一种建立粗糙数据模型的监督模糊聚类方法*

黄金杰[1+], 李士勇[2], 蔡云泽[1]

[1](上海交通大学 自动化系,上海 200030)

[2](哈尔滨工业大学 控制科学与工程系,黑龙江 哈尔滨 150001)

An Approach to Building Rough Data Model Through Supervised Fuzzy Clustering

HUANG Jin-Jie[1+], LI Shi-Yong[2], CAI Yun-Ze[1]

[1](Department of Automation, Shanghai Jiaotong University, Shanghai 200030, China)

[2](Department of Control Science and Engineering, Harbin Institute of Technology, Harbin 150001, China)

+ Corresponding author: Phn: +86-21-62826946, E-mail: huangjinjie@sjtu.edu.cn, http://www.sjtu.edu.cn

Received 2003-07-28; Accepted 2004-09-08

Huang JJ, Li SY, Cai YZ. An approach to building rough data model through supervised fuzzy clustering. *Journal of Software*, 2005,16(5):744−753. DOI: 10.1360/jos160744

Abstract: A new method for fast building the rough data model (RDM) by means of supervised fuzzy clustering in the product space of input and output variables is proposed. The approach incorporates the RDM's classification quality performance index with Gustafson-Kessel (GK) clustering algorithm and is of many good properties. The way to convert the fuzzy cluster models to rough data models by introducing the concept of putative membership degree of a data point to a fuzzy cluster is suggested. Hence, an efficient algorithm that can obtain RDMs by just iteratively computing two necessary condition equations is worked out. It minimizes the objective function and turns the multi-dimensional search process of the Kowalczyk's method to one dimensional search strategy (in terms of the number of clusters). This technique reduces the searching time greatly. Compared with the traditional rough set theory and the Kowalczyk's method, the approach has more powerful ability to handle data contaminated by noise and better generalization ability. Finally, different examples of data sets illustrate the effectiveness of the approach.

Key words: rough data model; rough set; supervised fuzzy clustering; Gustafson-Kessel algorithm; putative membership degree

摘 要: 提出了在输入-输出积空间中利用监督模糊聚类技术快速建立粗糙数据模型(rough data model,简称

* Supported by the National Grand Fundamental Research 973 Program of China under Grant No.2002cb312200 (国家重点基础研究发展规划(973)); the Natural Science Foundation of Heilongjiang Province of China under Grant No.F0316 (黑龙江省自然科学基金); the China Postdoctoral Science Foundation under Grant No.2004036321 (中国博士后科学基金)

作者简介: **黄金杰**(1967−),男,山东烟台人,博士,副教授,主要研究领域为粗糙集理论及应用,复杂系统智能建模与优化控制; **李士勇**(1943−),男,教授,博士生导师,主要研究领域为模糊控制,智能控制; **蔡云泽**(1975−),女,博士,讲师,主要研究领域为时滞系统的鲁棒控制,鲁棒滤波和小波滤波,信息融合.

RDM)的一种方法.该方法将RDM模型的分类质量性能指标与具有良好特性的Gustafson-Kessel(G-K)聚类算法结合在一起,并通过引入数据对模糊类的推定隶属度的概念,给出了将模糊聚类模型转化为粗糙数据模型的方法,从而设计出一种通过迭代计算使目标函数最小的两个必要条件方程来获取 RDM 模型的有效算法,将Kowalczyk方法的多维搜索过程变为以聚类数目为参数的一维搜索,极大地减少了寻优时间.与传统的粗糙集理论和 Kowalczyk 方法相比,提出的方法具有更好的数据概括能力和噪声数据处理能力.最后,通过不同的数据集实验测试,结果表明了该方法的有效性.

关键词: 粗糙数据模型;粗糙集;监督模糊聚类;GK 算法;推定隶属度

中图法分类号: TP18 **文献标识码**: A

粗糙集理论[1]作为一种研究数据表达、学习、归纳的理论方法,近年来得到各领域广泛的关注和应用,成为数据挖掘、知识发现、不确定性推理、概念自适应学习(adaptive learning of concepts)、粒计算(granular computing)等研究的一个有力工具[2−7].粗糙集研究的一个中心问题是分类分析[8].Pawlak 粗糙集模型由于是严格按照等价关系来分类的,所以它所处理的分类必须是精确的或称完全包含的[8].但在实际中,数据集经常含有噪声和不确定性,结果导致 Pawlak 模型的分类效果不佳,甚至失败[9].

为了解决这一问题,人们提出了粗糙集方法的各种扩展形式,其中将 Pawlak 粗糙集理论中的等价关系扩展为相似关系(similarity relation)[10]、覆盖(covers)[11]和一般关系(general relation)[12,13]等是最主要的形式.从集合的角度,Ziarko 引入了变精度粗糙集模型(VPM)[14],将完全精确的包含关系“软化”为某种程度上的包含,即如果等价类 C 中至少有$(1-\beta)$%的元素属于子集 X,则称类 C β-包含于 X,$0\leq\beta<0.5$.通过改变β,得到 X 的各种不同精度的近似集.Kowalczyk 从另一角度提出了粗糙数据模型(rough data model,简称 RDM)[15]的概念,试图通过建立论域中的等价类到决策类之间的映射关系 $type:C\rightarrow x$ 来描述某一子集概念 X.定义的映射关系 $type$ 不同,所得到的 X 的近似集也不同.这两种方法都取得了非常成功的应用,特别地,RDM 模型以其相对简单的算法和很高的模型质量而在 EUFIT’96 国际会议上的竞赛题目中赢得金奖.

然而,VPM 和 RDM 仍然具有局限性.首先,两种模型所涉及的概念和知识都是清晰的,而在实际问题中,人们涉及更多的往往是一些模糊概念和模糊知识,VPM 和 RDM 都难以有效地处理.其次,寻求最优的 VPM 或 RDM 模型是一个典型的 NP-hard 或 NP-full 问题,目前尚缺乏一种统一而有效的求解方法,大多仍采用某种复杂的启发式算法.而另一方面我们注意到,模糊集理论通过取值在[0,1]上的模糊隶属函数,能够对论域中任意不精确的概念进行精细的描述.因此,将粗糙集与模糊集相结合,以寻求更强的不确定性模型,已成为一个重要的研究方向[16,17].本文在 VPM 和 RDM 的基础上,提出一种在输入-输出积空间中通过有监督的 G-K 模糊聚类算法快速建立 RDM 模型的有效方法,将最优 RDM 模型的求解过程转化为分类目标函数的寻优过程,通过交替计算分类目标函数的两个必要条件等式,把 Kowalczyk 方法的多维搜索过程转变为以聚类数目 c 为参数的一维搜索,极大地减少了搜索时间,避免了 NP-hard 问题的求解;同时,通过引入输入数据对模糊类的推定隶属度 PM(putative membership degree)的概念,实现了从模糊聚类模型到清晰的 RDM 模型的转换,能够更好地描述论域中的不精确概念.特别地,由于采用了 G-K 聚类方法,本文所建立的 RDM 模型能够用较少数目的模糊类反映出数据集中具有超椭球、超平面或超线型的特征模式类,而且其搜索过程是与各模式类的不同形状动态相适应的,具有很强的数据概括能力,克服了 Kowalczyk 的 RDM 模型概括能力严重依赖于其对论域空间事先的网格划分的缺点.据此,我们在 Matlab6.1 环境下成功地开发了一套应用软件系统——SFRDMS(supervised fuzzy rough data model),该软件在初步的应用中已表现出良好的性能.

1 粗糙数据模型

考虑信息系统 $S=(U,C\cup d)$,其中,U 为论域对象的非空有限集合,C 和 d 分别为条件属性集和决策属性,且 $C\cap d=\varnothing$ 对于 $\forall a\in C\cup d$,有 $a:U\rightarrow V_a$,其中 V_a 是属性 a 的值集.设 R 是由条件属性 C 在论域 U 上定义的一个等价关系,R 将论域 U 中的元素划分成各不相交的等价类:$U/R=\{q_1,q_2,...,q_n\}$.

在 Kowalczyk 的粗糙数据模型(KRDM)[15]中,输入空间中同一等价类 q_i $(i=1,2,...,n)$ 中的元素可能属于决策空间中不同的决策类 y_j $(j=1,2,...,p)$.但是由于输入空间中同一等价类中的各个元素相互之间是不可分辨的,所以不得不将它们划归为同一决策类.这种输入空间与决策空间之间人为的对应关系,被定义为 Kowalczyk RDM 中的类型(type)函数,*type*: $q \to y$.KRDM 就是通过 *type* 函数以及一些相应的统计量来刻画决策空间各个决策类的,这些统计量主要包括:

- 类大小 *size*(q_i):输入空间中等价类 q_i 中的元素个数,*size*(q_i)=$|q_i|$;
- 类中正确划分元素数 *corr*(q_i):等价类 q_i 中属于决策类 *type*(q_i)的元素个数,即 $corr(q_i)=|q_i \cap type(q_i)|$;
- 类划分精度: $acc(q_i)=\dfrac{corr(q_i)}{size(q_i)}$;
- 累积模型精度: $cma[\%]=\left(\sum_{i=1}^{n} corr(q_i)\right)\Big/|U|$.

Table 1 A basic form of RDM
表 1 RDM 模型的基本形式

U/R	*type*(q)	*acc*[%]
q_1	y_1	acc_1
q_2	y_2	acc_2
⋮	⋮	⋮
q_n	y_p	acc_n

这些统计量的数值被用于按照一些用户确定的标准来对类 q_i $(i=1,2,...,n)$ 进行排序,于是信息表 $S=(U,C \cup d)$ 的粗糙数据模型(RDM)被定义为一个如下的三元组[15]:

$$T=\langle q, type, \leq \rangle \tag{1}$$

其中≤表示定义在 $q=\{q_1,q_2,...,q_n\}$ 上的一个线性的序关系.一个基本的 RDM 模型形式见表 1,其中 $acc_1 \geq acc_2 \geq ... \geq acc_n$.

在 RDM 模型中,用户可以根据实际问题的不同而定义不同的类型函数 *type*(·),从而更为灵活地完成分类.同时,RDM 模型形式简单,内部的划分情况一目了然,非常实用.

应用 RDM 模型的一般方法(称其为 Kowalczyk 方法)如下[15]:给定一个数据集及 RDM 模型的性能指标函数.首先,选用某种离散化方法(如等频率法等)将数据集的每一个连续属性离散化为较多数目的小区间;然后,确定用于建模的属性子集及子集中各个属性使用的离散值个数(数目较少),其离散区间是利用从小区间的分点中选取适当数目的分点获取的;最后,计算所有可能的 RDM 模型及其性能指标,从中选择最优的一个.显然,这种穷尽式的搜索策略是非常耗时的.引入诸如遗传算法(GA)等启发式搜索算法是在一定程度上解决此问题的一种有效途径[18].

2 利用 G-K 模糊聚类构造粗糙数据模型

Kowalczyk 的 RDM 模型(KRDM)是建立在对数据集论域空间事先的网格状划分基础上的,其模型分类精度取决于对数据集论域空间预先划分的精细程度.从本质上说,KRDM 是对数据空间的“硬划分”,界限分明,是有“粒度”的,无法很好地处理模糊概念.同时,其模型的数量将随着属性及属性离散区间的数目而呈指数级增加,产生了很大的计算负担.为了克服这些弱点,我们考虑采用模糊聚类技术来构造 RDM 模型.

2.1 Gustafson-Kessel(G-K)聚类算法

模糊聚类是一种辨识数据模式的重要工具,Gustafson-Kessel(G-K)算法是距离自适应动态聚类算法(adaptive distance dynamic clustering algorithm)的模糊推广,它可以有效地搜索超椭球、平面或线型的数据类[19].

在 G-K 算法中,n 维数据空间中点 x_k 到聚类中心 v_i 的距离是一个平方内积距离范数:

$$D^2_{ikM_i}=\|x_k-v_i\|^2_{M_i}=(x_k-v_i)^T M_i (x_k-v_i) \tag{2}$$

其中,$M_i=\det(F_i)^{1/n} F_i^{-1}$,$F_i$ 是第 i 个聚类中心的协方差矩阵,为正定对称矩阵.将数据集$\{x_1,...,x_N\}$划分成 c 个模糊类是通过最小化目标函数

$$J(X;U,V)=\sum_{i=1}^{c}\sum_{k=1}^{N}(u_{ik})^m \|x_k-v_i\|^2_{M_i} \tag{3}$$

来完成的,其中,U=$[u_{ik}]$是数据集的模糊划分矩阵,且满足

$$\sum_{i=1}^{c} u_{ik} = 1, \quad 1 \le k \le N, \ u_{ik} \in [0,1] \tag{4}$$

$m \in [1,\infty]$为一个加权指数,决定着所得分类的模糊程度(对于清晰模型,m=1;模糊模型 m>1;通常 m=2).Lagrange 乘子λ_k可以将目标函数(3)及其约束(4)转化成新的目标函数

$$\bar{J}(X;U,V,\lambda) = \sum_{i=1}^{c}\sum_{k=1}^{N}(u_{ik})^m D_{ikm_i}^2 + \sum_{k=1}^{N}\lambda_k\left[\sum_{i=1}^{c} u_{ik} - 1\right] \tag{5}$$

设 $D_{ikM_i}^2 > 0, \forall i,k$ 及 m>1,令 $\bar{J}$ 关于 U,V 和λ的梯度为 0,则可求得使式(3)取极小值的两个必要条件:

$$u_{ik} = \frac{1}{\sum_{j=1}^{c}(D_{ikM_i}/D_{jkM_j})^{2/(m-1)}}, 1 \le i \le c, 1 \le k \le N \tag{6}$$

$$v_i = \frac{\sum_{k=1}^{N}(u_{ik})^m x_k}{\sum_{k=1}^{N}(u_{ik})^m}, \quad 1 \le i \le c \tag{7}$$

所以,G-K 算法可以形式化如下:给定数据集 $\{x_k \mid k = 1,2,...,N\}$,选择分类数目 1<$c$<$N$,加权指数 m>1 和终止容许误差ε>0.随机初始化模糊划分矩阵 U,并使之满足式(4).对于 l=1,2,…,重复以下步骤:

Step 1.计算聚类中心:

$$v_i^{(l)} = \frac{\sum_{k=1}^{N}(u_{ik}^{(l-1)})^m x_k}{\sum_{k=1}^{N}(u_{ik}^{(l-1)})^m}, \quad 1 \le i \le c.$$

Step 2.计算类协方差矩阵:

$$F_i = \frac{\sum_{k=1}^{N}(u_{ik}^{(l-1)})^m (x_k - v_i^{(l)})(x_k - v_i^{(l)})^T}{\sum_{k=1}^{N}(u_{ik}^{(l-1)})^m}, \quad 1 \le i \le c.$$

Step 3.计算距离:

$$M_i = \det(F_i)^{1/n} F_i^{-1},$$

$$D_{ikM_i}^2 = (x_k - v_i^{(l)})^T M_i (x_k - v_i^{(l)}), 1 \le i \le c, \quad 1 \le k \le N.$$

Step 4.更新模糊划分矩阵:

$$u_{ik}^{(l)} = \frac{1}{\sum_{j=1}^{c}(D_{ikM_i}/D_{jkM_j})^{2/(m-1)}}, \quad 1 \le i \le c, 1 \le k \le N.$$

直到 $\left\|U^{(l)} - U^{(l-1)}\right\| < \varepsilon$ 终止.

2.2 模糊聚类在建立RDM中的应用

建立一个最优的 RDM 模型,根本上需要解决以下 3 个问题:

(1) 数据空间的划分

利用 G-K 模糊聚类对数据空间进行划分建立 RDM 模型,需要将 RDM 的分类质量指标融入 G-K 聚类算法的目标函数之中.试图将 RDM 的 *cma*[%]指标直接加到 G-K 算法的目标函数上,以获取使目标函数取极值的必要条件将是很困难的,这是由于 *cma*[%]引入了基于元素个数的自变量而使得目标函数的有关梯度难以计算.因此,我们不妨把类别数据作为输出变量,把其余的数据分量作为输入变量,首先研究一下在此输入-输出的笛卡尔积空间中对应于不同分类质量的 RDM 的模糊聚类的特性.

设 $X \in R^n$ 为输入数据向量,$y \in R$ 为类别数据,即输出变量,通常取值为整数.记 Z_k=[X_k^T ,y_k]^T,k 表示第 k 个数据点.

命题 1. 高质量的 RDM 模型意味着其每一类的分类精度都很高,即 $acc(C_i)$[%]≈100%.在聚类结果上表现为:

(i) 每一类 C_i 中各数据点 $z_{ik} = [x_{ik}^T, y_{ik}]^T$ 的类别数值 y_{ik} 的均值几乎等于其聚类中心 V_i=[v_{i1},v_{i2},…,$v_{i,n+1}$]^T 的

类别分量 $v_{i,n+1}$,i=1,2,…,c.

(ii) 在类 C_i 中,其各数据点 $z_{ik}=[x_{ik}^T, y_{ik}]^T$ 的类别数值 y_{ik} 几乎都相等.这使得 C_i 类的模糊协方差矩阵 F_i 具有如下的形式:

$$F_i=\begin{bmatrix} * & \cdots & * & 0 \\ \vdots & \vdots & \vdots & \vdots \\ * & \cdots & * & 0 \\ 0 & \cdots & 0 & 0 \end{bmatrix}_{(n+1)\times(n+1)} \tag{8}$$

即(a) C_i 中的类别变量 y_i 与输入数据各分量 x_{ij}(j=1,…,n)的协方差接近于 0,cov(x_{ij},y_i)≈0;(b) C_i 中的类别变量的方差接近于 0,

$$D(y_i)=cov(y_i,y_i)=F_i(n+1,n+1)\approx 0.$$

于是,高质量 RDM 模型对应的各分类 $C_i\,(1\le i\le c)$ 的高斯(Gaussian)型隶属函数是一些中心为聚类中心类别分量 $v_{i,n+1}$ 的"窄脉冲",如图 1 所示.

命题 2. 低质量的 RDM 模型意味着其大多数类的分类精度较低,即 $acc(C_i)$[%]<<1.在聚类结果上表现为:

(i) 大部分聚类中心 $V_i=[v_{i1},v_{i2},\ldots,v_{i,n+1}]^T$ 的类别分量 $v_{i,n+1}$ 与其类中多数数据点 $z_{ik}=[x_{ik}^T, y_{ik}]^T$ 的类别数值 y_{ik} 相差较大;

(ii) 同一类 C_i 中的各数据点 $z_{ik}=[x_{ik}^T, y_{ik}]^T$ 的类别数值 y_{ik} 相差较大,即 C_i 中的许多点实际上属于不同的类别.

这意味着:(a) C_i 中的类别变量 y_{ik} 与输入数据各分量 x_{ij}(j=1,…,n)的协方差 $cov(x_{ij},y_i)$>>0;(b) C_i 中的类别变量的方差 $D(y_i)=cov(y_i,y_i)=F_i(n+1,n+1)$>>0.因此,低质量 RDM 模型对应的各分类 $C_i\,(1\le i\le c)$ 的高斯型隶属函数如图 2 所示,图中存在着一些曲线平坦、散度较大的分类精度较低的类,如 C_2.

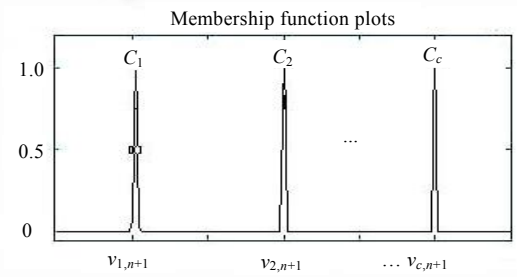

Fig.1 Gaussian membership function of C_i in the type $v_{i,n+1}$ of RDM of high classification quality (i=1,2,…,c)

图 1 高分类质量 RDM 模型中各类 C_i 在聚类中心类别分量 $v_{i,n+1}$ 上的 Gaussian 型隶属函数(i=1,…,c)

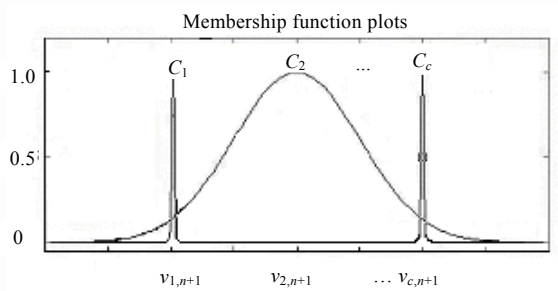

Fig.2 Gaussian membership function of C_i in the type $v_{i,n+1}$ of RDM of low classification quality (i=1,…,c)

图 2 低分类质量 RDM 模型中各类 C_i 在聚类中心类别分量 $v_{i,n+1}$ 上的 Gaussian 型隶属函数(i=1,…,c)

可见,在上述的输入-输出积空间中应用 G-K 模糊聚类,相应的 RDM 模型的分类质量指标实际上隐含在了各模糊类 C_i 的聚类中心类别分量 $v_{i,n+1}$ 之中,可以用于模糊聚类结果的监督.而且,经过积空间中 G-K 算法模糊聚类之后,实际上对数据集论域空间完成了与数据集中的原型(prototype,如椭球型、平面或线型)相适合的精细可调的模糊划分,具有 Kowalczyk 方法中的网格划分所不可比拟的优越性.原来的数据集连同它们的类别数值一起被分成了一组模糊类 C_i(i=1,2,…,c).

(2) 类型函数 $type(\cdot)$的定义

定义 1. 称积空间模糊类 C_i 的聚类中心 V_i 中与类别变量 y 相对应的分量 $v_{i,n+1}$ 为模糊类 C_i 的类型,i=1,2,…,c.

定义 2. 称由积空间的模糊聚类所给出的输入-输出空间之间的这种隐含的映射关系 $type$:$C_i\to v_{i,n+1}$ 为模糊类 C_i 的类型函数,即 $type(C_i)=v_{i,n+1}$,i=1,2,…,c.

定义 3. 称积空间模糊类 C_i 的协方差矩阵 F_i 中与类别变量 y 相对应的方差分量 $F_i(n+1,n+1)$为模糊类 C_i 的类型方差,记为 $D(v_{i,n+1})$,i=1,2,…,c.

在本文中,设定一个容许误差向量 tolSig2=[s_1,s_2,…,s_c],其中 s_i>0,i=1,…,c,c 为模糊类的个数.只有当聚类结果中所有各类 C_i 的类型方差 $D(v_{i,n+1})$满足

$$D(v_{i,n+1})<\text{tolSig2}(i),i=1,2,\ldots,c$$

时,其对应的模糊划分才可被接受用以构建 RDM 模型,以保证 RDM 模型的性能.因此,称这种在 RDM 分类质量指标指导下的、类别数据直接参入的积空间模糊聚类为监督模糊聚类.

(3) 搜索策略的确定

建立的 RDM 模型是否“最好”,取决于所得的模糊类 $C_i(i=1,2,\ldots,c)$是否是对数据集空间的“最优”划分.只有当模糊类的聚类原型(prototype)和数目与数据集中实际存在的数据类原型和数目相符合或接近时,聚类算法才能够将这些数据类原型正确地或接近正确地辨识出来,实现数据集空间的最优划分.由于 G-K 算法对数据类原型具有一定程度的自适应性,因此,最优 RDM 模型的求取可以依据聚类数目 c 的搜索而进行.

具体的搜索策略如下:开始时先指定一个小的聚类数目 c,然后调用 G-K 算法,并用所得到的模糊类构建相应的 RDM 模型;如果所得到 RDM 模型的性能指标较前一次的有所提高,则增加聚类数目 c,重复使用 G-K 算法,直到所得 RDM 的性能指标开始下降或较此前“最好”指标连续降低时为止.这样就将 Kowalczyk 方法在所有所选属性上的多维搜索过程转变为聚类数目上的一维搜索,从而极大地减少了 RDM 的搜索数目,缩短了计算时间.

2.3 模糊聚类模型转变为RDM模型

通过数据集在输入变量-类别变量的乘积空间中聚类所获得的 c 个模糊类 $C_i(i=1,\ldots,c)$构成该乘积空间中一个模糊聚类模型,每个模糊类 $C_i(1\le i\le c)$ 都是一个多维(n+1 维)模糊集合,其隶属函数可以通过与其聚类中心 $V_i=[v_{i1},v_{i2},\ldots,v_{i,n+1}]^T$ 之间的距离的倒数形式给出.

定义 4. 在积空间的模糊聚类模型中,任意数据点 $x\in R^n$,对模糊类 $C_i(1\le i\le c)$ $(\in R^{n+1})$ 的隶属度(degree of membership),记为 PM(x,C_i),可以通过下式来计算:

$$\text{PM}(x,C_i)=\frac{(D^2_{iiM_i})^{-1/(m-1)}}{\sum_{j=1}^{c}(D^2_{jiM_j})^{-1/(m-1)}} \tag{9}$$

这里,$D^2_{jim_j}$ 是积空间中数据点 $z'=[x^T,v_{in+1}]^T$ 与模糊类 C_j 的中心 $V_j=[v_{j1},\ldots,v_{jn},v_{jn+1}]^T$ 之间的内积距离.

$$D^2_{jiM_j}=(z'-V_j)^T M_j(z'-V_j) \tag{10}$$

$$M_j=\det(F_j)^{1/n}F_j^{-1} \tag{11}$$

F_j 为模糊聚类模型中类 C_j 的协方差矩阵.

由式(10)可以看出,在积空间中判定一输入数据 $x\in R^n$ 对模糊类 C_i 的隶属程度,首先是假定数据 x 属于 C_i 类,将其在积空间中对应的数据向量扩充为 $z'=[x^T,v_{i,n+1}]^T$,然后利用内积距离通过式(9)计算而得.所以,我们称 $\text{PM}(x,C_i)$ 为输入数据 $x\in R^n$ 对模糊类 $C_i(\in R^{n+1})$ 的推定隶属度(putative membership degree).推定隶属度能够有效地将类别属性差异 $(v_{i,n+1}-v_{j,n+1})$, $i\ne j$, $i,j=1,\ldots,c$ 对聚类和隶属判别的影响考虑进来,从而使得判别更符合积空间模糊聚类模型的要求,所得结果也要比仅在输入空间中判定更为合理和准确.

根据数据 $x\in R^n$ 对各模糊类 C_i $(i=1,2,\ldots,c)$ 的推定隶属度,可以确定 x 的类型 $type(x)$.首先,对各模糊类 C_i $(i=1,2,\ldots,c)$ 设置一个阈值常量(threshold constant),记为 $\text{TH}=[th_1,th_2,\ldots,th_c]$.只有当 $\text{PM}(x,C_i)>\text{TH}(i)$ 时,类 C_i 的类型 $v_{i,n+1}$ 才被考虑作为 $type(x)$的一个候选者,用以抑制噪声和提高算法效率;然后根据所有入选的模糊类,取推定隶属度超过阈值且为最大的类的类型作为 x 的最终类型,即

$$type(x)=type(C_j)=v_{j,n+1} \tag{12}$$

其中,C_j 满足 $\text{PM}(x,C_j)=\max(\text{PM}(x,C_i)\,|\,\text{PM}(x,C_i)>\text{TH}(i),i=1,\ldots,c)$.

对于训练数据集的模糊聚类模型,可以通过每一数据的类型与其类别真值的比较,判断分类的正误,从而计算各模糊类的分类精度等指标,建立其相应的 RDM 模型.

至此,利用有监督的 G-K 模糊聚类建立 RDM 模型的算法(supervised fuzzy rough data model,简称 SFRDM)

总结如下.

给定训练数据集$\{z_k|k=1,2,\ldots,N\}$,其中$z_k=[x_k^T,y_k]^T$,x_k为输入数据,y_k为类别数据.

Step 1. 求取数据集积空间中的最优模糊聚类模型.

Step 1.1. 设定聚类数目初值c=1,选取 tolSig2>0,tolSig2$\in R^C$.

Step 1.2. 对数据集$\{z_k\,|\,k=1,2,...,N\}$调用 G-K 算法进行模糊划分,计算c个模糊特征类C_i的类型$v_{i,n+1}$和类型方差$D(v_{i,n+1})$,i=1,2,…,c.

Step 1.3. 如果对于所有i=1,2,…,c,满足$D(v_{i,n+1})<\text{tolSig2}(i)$,则转 Step 2;否则,$c$:=$c$+1,返回 Step 1.2.

Step 2. 计算相应的 RDM 模型.

Step 2.1. 设置阈值常量 TH>0, TH$\in R^c$.

Step 2.2. 对于k=1,2,…,N,

Step 2.2.1. 利用式(9)计算x_k对每一类C_i的推定隶属度$\text{PM}(x_k,C_i)$,i=1,2,…,c;

Step 2.2.2. 求x_k的所有推定隶属度超过其相应阈值的模糊特征类集,$\boldsymbol{C}_b=\{C_i\,|\,\text{PM}(x_k,C_i)>\text{TH}(i),\ i=1,2,...,c\}$;

Step 2.2.3. 如果C_b=∅,则x_k无法被分类;否则,计算C_m,满足$\text{PM}(x_k,C_m)=\max\{\text{PM}(x_k,C_j)\,|\,C_j\in\boldsymbol{C}_b\}$,计算$type(x_k)=type(C_m)$;

Step 2.2.4. 如果$type(x_k)$=y_k,则x_k被正确分类;否则,x_k被错误分类.

Step 2.3. 按最大隶属度原则确定各类C_i中的元素,并计算相应的统计量,如类分类精度acc[%](i)等,i=1,2,…,c.

Step 2.4. 按各类C_i的分类精度acc[%](i)由高到低排序,i=1,2,…,c,建立相应的 RDM 模型,计算累积模型分类精度cma[%].

Step 3. 如果当前 RDM 的cma[%]>前一个 RDM 的cma[%],则c:=c+1,转 Step 1.2;否则,输出前一个 RDM 模型.

Step 4. 结束.

3 仿真研究

例 1:考虑如图 3 所示的具有线性原型(prototype)的交叉数据集:

$$C_1\text{: } y_1=2x_1+\varepsilon,\quad 0\le x_1\le 0.5,\qquad C_2\text{: } y_2=1-x_2+\varepsilon,\quad 0\le x_2\le 1.$$

其中,数据噪声$\varepsilon\sim N(0,\sigma^2)$,$\sigma=0.08$,样本数据采样间隔为$\Delta x$=0.01,先产生 152 个数据点作为训练数据集,另产生 152 个数据点作为测试集.

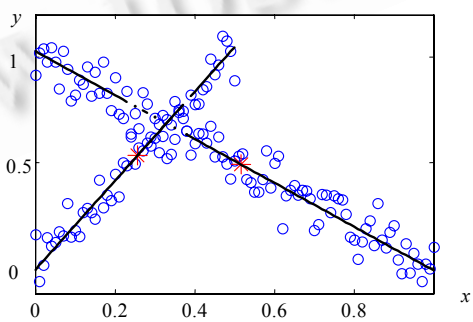

Fig. 3 Scatter plots of Example 1

图 3 例 1 的数据散点图

应用本文方法(SFRDM)能够较好地识别出两种线性原型,并且所建立的 RDM 模型的累积分类精度为 93.42%,测试集上的有效累积分类精度达 95.395%;对同样的数据集应用 Kowalczyk 方法(KRDM),对x,y输入变量,首先利用等频率离散化方法各划分成 10 个小区间,然后从其对应的(9,9)个分点中分别取(1,1)(2,1)(1,2)

(2,2)(4,4)个分点,建立相应的 RDM 模型,结果见表 2.

Table 2 Some main results of the methods on data set of Example 1

表 2 例 1 数据集上各方法的主要结果

	SFRDM	KRDM			
Clusters or cut points	c=2	c=4 (1,1)	c=6 (2,1) or (1,2)	c=9 (2,2)	c=18 (4,4)
cma[%]	93.421	93.421	93.421	94.737	94.737
Time (s)	0.17	1.32	5.76	24.55	410.85
Valid [%]	95.395	88.616	88.616	90.132	90.132

SFRDM 与 KRDM 对例 1 数据空间的划分如图 4、图 5 所示.可以看出:(1) SFRDM 对输入空间的划分方向是由聚类的数据原型(prototype)决定的,不一定与坐标轴平行;而 KRDM 对输入空间的划分是阶梯状的,其各段方向一定与坐标轴平行.(2) SFRDM 对输入空间的划分是平滑过渡的,符合受噪声影响的数据分类的特点和规律,更客观地反映了实际问题;而 KRDM 对输入空间的划分是界线分明、截然分割的,在图形上表现为隶属函数值在分界线上是垂直下降的,是一种“硬”划分,不符合噪声数据的特点和区分规律.(3) SFRDM 能够从整体上识别出聚类数据中的线性原型;而 KRDM 只能从局部以细碎的格点去近似某一数据类.格点矩形块越小,数据划分精度越高,需要的分点数就越多.而分点数目的增加,将导致所要计算的 RDM 模型数目呈指数级增加,造成很大的计算负担.

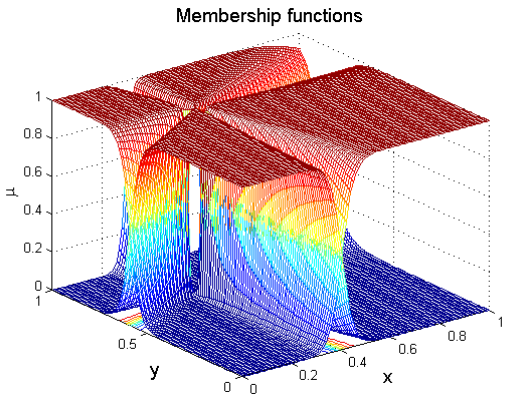

Fig.4 The partition of data space in Example 1 with SFRDM

图 4 SFRDM 对例 1 数据空间的划分

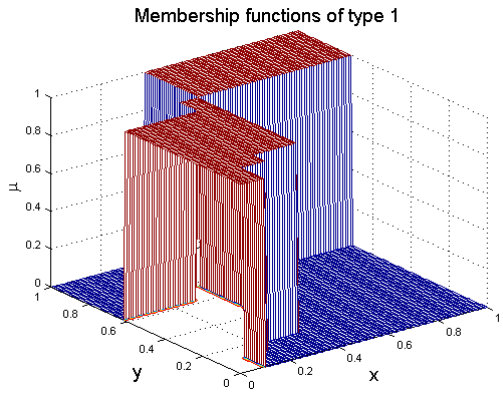

Fig.5 The partition of data space in Example 1 with KRDM (cut points(4,4))

图 5 KRDM 对例 1 数据空间的划分(分点数(4,4))

例 2:UCI 数据库[20]是机器学习领域中一个著名的数据库.我们从中选取 5 个典型的数据集(见表 3)对本文的方法进行测试和比较.分别对每一个数据集随机地选取约 2/3 的事例样本作为训练集,剩下的事例样本作为测试集.测试集上的正确分类样本百分比记为有效精度 valid[%].

分别将 SFRDM 方法和 Kowalczyk 方法应用于上述数据集,每个数据集随机运行 10 次,取其平均值作为计算结果,见表 4.

在 Kowalczyk 方法中,每个属性首先被用等频率法划分成 10 个小区间,然后从每个属性的 9 个分点中选取 1 或 2 个分点将连续属性离散化,建立相应的 KRDM.由于部分数据集的属性数目较多而使 Kowalczyk 的穷尽式搜索方法因搜索时间太长(可达几十小时)而难以进行,文中给出了其使用遗传算法的 Kowalczyk 方法计算的结果.

从表 3 和表 4 中可以看出,SFRDM 可以较好地用于数据规模较大的数据集模式分类,但需要指出的是,由于 SFRDM 中采用的 G-K 聚类算法缺乏对类间样本容量的估计信息,所以,SFRDM 仅能有效地发现样本数量近似相当的各模式类,而对于样本容量差异较大的模式类数据集,其识别能力相对较弱.例如对数据集 Car,就把样本容量很小的两个相近的模式类“good”和“very good”(分别占总数据量的 4.00%和 3.76%)划分为一类了;另外,SFRDM 对高维数据集也具有较好的适应性,而 KRDM 却由于需要搜索的 RDM 模型个数随属性个数(即维

数)的增加而急剧增加,以至于无法进行,必须依靠 GA 等启发式搜索方法对之加以改造.

Table 3 Description of data sets used in the experiments

表 3 实验数据集的构成描述

Data sets	Size (training set/testing set)	Attributes num.	Classes num.
Iris	150 (100/50)	4	3
Car	1 728 (1128/600)	6	4
Nursery	12 960 (10000/2960)	8	5
Wine	178 (128/50)	13	3
Satellite	6 435 (4435/2000)	36	6

Table 4 Experimental results of the methods on the data sets of Example 2

表 4 例 2 数据集上各方法的实验结果

	SFRDM				KRDM				KRDM+GA			
	c	Cma [%]	Time (s)	Valid [%]	*c*	Cma [%]	Time (s)	Valid [%]	*c*	Cma [%]	Time (s)	Valid [%]
Iris	3	98.00	0.86	100.0	9	96.00	91.06	96.00	8	95.00	3.87	94.00
Car	4	85.02	0.30	85.00	32	80.76	170.58	82.33	32	80.76	111.04	82.33
Nursery	10	86.92	22.69	86.15	256	82.85	51541.3	81.25	256	82.85	2409.57	81.25
Wine	5	92.19	0.27	100.0	-	-	-	-	60	100.0	83.66	92.00
Satellite	13	87.71	85.45	87.15	-	-	-	-	177	87.49	2386.4	85.85

注:*c* 表示聚类数目

总之,从上述两个例子可以发现:(1) 本文方法建立的 RDM 模型,与 Kowalczyk 方法及使用 GA 的 Kowalczyk方法建立的RDM模型相比,其累积模型分类精度 *cma*[%]没有显著意义上的差别,但本文方法具有更高的求解效率,使用的时间更少;(2) 当将 3 种方法所获得的 RDM 模型应用于测试数据集时,发现本文方法建立的RDM模型具有更高的分类正确率,即模型有效精度,这意味着本文方法具有更强的数据概括能力;(3) 本文方法具有更大的"柔性",可以依据数据集的特性,通过调节各模糊类推定隶属度的阈值分量 TH,总可以求得一个合适的 RDM 模型,从而更好地处理噪声数据的影响.

4 结 论

本文提出了一种利用在输入-输出积空间中的监督 G-K 模糊聚类算法快速建立粗糙数据模型(RDM)的方法,并研究了其在数据分类方面的应用.基于本文思想,我们已在 Matlab6.1 环境下研制开发了一套应用软件系统——SFRDMS.该软件对多个实例数据集的应用结果表明,SFRDM 方法是一个有效的数据分析工具,它不仅能够用于构建高质量的数据分类模型,而且能够以高效的搜索策略发现数据中最优的特征模式.特别地,通过在输入数据-类别积空间中引入输入数据对模糊特征类的推定隶属度的概念,相对于传统的粗糙集理论和Kowalczyk方法,使SFRDM方法更具灵活性,具有更强的处理噪声数据的能力.从这一意义上来说,SFRDM方法实际上是一种建立粗糙数据模型的"软"技术.

References:

[1] Pawlak Z. Rough Set: Theoretical Aspects of Reasoning about Data. Boston: Kluwer Publishers, 1991.

[2] Skowron A, Peters J F. Rough sets: Trends and challenges. In: Wang G, Liu Q, Yao Y, Skowron A, eds. Rough Sets, Fuzzy Sets, Data Mining and Granular Computing. LNAI 2639, Berlin, Heidelberg: Springer-Verlag, 2003. 25−34.

[3] Tsumoto S. Mining diagnostic rules from clinical databases using rough sets and medical diagnostic model. Information Sciences, 2004,162(2):65−80.

[4] Peters JF, Skowron A. A rough sets approach to knowledge discovery. International Journal of Intelligent Systems, 2002, 17(2):109−112.

[5] Huang C-C, Tseng T-L. Rough set approach to case-based reasoning application. Expert Systems with Applications, 2004, 26(3):369−385.

[6] Polkowski L. Toward rough set foundations-mereological approach. In: Tsumoto S, Slowinski R, Komorowski HJ, Grzymala-Busse JW, eds. Rough Sets and Current Trends in Computing. LNAI 3066, Berlin, Heidelberg: Springer-Verlag, 2004. 8−25.

[7] Peters JF, Skowron A, Synak P, Ramanna S. Rough sets and information granulation. LNCS 2715, Heidelberg: Springer-Verlag, 2003. 370−377.

[8] Zhang WX, Wu WZ, Liang JY, Li DY. Rough Set Theory and Methodology. Beijing: Science Press, 2001 (in Chinese).

[9] Han JC, Hu XH, Nick C. Supervised learning: A generalized rough set approach. In: Ziarko W, Yao Y, eds. Rough Sets and Current Trends in Computing. LNAI 2005, Heidelberg: Springer-Verlag, 2001. 322−329.

[10] Slowinski R, Vanderpooten D. A generalized definition of rough approximations based on similarity. IEEE Trans. on Knowledge and Data Engineering, 2000,12(2):331−336.

[11] Inuiguchi M, Tanino T. On rough sets under generalized equivalence relations. In: Terano T, Nishida T, Namatame A, Tsumoto S, Ohsawa Y, Washio T, eds. New Frontiers in Artificial Intelligence: Joint JSAI 2001 Workshop Post-Proc. LNAI 2253, Heidelberg: Springer-Verlag, 2001. 295−300.

[12] Yao YY. Generalized rough set models. In: Polkowski L, Skowron A, eds. Rough Sets in Knowledge Discovery 1─Methodology and Applications. Heidelberg: Physica-Verlag, 1998. 287−318.

[13] Yao YY. On generalizing rough set theory. In: Wang G, Liu Q, Yao Y, Skowron A, eds. Rough Sets, Fuzzy Sets, Data Mining and Granular Computing. LNAI 2639, Berlin, Heidelberg: Springer-Verlag, 2003. 44−51.

[14] Ziarko W. Variable precision rough sets model. Journal of Computer and System Sciences, 1993,46(1):39−59.

[15] Kowalczyk W. Rough data modeling: A new technique for analyzing data. In: Polkowski L, Skowron A, eds. Rough Sets in Knowledge Discovery 1: Methodology and Applications. Heidelberg: Physica-Verlag, 1998. 400−421.

[16] Pal SK, Skowron A. Rough-Fuzzy hybridization: A new trend in decision-making. Singapore: Springer-Verlag, 1999.

[17] Inuiguchi M. Generalizations of rough sets: From crisp to fuzzy cases. In: Tsumoto S, Slowinski R, Komorowski HJ, Grzymala-Busse JW, eds. Rough Sets and Current Trends in Computing. LNAI 3066, Berlin, Heidelberg: Springer-Verlag, 2004. 26−37.

[18] Eiben AE, Euverman TJ, Kowzlczyk W, Slisser F. Modeling customer retention with statistical techniques, rough data models, and genetic programming. In: Pal SK, Skowron A, eds. Rough-Fuzzy Hybridization: A New Trend in Decision-Making. Singapore: Springer-Verlag, 1999. 330−345.

[19] Höppner F, Klawonn F, Kruse R, Runkler T. Fuzzy Cluster Analysis. Chichester: John Wiley & Sons Ltd., 1999.

[20] Merz CJ, Murphy PM. UCI Repository of machine learning databases. Irvine: Department of Information and Computer Science, University of California, 2004. http://www.ics.uci.edu/~mlearn/MLRepository .html

附中文参考文献:

[8] 张文修,吴伟志,梁吉业,李德玉.粗糙集理论与方法.北京:科学出版社,2001.