计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200400043

基于密度峰值的加权犹豫模糊聚类算法

张 煜 陆亿红 黄德才
浙江工业大学计算机科学与技术学院 杭州 310023
(onlyyousee6@163.com)

摘 要 由于人们对事物认知的局限性和信息的不确定性,在对决策问题进行聚类分析时,传统的模糊聚类不能有效解决实际场景中的决策问题,因此有学者提出了有关犹豫模糊集的聚类算法。现有的层次犹豫模糊 K 均值聚类算法没有利用数据集本身的信息来确定距离函数的权值,且簇中心的计算复杂度和空间复杂度都是指数级的,不适用于大数据环境。针对上述问题,文中提出了一种基于密度峰值思想的加权犹豫模糊聚类算法(WHFDP),首先给出了犹豫模糊元素集的补齐方法,并结合变异系数理论给出了新的距离函数权重计算公式,然后利用密度峰值选取簇中心,不仅降低了簇中心计算的复杂度,而且提高了对不同规模以及任意形状数据集的适应性,算法的时间复杂度和空间复杂度也降为多项式级,最后采用典型数据集进行仿真实验,证明了所提算法的有效性。

关键词:数据挖掘;聚类算法;犹豫模糊集;密度峰值;变异系数
中图法分类号 TP391

Weighted Hesitant Fuzzy Clustering Based on Density Peaks

ZHANG Yu,LU Yi-hong and HUANG De-cai
College of Computer Science and Technology,Zhejiang University of Technology,Hangzhou 310023,China

Abstract Due to cognitive limitations and the information uncertainty,traditional fuzzy clustering cannot effectively solve the decision-making problems in a real-life scenario when cluster analysis is carried out on the decision problem. Therefore,hesitant fuzzy sets(HFSs) clustering algorithms were proposed. The conception of hesitant fuzzy sets is evolved from fuzzy sets which are applied to fuzzy linguistic approach. The distance function of the hierarchical hesitant fuzzy K-means clustering algorithm has the same weight since the datasets information is seldom considered,and the computational complexity for computing the cluster center is exponential which is unavailable in the big data environment. In order to solve the above problems,this paper presents a novel clustering algorithm for hesitant fuzzy sets based on density peaks,called WHFDP. Firstly,a new method for extending the short hesitant fuzzy elements set to calculate the distance between two HFSs is proposed and a new formula for calculating the weight of distance function combined with the coefficient of variation is given. In addition,the computational complexity for computing the cluster center is reduced by using density peaks clustering method to select cluster center. Meanwhile,the adaptability to data sets with different sizes and arbitrary shapes is also improved. The time complexity and space complexity of the algorithm are reduced to polynomial level. Finally,typical data sets are used for simulation experiments,which prove the effectiveness of the new algorithm.

Keywords Data mining,Clustering algorithm,Hesitant fuzzy sets,Density peaks,Coefficient of variation

1 引言

聚类分析作为机器学习的重要组成部分,一直以来都是专家和学者们重点关注的一个热门领域。聚类指根据数据的属性或特征将相似的数据对象聚成一个簇,以便发现世界上已存在的但未被发现的知识、原理或规则。其在图像处理、数据挖掘、社交网络、模式识别等领域的科学实验和工程实践中有着重要的应用价值[1-3],特别是在数据挖掘中有着不可取代的重要地位。随着互联网络的信息化和高速发展,越来越多的数据被存储在电子存储器中,形成了规模庞大的数据集群,而我们很难直接从这些已有的数据中发现可以用来做支持决策的信息。对于那些没有标签的数据,通过聚类可以发现其中的关系,而这样的聚类分析往往能够帮助人们获得具有前瞻性价值的知识。

传统的聚类算法有很多种,如基于划分的 K-means 聚类算法、基于层次的 AGNES 聚类算法、基于密度的 DBSCAN

到稿日期:2020-04-10 返修日期:2020-08-08 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:浙江省公益技术应用项目(LGG19E090001)
This work was supported by the Zhejiang Public Welfare Technology Research Project(LGG19E090001).
通信作者:陆亿红(lyh@zjut.edu.cn)

聚类算法、基于网格的 STING 聚类算法以及基于模型的 SOM 聚类算法等[4]。这些聚类算法各有优点，同时也存在一些不足之处。例如，K-means 聚类算法依赖于初始聚类中心，容易受到噪声的影响且无法检测到非球形簇，导致聚类结果较差，而 DBSCAN 聚类算法能够识别不同形状的簇，但对参数选取敏感。因此，Strehl 等于 2002 年提出了基于图划分的聚类集成思想[5]。聚类集成可以选取多种聚类算法或采用不同特征子集等方法生成基聚类，再通过集成函数得出最终的聚类结果，从而提高聚类的精确率和鲁棒性。Fred 等提出了一种证据积累的聚类集成算法[6]，该算法将每个基聚类结果视为数据组织的独立证据构建共协矩阵，并采用凝聚层次算法得到最终的聚类结果。Iamon 等提出了基于链接的相似度评估聚类集成算法[7]，在计算共协矩阵时加入簇与簇之间的关系信息，并将图分割技术应用于加权二部图以得到最终的聚类结果。Tang 等提出了利用互信息加权投票选择基聚类成员[8]的聚类集成算法，Huang 等提出了超尺度谱聚类(U-SPEC)算法及其集成算法[9]，并针对稀疏亲和度子矩阵的构造提出了混合代表选择策略，为了增强算法的鲁棒性，集成多个 U-SPEC 作为基聚类并合并生成新的二分图，并通过对数据对象和基簇进行图节点划分得出聚类结果。

上述聚类方法适用于在确定性数据集下的问题分析，然而在现实生活中进行决策判断时，决策者给出某元素在集合中的隶属程度是犹豫不确定的，或是多个专家对其有多个不同的判定结果，Zadeh 最早提出了将模糊集的概念用于模糊语言信息中[10]，灵活地反映了决策者的真实偏好，其将语言中的词或句子转换为精确的数值，达到量化决策分析结果的目的，该方法已被广泛应用在医疗诊断[11]、机械引擎测评[12]、燃料选择[13]等决策问题中。Atanassov[14]，Mizumoto[15]和 Torra[16]等对模糊集进行了进一步的研究扩展，如提出了直觉模糊集[14]、2-型模糊集[15]和犹豫模糊集[16]等概念。Mizumoto 等[15]于 1976 年提出了 2-型模糊集，随后 Yager 等提出了模糊多集[17]，即集合中的一个元素可以出现多次且其隶属度值可以不同。Miyamoto 将模糊多集中的隶属度降序排列，提出了并集和交集的运算规则[18]。Atanassov 提出了直觉模糊集(Intuitionistic Fuzzy Set，IFS 或 A-IFS)[14]。直觉模糊集与模糊集的不同之处在于模糊集中的每个元素都有一个确定的隶属度，而直觉模糊集允许该隶属度包含一定的犹豫度，即直觉模糊集包含隶属度和非隶属度两部分。

上述模糊集不能完全适用于群体决策时产生的多个隶属度的情况，因此 Torra 等[16]于 2009 年提出了犹豫模糊集的概念，并将模糊集中的每个元素由一个确定的隶属度拓展到单位区间内的隶属度集合，可以在不确定的情况下最大程度地保证原始数据的完整性，实现了在群体决策分析时，不同专家根据自己的知识理论对同一个事物的某个特性可能存在不同的评判结果，或是一个专家对某个元素的隶属度无法给出确定结果而产生犹豫不定的情况。Yao 等[19]于 2018 年扩展了犹豫模糊集，令集合元素包含模糊性、直觉性和犹豫性的犹豫直觉模糊集。

本文在文献[20]的基础上提出了新的犹豫模糊聚类算法。其首先给出了更符合解决实际问题的犹豫模糊元素的扩充方法，并且结合数据集本身的信息提出了新的犹豫模糊对象加权距离的权重计算方法。其次，对于在聚类过程中指数级的时间复杂度和空间复杂度，本文结合密度峰值算法，优化了聚类簇中心的计算，避免了文献[20]中复杂度随数据集的增加呈指数级增长的问题，使犹豫模糊集的聚类算法能够适用于大规模的数据集，具有实际应用价值。本文第 2 节对犹豫模糊集的相关定义进行了简要的介绍；第 3 节提出了新的多属性决策的聚类算法和复杂度分析；第 4 节进行了实验分析；最后总结全文。

2 相关定义

2.1 犹豫模糊集的相关定义

由于人们对于事物特性的描述通常不能用一个精确的数字来衡量其隶属程度，为了量化隶属度，给出事物的某一特性所属的类，Zadeh[10]最早提出了模糊集的概念，定义如下。

定义 1[10] 设 X 为一个给定的非空对象集合，则定义模糊集 A 是由集合 X 通过函数 f 映射所得的一个新对象集合，即 $A\subset X$)，$f_A(x)\in[0,1]$的值表示集合 A 中元素 $x\in X$ 的隶属程度，简称隶属度。模糊集为集合中的元素赋予一个范围在 0～1 之间的隶属度，并给出了模糊集的并集、交集、补集等运算规则和性质。Atanassov[14]提出的直觉模糊集对元素的隶属度进行了扩展，定义如下。

定义 2[14] 设 X 为一个非空集合，则集合 X 的一个直觉模糊集 IFS 可表示为：

$$A=\{\langle x,\mu_A(x),\upsilon_A(x)\rangle\mid x\in X\}$$

其中，$\mu_A(x)$为元素 x 在集合 A 中的隶属度，$\upsilon_A(x)$为元素 x 在集合 A 中的确定不属于隶属度，$\mu_A(x)$和$\upsilon_A(x)$的取值均为$[0,1]$之间的实数且 $0\leqslant\mu_A(x)+\upsilon_A(x)\leqslant 1$。$\pi_A=1-\mu_A(x)-\upsilon_A(x)$表示元素 x 属于 A 的犹豫度，当 IFS 的犹豫度 $\pi_A=0$ 时，其与模糊集同理。

Torra 等[16]给出了最早犹豫模糊集的定义，定义如下。

定义 3[16] 设 $M=\{\mu_1,\mu_2,\cdots,\mu_k\}$是包含 k 个隶属度函数的集合，则与 M 相关的犹豫模糊集h_M可表示为：

$$h_M(x)=\{\mu_1(x),\mu_2(x),\cdots,\mu_k(x)\}$$

其中，元素 $x\in X$ 在 μ 函数的映射下得到关于 M 的犹豫模糊集。随后 Xia 等[21]总结出了犹豫模糊集 HFS 的如下表达式：

$$A=\{\langle x,h_A(x)\rangle\mid x\in X\}$$

其中，集合$h_A(x)$表示元素 x 属于集合 A 的所有可能的隶属度集合，集合中元素个数大于等于 1，其取值也均为$[0,1]$之间的实数。若集合$h_A(x)=\{0\}(x\in X)$，则定义该集合为空集；若集合$h_A(x)=\varnothing$，则表示该集合无实际意义。$h_A(x)$被称为一个犹豫模糊元素 HFE，是犹豫模糊集 HFS 的基本单位。犹豫模糊元素 HFE 设定其下界为$h^-(x)=\min h(x)$，上界为$h^+(x)=\max h(x)$。

定义 4[16] 任意给定犹豫模糊元 h，h_A和h_B，则：

1) $h^\lambda=\bigcup_{\gamma\in h}\{\gamma^\lambda\}$

2) $\lambda h=\bigcup_{\gamma\in h}\{1-(1-\gamma)^\lambda\}$

3) $h^c=\bigcup_{\gamma\in h}\{1-\gamma\}$

4) $h_A \cup h_B = \{h \in (h_A \cup h_B) \mid h \geqslant \max(h_A^-, h_B^-)\}$

5) $h_A \cap h_B = \{h \in (h_A \cup h_B) \mid h \leqslant \min(h_A^+, h_B^+)\}$

其中，λ 为常数，γ 为隶属度值，其取值均为[0,1]之间的实数。两个犹豫模糊元素的并集中的元素需满足最小值大于两个集合下界的最大值，交集的元素最大值需小于两个集合上界的最小值。

通常设数据集 $S=\{X_1, X_2, \cdots, X_n\}$，包含 n 个数据对象，数据对象 $X_i(i=1,2,\cdots,n)$ 有 d 个属性特征，因此数据对象也可以表示为 $X_i=\{x_{i1}, x_{i2}, \cdots, x_{id}\}(j=1,2,\cdots,d)$。为了便于后续的计算，我们重新定义了犹豫模糊元素和犹豫度等概念，并给出了相应的公式，定义如下。

定义 5　设包含 d 个属性的数据对象 X 的犹豫模糊集为 $H(X)=\{h(x_1), h(x_2), \cdots, h(x_d)\}$，$h(x)$ 为犹豫模糊元素集，即通过函数 h 使得数据对象 X 的属性 $x_j(j=1,2\cdots,d)$ 映射得到一个隶属度集合，且集合中元素均为 0～1 之间的实数，其中 $h(x_1)$ 表示数据对象 X 关于第一个属性特征的犹豫模糊元素集，则 $H(X)$ 为关于 X 的一个犹豫模糊集。

定义 6　设两个犹豫模糊集 $A=H(X_a)$ 和 $B=H(X_b)(a,b=1,2,\cdots,n)$，如果满足如下条件：1) $0 \leqslant dist(A,B) \leqslant 1$；2)当且仅当 $A=B$ 时，$dist(A,B)=0$；3) $dist(A,B)=dist(B,A)$；4) $dist(A,B) \leqslant dist(A,C)+dist(B,C)$。则其为犹豫模糊集 A 与 B 之间的距离 $dist(A,B)$。

为了计算犹豫模糊集之间的距离，首先定义犹豫度 $\lambda_x=|h(x)|$，表示在属性 x 的犹豫模糊元素集中隶属度值的个数，并且需要保证数据对象的每个属性的犹豫度相同，即 $\lambda_{x_{ai}}=\lambda_{x_{bj}}$，也可以用 $l(H(X_i))=\sum_{j=1}^{d}\lambda_{x_{ij}}(j=1,2,\cdots,d)$ 表示犹豫模糊集所有属性的犹豫度之和，即在计算犹豫模糊集距离时保证 $l(H(X_a))=l(H(X_b))$。当 $l(H(X_a))<l(H(X_b))$ 时，计算每个属性犹豫模糊元素集的犹豫度 $\lambda_{x_{ij}}$。若 $\lambda_{x_{ai}}<\lambda_{x_{bj}}$，则在 $h(x_{aj})$ 中补充 $\lambda_{x_{bj}}-\lambda_{x_{aj}}$ 个隶属度值，使得 $\lambda_{x_{aj}}=\lambda_{x_{bj}}$。

在对模糊集进行补充时，对于补充的隶属度值没有特定标准，原则上可以补充 0～1 之间的任意实数。通常情况下，为了保证原犹豫模糊元素集的稳定性，需要补充犹豫模糊元素集中出现次数最多的隶属度值，但是当集合中的元素个数较少时，每个元素极有可能只出现一次，导致无法筛选出高频元素，因此提出根据决策者的风险偏好来增加原犹豫模糊元素集中的属度值[20]，乐观的决策者期望得到理想的结果，因此增加原犹豫模糊元素集中的最大隶属度值，而悲观的决策者则期望得到不利的结果，可能增加原犹豫模糊元素集中的最小隶属度值甚至增加零。例如，有 $h(x_{aj})=\{0.2,0.5\}$，$h(x_{bj})=\{0.2,0.3,0.5,0.8\}$，为了计算集合的相似度，需要扩充 $h(x_{aj})$，乐观的决策者可扩充成 $h(x_{aj})=\{0.2,0.5,0.5,0.5\}$，悲观的决策者则扩充成 $h(x_{aj})=\{0.2,0.2,0.2,0.5\}$。

虽然通过增加不同的值来扩展较短的犹豫模糊元素集可能会产生不同的计算结果，但这是符合决策分析要求的，因为决策者的风险偏好会直接影响最终的决策结果，同样的情况也可以在许多现有的参考文献中找到[22-23]。在文献[20]中，Chen 等假设决策者都是悲观的，并用原犹豫模糊元素集的最小隶属度值来扩充集合。本文采用均值补充法，避免隶属度集合偏向某个专家的决策结果而导致其他专家的决策评判失去价值。在对犹豫模糊集进行补充后，根据用于计算确定数据集距离的汉明公式和欧氏距离公式，对于任意犹豫模糊集 A 和 B，给出了犹豫加权距离公式[24]。

$$d(A,B)=\left[\sum_{j=1}^{d}\omega_j\left(\frac{1}{\lambda_j}\sum_{\sigma=1}^{\lambda_j}|h_\sigma(x_{aj})-h_\sigma(x_{bj})|^\alpha\right)\right]^{\frac{1}{\alpha}} \tag{1}$$

其中，$\lambda_j=\max(\lambda_{x_{aj}}, \lambda_{x_{bj}})$，$h_\sigma(x)$ 表示 x 的犹豫模糊元素集中第 σ 位的值。当 $\alpha=1$ 时，其为犹豫加权汉明距离公式；当 $\alpha=2$ 时，其为欧氏加权距离公式。数据对象的属性权重 $\omega_j \in [0,1]$ 且 $\sum_{j=1}^{d}\omega_j=1$。决策者可以根据属性的重要程度对权重进行人工设置，在没有特定要求时也可以设权重 $\omega_j=1/d$，即每个属性具有相同的权重。

2.2　HHFK 算法

在聚类算法中，K-means 算法是无监督学习中被广泛应用的算法之一，但其也存在不足之处。由于算法中初始簇中心是随机选取的，当初始簇中心选取不佳或簇与簇之间的距离较近时，容易陷入局部最优解，产生簇重叠的现象，并且 K-means 算法对噪声点非常敏感，从而导致最终的聚类结果较差。因此，文献[20]提出了层次犹豫模糊 K 均值算法（Hierarchical Hesitant Fuzzy K-Means Clustering Algorithm，HHFK），改进了 K-means 算法初始簇中心选取的方法。该算法采用自底向上构建树结构的凝聚型层次聚类来确定初始的簇中心，降低了 K-means 算法迭代时产生的时间损耗。在利用式(1)计算模糊集之间的距离时，不断将数据对象分配到与其最相似的簇中，直到聚类中心不再变化或达到最大迭代次数时停止。其中，计算犹豫模糊集的运算规则与犹豫模糊元素集相似。

定义 7[20-21]　设任意两个犹豫模糊集 $A=\{a_1, a_2, \cdots, a_d\}$ 和 $B=\{b_1, b_2, \cdots, b_d\}$，且对应属性的元素集犹豫度相同，即 $\lambda_{a_j}=\lambda_{b_j}(j=1,2,\cdots,d)$，则具有以下运算规则：

1) $\lambda A=\bigcup_{\gamma \in a_j}\{\lambda\gamma\}$；

2) $A^\lambda=\bigcup_{\gamma \in a_j}\{\gamma^\lambda\}$；

3) $A \oplus B=\{a_1 \oplus b_1, a_2 \oplus b_2, \cdots, a_n \oplus b_n\}$，$a_j \oplus b_j=\bigcup_{\gamma_1 \in a_j, \gamma_2 \in b_j}\{\gamma_1+\gamma_2-\gamma_1\gamma_2\}$；

4) $A \otimes B=\{a_1 \otimes b_1, a_2 \otimes b_2, \cdots, a_n \otimes b_n\}$，$a_j \otimes b_j=\bigcup_{\gamma_1 \in a_j, \gamma_2 \in b_j}\{\gamma_1\gamma_2\}$。

数据对象犹豫模糊集之间的运算是由其对应属性的犹豫模糊元素集运算的集合，其中不涉及不同属性间犹豫模糊元素集的交叉运算。另外，在犹豫模糊集的“加法”⊕和“乘法”⊗运算中，在对对应犹豫模糊元素集进行“加法”和“乘法”运算时，需要对两个集合的元素进行全排列，若两个犹豫模糊元素集的犹豫度均为 4，则计算后得到的集合犹豫度为 4×4，是原犹豫模糊元素集犹豫度的二次方。

定义 8[20]　令 $X_i=\{x_{i1}, x_{i2}, \cdots, x_{id} \mid x_{ij} \in X_i\}(j=1,2,\cdots,d)$ 为数据集 S 的一组犹豫模糊集，即由 m 个数据对象形成的簇的中心计算公式如下：

$$f(X_1,X_2,\cdots,X_m)=\frac{1}{m}(X_1\oplus X_2\oplus\cdots\oplus X_m) \tag{2}$$

结合定义 7,可以得出簇心的计算公式如下:

$$f(X_1,X_2,\cdots,X_m)=\bigcup_{\gamma_1\in x_{1j},\gamma_2\in x_{2j},\cdots,\gamma_m\in x_{mj}}\{1-\prod_{j=1}^{m}(1-\gamma_j)^{\frac{1}{m}}\} \tag{3}$$

由式(2)和式(3)可以看出,当 m 个犹豫模糊集形成一个簇心时,计算的时间复杂度为 $O(\lambda^m)$,因此在算法不断循环迭代的过程中,指数级的时间复杂度是非常耗时的,导致算法的性能差,不适用于解决现实生活中庞大且高维的数据问题。

3 WHFDP 算法

3.1 WHFDP 算法的思想

本节针对 HHFK 算法存在的缺陷,优化了犹豫模糊集的运算,提出了结合密度峰值思想的加权犹豫模糊聚类算法(Weighted Hesitant Fuzzy Clustering Algorithm Based on Density Peaks,WHFDP),使犹豫模糊环境下的聚类分析有更好的效果。

在对数据集执行聚类算法之前,首先需要对所有数据对象的犹豫模糊元素集进行校验填充,保证各属性特征的犹豫度相同。犹豫模糊元素集在决策分析中表示数据对象的属性特征在某一类的隶属度程度,其取值越接近于 1,表示该属性越符合决策者或专家的期望。例如,在对学生进行综合评定时,教师 A 和教师 B 给出学生 C 在科研能力这一项的分数为 0.3 和 0.99,可以理解为学生 C 的科研能力非常符合教师 B 的期望,而不太符合教师 A 的期望。文献[20]提出采用犹豫模糊元素集的最小值来扩充集合,然而在方差较大的集合中,通过最小值或最大值扩充犹豫度小的集合时,隶属度集合偏向一个专家的决策结果,这会影响到原数据集合的稳定性,使已有的专家决策的评判失去价值。如两个犹豫模糊元素集 $h(x)=\{0.2,0.8\}$ 和 $h(y)=\{0.2,0.6,0.7,0.8,0.9\}$,由于 x 的犹豫度小于 y 的犹豫度,需要对犹豫模糊元素集 $h(x)$ 进行扩充,当决策者悲观地扩充 x 的犹豫模糊元素集为 $h'(x)=\{0.2,0.2,0.2,0.8\}$ 时,特征 x 的隶属期望值会大大降低。因此,我们应当采取保守的态度,尽可能地保证数据集的原始性,这也是提出模糊集的一个原因所在。因此,我们给出了如下补充犹豫模糊元素集的方法。

定义 9 设数据对象 A 和 B 的第 j 个属性的犹豫模糊元素集为 $h(x_{aj})=\{\gamma_{aj_1},\gamma_{aj_2},\cdots,\gamma_{aj_w}\}$ 和 $h(x_{bj})=\{\gamma_{bj_1},\gamma_{bj_2},\cdots,\gamma_{bj_z}\}$,若 $\lambda_{x_{aj}}<\lambda_{x_{bj}}$,则利用式(4)计算得出 $h(x_{aj})$ 的均值 $\bar{\gamma}'$,增加 $z-w$ 个 $\bar{\gamma}'$ 以扩充 $h(x_{aj})$,使得 $\lambda_{x_{aj}}=\lambda_{x_{bj}}$。其中,均值的公式可以表示为:

$$\bar{\gamma}'=\frac{\sum_{k=1}^{w}\gamma_{aj_k}}{w} \tag{4}$$

结合式(1),本文采用犹豫加权汉明距离,公式如下:

$$d_{ham}(A,B)=\sum_{j=1}^{d}\omega_j\left[\frac{1}{\lambda_j}\sum_{\sigma=1}^{\lambda_j}|h_\sigma(x_{aj})-h_\sigma(x_{bj})|\right] \tag{5}$$

另外,决策者可以根据数据对象各属性的重要程度对权重进行人工设置。文献[17]提出在没有特定要求的情况下,可以设置平均权重 $\omega=(1/d,\cdots,1/d)^{\mathrm{T}}$。本文根据数据本身的信息结合变异系数理论提出了新的权重计算的方法,首先计算数据集中属性 j 所有隶属度的平均值 $\bar{\gamma}_j$ 及其标准差 S_j,公式如下:

$$\bar{\gamma}_j=\frac{\sum_{i=1}^{n}\sum_{k=1}^{\lambda_{x_{ij}}}\gamma_{ij_k}}{\sum_{i=1}^{n}\lambda_{x_{ij}}} \tag{6}$$

$$S_j=\left[\frac{\sum_{i=1}^{n}\sum_{k=1}^{\lambda_{x_{ij}}}(\gamma_{ij_k}-\bar{\gamma}_j)^2}{n}\right]^{\frac{1}{2}} \tag{7}$$

$$CV_j=\frac{S_j}{\bar{\gamma}_j} \tag{8}$$

将计算得到的 $\bar{\gamma}_j$ 和 S_j 带入式(8),即变异系数公式。若属性的变异程度越大,说明该属性的隶属度取值越不稳定,专家对该属性在决策中的分歧就越大,因此需要减小该属性的权重,则对 CV_j 取倒数得出 CV'_j,最终得出 ω_j,公式如下:

$$CV_j'=\frac{1}{CV_j} \tag{9}$$

$$\omega_j=\frac{CV_j'}{\sum_{j=1}^{d}CV_j'} \tag{10}$$

在 HHFK 算法中,利用凝聚型层次聚类计算初始簇心时,需要不断合并簇并利用式(3)更新簇中心,导致计算的时间复杂度随簇中数据对象的增加呈指数型增长。为了使犹豫模糊集能够有效应用于大数据的环境,本文采用以数据对象为簇心的密度峰值聚类算法(Density Peaks Clustering,DPC)[25]来优化簇心的计算过程,同时避免了 K-means 算法对噪声点敏感的问题,提高了对非球型簇的识别能力。DPC 算法的主要思想是簇中心的局部密度高于其邻近点的局部密度,且局部密度高的簇心之间的距离大于簇心到其簇内的对象的距离。

定义 10[25] 设数据集 S 的数据对象为 X_i,其局部密度 ρ_i 可表示与 X_i 距离小于 d_c 的数据对象的个数,公式如下:

$$\rho_i=\sum_{j\neq i}\chi(d_{ij}-d_c) \tag{11}$$

$$\chi(x)=\begin{cases}0, & x\geqslant 0\\ 1, & x<0\end{cases} \tag{12}$$

在数据量较大时,为了降低数据集内数据对象局部密度值相等的概率,采用高斯核公式代替式(11),公式如下:

$$\rho_i=\sum_{j\neq i}\exp\left(-\left(\frac{d_{ij}}{d_c}\right)^2\right) \tag{13}$$

$$\rho_i=\sum_{j\in KNN_i}\exp(-d_{ij}) \tag{14}$$

其中,d_{ij} 表示数据对象 X_i 的犹豫模糊集与数据对象 X_j 的犹豫模糊集之间的距离,利用式(5)计算得出。截断距离 d_c 的取值通常使得截断距离内包含 1%~2%的数据样本量,然而当数据集样本较少时,可能产生 d_c 内无数据点的情况,从而影响局部密度 ρ_i 的计算结果,因此可以通过式(14)改进 d_c 选取困难的情况,可选取近邻数 K 为数据样本量的 1%~2%,从而减小 d_c 对数据局部密度的影响[26-27]。

$$\delta_i=\begin{cases}\min\limits_{j:\rho_i<\rho_j}(d_{ij}), & \exists j\rho_i<\rho_j\\ \max\limits_{j}(d_{ij}), & \text{otherwise}\end{cases},j=1,2,\cdots,n \tag{15}$$

由式(15)可得，当数据对象具有最大局部密度时，则其相对距离δ_i为其到距其最远的数据对象的距离。其他数据对象的相对距离δ_i表示与其距离最近且局部密度大于X_i的对象之间的距离。通过$\rho-\delta$决策图，从右上角选取ρ值和δ值均大的数据点作为簇中心，由于通过决策图容易出现难以确定的情况，因此通过计算$\tau_i=\rho_i\times\delta_i$值进行排序来选择簇心。

根据上述思想，算法的基本步骤如算法1所示。

算法1　WHFDP算法

输入：数据集S，参数d_c或K

输出：数据对象的聚类结果

step 1　根据式(10)计算各属性的权重ω_j。

step 2　计算数据对象各属性的犹豫度λ_j，根据式(4)补齐犹豫模糊元素集。

step 3　根据式(5)计算数据对象之间的距离。

step 4　根据式(13)或式(14)计算每个数据对象的ρ值，根据式(15)计算每个数据对象的δ值，通过$\rho-\delta$决策图或τ值排序确定簇中心。

step 5　分配簇心以外的数据对象到距离其最近且局部密度大于它的数据对象所在的簇中。

3.2　WHFDP算法的复杂度分析

本节对WHFDP算法进行时间复杂度和空间复杂度分析，并与HHFK算法进行对比分析，结果表明新算法在性能上有所提高。

在WHFDP算法中，存储距离矩阵的空间复杂度为$O(n^2)$，存储权重的空间复杂度为$O(d)$，更新犹豫模糊元素集的空间复杂度为$O(\bar{\lambda})$，因此算法的空间复杂度近似于$O(n^2)$。计算距离矩阵的时间复杂度为$O(n^2)$，计算权重的复杂度为$O(n)$，分配数据对象的时间复杂度为$O(n^2)$，因此算法的时间复杂度近似为$O(n^2)$。

在HHFK算法中，计算初始簇中心时的空间复杂度为$O(\lambda^n)$，计算对象到簇中心距离的空间复杂度为$O(kn)$，因此算法的空间复杂度近似于$O(\lambda^n)$。若n个数据对象凝聚成一个簇时，计算初始簇中心的时间复杂度为$O(\lambda^n)$，分配数据对象到k个簇迭代t次的时间复杂度为$O(ktn)$，因此算法的时间复杂度近似为$O(\lambda^n)$。

表1　复杂度的比较

Table 1　Comparison of complexity

	WHFDP	HHFK
时间复杂度	$O(n^2)$	$O(\lambda^n)$
空间复杂度	$O(n^2)$	$O(\lambda^n)$

从上述的复杂度分析可以看出，当数据集的所有犹豫度λ均为1或是数据对象个数$n\leqslant 2$时，HHFK算法才优于本文的WHFDP算法。而且HHFK算法复杂度为指数级，通常情况下n远大于λ，几乎不适用于解决现实问题，因此在计算实际数据集的过程中，本文算法在时间和空间上的开销都远低于HHFK算法。

4　仿真实验分析

本节用实例对上述算法进行了说明，如表2所列，实例由8个犹豫模糊对象组成，形成两个簇，且每个簇包含4个数据对象，每个数据对象的属性维度为3。

表2　犹豫模糊集

Table 2　Hesitant fuzzy sets

	$\boldsymbol{h}(\boldsymbol{x}_1)$	$\boldsymbol{h}(\boldsymbol{x}_2)$	$\boldsymbol{h}(\boldsymbol{x}_3)$
$\boldsymbol{X}_1$	{0.13,0.2,0.27}	{0.11,0.29}	{0.3}
$\boldsymbol{X}_2$	{0.26,0.34}	{0.18,0.22}	{0.18,0.33,0.39}
$\boldsymbol{X}_3$	{0.2}	{0.21,0.33,0.36}	{0.25,0.32,0.33}
$\boldsymbol{X}_4$	{0.22,0.38}	{0.17,0.36,0.37}	{0.23,0.37}
$\boldsymbol{X}_5$	{0.7}	{0.64,0.76}	{0.65,0.71,0.74}
$\boldsymbol{X}_6$	{0.73,0.87}	{0.67,0.73}	{0.62,0.69,0.79}
$\boldsymbol{X}_7$	{0.57,0.75,0.78}	{0.75,0.79,0.86}	{0.62,0.78}
$\boldsymbol{X}_8$	{0.7,0.75,0.95}	{0.78,0.82}	{0.6,0.8}

首先根据式(6)－式(10)计算每个属性的权重，如计算第一个属性的权重ω_1，代入式(6)求得所有对象有关属性x_1犹豫模糊集的平均值$\bar{\gamma}_1=0.5176$，代入式(7)求得标准差$S_1=0.2763$，将平均值和标准差的结果代入式(8)求得变异系数$CV_1=0.5337$，同理可求得$CV_2=0.5391$，$CV_3=0.4248$。对求得所有属性的CV值取倒数并代入式(10)，最终得出权重$\omega_1=0.3080$，$\omega_2=0.3049$，$\omega_3=0.387$。

根据犹豫模糊加权距离公式，即式(5)，计算数据对象之间的距离，需要比较犹豫模糊集各属性的犹豫模糊元素集的犹豫度是否相同，对犹豫度较小的犹豫模糊元素集进行补齐。以数据对象X_1与X_2之间的距离为例，X_1和X_2的第1个属性的犹豫模糊元素集的犹豫度$\lambda_{x_{11}}=3$，$\lambda_{x_{21}}=2$，由于$\lambda_{x_{11}}>\lambda_{x_{21}}$，则对$h(x_{21})$补充$\lambda_{x_{11}}-\lambda_{x_{21}}=1$个犹豫模糊元素，代入式(4)求得$\gamma'=0.3$，则补齐后的犹豫模糊元素集为$h'(x_{21})=\{0.26,0.34,0.3\}$，同理补齐$h'(x_{13})=\{0.3,0.3,0.3\}$，扩充后的犹豫模糊集如表3所列。

表3　补充后的犹豫模糊集X_1和X_2

Table 3　Hesitant fuzzy sets X_1 and X_2 after replenshment

	$\boldsymbol{h}'(\boldsymbol{x}_1)$	$\boldsymbol{h}(\boldsymbol{x}_2)$	$\boldsymbol{h}'(\boldsymbol{x}_3)$
$\boldsymbol{X}_1$	{0.13,0.2,0.27}	{0.11,0.29}	{0.3,0.3,0.3}
$\boldsymbol{X}_2$	{0.26,0.34,0.3}	{0.18,0.22}	{0.18,0.33,0.39}

在补齐犹豫模糊集之后，计算对应犹豫模糊元素集之间的距离$d(x_{11},x_{21})=0.1$，$d(x_{12},x_{22})=0.07$，$d(x_{13},x_{23})=0.08$，$d_{ham}(X_1,X_2)=\omega_1*d(x_{11},x_{21})+\omega_2*d(x_{12},x_{22})+\omega_3*d(x_{13},x_{23})=0.0831$。采用同样的方式补齐其他犹豫模糊集与数据对象犹豫模糊集之间的距离，根据犹豫模糊集的性质可得$d(X_i,X_i)=0$，由犹豫模糊集之间距离的对称性将距离结果简化为上三角矩阵的形式，计算结果如表4所列。

表4　犹豫模糊集之间的距离

Table 4　Distance between hesitant fuzzy sets

	$\boldsymbol{X}_1$	$\boldsymbol{X}_2$	$\boldsymbol{X}_3$	$\boldsymbol{X}_4$	$\boldsymbol{X}_5$	$\boldsymbol{X}_6$	$\boldsymbol{X}_7$	$\boldsymbol{X}_8$
$\boldsymbol{X}_1$	0	0.0831	0.0578	0.0884	0.4613	0.4921	0.4917	0.5225
$\boldsymbol{X}_2$		0	0.0794	0.0681	0.4305	0.4613	0.4609	0.4917
$\boldsymbol{X}_3$			0	0.0518	0.4308	0.4616	0.4613	0.4921
$\boldsymbol{X}_4$				0	0.4000	0.4308	0.4305	0.4613
$\boldsymbol{X}_5$					0	0.0528	0.0752	0.0845
$\boldsymbol{X}_6$						0	0.0845	0.0897
$\boldsymbol{X}_7$							0	0.0507
$\boldsymbol{X}_8$								0

计算得出距离后，则得出数据对象的ρ_i值和δ_i值，并计算$\tau_i=\rho_i\times\delta_i$。在该数据集中，可以计算得出$\tau_3>\tau_7>\tau_4>\tau_5>$

$\tau_8>\tau_6>\tau_1>\tau_2$，其中$\tau_3=0.3066$，$\tau_7=0.2329$，远大于$\tau_4=0.05$，因此可以判断该数据集有两个簇中心，分别为$X_3$和$X_7$，然后分配簇心以外的数据对象，最后得出$C_1=\{X_1,X_2,X_3,X_4\}$，$C_2=\{X_5,X_6,X_7,X_8\}$。

文献[20]中 HHFK 算法则设定合并的簇数为 2，即得到表 5 最后一行的聚类结果，将其结果代入式(3)得到簇心C_1和C_2，作为 K-means 算法的初始簇中心，比较簇心到各点的距离，通过迭代最终得到的聚类结果为$C_1=\{X_1,X_2,X_3,X_4\}$，$C_2=\{X_5,X_6,X_7,X_8\}$，与本文算法的结果一致。

为了验证所提算法的聚类效果，选取常用的合成数据集进行实验，如图 1 所示。为了将数据集应用在新算法中，需要对其进行改造。首先将数据归一化，保证隶属度在 0～1 之间，然后对所有数据对象的属性值进行犹豫模糊化，生成随机数以原来数值x_{ij}为基准、设置浮动范围为0～0.2。通过 Matlab 的 rand 函数随机生成一个范围在[－0.1，0.1]之间的随机数 ε，则犹豫模糊元素的值为$x_{ij}\pm\varepsilon$。如果生成犹豫模糊元素超出了 0～1 的范围，则重新生成该数值，并分别取犹豫度 λ=2，λ=4 和 λ=8 的犹豫模糊元素集进行聚类，并对结果进行比较，结果如表 6 所列。

表 5 初始化的聚类结果

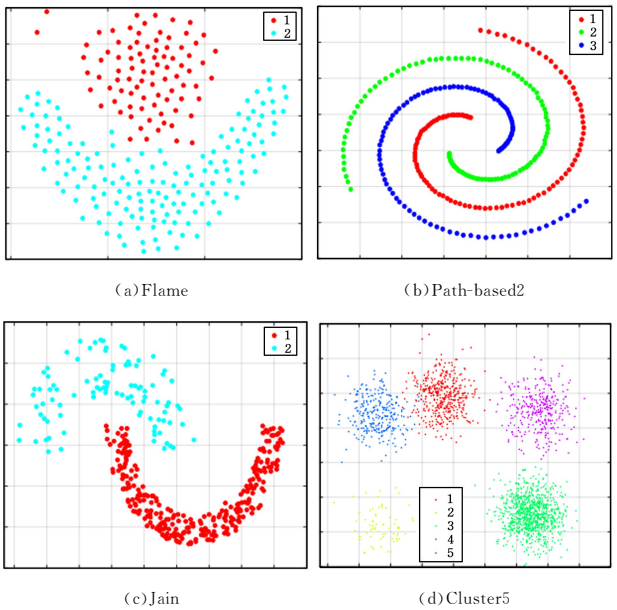
Table 5 Clustering results of initialization

簇数	簇与簇的合并过程
8	$\{X_1\},\{X_2\},\{X_3\},\{X_4\},\{X_5\},\{X_6\},\{X_7\},\{X_8\}$
7	$\{X_1\},\{X_2\},\{X_3\},\{X_4\},\{X_5\},\{X_6\},\{X_7,X_8\}$
6	$\{X_1\},\{X_2\},\{X_3\},\{X_4\},\{X_5,X_6\},\{X_7,X_8\}$
5	$\{X_1\},\{X_2\},\{X_3,X_4\},\{X_5,X_6\},\{X_7,X_8\}$
4	$\{X_1\},\{X_2,X_3,X_4\},\{X_5,X_6\},\{X_7,X_8\}$
3	$\{X_1\},\{X_2,X_3,X_4\},\{X_5,X_6,X_7,X_8\}$
2	$\{\boldsymbol{X_1},\boldsymbol{X_2},\boldsymbol{X_3},\boldsymbol{X_4}\},\{\boldsymbol{X_5},\boldsymbol{X_6},\boldsymbol{X_7},\boldsymbol{X_8}\}$

表 6 WHFDP 的聚类结果

Table 6 Clustering results of WHFDP algorithm

Datasets	λ=1(DPC)			λ=2			λ=4			λ=8		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
Flame	0.713	0.413	0.327	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
Path-based2	**1.000**	**1.000**	**1.000**	0.439	0.030	0.020	0.532	0.169	0.121	0.631	0.318	0.250
Jain	0.861	0.507	0.515	0.925	0.653	0.715	0.930	0.669	0.733	**0.936**	**0.685**	**0.752**
Cluster5	0.993	0.974	0.988	**0.994**	0.977	**0.990**	**0.994**	0.977	**0.990**	**0.994**	**0.978**	**0.990**



(a)Flame (b)Path-based2

(c)Jain (d)Cluster5

图 1 人工可视化数据集

Fig. 1 Visualized synthetic datasets

实验中的截断距离均取包含 2%的样本量，表中 λ=1 即是在非犹豫模糊环境下密度峰值聚类(DPC)算法的结果。由于犹豫度是随机生成的，对不同犹豫度的犹豫模糊集各重复 50 次实验得出的最优结果进行比较，采用精确度(Accuracy，ACC)、标准化互信息(Normalized Mutual Information，NMI)和调整兰德系数(Adjusted Rand Index，ARI)3 个指标评价聚类效果。从表 6 中可以看出，算法在数据集 Flame，Jain 和 Cluster5 上的聚类结果较为理想，在犹豫模糊数据集情况下的聚类结果优于确定性数据集的聚类结果，但是螺旋型数据集 Path-based2 下的聚类结果与精确数据集下的聚类结果的差距较大。由于不同簇之间的数据对象的距离都较近，且在犹豫模糊的情况下，数据对象在某一维度或某几个维度不断靠近另一个簇的数据对象，导致二者距离过近，从而容易形成球型簇，这种情况是符合认知的。

另外，从实验结果可以看出，随着犹豫模糊集犹豫度的增加，不会对聚类结果产生负向效果，而且在结构复杂的数据集中，随着犹豫度的增加，还能够提高聚类结果的准确性。因此，将犹豫模糊加权密度峰值算法应用于群体决策时，随着决策意见的增加会对决策结果产生积极影响。通过犹豫模糊集保留了原始的不精确数据，减少了数据的预处理，或是将确定性数据集进行适当的模糊处理，可以获得准确率更高的聚类结果。

结束语 本文基于犹豫模糊集提出了一种犹豫模糊加权聚类算法，优化了原始犹豫模糊聚类中簇中心的选取方式，降低了算法的复杂度，同时给出了新的加权距离公式，使其能够更适用于解决现实的决策问题。通过实验分析，结合密度峰值算法使聚类在球型簇和凹型簇都有更好的识别能力，但其对螺旋状数据集的识别能力不足，下一步的研究可以通过引入信息熵等思想来改进犹豫模糊集之间相异度的计算方式，以提高不同簇形的识别能力，进一步降低算法复杂度，并拓展该算法使其适用于直觉模糊集合等不确定数据集中。

参 考 文 献

[1] XIA Z H, WANG X H, SUN X M, et al. Steganalysis of LSB matching using differences between nonadjacent pixels[J]. Multimedia Tools and Applications, 2016, 75(4): 1947-1962.

[2] ANWAR T, LIU C F, VU H L, et al. Partit-ioning road networks using density peak graphs: Efficiency vs. accuracy[J]. Information Systems, 2017, 64(C): 22-40.

[3] AHN C S,OH S Y. Robust vocabulary reco-gnition clustering model using an average estimator least mean square filter in noisy environments[J]. Personal & Ubiquitous Computing, 2013,18(6):1295-1301.

[4] JIN J G. Review of clustering method[J]. Computer Science, 2014,41(11A):288-293.

[5] STREHL A,GHOSH J. Cluster ensembles:a knowledge reuse framework for combining partitionings[J]. Journal of Machine Learning Research,2002,3(3):583-617.

[6] FRED A L N,JAIN A K. Combining multiple clusterings using evidence accumulation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2005,27(6):835-850.

[7] IAMON N,BOONGOEN T,GARRETT S,et al. A link-based cluster ensemble approach for categorical data clustering[J]. IEEE Transactions on Knowledge & Data Engineering,2012, 24(3):413-425.

[8] TANG W,ZHOU Z H. Bagging-based selective clusterer ensemble[J]. Journal of Software,2005,16(4):496-502.

[9] HUANG D,WANG C D,WU J S,et al. Ultra-Scalable Spectral Clustering and Ensemble Clustering[J]. IEEE Transactions on Knowledge and Data Engineering,2020,32(6):1212-1226.

[10] ZADEH L A. The concept of a linguistic variable and its application to approximate reasoning[J]. Information Science, 1975, 8(3):199-249.

[11] LIAO H C,XU Z S,ZENG X J,et al. Qualitative decision making with correlation coefficients of hesitant fuzzy linguistic term sets[J]. Knowledge Based Systems,2015,76:127-138.

[12] MENG F,CHEN X,ZHANG Q. Multi-attribute decision analysis under a linguistic hesitant fuzzy environment[J]. Information Sciences,2014,267:287-305.

[13] YAVUZ M,OZTAYSI B,ONAR S C,et al. Multi-criteria evaluation of alternative-fuel vehicles via a hierarchical hesitant fuzzy linguistic model[J]. Expert Systems with Applications, 2015, 42(5):2835-2848.

[14] ATANASSOV K T. Intuitionistic fuzzy sets[J]. Fuzzy Sets and Systems,1986,20(1):87-96.

[15] MIZUMOTO M,TANAKA K. Some properties of fuzzy sets of type 2[J]. Information and Control,1976,31(4):312-340.

[16] TORRA V,NARUKAWA Y. On Hesitant fuzzy sets and decision[C]// IEEE International Conference on Fuzzy Systems. 2009:1378-1382.

[17] YAGE R,RONALD R. On the theory of bags[J]. International Journal of General System,1986,13(1):23-37.

[18] MIYAMOTO S. Information clustering based on fuzzy multisets [J]. Information Processing and Management,2003,39(2):195-213.

[19] YAO D B,WANG C C. Hesitant intuitionistic fuzzy entropy/cross-entropy and their applications[J]. Soft Computing,2018, 22(9):2809-2824.

[20] CHEN NA,XU Z S,XIA M M. Hierarchical hesitant fuzzy K-means clustering algorithm[J]. Applied Mathematics-A Journal of Chinese Universities,2014,29(1):1-17.

[21] XIA M M,XU Z S. Hesitant fuzzy information aggregation in decision making[J]. International Journal of ApproximateReasoning,2011,52(3):395-407.

[22] MERIGO J M,CASANOVAS M. Induced aggregation operators in decision making with the Dempster-Shafer belief structure [J]. International Journal of Intelligent Systems, 2009, 24(8):934-954.

[23] LIU H W,WANG G J. Multi-criteria decision making methods based on intuitionistic fuzzy sets[J]. European Journal of Operational Research,2007,179(1):220-233.

[24] XU Z S,XIA M M. Distance and similarity measures for hesitant fuzzy sets[J]. Information Science,2011,181(11):2128-2138.

[25] RODRIGUEZ A,LAIO A. Clustering by fast search and find of density peaks[J]. Science,2014,344(6191):1492-1496.

[26] DU M J,DING S F,JIA H. Study on density peaks clustering based on k-nearest neighbors and principal component analysis [J]. Knowledge Based Systems,2016,99:135-145.

[27] XIE J Y,GAO H C,LIU X,et al. Robust clustering by detecting density peaks and assigning points based on fuzzy weighted K-nearest neighbors[J]. Information Sciences,2016,354(c):19-40.

ZHANG Yu, born in 1994, postgraduate, is a member of China Computer Federation. Her main research interests include data mining and so on.

LU Yi-hong, born in 1968, master, associate professor, is a member of China Computer Federation. Her main research interests include software theory and data mining.