a2)和图3(b2)表示2维仿真数据聚类结果。从图中可以看出，只加入距离度量的集成模型产生了误划分，将数据集分成4类，聚类效果不理想。而加入距离决策的集成模型聚类准确率较高，由于加入信息矩阵$\boldsymbol{L}$作决策，判断数据样本x_a和x_b是否被分到某一类中，保证距离度量的准确性，模型的RI为0.866, NMI为0.742。

4.2.3 对比实验分析　对比实验方法包括：单一的FCM(Fuzzy C-means)[6], CSP (Cluster-based Similarity Partitioning)[7], HGP (Hyper Graph Partitioning)[7], MCL (Meta-CLustering)[7]以及本文构造的集成模型。在9组数据集上的RI值比较见表3，在9组数据集上的NMI值比较见表4。



(a1) 决策图 (a2) 2 维仿真聚类结果

(a) 加入距离决策的集成模型



(b1) 决策图 (b2) 2 维仿真聚类结果

(b) 只加入距离度量的集成模型

图 3 基于不同距离度量的聚类模型聚类效果

表 3 5种聚类方法的RI比较结果

序号	FCM		CSP		HGP		MCL		本文	
	均值	标准差	均值	标准差	均值	标准差	均值	标准差	均值	标准差
1	0.579	0.068	0.638	0.012	0.607	0.007	0.613	0.016	0.646	0.003
2	0.723	0.047	0.833	0.013	0.812	0.019	0.841	0.021	0.852	0.010
3	0.563	0.057	0.628	0.015	0.617	0.017	0.613	0.007	0.662	0.014
4	0.625	0.039	0.729	0.016	0.746	0.006	0.731	0.004	0.758	0.009
5	0.786	0.034	0.832	0.023	0.822	0.015	0.852	0.016	0.887	0.006
6	0.732	0.041	0.856	0.011	0.859	0.017	0.826	0.008	0.863	0.013
7	0.628	0.037	0.646	0.007	0.638	0.014	0.637	0.011	0.641	0.015
8	0.547	0.022	0.627	0.021	0.593	0.014	0.616	0.005	0.635	0.007
9	0.715	0.033	0.816	0.004	0.849	0.005	0.834	0.010	0.866	0.008
平均值	0.655	0.042	0.734	0.014	0.727	0.013	0.729	0.011	0.757	0.009

表 4 5种聚类方法的NMI比较结果

序号	FCM		CSP		HGP		MCL		本文	
	均值	标准差	均值	标准差	均值	标准差	均值	标准差	均值	标准差
1	0.412	0.043	0.483	0.004	0.488	0.007	0.474	0.006	0.516	0.002
2	0.536	0.028	0.576	0.008	0.619	0.007	0.577	0.012	0.634	0.016
3	0.512	0.063	0.568	0.006	0.567	0.003	0.544	0.008	0.574	0.002
4	0.499	0.052	0.582	0.011	0.612	0.005	0.594	0.008	0.625	0.004
5	0.692	0.025	0.783	0.016	0.808	0.015	0.747	0.012	0.813	0.014
6	0.645	0.026	0.721	0.007	0.736	0.004	0.712	0.009	0.742	0.003
7	0.525	0.020	0.585	0.011	0.564	0.007	0.579	0.002	0.591	0.004
8	0.118	0.016	0.135	0.004	0.136	0.002	0.133	0.005	0.142	0.006
9	0.632	0.034	0.724	0.012	0.716	0.011	0.712	0.008	0.742	0.010
平均值	0.508	0.034	0.573	0.009	0.583	0.007	0.564	0.008	0.598	0.007

由表3和表4第1列的实验结果可以看出，FCM的RI和NMI值较低，且标准差较大，这说明FCM对初始聚类中心点的选取较为敏感，一旦选取不当，则造成最后的聚类准确率降低。与FCM相比，4种聚类集成模型(CSP, HGP, MCL，本文)都可以维持较高的聚类准确率，其中，RI平均值在0.7左右，NMI平均值在0.5左右，且模型的标准差较低，说明集成模型的稳定性较高，由于集成多次FCM基聚类算法，因此可以较好地解决FCM对初始聚类中心选取敏感的问题，特别是本文提出的DP集成模型，利用模糊聚类隶属度关系构建累积距离矩阵，采用软化分的方式计算累积距离，防止由于一次初始值而导致的误划分，保证了集成模型的聚类准确性。

与CSP, HGP, MCL这3种聚类集成相比，本文利用距离决策方法将FCM算法得到的隶属度关系转化为累积距离矩阵，判断各个数据样本之间的距离，并将其引入到DP算法中，利用各个数据样本间的密度峰值判断聚类集成中心点，从而获取最终聚类结果。根据表3和表4的实验结果可以看出，本文提出的基于改进DP算法的聚类集成模型的RI值和NMI值在9个UCI数据集上均能取得最优结果，在9组数据集上的RI平均值为0.712, NMI的平均值为0.522，也明显优于其他3种集成方法，说明本文模型的有效性。

4.3 基于DP的集成聚类性能分析

基于DP的聚类集成决策如图4所示。图4中，横、纵坐标分别表示数据样本x_i的局部密度ρ_i、与具有更高局部密度的数据样本点的距离δ_i, DP算法根据ρ_i和δ_i大小判断聚类中心点，然后将其他数据样本根据密度距离进行划分类簇。从图4可以看出，在9组数据集中，基于本文提出的距离决策方

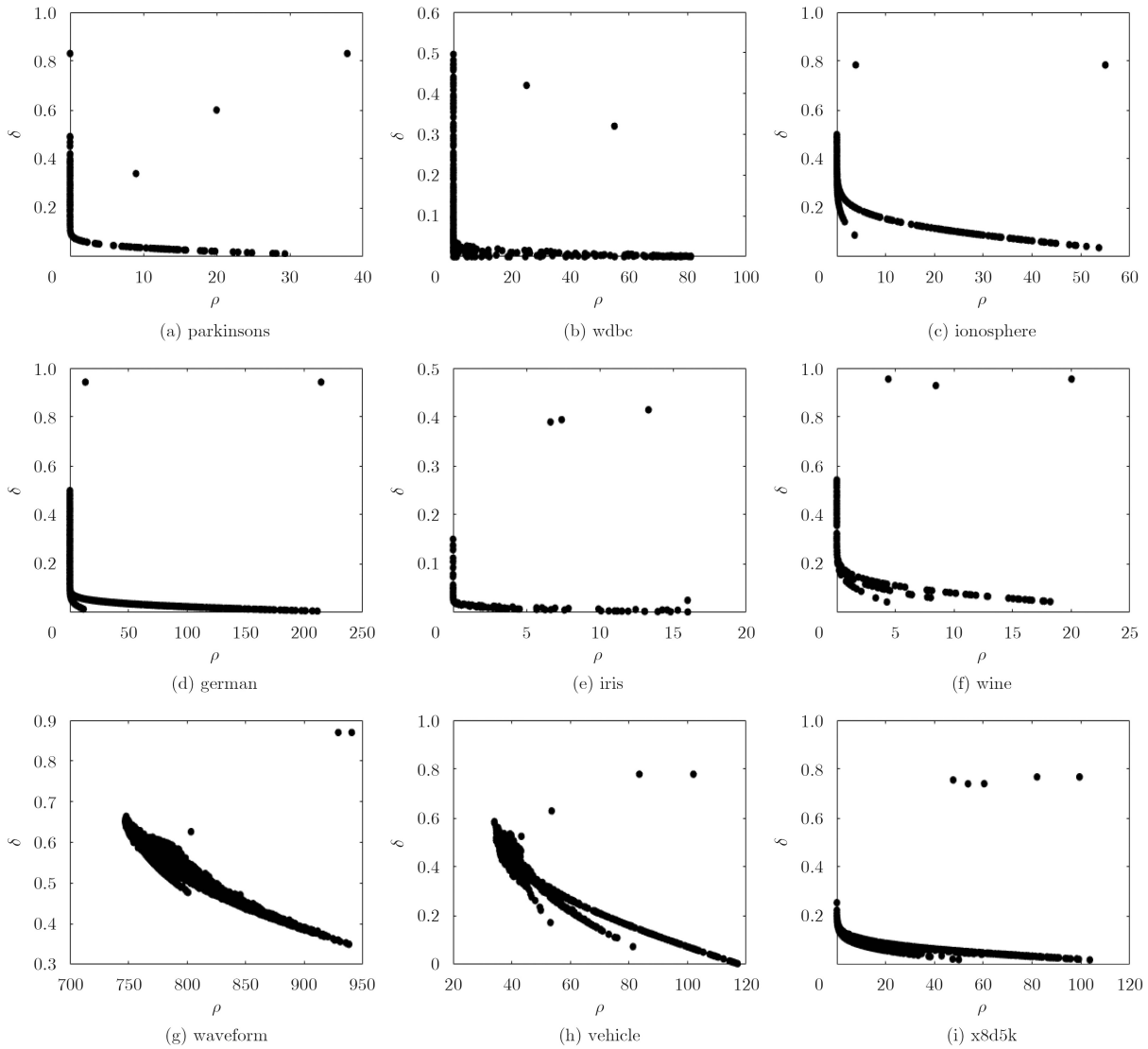


图 4 基于距离的DP集成决策图

法构建的距离矩阵可以较好地划分出聚类中心，如：在x8d5k数据集中，将具有较高局部密度和距离的样本点确定为中心，具体RI和NMI指标值见表3和表4。

在图4(a)的parkinsons数据集中，若将具有较高局部密度和距离的数据点确定为聚类中心，则该聚类中心为3个，而实际聚类类别为2。虽然聚类中心点c_3(在图2(a)中，ρ_i=37, δ_i=0.82的数据点)与其他样本的距离较大，但处理剩余数据样本难免会发生误划分，导致聚类准确率降低，因此在今后的工作中，将重点对DP算法的距离矩阵的构造以及聚类中心的选取方式做进一步的研究，提高模糊聚类集成模型的精度。

5 结束语

距离决策下的模糊聚类集成模型利用FCM 算法作为基聚类对数据样本进行多次聚类，对应得到多个隶属度矩阵，并利用提出的距离决策方法，将得到的隶属度矩阵转化为一个累积距离关系矩阵，判断各个数据样本之间的距离。然后将样本间的距离信息引入密度峰值算法中，利用改进的DP算法进行聚类集成以获取最终聚类结果。通过对UCI机器学习库中选择9组数据集进行实验对比分析，实验结果表明，与单一FCM聚类算法和同类集成模型比较，本文利用距离决策下的模糊聚类集成模型处理数据集，可以获得较高的聚类准确率，综合性能更佳。当今的智慧城市建设中，以移动设备和互联网为基础设施，通过获取丰富的城市数据资源，对城市建设、规划进行分析。本文设计的聚类集成模型能够为企业、商户合理地选择商业地址，根据大量移动设备用户的属性例如年龄、性别、学历、职位、收入等信息结合用户经常所处的位置信息为企业选择建址地点。

参 考 文 献

[1] MEI Jianping, WANG Yangtao, CHEN Lihui, *et al*. Large scale document categorization with fuzzy clustering[J]. *IEEE Transactions on Fuzzy Systems*, 2017, 25(5): 1239–1251. doi: 10.1109/TFUZZ.2016.2604009.

[2] 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.

[3] 叶茂, 刘文芬. 基于快速地标采样的大规模谱聚类算法[J]. 电子与信息学报, 2017, 39(2): 278–284. doi: 10.11999/JEIT160260.
YE Mao and LIU Wenfen. Large scale spectral clustering based on fast landmark sampling[J]. *Journal of Electronics & Information Technology*, 2017, 39(2): 278–284. doi: 10.11999/JEIT160260.

[4] 周林, 平西建, 徐森, 等. 基于谱聚类的聚类集成算法[J]. 自动化学报, 2012, 38(8): 1335–1342. doi: 10.3724/SP.J.1004.2012.01335.
ZHOU Lin, PING Xijian, XU Sen, *et al*. Cluster ensemble based on spectral clustering[J]. *Acta Automatica Sinica*, 2012, 38(8): 1335–1342. doi: 10.3724/SP.J.1004.2012.01335.

[5] 张敏, 于剑. 基于划分的模糊聚类算法[J]. 软件学报, 2004, 15(6): 858–868. doi: 10.13328/j.cnki.jos.2004.06.008.
ZHANG Min and YU Jian. Fuzzy partitional clustering algorithms[J]. *Journal of Software*, 2004, 15(6): 858–868. doi: 10.13328/j.cnki.jos.2004.06.008.

[6] BEZDEK J C, HATHAWAY R J, SABIN M J, *et al*. Convergence theory for fuzzy c-means: Counter-examples and repairs[J]. *IEEE Transaction on Systems, Man, and Cybernetics*, 1987, 17(5): 873–877. doi: 10.1109/TSMC.1987.6499296.

[7] STREHL A and GHOSH J. Cluster ensembles a knowledge reuse framework for combining multiple partitions[J]. *The Journal of Machine Learning Research*, 2003, 3(3): 583–617. doi: 10.1162/153244303321897735.

[8] GOSWAMI J P and MAHANTA A K. A genetic algorithm based ensemble approach for categorical data clustering[C]. Proceedings of the 2015 Annual IEEE India Conference (INDICON), New Delhi, India, 2015: 1–6.

[9] BANERJEE B, BOVOLO F, BHATTACHARYA A, *et al*. A new self-training-based unsupervised satellite image classification technique using cluster ensemble strategy[J]. *IEEE Geoscience and Remote Sensing Letters*, 2015, 12(4): 741–745. doi: 10.1109/LGRS.2014.2360833.

[10] HAO Zhifeng, WANG Lijuan, CAI Ruichu, *et al*. An improved clustering ensemble method based link analysis[J]. *World Wide Web*, 2015, 18(2): 185–195. doi: 10.1007/s11280-013-0208-6.

[11] ZHONG Caiming, YUE Xiaodong, ZHANG Zehua, *et al*. A clustering ensemble: two-level-refined co-association matrix with path-based transformation[J]. *Pattern Recognition*, 2015, 48(8): 2699–2709. doi: 10.1016/j.patcog.2015.02.014.

[12] 褚睿鸿, 王红军, 杨燕, 等. 基于密度峰值的聚类集成[J]. 自动化学报, 2016, 42(9): 1401–1412. doi: 10.16383/j.aas.2016.c150864.
CHU Ruihong, WANG Hongjun, YANG Yan, *et al*. Clustering ensemble based on density peaks[J]. *Acta Automatica Sinica*, 2016, 42(9): 1401–1412. doi: 10.16383/j.aas.2016.c150864.

[13] RODRIGUEZ A and LAIO A. Clustering by fast search and find of density peaks[J]. *Science*, 2014, 344(6191): 1492–1496. doi: 10.1126/science.1242072.

[14] ZHOU Zhihua and TANG Wei. Clusterer ensemble[J]. *Knowledge-Based Systems*, 2006, 19(1): 77–83. doi: 10.1016/j.knosys.2005.11.003.

[15] GAN Guojun, YANG Zijiang, and WU Jianhong. A genetic K-modes algorithm for clustering categorical data[J]. *Springer Berlin Heidelberg*, 2005, 36(2): 728–728. doi: 10.1007/11527503_23.

费博雯：女，1991年生，博士生，研究方向为数据挖掘与智能数据处理.

邱云飞：男，1976年生，博士，教授，主要研究方向为数据挖掘与智能数据处理.

刘万军：男，1959年生，硕士，教授，主要研究方向为图像与视觉信息计算、运动目标检测与跟踪.

刘大千：男，1992年生，博士生，研究方向为图像与视觉信息计算、运动目标检测与跟踪.