# 基于阴影集的粗糙聚类阈值选择

郭晋华 苗夺谦 周 杰
(同济大学计算机科学与技术专业 上海 201804)

**摘 要** 粗糙聚类思想自提出以来,在软划分聚类方面取得了广泛应用,但其阈值参数常主观确定,未能考虑数据集本身的特性。基于阴影集(Shadowed Sets)的优化理论给出了一种客观的阈值选择方法,并将其应用于粗糙模糊C均值聚类算法。人工数据与UCI数据实验结果表明了所提方法的有效性。

**关键词** 阴影集,粗糙聚类,粗糙模糊聚类

## Shadowed Sets Based Threshold Selection in Rough Clustering

GUO Jin-hua MIAO Duo-qian ZHOU Jie
(Department of Computer Science and Technology, Tongji University, Shanghai 201804, China)

**Abstract** Rough set-based clustering has been applied widely in soft clustering since proposed, but the threshhold is often given subjectively, fails to consider the characteristics of the data set itself. This study based on shadowed sets optimization theory gave an objective thereshold selection method, and applied it to rough fuzzy C-Means clustering algorithm. Artificial data and UCI data experimental results show the effectiveness of the proposed method.

**Keywords** Shadowed sets, Rough clustering, Rough-fuzzy clustering

## 1 引言

传统的聚类分析是一种硬划分,将每个待识别的对象严格的划分到某个类中,类别界限较分明。实际应用中,大多数对象并没有严格的属性,它们在形态和类属方面存在着一定的过渡,即某个对象对于类并非只有属于和不属于两种状态,适合进行软划分。

粗糙集理论是一种刻画不确定性和不完整性知识的数学工具,由波兰数学家在20世纪80年代初首先提出。粗糙集的核心思想就是用上近似和下近似来近似刻画不确定信息[1]。Lingras[2]最早将粗糙集理论引入到聚类算法中,提出了粗糙C均值聚类算法RCM。该算法将确定属于某类的对象放到该类的下近似中,而将可能属于某类的对象放到该类的上近似中。粗糙C均值聚类算法能很好地处理不确定性对象,已经被广泛应用到了生物信息学等领域。

由于粗糙C均值聚类算法只应用距离差来划分聚类,没有考虑全局性,于是Mitra[3]最早将模糊聚类的隶属度引入粗糙C均值中,提出了粗糙模糊C均值聚类算法RFCM。粗糙模糊C均值对边界元素的分辨率相对粗糙C均值高。

但是这些算法在划分聚类的时候都需要人为给定一个阈值,为了选取一个比较合适的阈值,一般都要经过多次实验试探,这样就增加了算法的代价,效率不高。基于这一点,本文提出基于阴影集的动态阈值确定方法。

## 2 粗糙聚类算法

### 2.1 粗糙C均值聚类

在粗糙集中,集合 $X$ 关于属性集 $R$ 的下近似被定义为式(1),表示肯定属于 $X$ 的对象集合;上近似被定义为式(2),表示可能属于 $X$ 的对象集合。

$$\underline{R}(X)=\bigcup\{[X]_R\mid[X]_R\subseteq X\}=\{x\in U\mid[X]_R\subseteq X\} \quad (1)$$

$$\overline{R}(X)=\bigcup\{[X]_R\mid[X]_R\cap X\neq\varnothing\}=\{x\in U\mid[X]_R\cap X\neq\varnothing\} \quad (2)$$

粗糙集的核心思想就是用上近似和下近似来描述不确定事物[1],这种性质很适合进行软划分。Lingras将粗糙集理论引入到聚类算法中,把基于划分的经典K-均值算法扩展为粗糙C均值聚类算法,并且规定上近似与下近似满足以下性质[2]。

(I) 一个对象最多只能属于一个类的下近似;

(II) 如果一个对象属于某个类的下近似,那么它也属于这个类的上近似;

(III) 如果一个对象不属于任何类的下近似,那么它属于两个或两个以上类的上近似。

设数据集 $X=\{x_1,x_2,\cdots,x_n\}$,$\overline{C}_i$ 和 $\underline{C}_i$ 分别代表类 $C_i$ 的上近似集合和下近似集合,$C_i^B$ 代表类 $C_i$ 的边界区域集合,$C_i^B=\overline{C}_i-\underline{C}_i$。$\underline{C}_i$ 中的对象一定属于类 $C_i$,$\overline{C}_i$ 中的对象可能属于类 $C_i$。$C_i$ 的均值按式(3)进行计算:

$$V_i=\begin{cases}\omega_{low}\dfrac{\sum\limits_{x_k\in\underline{C}_i}x_k}{|\underline{C}_i|}+\omega_{up}\dfrac{\sum\limits_{x_k\in C_i^B}x_k}{|C_i^B|}, & \text{若}\underline{C}_i\neq\varnothing\wedge C_i^B\neq\varnothing\\ \dfrac{\sum\limits_{x_k\in\underline{C}_i}x_k}{|\underline{C}_i|}, & \text{若}\underline{C}_i\neq\varnothing\wedge C_i^B=\varnothing\\ \dfrac{\sum\limits_{x_k\in C_i^B}x_k}{|C_i^B|} & \text{若}\underline{C}_i=\varnothing\wedge C_i^B\neq\varnothing\end{cases} \quad (3)$$

到稿日期:2011-04-11 返修日期:2011-07-01 本文受国家自然科学基金项目(60970061,61085056)资助。

郭晋华(1986-),女,硕士,主要研究方向为数据挖掘、聚类分析,E-mail:lantianqingqinghua@163.com;苗夺谦(1964-),教授,博士生导师,主要研究方向为粗糙集理论、模式识别、人工智能等;周 杰(1982-),男,博士生,主要研究方向为粗糙集理论、数据挖掘等。

式中，$|\underline{C}_i|$是类$C_i$中下近似集合的对象个数，$|C_i^B|$是类$C_i$的边界区域的对象个数。$\omega_{low}$是下近似的权重，$\omega_{up}$是上近似的权重，且$\omega_{low}+\omega_{up}=1$。聚类过程中每个类的下近似对该类具有重要的贡献，故$\omega_{low}$值一般较大。

**算法 1** 粗糙 C 均值聚类算法

Step 1 给定$\omega_{up}$、$\omega_{low}$、阈值$\varepsilon$、聚类数$C$，初始化聚类中心$V$；

Step 2 令$d_{ik}$表示对象$x_k$到所有类的最小距离，$d_{jk}$表示次小距离，如果$d_{jk}-d_{ik}\leqslant\varepsilon$，那么$x_k\in\overline{C}_i$并且$x_k\in\overline{C}_j$，否则$x_k\in\underline{C}_i$；

Step 3 使用式(3)重新计算聚类中心$V$；

Step 4 重复 Step 2—Step 3 直到算法收敛。

### 2.2 粗糙模糊 C 均值聚类

粗糙 C 均值聚类算法只应用距离差来划分聚类，没有考虑全局性，Mitra[3]将模糊 C 均值[4]中的隶属度概念引入到粗糙 C 均值算法中，聚类中心的计算变为式(4)：

$$V_i=\begin{cases}\omega_{low}\dfrac{\sum\limits_{x_k\in\underline{C}_i}\mu_{ik}^m x_k}{\sum\limits_{x_k\in\underline{C}_i}\mu_{ik}^m}+\omega_{up}\dfrac{\sum\limits_{x_k\in C_i^B}\mu_{ik}^m x_k}{\sum\limits_{x_k\in C_i^B}\mu_{ik}^m}, & \text{若}\underline{C}_i\neq\varnothing\wedge C_i^B\neq\varnothing\\ \dfrac{\sum\limits_{x_k\in\underline{C}_i}\mu_{ik}^m x_k}{\sum\limits_{x_k\in\underline{C}_i}\mu_{ik}^m}, & \text{若}\underline{C}_i\neq\varnothing\wedge C_i^B=\varnothing\\ \dfrac{\sum\limits_{x_k\in C_i^B}\mu_{ik}^m x_k}{\sum\limits_{x_k\in C_i^B}\mu_{ik}^m}, & \text{若}\underline{C}_i=\varnothing\wedge C_i^B\neq\varnothing\end{cases}\tag{4}$$

式中，隶属度$\mu_{ik}$的计算采用模糊 C 均值算法中的方式，其描述如式(5)：

$$\mu_{ij}=\left(\sum_{k=1}^{c}\left(\frac{d_{ij}}{d_{kj}}\right)^{\frac{2}{m-1}}\right)^{-1}\tag{5}$$

式中，$d_{ij}^2=\|x_j-v_i\|^2$。

**算法 2** 粗糙模糊 C 均值聚类算法

Step 1 初始化聚类中心$V$；

Step 2 对于所有数据对象和所有类利用式(5)计算隶属度；

Step 3 令$\mu_{ik}$表示对象$x_k$到所有类的最大的隶属度，$\mu_{jk}$表示次大的隶属度，如果$\mu_{ik}-\mu_{jk}\leqslant\varepsilon$，那么$x_k\in\overline{C}_i$并且$x_k\in\overline{C}_j$，否则$x_k\in\underline{C}_i$；

Step 4 使用式(4)重新计算聚类中心$V$；

Step 5 重复 Step 2—Step 4 直到算法收敛。

## 3 基于阴影集的粗糙聚类算法

本节介绍基于阴影集的粗糙聚类算法，依靠阴影集来动态确定划分聚类的阈值[5]。

### 3.1 阴影集

阴影集的概念是 Pedrycz[5]提出来的。给定模糊集$f(X)$，其阴影集的结构如图 1 所示。如果$f(x)>1-\alpha$，则令$f(x)=1$；如果$f(x)<\alpha$，则令$f(x)=0$，这样就相当于将数据集$X$映射到 0，1 和[0,1]，即$f:X\to\{0,1,[0,1]\}$，我们称之为阴影集。阴影集和粗糙集有相似之处，$f(x)=0$的数据对象对应着粗糙集中的负域，$f(x)=1$的数据对象对应于下近似，$f(x)=[0,1]$对应于上近似。阴影集是连接模糊集与粗糙集的桥梁。

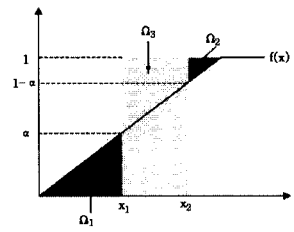


图 1 阴影集的结构

$f(x)$减小的区域：

$$\Omega_1=\int_{x;f(x)<\alpha}f(x)\mathrm{d}x\tag{6}$$

$f(x)$变大的区域：

$$\Omega_2=\int_{x;f(x)>1-\alpha}(1-f(x))\mathrm{d}x\tag{7}$$

阴影形成的区域：

$$\Omega_3=\int_{x;\alpha\leqslant f(x)\leqslant 1-\alpha}\mathrm{d}x\tag{8}$$

阴影集中的$\alpha$值通过最小化目标函数(9)来确定，最优的$\alpha$值使目标函数(9)取到最小值。

$$\nu(\alpha)=|\Omega_1+\Omega_2-\Omega_3|,\alpha\in[0,0.5)\tag{9}$$

### 3.2 基于阴影集的粗糙模糊 C 均值聚类

将所有数据对象到每个类的隶属度$\mu_i(i=1,2,\cdots,C)$当成模糊集，利用阴影集来确定阈值$\alpha_i$。

**算法 3** 基于阴影集的粗糙模糊 C 均值聚类算法

Step 1 初始化聚类中心$V$；

Step 2 利用式(5)计算隶属度$\mu_i(i=1,2,\cdots,C)$；

Step 3 利用阴影集最优化目标函数的方法，对于$\mu_i(i=1,2,\cdots,C)$确定最优的$\alpha_i$，$\alpha_i=\mathrm{argmin}(\nu_i)$，其中

$$\nu_i=\Big|\sum_{k;\mu_{ik}<\alpha}\mu_{ik}+\sum_{k;\mu_{ik}>1-\alpha}(1-\mu_{ik})-\mathrm{card}\{x_k|\alpha\leqslant\mu_{ik}\leqslant 1-\alpha\}\Big|\tag{10}$$

Step 4 根据$\alpha_i$的值来确定类的下近似区域和上近似区域：

$$\underline{C}_i=\{x_k|\mu_{ik}>1-\alpha_i\}\tag{11}$$

$$\overline{C}_i=\{x_k|\mu_{ik}>\alpha_i\}\tag{12}$$

所以边界区域

$$C_i^B=\overline{C}_i-\underline{C}_i=\{x_k|\alpha_i\leqslant\mu_{ik}\leqslant 1-\alpha_i\}\tag{13}$$

Step 5 利用式(4)更新聚类中心；

Step 6 重复 Step 2—Step 5 直到算法收敛。

## 4 实验

### 4.1 人造数据集

人造数据是一个二维的数据集，具体如图 2 所示。实验中的模糊因子$m=2$，下近似权重$\omega_{low}=0.90$，相应的上近似权重$\omega_{up}=1-\omega_{low}=0.10$。其实验比较结果如表 1 所列，其中 PBM[6]的值越大，DB[7]和 XB[8]的值越小，聚类结果越好。实验结果如图 3—图 6 所示。

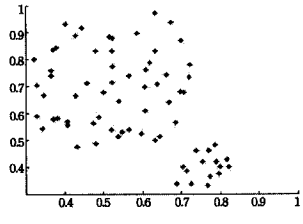

图 2 人造数据集

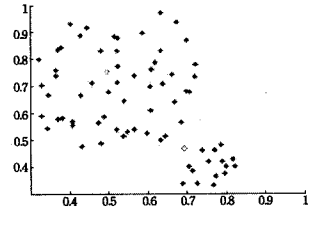

图 3 FCM

(下转第 227 页)

chical quantitative attributes[J]. Expert Systems with Applications,2009,36(3):6790-6799

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

[12] Zhou Lei,Wu Wei-zhi. On generalized intuitionistic fuzzy rough approximation operators[J]. Information Sciences, 2008, 178(11):2448-2465

[13] Zhou Lei,Wu Wei-zhi. On characterization of intuitionistic fuzzy rough sets based on intuitionistic fuzzy implicators [J]. Information Sciences,2009,179(7):883- 898

[14] 徐小来,雷英杰,谭巧英. 基于直觉模糊三角模的直觉模糊粗糙集[J]. 控制与决策,2009,23(8):900-904

[15] 樊雷,雷英杰. 基于直觉模糊粗糙集的属性约简研究[J]. 计算机工程与科学,2008,30(7):79-81

[16] 黄兵,胡作进,周献中. 优势模糊粗糙模型及其在审计风险评估中的应用[J]. 控制与决策,2009,24(6):899- 902

[17] 黄兵,周献中,史迎春. 优势-模糊目标 VPRSM 及应用[J]. 计算机科学,2010,37(3):227-229,241

(上接第 210 页)

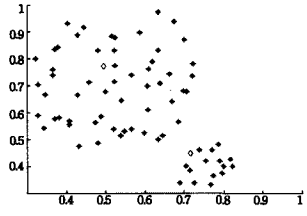

图 4 RCM

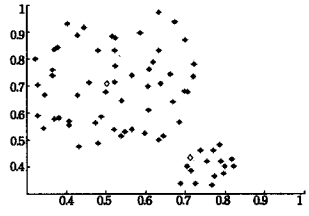

图 5 RFCM

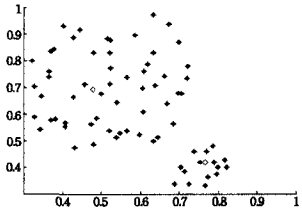

图 6 SRFCM

从图 3—图 6 中的聚类中心可以直观地得到各种算法的聚类效果,很明显基于阴影集的粗糙模糊 C 均值算法的聚类中心最真实,所以聚类效果最好。模糊 C 均值算法的聚类中心偏离得较厉害。

评价指标的计算如表 1 所列,3 种评价指标都显示基于阴影集的粗糙模糊 C 均值算法的聚类效果最好,这与从图 3—图 6 分析得到的结论一致。

表 1 评价指标

| 算法 | PBM | DB | XB |
|---|---|---|---|
| FCM | 3.485372 | 0.910035 | 0.183324 |
| RCM | 3.570061 | 0.894452 | 0.169025 |
| RFCM | 3.846634 | 0.748903 | 0.157743 |
| SRFCM | **4.370158** | **0.696433** | **0.139605** |

### 4.2 UCI 数据集

本文采用 UCI[9] 中的 Wine,Balance 和 Ionosphere 3 个数据集来进行对比实验,实验结果如表 2—表 4 所列。

表 2 Wine($C$=3)

| 算法 | PBM | DB | XB |
|---|---|---|---|
| FCM | 1.980322 | 3.076901 | 7.094532 |
| RCM | 2.531437 | 2.795603 | 2.906314 |
| RFCM | 3.811251 | 1.475322 | 1.657411 |
| SRFCM | **4.217045** | **1.036742** | **1.225482** |

表 3 Balance($C$=3)

| 算法 | PBM | DB | XB |
|---|---|---|---|
| FCM | 0.001333 | 41.648942 | 110.648942 |
| RCM | 1.037491 | 1.961526 | 0.506262 |
| RFCM | 1.112036 | 1.658627 | 0.357915 |
| SRFCM | **1.200132** | **1.319002** | **0.245924** |

表 4 Ionosphere($C$=2)

| 算法 | PBM | DB | XB |
|---|---|---|---|
| FCM | 0.424221 | 2.468362 | 1.031237 |
| RCM | 0.723741 | 1.521138 | 0.613359 |
| RFCM | 0.932734 | 1.473451 | 0.513922 |
| SRFCM | **1.013672** | **1.420057** | **0.392257** |

从表 2—表 4 可以得出,基于阴影集的粗糙模糊 C 均值算法的聚类结果得到了最大的 PBM 值、最小的 DB 值和 XB 值,3 种评价指标都显示基于阴影集的粗糙模糊 C 均值算法具有最好的聚类效果,体现出了本文基于阴影集动态确定划分聚类阈值的优点。

**结束语** 本文针对粗糙聚类算法中划分近似区域的阈值参数确定进行了探讨,提出了基于阴影集理论动态确定阈值的粗糙模糊聚类算法。采用 UCI 数据与人造数据进行实验分析,通过 PBM、DB 和 XB 等评价指标的对比,说明了本文所提算法的有效性。所提算法的理论分析将是未来主要研究工作之一。

## 参 考 文 献

[1] Pawlak Z. Rough sets[J]. International Journal of Information Computer Sciences,1982,11:145-172

[2] Lingras P, West C. Interval set clustering of web users with rough $k$-means[J]. Journal of Intelligent Information Systems, 2004,23(1):5-16

[3] Mitra S,Banka H,Pedrycz W. Rough-Fuzzy collaborative clustering[J]. IEEE Transactions on Systems,Man,and Cybernetics-Part B:Cybernetics,2006,36(4):795-805

[4] Bezdek J C. Pattern recognition with fuzzy objective function algorithms[M]. New York:Plenum,1981

[5] 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

[6] Pakhira M K, Bandyopadhyay S, Maulik U. Validity index for crisp and fuzzy clusters[J]. Pattern Recognition,2004,37:487-501

[7] Davies D L, Bouldin D W. A cluster separation measure[J]. IEEE Trans,Pattern Anal. Mach. Intell. ,1979,1:224-227

[8] Xie X L, Beni G. A validity measure for fuzzy clustering[J]. IEEE Trans. Pattern Anal. Mach. Intell,1991,13(8):841-846

[9] Blake C L,Merz C J. UCI repository of learning databases[OL]. http://www.ics.uci.edu/~mlearn/ MLRepository.html