基于阴影集的多粒度三支聚类集成

姜春茂,赵书宝
(哈尔滨师范大学计算机科学与信息工程学院,黑龙江哈尔滨 150025)

摘 要: 聚类集成旨在通过融合多个不同的基聚类结果得到一个统一的类簇划分.针对现实环境中的模糊和不确定性数据,本文提出了一种基于阴影集的多粒度三支聚类集成算法.算法首先使用FCM聚类产生一组有差异性的基聚类成员,并通过阴影集构造三支聚类.然后引入多粒度粗糙集构建了四个近似集合,将每一个类簇划分为一个核心域和三个边界域.最后对边界域中的数据依次划分到核心域中,无法划分的对象则留在边界域,最终得到了三支聚类集成的结果.实验结果表明,本算法在准确率、调整兰德系数和归一化互信息方面,与多种现有的聚类集成算法相比得到了更好的聚类集成结果.

关键词: 聚类集成;三支聚类;多粒度;模糊c均值(FCM);阴影集
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112(2021)08-1524-09
电子学报URL: http://www.ejournal.org.cn **DOI:** 10.12263/DZXB.20200626

Multi-granulation Three-Way Clustering Ensemble Based on Shadowed Sets

JIANG Chun-mao, ZHAO Shu-bao
(*College of Computer Science and Information Engineering, Harbin Normal University, Harbin, Heilongjiang* 150025, *China*)

Abstract: The purpose of clustering ensemble is to find a unified partition of objects by fusing a set of clustering results. This paper proposes a multi-granulation three-way clustering ensemble algorithm based on shadowed sets to deal with the fuzzy and uncertainty data in the actual world. First, the algorithm generates a set of clustering members using the fuzzy c-means algorithm, and then the membership degree is mapped into three regions to construct three-way clustering. Second, the multi-granulation rough sets are used to construct four different approximate regions. Each cluster contains a core region and three boundary regions. Finally, the shadowed set is used to classify objects in boundary regions sequentially. Objects that cannot be divided are left in the boundary region. The experimental results show the algorithm obtains better clustering ensemble results in accuracy, adjust rand index, and normalized mutual information compared to multiple existing clustering ensemble algorithms.

Key words: clustering ensemble; three-way clustering; multi-granulation; fuzzy c-means (FCM); shadowed sets

1 引言

聚类分析旨在根据某种相似性度量使得相似的样本分为一类,不相似的样本分为不同的类别.这使得同一类之间的样本保持较高的相似性,而不同的类之间的样本则保持较高的相异性.由于数据的复杂性和多样性,单一的聚类算法通常难以精确描述类簇的结构特征.即使同一个聚类算法在参数不同时,聚类结果也可能会存在差别.因此聚类集成[1~6]被提出以获得更好的稳定性、鲁棒性和准确率.

聚类集成一般分为两个阶段,即基聚类生成阶段和一致性集成阶段.其中基聚类生成阶段采用不同的聚类算法或者不同参数下的同一聚类算法生成多个具有差异性的基聚类结果.一致性集成阶段则通过设计一致性函数如基于共协矩阵的方法[5],基于图的方法[1],基于投票的方法[4],基于证据积累的方法[2]和基于概率积累的方法[6]等将多个不同的基聚类结果融合为一个统一的类簇划分.

由于先验信息的缺乏,聚类集成的研究相比于分类集成更加困难,如何设计优质的一致性函数,融合多

收稿日期:2020-06-29;修回日期:2020-12-21;责任编辑:李勇锋
基金项目:黑龙江省自然科学基金(No.LH2020F031)

个不同的基聚类结果,从而得到一个更为稳定和鲁棒的类簇划分一直是聚类集成研究的热点问题. 事实上,聚类集成通过多个基聚类成员,分别从不同的视角对类簇结构进行描述,是多个局部视角的组合. 这是一种典型的多粒度计算思想. 多粒度计算通过多个粒度空间对数据进行建模和分析,而多个粒度空间则可能是人们在多个层次、多个视角或多个尺度对数据进行描述的结果. 从多个粒度空间对数据进行建模和分析有助于发现数据隐含的特征,获得更加合理和满意的结果. Qian[7]借鉴人类的多粒度认知机理,提出了多粒度粗糙集模型. 该模型能够对复杂数据进行有效处理以获得高质量的分析结果.

传统的聚类算法中,对象与类簇之间的关系是明确的,即对象确定属于或者确定不属于某个类簇. 这种聚类算法又称为硬聚类算法或二支聚类算法. 在处理类簇边界模糊和不确定的数据时,硬聚类的结果难以精确描述类簇的结构特征. 当信息不充分时将对象强制划分到某一类簇容易带来较高的误分类代价. Yao将三支决策[8~13]的思想引入到聚类分析中并提出了三支聚类算法[14~19]. 不同于传统的硬聚类,三支聚类将对象与类簇之间的关系分为三种,即对象确定属于该类簇,对象确定不属于该类簇和不确定对象是否属于该类簇. 三支聚类能够很好地描述在某些数据上对象与类簇之间缺乏明确的归属关系现象,能够更加精确地描述类簇的结构特征.

为解决聚类集成中面对的模糊和不确定数据,本文借鉴多粒度粗糙集和三支聚类的优势,提出了一种基于阴影集的多粒度三支聚类集成算法(MTWCES). 算法以FCM为基聚类,FCM的聚类结果中对象与类簇之间的关系通过隶属度矩阵呈现. 通过阴影集对FCM聚类的结果进行处理. 将对象划分到类簇的核心域、阴影域和排外域中,这是一种典型的三支聚类结果. 位于核心域的对象表示确定属于该类簇,位于阴影域的对象表示可能属于该类簇,位于排外域的对象表示确定不属于该类簇. 通过引入多粒度粗糙集的相关思想构造了两种多粒度粗糙集模型,即乐观多粒度粗糙集和悲观多粒度粗糙集. 通过两种多粒度粗糙集模型构建了四个近似集合. 通过四个近似集合对数据进行划分,对每个类簇均构造了一个核心域和三个边界域,且不同区域的数据对类簇的贡献度不同,即不同区域的数据对类簇的重要性存在一定的偏序关系. 最后使用核心域中的对象计算类簇中心,对三个边界域中的对象通过阴影集依次划分到核心域,而无法分配的对象则保留在边界域中. 在UCI真实数据集上与多种聚类集成算法进行比较,验证了本文所提出的算法是有效的.

2　基础知识

2.1　阴影集

阴影集[20,21]由Witold Pedrycz于1998年首次提出. 阴影集由模糊集演化而来,为了消除模糊集中的歧义信息并捕捉数据的分布,阴影集通过精确的隶属度值识别类簇边界模糊的现象,其能够有效发现与类簇关系不确定的数据,加强了结果的可解释性. 作为模糊集与粗糙集连接的桥梁,阴影集将模糊结果以粗糙集的形式呈现.

定义1[20]　在论域U中,给定一对阈值(α,β),并且有$0 \leqslant \beta < \alpha \leqslant 1$. 阴影集将论域$U$中所有的对象根据隶属度$\mu_A(x)$映射到集合$\{0,[0,1],1\}$中,即

$$S_{(\alpha,\beta)}(\mu_A(x)) = \begin{cases} 0, & \mu_A(x) \leqslant \beta \\ [0,1], & \beta < \mu_A(x) < \alpha \\ 1, & \mu_A(x) \geqslant \alpha \end{cases} \tag{1}$$

其中隶属度$\mu_A(x)$表示对象x隶属度概念A的程度. 如果隶属度函数$\mu_A(x)$大于或等于阈值α,则通过提升操作将对象x的隶属度$\mu_A(x)$提升到1. 如果对象x的隶属度$\mu_A(x)$小于或等于阈值β,则通过降低操作将隶属度$\mu_A(x)$降低到0. 如果对象x的隶属度$\mu_A(x)$在α和β之间,则将对象划分到阴影区域.

Witold Pedrycz的阴影集模型自提出以来得到了广泛的发展和应用. Mitra[22]结合模糊聚类和粗糙聚类的相关优势,提出了一种基于阴影集的c-means聚类算法. Yue[23]针对邻域分类模型对处理不确定对象的不足提出了一种基于阴影邻域的三支分类算法. 为减少模糊集及其诱导的阴影集之间存在较大的不确定性,Zhang[24]通过模糊熵提出了一种区间阴影集模型.

2.2　多粒度粗糙集

在决策过程中,如果决策者仅考虑自己粒度空间的决策结果,而不反对其他粒度空间的决策结果,那么相应的决策为乐观决策,其对应的多粒度粗糙集称为乐观多粒度粗糙集. 定义如下:

定义2[7]　给定信息系统$IS = \{U, A, V, f\}$,其中$A_1, A_2, \cdots, A_m \subseteq A$和$X \subseteq U$. $[x]_{A_i}$表示在属性集A_i下为等价关系的一组对象的集合. 目标概念X关于$A_1, A_2, \cdots, A_m$的乐观多粒度粗糙集下近似和乐观多粒度粗糙集上近似记为$\underline{\sum_{i=1}^{m}} A_i^o(X)$和$\overline{\sum_{i=1}^{m}} A_i^o(X)$,其中

$$\underline{\sum_{i=1}^{m}} A_i^o(X) = \left\{ x \in U \middle| [x]_{A_1} \subseteq X \vee [x]_{A_2} \subseteq X \vee \cdots \vee [x]_{A_m} \subseteq X \right\} \tag{2}$$

$$\overline{\sum_{i=1}^{m}}A_i^o(X)=\sim\underline{\sum_{i=1}^{m}}A_i^o(\sim X) \tag{3}$$

乐观多粒度粗糙集中的边界域为

$$BND\left(\sum_{i=1}^{m}A_i^o(X)\right)=\overline{\sum_{i=1}^{m}}A_i^o(X)-\underline{\sum_{i=1}^{m}}A_i^o(X) \tag{4}$$

在决策过程中,如果决策者不仅根据自己的粒度空间进行决策,同时考虑其他决策者粒度空间的决策结果,即使用所有决策者共同满意的方案进行决策,那么相应的决策为悲观决策,其对应的多粒度粗糙集称为悲观多粒度粗糙集. 定义如下:

定义 3[7] 给定信息系统 $IS=\{U,A,V,f\}$,其中 $A_1,A_2,\cdots,A_m\subseteq A$ 和 $X\subseteq U$. $[x]_{A_i}$ 表示在属性集 A_i 下为等价关系的一组对象的集合. 目标概念 X 关于 $A_1,A_2,\cdots,A_m$ 的悲观多粒度粗糙集下近似和悲观多粒度粗糙集上近似分别记为 $\underline{\sum_{i=1}^{m}}A_i^p(X)$ 和 $\overline{\sum_{i=1}^{m}}A_i^p(X)$,其中

$$\underline{\sum_{i=1}^{m}}A_i^p(X)=\left\{x\in U\left|[x]_{A_1}\subseteq X\wedge[x]_{A_2}\subseteq X\wedge\cdots\wedge[x]_{A_m}\subseteq X\right.\right\} \tag{5}$$

$$\overline{\sum_{i=1}^{m}}A_i^p(X)=\sim\underline{\sum_{i=1}^{m}}A_i^p(\sim X) \tag{6}$$

悲观多粒度粗糙集的边界域定义为

$$BND\left(\sum_{i=1}^{m}A_i^p(X)\right)=\overline{\sum_{i=1}^{m}}A_i^p(X)-\underline{\sum_{i=1}^{m}}A_i^p(X) \tag{7}$$

Sun[25]研究了双论域上的多粒度决策粗糙集模型并将其运用到了群决策中. Lin[26]研究了多粒度粗糙集模型与证据理论之间的关系,并提出了一种基于证据理论的多粒度信息融合策略. Sang[27]提出了一种多源信息系统上的多粒度双量化决策理论粗糙集模型. Ju[28]将多粒度计算方法引入到代价敏感粗糙集,提出了一种基于多粒度方法的代价敏感粗糙集模型.

2.3 三支聚类

传统的聚类算法是一种硬聚类或二支聚类的结果,对象和类簇之间的关系是明确的,即对象确定属于该类簇或确定不属于该类簇. 在对象与类簇之间缺乏明确的归属关系时,硬聚类难以精确刻画类簇的结构特征. 给定一组数据 $U=\{x_1,x_2,\cdots,x_n\}$,三支聚类的每个类簇表示为 $C_i=\{Co(C_i),Fr(C_i)\}$,即类簇 C_i 由核心域 $Co(C_i)$ 和边界域 $Fr(C_i)$ 两个子集组成. 类簇 C_i 的琐碎域表示为 $Tr(C_i)=U-Co(C_i)-Fr(C_i)$. 类簇 C_i 的琐碎域表示由确定不属于类簇 C_i 的对象组成的集合. 类簇 C_i 的三个域满足如下条件:

(1) $Co(C_i)=\varnothing, i=1,2,\cdots,k$

(2) $\bigcup_{i=1}^{k}\left(Co(C_i)\cup Fr(C_i)\right)=OB$

(3) $Co(C_i)\cap Co(C_j)=\varnothing$

上述三个条件说明任意一个类簇的核心域不为空,所有类簇的核心域和边界域的并集为论域 OB,任意两个类簇的核心域的交集为空. 三支聚类能够很好地描述在某些复杂数据中,对象与类簇之间缺乏明确归属关系的现象.

Yu[14]首次将三支决策的思想引入到聚类分析中,并提出了三支聚类理论. 通过引入一致性低秩矩阵表示所有视图共享的潜在类簇结构,Yu[16]提出了一种基于低秩表示的多视图三支聚类算法. Wang[29]借鉴数学形态学的收缩和扩张的思想,提出了一种基于数学形态学的三支聚类算法,该算法可将现有的二支聚类结果扩展为三支聚类结果. Yu[30]将证据理论引入到三支聚类中,提出了一种基于证据理论的密度峰值聚类算法. 为提高云平台能源利用效率,Jiang[31]提出了一种基于三支聚类的云任务调度算法.

3 基于阴影集的多粒度三支聚类集成

本节首先描述基于阴影集的多粒度三支聚类集成的基本思想,然后详细描述了算法的各个步骤.

3.1 多粒度三支聚类集成的基本思想

图 1 给出了多粒度三支聚类集成的基本框架. 在基于阴影集的多粒度三支聚类集成中,著名的模糊聚类算法 FCM 被选择作为基聚类,FCM 聚类通过隶属度矩阵呈现对象与类簇之间的关系. 为构造多粒度三支聚类集成模型,首先通过阴影集对每一个基聚类的结果进行处理,并构造三支聚类. 根据阴影集的三种操作将隶属度映射到集合 $\{0,[0,1],1\}$ 中,其中隶属度为 1 的对象被划分到核心域,表示确定属于该类簇. 隶属度为 $[0,1]$ 的对象被划分到阴影域,表示可能属于该类簇,隶属度为 0 的对象被划分到排外域,表示确定不属于该类簇. 通过引入多粒度粗糙集对基聚类的结果进行处理,在一组基聚类上构造了两种多粒度粗糙集模型,即悲观多粒度粗糙集和乐观多粒度粗糙集,从而得到四个近似集合. 通过四个近似集合对每一个类簇 C_i 均获得了核心域 $Co(C_i)$ 和三个边界域 $Fr_1(C_i),Fr_2(C_i)$ 和 $Fr_3(C_i)$. 核心域和边界域的数据存在一定的偏序关系,位于核心域的数据通过多个角度的观察得到,表示确定属于该类簇,构成该类簇的核心结构. 位于边界域的数据则是对类簇结构的补充. 通过对核心域 $Co(C_i)$ 的数据计算类簇中心,然后对 $Fr_1(C_i),Fr_2(C_i)$ 和 $Fr_3(C_i)$

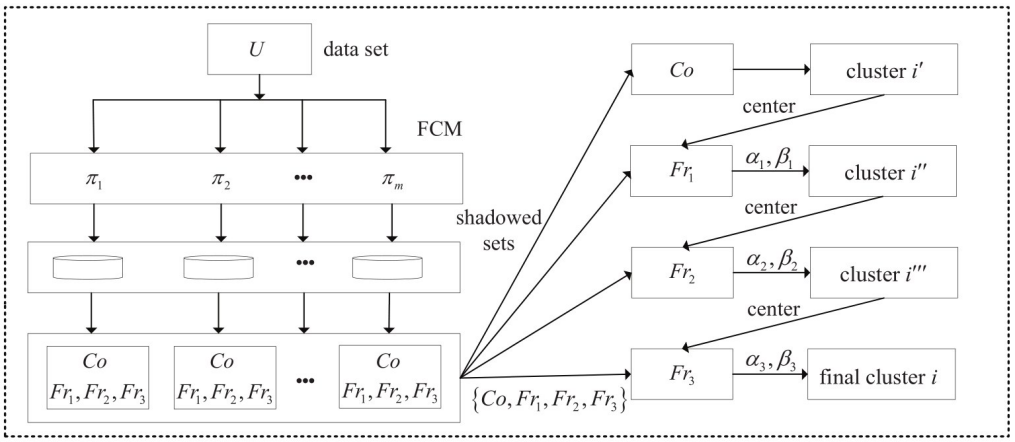

图1　基于阴影集的多粒度三支聚类集成基本框架

三个区域的数据依次进行分类,无法分类的数据则划分到边界域中. 集成结果以三支聚类的形式呈现. 该算法将多粒度粗糙集引入聚类集成,通过多个视角对样本进行描述以求精确刻画数据的结构特征,而三支聚类的引入则有效避免二支聚类引起的误分类风险.

3.2　基聚类生成

给定一组数据 $U=\{x_1,x_2,\cdots,x_n\}$,其中 $x_i=\{x_{i1},x_{i2},\cdots,x_{id}\}$ 表示对象有 d 维. 在 U 的特征子集上进行FCM聚类,在 m 个基聚类中每次选取不同的特征子集,对FCM固定类簇数并随机初始化隶属度函数,因此得到 m 个具有差异性的基聚类集合 $\Pi=\{\pi_1,\pi_2,\cdots,\pi_m\}$,其中 $\pi_i=\{C_{i1},C_{i2},\cdots,C_{ik}\}$.

3.3　基聚类标签匹配

由于聚类分析是一种典型的无监督学习方式,标签缺乏先验的类别信息,因此会导致不同的基聚类结果中标签不匹配的现象,如基聚类成员 $\pi_1=\{1,1,1,2,2,3,3,3\}$ 和 $\pi_2=\{2,2,2,3,3,1,1,1\}$,两种聚类结果虽然表现形式不同,但却是同一个类簇划分的结果. Zhou[4]指出具有对应关系的两个类簇所覆盖的共同对象的个数应该是最多的,从而构建了重叠矩阵,选择具有最大覆盖对象个数的类簇建立对应关系,解决标签匹配的问题. 本文采用Zhou所提出的方法,对基聚类成员的标签进行匹配.

3.4　通过阴影集构造三支聚类

FCM聚类算法通过最小化类内间距不断更新隶属度,当所有的对象对所有的类的隶属度都确定后,生成对象与类簇的隶属度关系矩阵. 在接下来进行聚类集成时,如果仅根据隶属度大小选取最大的作为该对象所属的类别,容易引起一些关键信息的损失,如对于类簇边界模糊的数据来说,如果数据对两个类簇的隶属度之间的差别很小,而仅根据隶属度最大确定类簇,则可能产生较高的误分类风险. 阴影集通过精确隶属度值识别数据中的模糊现象,将隶属度映射到集合 $\{0,[0,1],1\}$ 简化了数据中的模糊关系. 阴影集在保留了数据中的模糊现象的同时,去除了冗余的歧义信息. 因此本文采用阴影集对FCM聚类的结果进行处理,从而构造三支聚类. 在三支聚类中,每一个类簇的结构均由两部分组成,分别是核心域和阴影域,其中核心域中的对象确定属于该类簇而阴影域中的对象可能属于该类簇. 具体如下.

给定一组数据 $U=\{x_1,x_2,\cdots,x_n\}$,通过FCM将论域 U 中的对象划分为 k 个类簇 $C=\{C_1,C_2,\cdots,C_k\}$,类簇中心为 $m=\{m_1,m_2,\cdots,m_k\}$. 对象 x 与类簇 C_j 之间的关系通过隶属度 $\mu_{C_j}(x)$ 表示. 给定一对阈值 (α,β) 并且有 $0\leqslant\beta<\alpha\leqslant1$,通过阴影集的一对阈值 α 和 β 将所有的模糊隶属度划分到集合 $\{0,[0,1],1\}$ 中. 对于隶属度高于阈值 α 的对象,通过提升操作将其隶属度提升至1,表示确定属于该类簇. 对于隶属度低于阈值 β 的对象,通过降低操作将其隶属度降低至0,表示确定不属于该类簇. 隶属度在 α 和 β 之间的对象其隶属度赋值为 $[0,1]$ 表示可能属于该类簇. 最终对每一个类簇得到如下三个域,即核心域、阴影域和排外域:

$$\begin{aligned}core(C_j)&=\{\mu_{C_j}(x)=1\,|\,\mu_{C_j}(x)\geqslant\alpha\}\\shadowed(C_j)&=\{\mu_{C_j}(x)=[0,1]\,|\,\beta<\mu_{C_j}(x)<\alpha\}\\exclusion(C_j)&=\{\mu_{C_j}(x)=0\,|\,\mu_{C_j}(x)\leqslant\beta\}\end{aligned}\tag{8}$$

位于核心域的对象表示确定属于该类簇,位于阴影域的对象表示可能属于该类簇,而位于排外域的对象表

示确定不属于该类簇. 因此类簇的下近似和上近似可以表示为

$$\underline{apr}\left(C_j\right)=core\left(C_j\right)=\left\{x\in U\left|\mu_{C_j}(x)=1\right.\right\} \tag{9}$$

$$\begin{aligned}\overline{apr}\left(C_j\right)&=core\left(C_j\right)\cup shadowed\left(C_j\right)\\&=\left\{x\in U\left|\mu_{C_j}(x)=1\right.\right\}\cup\\&\quad\left\{x\in U\left|\mu_{C_j}(x)=[0,1]\right.\right\}\end{aligned} \tag{10}$$

为了计算最优阈值(α,β),Zhang[32]将博弈论引入到阴影集,通过求解纳什均衡获得最优阈值. Witold Pedrycz[20]将阴影集中提升和降低操作变化的隶属度之和与阴影集中对象数量之差的绝对值作为优化函数,通过求解优化函数的最小值获得最优阈值. Deng[33]将贝叶斯风险决策引入到阴影集,寻找决策风险最小时的阈值. 本文采用Witold Pedrycz所提出的方法,通过求解如下优化函数获得最优阈值α和β. 对于任意类簇C_j,$\mu_{C_j}(x)$表示对象x对C_j的隶属度,通过分析对象与类簇的隶属度关系,构建优化函数:

$$\alpha_j,\beta_j=\arg\min_{\alpha}\left(O_j\right) \tag{11}$$

其中

$$\begin{aligned}O_j=\Bigg|&\sum_{\mu_{C_j}(x)\geqslant\alpha_j}\left(\mu_{\max}(x)-\mu_{C_j}(x)\right)+\sum_{\mu_{C_j}(x)\leqslant\beta_j}\mu_{C_j}(x)\\&-card\left\{x\in U\left|\beta_j<\mu_{C_j}(x)<\alpha_j\right.\right\}\Bigg|\end{aligned}$$

3.5 多粒度模型的构建

给定一组基聚类$\Pi=\left\{\pi_1,\pi_2,\cdots,\pi_m\right\}$,在每一个基聚类上均通过阴影集构造类簇的下近似集合和上近似集合. 借鉴多粒度粗糙集的思想,构建了两种多粒度粗糙集模型.

在m次基聚类中,如果对象x每次都被划分到类簇C_j的下近似中,则对象x属于类簇C_j的悲观多粒度粗糙集下近似. 悲观多粒度粗糙集下近似表示为

$$\begin{aligned}\underline{\sum_{i=1}^{m}}\pi_i^p\left(C_j\right)=\Big\{x\in U\Big|&core_{\pi_1}\left(C_j\right)\wedge core_{\pi_2}\left(C_j\right)\wedge\cdots\\&\wedge core_{\pi_m}\left(C_j\right)\Big\}\end{aligned} \tag{12}$$

悲观多粒度粗糙集上近似表示为

$$\overline{\sum_{i=1}^{m}}\pi_i^p\left(C_j\right)=\sim\underline{\sum_{i=1}^{m}}\pi_i^p\left(\sim C_j\right) \tag{13}$$

在m次FCM聚类中,如果至少有一次对象x被划分到类簇C_j的下近似中,则对象x属于类簇C_j的乐观多粒度粗糙集下近似. 乐观多粒度粗糙集下近似表示为

$$\begin{aligned}\underline{\sum_{i=1}^{m}}\pi_i^o\left(C_j\right)=\Big\{x\in U\Big|&core_{\pi_1}\left(C_j\right)\vee core_{\pi_2}\left(C_j\right)\vee\cdots\\&\vee core_{\pi_m}\left(C_j\right)\Big\}\end{aligned} \tag{14}$$

乐观多粒度粗糙集上近似表示为

$$\overline{\sum_{i=1}^{m}}\pi_i^o\left(C_j\right)=\sim\underline{\sum_{i=1}^{m}}\pi_i^o\left(\sim C_j\right) \tag{15}$$

根据上述四个近似集合,对每一个类簇构造了一个核心域和三个边界域:

$$\begin{aligned}Co\left(C_j\right)&=\underline{\sum_{i=1}^{m}}\pi_i^p\left(C_j\right)\\Fr_1\left(C_j\right)&=\underline{\sum_{i=1}^{m}}\pi_i^o\left(C_j\right)-\underline{\sum_{i=1}^{m}}\pi_i^p\left(C_j\right)\\Fr_2\left(C_j\right)&=\overline{\sum_{i=1}^{m}}\pi_i^o\left(C_j\right)-\underline{\sum_{i=1}^{m}}\pi_i^o\left(C_j\right)\\Fr_3\left(C_j\right)&=\overline{\sum_{i=1}^{m}}\pi_i^p\left(C_j\right)-\overline{\sum_{i=1}^{m}}\pi_i^o\left(C_j\right)\end{aligned} \tag{16}$$

其中,$Fr_1\left(C_j\right)$为乐观多粒度粗糙集下近似与悲观多粒度粗糙集下近似的差集,$Fr_2\left(C_j\right)$为乐观多粒度粗糙集上近似与乐观多粒度粗糙集下近似的差集,$Fr_3\left(C_j\right)$为悲观多粒度粗糙集上近似与乐观多粒度粗糙集上近似的差集.

四个集合存在偏序关系:

$$Co\left(C_j\right)>Fr_1\left(C_j\right)>Fr_2\left(C_j\right)>Fr_3\left(C_j\right) \tag{17}$$

其表示不同集合中的对象对类簇C_j的贡献度不同,从$Co\left(C_j\right)$中的对象到$Fr_3\left(C_j\right)$中的对象对类簇C_j的贡献度依次降低. 其中$Co\left(C_j\right)$中的对象对C_j的贡献度最高,$Fr_3\left(C_j\right)$中的对象对C_j的贡献度最低,对类簇的影响最小.

3.6 边界域的划分

下面介绍对三个边界域的处理. 使用$Co\left(C_j\right)$中的数据计算类簇中心v_j,类簇中心的计算公式如下

$$v_j=\frac{\sum_{k=1}^{N}x_k}{\left|C_j\right|} \tag{18}$$

对三个边界域中的数据根据其偏序关系依次进行分类. 首先计算$Fr_1\left(C_j\right)$中的对象到类簇中心的隶属度矩阵,然后通过阴影集获得最优的阈值α_1和β_1,根据隶属度与阈值α_1和β_1之间的关系对$Fr_1\left(C_j\right)$中的数据进行分类. 如果隶属度大于α_1则被划分到类簇的核心域中,如果隶属度在α_1和β_1之间则暂不进行划分. 此时将未被划分的对象并入$Fr_2\left(C_j\right)$,重新计算隶属度并获

得新的最优阈值α_2和β_2,根据隶属度与阈值α_2和β_2之间的关系对Fr_1中的数据进行分类,如果隶属度大于α_2则被划分到类簇的核心域中,如果隶属度在α_2和β_2之间则暂不进行划分. 最后,采取同样的方式对$Fr_3(C_j)$中的数据进行划分. 最终得到三支聚类集成的结果,每个类簇C_j均由核心域$Co(C_j)$和边界域$Fr(C_j)$组成. 其中$Co(C_j)$中的对象表示确定属于C_j,$Fr(C_j)$中的对象表示可能属于C_j.

3.7 多粒度三支聚类集成算法描述

算法的第一步通过FCM算法生成多个软聚类结果. 第二步将阴影集引入到FCM,构建了基于阴影集的三支聚类模型,其中类簇的核心域作为类簇的下近似,而核心域和阴影域的并集为上近似. 对每个FCM的聚类结果,均通过阴影集构造下近似和上近似. 第三步引入多粒度粗糙集构建面向聚类集成的两种多粒度粗糙集,即乐观多粒度粗糙集和悲观多粒度粗糙集. 对每个类簇均得到一个核心域$Co(C_j)$和三个边界域$Fr_1(C_j)$,$Fr_2(C_j)$和$Fr_3(C_j)$. 第四步对三个边界域的对象依次进行分类,而无法划分的对象则留在边界域中. 最终所有的对象都被划分到类簇的三个不同的区域,分别是核心域、边界域和琐碎域. 多粒度三支聚类集成的关键步骤如算法1所示.

算法1　Multi-granulation three-way clustering ensemble based on shadowed sets

input: a set of objects $U=\{x_1,x_2,\cdots,x_n\}$.
output: the result of three-way clustering ensemble.
generate a set of clustering members $\Pi=\{\pi_1,\pi_2,\cdots,\pi_m\}$
for π_i in Π:
    calculate the optimal thresholds α_i and β_i according to shadowed sets.
    for x in U:
        if $\mu_{C_i}(x)\geqslant\alpha_i$:
            divide x into $core(C_i)$.
        elseif $\beta_i<\mu_{C_i}(x)<\alpha_i$:
            divide x into $shadowed(C_i)$.
    // determine the lower approximate region and upper approximate region for each cluster in π_i
    $\underline{apr}(C_j)=core(C_j)$
    $\overline{apr}(C_j)=core(C_j)\cup shadowed(C_j)$
for C_i in C:
    $\underline{\sum_{i=1}^{m}\pi_i^p(C_j)}=\bigcap_{i=1}^{m}\underline{apr^{\pi_i^p}}(C_j)$,
    $\overline{\sum_{i=1}^{m}\pi_i^p(C_j)}=\sim\underline{\sum_{i=1}^{m}\pi_i^p}(\sim C_j)$,
    $\underline{\sum_{i=1}^{m}\pi_i^o(C_j)}=\bigcup_{i=1}^{m}\underline{apr^{\pi_i^o}}(C_j)$,
    $\overline{\sum_{i=1}^{m}\pi_i^o(C_j)}=\sim\underline{\sum_{i=1}^{m}\pi_i^o}(\sim C_j)$,
    obtain $Co(C_i)$, $Fr_1(C_i)$, $Fr_2(C_i)$ and $Fr_3(C_i)$.
    for each $Fr_j(C_i)$:
        calculate new centers,
        calculate optimal thresholds,
        divide object in $Fr_j(C_i)$ using the shadowed sets.

4 实验分析

4.1 数据集及评价指标

本文通过8个真实的UCI数据集评估所提出算法的有效性,表1给出了相关数据集的详细信息,包括样本个数、属性数和类别数.

为评估所提出算法的性能,本文通过三种聚类评估指标,分别是准确率[34](ACC)、调整兰德指数[35](ARI)和归一化互信息[1](NMI)对所提出的算法的性能进行分析. 三种有效性度量指标ACC, NMI和ARI均为正向指标,即数值越大,聚类效果越好.

表1　实验数据集

Datasets	Attributes	Samples	Classes
Iris	4	150	3
Wine	13	178	3
Ionosphere	34	351	2
Wdbc	30	569	2
Heart-statlog	13	270	2
Banknote	4	1372	2
Vehicle	18	846	4
Sonar	60	208	2

4.2 对比实验及结果分析

为验证本文所提出的基于阴影集的多粒度三支聚类集成算法的有效性,本文选取了一些具有代表性的聚类集成算法,如经典的Voting[4]算法,基于超图划分的CSPA[1]算法,基于共享近邻的WSNNG[36]算法,基于证据积累的EA[2]算法,基于投票的三支聚类集成算法TWCE[37]和基于粗糙集的增量集成学习算法IFCERS[38].

本实验设置模糊化指数m的取值范围为1.5~2.5. 在数据集的特征子集上运行20次FCM聚类得到20个有差异性的聚类结果作为基聚类并进行集成,该过程重复20次,最终结果取其平均值. 在8个不同的UCI数据集上,计算了所提出的算法与其他6个不同的对比算

法在准确率、调整兰德系数和归一化互信息三个指标上的实验结果,具体结果如表2~4所示.7种不同的聚类集成算法在各个数据集上的聚类结果的最优值均用粗体进行标识,从表2中可以发现在8个不同的数据集上,本文所提出的算法在7个数据集上均获得了最高的聚类准确率,且平均准确率为7种不同的算法的最高值.相应的在表3和表4中可以发现在ARI和NMI两个指标的比较中,同样获得了最优值.因此可以验证本文所提出的算法能够获得更好的聚类集成结果.

表2 不同聚类算法的平均准确率

Datasets	Voting	TWCE	IFCERS	CSPA	EA	WSNNG	MTWCES
Iris	0.8973	0.8973	0.9030	0.8686	0.8809	0.8893	**0.9419**
Wine	0.6853	0.6853	0.6912	0.6460	0.6853	0.6460	**0.8525**
Ionosphere	0.7064	**0.7069**	0.6876	0.6860	0.7061	0.6829	0.6835
Wdbc	0.8541	0.8558	0.8521	0.6910	0.8541	0.6785	**0.8684**
Heart-statlog	0.5962	0.6125	0.5887	0.6103	0.5959	0.6048	**0.6403**
Banknote	0.6048	0.6200	0.6182	0.6335	0.6260	0.6358	**0.6429**
Vehicle	0.4515	0.4514	0.4483	0.3807	0.4514	0.4057	**0.4893**
Sonar	0.5528	0.5536	0.5596	0.5413	0.5528	0.5355	**0.5616**
Average	0.668	0.6728	0.6685	0.6321	0.6690	0.6348	**0.7100**

表3 不同聚类集成算法的调整兰德系数

Datasets	Voting	TWCE	IFCERS	CSPA	EA	WSNNG	MTWCES
Iris	0.7384	0.7384	0.7583	0.6855	0.7196	0.7259	**0.8562**
Wine	0.3539	0.3539	0.3656	0.2679	0.3539	0.2679	**0.6731**
Ionosphere	0.1679	**0.1688**	0.1364	0.1363	0.1674	0.1316	0.1099
Wdbc	0.4914	0.4966	0.4855	0.1453	0.4914	0.1262	**0.5233**
Heart-statlog	0.0333	0.0575	0.0318	0.0453	0.0330	0.0407	**0.0785**
Banknote	0.0432	0.0577	0.0666	0.0715	0.0632	0.0739	**0.0838**
Vehicle	0.1186	0.1187	0.1202	0.0761	0.1187	0.0856	**0.1941**
Sonar	0.0087	0.0090	0.0128	0.0026	0.0085	0.0026	**0.0137**
Average	0.2444	0.2500	0.2471	0.1788	0.2444	0.1818	**0.3165**

表4 不同聚类集成算法的平均归一化互信息

Datasets	Voting	TWCE	IFCERS	CSPA	EA	WSNNG	MTWCES
Iris	0.7580	0.7580	0.7630	0.7052	0.7540	0.7346	**0.8489**
Wine	0.4167	0.4167	0.4330	0.2863	0.4167	0.2863	**0.7461**
Ionosphere	0.1290	**0.1297**	0.0820	0.1156	0.1287	0.1115	0.0603
Wdbc	0.4671	0.4711	0.4628	0.1201	0.4671	0.1038	**0.4931**
Heart-statlog	0.0235	0.0415	0.0226	0.0358	0.0232	0.0324	**0.0529**
Banknote	0.0275	0.0375	0.0465	0.0536	0.0443	0.0553	**0.0558**
Vehicle	0.1784	0.1787	0.1742	0.1118	0.1787	0.1246	**0.2537**
Sonar	0.0105	0.0105	0.0270	0.0024	0.0105	0.0024	**0.0286**
Average	0.2513	0.2554	0.2513	0.1788	0.2529	0.1813	**0.3174**

4.3 基聚类规模与集成质量关系分析

为分析基聚类的规模对聚类集成结果的影响,本文设置了以下实验.在8个不同的UCI数据集上,分析了所提出的MTWCES算法在不同基聚类规模下的集成效果.基聚类规模设置为5,10,15,20,25,30,35和40.在不同的基聚类规模下均进行了20次重复实验,最终结果取其平均值,分析聚类准确率的变化,具体结果如图2所示.从图中可以看出,随着集成规模的不断增加,所提出的算法在除Wine之外的各个数据集上均无较大的波动.随着基聚类规模的增加,Wine数据集的聚类准确率有轻微的上升.但整体来看,集成的质量对基聚类规模的变化并不敏感.因此基聚类规模的变化并不会对MTWCES算法的聚类结果带来显著性的改变.

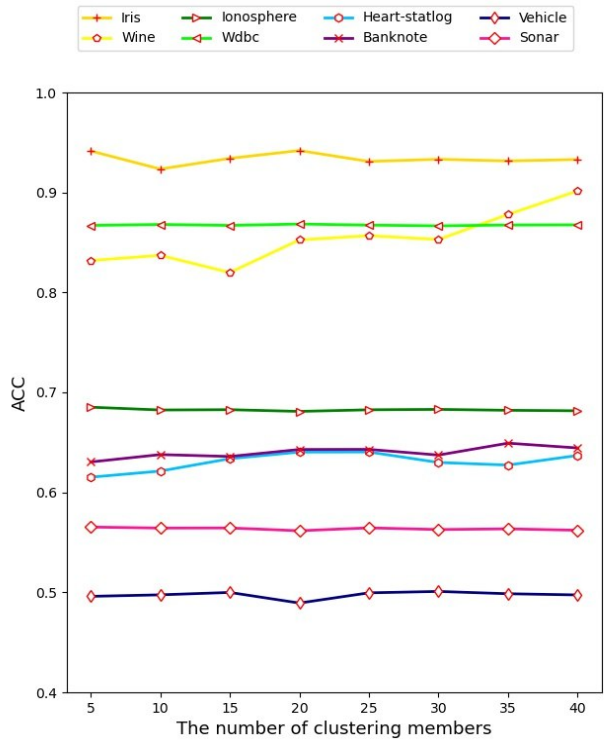

图2　8个数据集上的平均准确率

5　总结

本文提出了一种基于阴影集的多粒度三支聚类集成算法,该算法首先通过FCM聚类产生一组具有差异性的基聚类成员,并通过阴影集构造三支聚类. 然后引入多粒度粗糙集构建了四个近似集合,对每个类簇均获得了一个核心域和三个边界域. 最后使用阴影集对三个边界域中的数据依次进行分类,而无法划分的数据则留在边界域中. 为验证所提出算法的有效性,本文选取了8个UCI数据集进行实验,结果表明所提出的算法在准确率、调整兰德系数和归一化互信息方面,相比于其余六种对比算法取得了更好的性能.

本文所提出的算法中,初始核心域所包含的对象为所有决策者共同满意的对象,即基聚类成员中每一次聚类都被划分到核心域中的对象,如果基聚类中存在个别质量较差的结果,容易导致初始被划分到核心域的数据量较少,因此集成的结果可能会受到基聚类中个别低质量聚类成员的影响. 在未来的工作中对算法的改进主要从以下两个方面考虑,一是通过评估函数来评估基聚类的质量,从而去除一些低质量的基聚类成员. 二是放松约束条件,降低低质基聚类的影响.

参 考 文 献

[1] Strehl A, Ghosh J. Cluster ensembles - a knowledge reuse framework for combining multiple partitions[J]. Journal of Machine Learning Research, 2003, 3(3): 583 – 617.

[2] Fred A L, Jain A K. Combining multiple clusterings using evidence accumulation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(6): 835 – 850.

[3] Topchy A, Jain A K, Punch W F, et al. Clustering ensembles: models of consensus and weak partitions[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(12): 1866 – 1881.

[4] Zhou Z H, Tang W. Clusterer ensemble[J]. Knowledge Based Systems, 2006, 19(1): 77 – 83.

[5] Zhong C M, Hu L Y, Yue X D, et al. Ensemble clustering based on evidence extracted from the co-association matrix [J]. Pattern Recognition, 2019, 92: 93 – 106.

[6] Wang X, Yang C Y, Zhou J, et al. Clustering aggregation by probability accumulation[J]. Pattern Recognition, 2009, 42(5): 668 – 675.

[7] Qian Y H, Liang J Y, Yao Y Y, et al. MGRS: A multi-granulation rough set[J]. Information Sciences, 2010, 180(6): 949 – 970.

[8] Yao Y Y. Tri-level thinking: models of three-way decision [J]. International Journal of Machine Learning and Cybernetics, 2020, 11(5): 947 – 959.

[9] Yao Y Y. Three-way decisions and cognitive computing[J]. Cognitive Computation, 2016, 8(4): 543 – 554.

[10] Yao Y Y. Set-theoretic models of three-way decision[J]. Granular Computing, 2021, 6(1): 133 – 148.

[11] Jiang C M, Yao Y Y. Effectiveness measures in movement-based three-way decisions[J]. Knowledge-Based Systems, 2018, 160: 136 – 143.

[12] Jiang C M, Guo D D, Duan Y, et al. Strategy selection under entropy measures in movement-based three-way decision[J]. International Journal of Approximate Reasoning, 2020,119: 280 – 291.

[13] Jiang C M, Zhao S B. Action strategy analysis in probabilistic preference movement-based three-way decision[J]. Mathematical Problems in Engineering, 2020, 2020: 1 – 13.

[14] Yu H. A framework of three-way cluster analysis[A]. International Joint Conference on Rough Sets[C]. Springer, 2017. 300 – 312.

[15] Yu H, Zhang C, Wang G Y, et al. A tree-based incremental overlapping clustering method using the three-way decision theory[J]. Knowledge-Based Systems, 2016, 91: 189 – 203.

[16] Yu H, Wang X C, Wang G Y, et al. An active three-way clustering method via low-rank matrices for multi-view da-

ta[J]. Information Sciences, 2020, 507: 823 – 839.
[17] Yu H, Chang Z H, Wang G Y, et al. An efficient three-way clustering algorithm based on gravitational search[J]. International Journal of Machine Learning and Cybernetics, 2020, 11(5): 1003 – 1016.
[18] Wang P X, Shi H, Yang X B, et al. Three-way k-means: integrating k-means and three-way decision[J]. International Journal of Machine Learning and Cybernetics, 2019, 10 (10): 2767 – 2777.
[19] Afridi M K, Azam N, Yao J T, et al. A three-way clustering approach for handling missing data using GTRS[J]. International Journal of Approximate Reasoning, 2018: 11 – 24.
[20] Pedrycz W. Shadowed sets: representing and processing fuzzy sets[J]. IEEE Transactions on Systems, Man and Cybernetics, 1998, 28(1): 103 – 109.
[21] Pedrycz W. From fuzzy sets to shadowed sets: Interpretation and computing[J]. International Journal of Intelligent Systems, 2009, 24(1): 48 – 61.
[22] Mitra S, Pedrycz W, Barman B, et al. Shadowed c-means: Integrating fuzzy and rough clustering[J]. Pattern Recognition, 2010, 43(4): 1282 – 1291.
[23] Yue X D, Zhou J, Yao Y Y. Shadowed neighborhoods based on fuzzy rough transformation for three-way classification[J]. IEEE Transactions on Fuzzy Systems, 2020, 28 (5): 978 – 991.
[24] Zhang Q H, Chen Y H, Yang J, et al. Fuzzy entropy: A more comprehensible perspective for interval shadowed sets of fuzzy sets[J]. IEEE Transactions on Fuzzy Systems, 2020, 28(11): 3008 – 3022.
[25] Sun B Z, Ma W M, Xiao X, et al. Three-way group decision making based on multigranulation fuzzy decision-theoretic rough set over two universes[J]. International Journal of Approximate Reasoning, 2017, 81: 87 – 102.
[26] Lin G P, Liang J Y, Qian Y H, et al. An information fusion approach by combining multigranulation rough sets and evidence theory[J]. Information Sciences, 2015, 314: 184 – 199.
[27] Sang B B, Yang L, Chen H M, et al. Generalized multigranulation double-quantitative decision-theoretic rough set of multi-source information system[J]. International Journal of Approximate Reasoning, 2019, 115: 157 – 179.
[28] Ju H R, Li H X, Yang X B, et al. Cost-sensitive rough set: A multi-granulation approach[J]. Knowledge-Based Systems, 2017, 123: 137 – 153.
[29] Wang P X, Yao Y Y. CE3: A three-way clustering method based on mathematical morphology[J]. Knowledge-Based Systems, 2018, 155: 54 – 65.
[30] Yu H, Chen L Y, Yao J T. A three-way density peak clustering method based on evidence theory[J]. Knowledge-Based Systems, 2021, 211: 106532.
[31] Jiang C M, Duan Y, Yao J, et al. Resource-utilization-aware task scheduling in cloud platform using three-way clustering[J]. Journal of Intelligent and Fuzzy Systems, 2019, 37(4): 5297 – 5305.
[32] Zhang Y, Yao J T. Game theoretic approach to shadowed sets: A three-way tradeoff perspective[J]. Information Sciences, 2020, 507: 540 – 552.
[33] Deng X F, Yao Y Y. Decision-theoretic three-way approximations of fuzzy sets[J]. Information Sciences, 2014, 279: 702 – 715.
[34] Huang J, Nie F P, Huang H, et al. Robust manifold nonnegative matrix factorization[J]. ACM Transactions on Knowledge Discovery from Data, 2014, 8(3): 1 – 21.
[35] Lawrence H, Phipps A. Comparing partitions[J]. Journal of Classification, 1985, 2(1): 193 – 218.
[36] Ayad H, Kamel M S. Finding natural clusters using multi-clusterer combiner based on shared nearest neighbors[A]. International Workshop on Multiple Classifier Systems[C]. GER: Springer, 2003. 166 – 175.
[37] Yu H, Zhou Q F. A cluster ensemble framework based on three-way decisions[A]. Rough Sets and Knowledge Technology[C]. GER: Springer, 2013. 302 – 312.
[38] Hu J, Li T R, Luo C, et al. Incremental fuzzy cluster ensemble learning based on rough set theory[J]. Knowledge-Based Systems, 2017, 132: 144 – 155.

作者简介

姜春茂 男,1972年生于黑龙江省哈尔滨市,工学博士,哈尔滨师范大学教授,硕士生导师,主要研究方向为云计算、嵌入式计算、三支决策理论.
E-mail:hsdrose@126.com

赵书宝(通讯作者) 男,1996年生于河南省周口市,硕士研究生,主要研究方向为云计算、机器学习、数据挖掘、三支决策理论.
E-mail:machinelearner@126.com