DOI:10.11721/cqnuj20140117

基于灰色关联度的变精度粗糙集模型*

王金山，王　磊
(解放军陆军军官学院 数学教研室，合肥 230031)

摘要:为了解决经典粗糙集理论无法处理不精确分类以及连续值属性离散化造成信息损失等问题,本文提出了基本灰色关联度的变经度粗糙集模型($\underline{\gamma'}_{\beta}(X)$,$\overline{\gamma'}_{\beta}(X)$)。首先利用灰色关联度定义相似并进一步定义粗糙集并研究了模型的部分性质和定理(当 $0<t_1\leqslant t_2<1$ 时,有 $\gamma'^{t_2}(x)\subseteq\gamma'^{t_1}(x)$、$\underline{\gamma'}_{\beta}^{t_1}(X)\subseteq\underline{\gamma'}_{\beta}^{t_2}(X)$及$\overline{\gamma'}_{\beta}^{t_2}(X)\subseteq\overline{\gamma'}_{\beta}^{t_1}(X)$成立),然后提出了一种基于重要度 $SIG_B(c)$的约简算法来计算最小约简,最后通过一个实例说明模型是有效和可行的。

关键词:连续值属性决策表;变经度粗糙集模型;灰色关联度;属性约简;重要度
中图分类号:O29　　**文献标志码**:A　　**文章编号**:1672-6693(2014)01-0080-04

粗糙集理论[1]是由波兰学者 Z. Pawlak 于 1982 年提出的一种处理不精确、不确定与不完全数据的新的数学方法,可以对数据进行分析和推理,从中发现隐含的知识,揭示潜在的规律。经典粗糙集模型是建立在等价关系基础上,能够处理离散属性值问题,但是现实问题中得到的数据往往是连续的。在使用经典粗糙集模型处理连续值属性问题时首先需要对属性进行离散化,但是属性离散化会造成一定程度的信息损失[2]。同时,经典粗糙集模型处理的分类必须是完全正确的或肯定的,而 Ziarko 教授于 1993 年提出的变精度粗糙集模型[3]允许一定程度上的错误分类。本文建立了基于灰色关联度的变精度粗糙集模型并研究了相关性质定理,然后给出了基于重要度的约简算法,最后通过一个实例说明模型建立和约简的具体过程。

1 基本概念

首先本节给出连续值属性信息系统、灰色关联度及关联类的定义。

定义 1[4]　设 $S=(U,A,V,f)$表示一个连续值属性信息系统,其中论域$U=\{x_1,x_2,\cdots,x_n\}$是非空对象集合;$A=C\cup d$ 是非空属性集合,其中 C 为条件属性集,对于任意的 $c\in C$ 为连续值条件属性,属性 d 为离散型决策属性;$V=\bigcup\limits_{a\in A}V_a$,$V_a$ 是属性 a 的值域;f: $U\times A\to V$ 是信息函数, 它对每个对象的每个属性赋予一个信息值,即 $\forall a\in A,x\in U$,有 $f(x,a)=V_a$。

定义 2　设连续值属性信息系统 $S=(U,A,V,f)$,对象集合 $X\subseteq U$,属性集合 $B\subseteq A$。设对象 $x_i,x_j\in U$,则定义 x_i,x_j 关于属性集 B 的灰色关联度[5]为 $\gamma(x_i,x_j)=\dfrac{1}{|B|}\sum\limits_{k=1}^{|B|}\gamma_{ij}(k)$。$\gamma(x_i,x_j)$简记为 γ_{ij},满足灰色关联四公理。其中 $\gamma_{ij}(k)$称为对象 x_i,x_j 关于属性 a_k 的关联系数。设分辨系数 $\xi\in(0,1)$,则关联系数定义为

$$\gamma_{ij}(k)=\frac{\min\limits_{z\in U}\min\limits_{a\in B}|\delta(x_i,z,a)|+\xi\max\limits_{z\in U}\max\limits_{a\in B}|\delta(x_i,z,a)|}{|\delta(x_i,x_j,a)|+\xi\max\limits_{z\in U}\max\limits_{a\in B}|\delta(x_i,z,a)|}$$

其中 $\delta(x_i,z,a)=f(x_i,a)-f(z,a)$,$\delta(x_i,x_j,a)=f(x_i,a_k)-f(x_j,a_k)$。

计算所有的 $\gamma_{ij}(i\leqslant j,\ i,\ j=1,\ 2,\cdots,|U|)$,可以得到一个 $|U|$ 阶上三角矩阵 H_B,称为灰色关联矩阵

$$H_B=\begin{bmatrix}\gamma_{11} & \gamma_{12} & \cdots & \gamma_{1|U|}\\ & \gamma_{22} & \cdots & \gamma_{2|U|}\\ & & \ddots & \vdots\\ & & & \gamma_{|U||U|}\end{bmatrix}$$

,其中,$\gamma_{ii}=1,i=1,2,\cdots,|U|$。

* **收稿日期**:2011-11-23　**修回日期**:2013-09-09　**网络出版时间**:2014-01-16　08:16
资助项目:军队科研基金项目(No. 011027)
作者简介:王金山,男,教授,研究方向为预测与决策分析,E-mail: wjs586@126.com
网络出版地址:http://www.cnki.net/kcms/detail/50.1165.N.20140116.0816.021.html

定义 3　设 $t\in[0,1]$，则 $\gamma_{ij}\geqslant t$ 记作 $x_i\gamma_B^t x_j$。对象 x 关于属性集 B 的 t 关联类定义为 $\gamma_B^t(x)=\{y\in U\mid x\gamma_B^t y\}$。属性集 B 的关联类集合 U/γ_B 定义为 $U/\gamma_B=\{\gamma_B^t(x)\mid x\in U\}$。$U/\gamma_B$ 表示所有 x 关于属性集 B 的关联类构成的集合。在不造成误解的前提下，x 关于属性集 B 的 t 关联类 $\gamma_B^t(x)$ 记为 $\gamma^t(x)$，灰色关联关系记为 γ^t。

性质 1　$t=0$ 时，$\forall x\in U,\gamma^t(x)=U$，此时整个论域 U 是一个相似类。

2 基于灰色关联度的 VPRS 模型

本节给出多数包含关系的定义，在此基础上定义了基于灰色关联关系的 β 上、下近似集，建立了变精度粗糙集模型并研究了部分性质、定理。

定义 4　设 X、Y 是论域 U 的非空子集，多数包含关系定义为 $Y\overset{\beta}{\supseteq}X\Leftrightarrow c(X,Y)\leqslant\beta,0\leqslant\beta<0.5$。其中，$c(X,Y)$ 为集合 X 关于集合 Y 的相对错误分类率：$c(X,Y)=\begin{cases}1-|X\cap Y|/|X|,&|X|>0\\0,&|X|=0\end{cases}$。多数包含关系的实际意义是：集合 X 与集合 Y 交集的元素数目除以集合 X 中元素数目要大于 $1-\beta$。显然，$Y\supseteq X\Leftrightarrow c(X,Y)=0$。

定义 5　设 (U,γ^t) 为近似空间，其中 U 为论域，γ^t 为灰色关联关系，$\gamma_B^t(x)$ 为对象 x 的 t 关联类。对于对象集合 $X\subseteq U$，定义 X 基于灰色关联关系 γ^t 的 β 下近似集：

$$\underline{\gamma}_\beta^t(X)=\{x\in U\mid X\overset{\beta}{\supseteq}\gamma^t(x)\}\text{或者}\underline{\gamma}_\beta^t(X)=\{x\in U\mid c(\gamma^t(x),X)\leqslant\beta\}$$

$\underline{\gamma}_\beta^t(X)$ 也称为 X 基于 γ^t 的 β 正域，记为 $POS_\beta^{\gamma^t}(X)$。

定义 X 基于 γ^t 的 β 上近似集：$\overline{\gamma}_\beta^t(X)=\{x\in U\mid c(\gamma^t(x),X)<1-\beta\}$。

定义 X 基于 γ^t 的 β 边界域：$BN_\beta^{\gamma^t}(X)=\{x\in U\mid\beta<c(\gamma^t(x),X)<1-\beta\}$。

定义 X 基于 γ^t 的 β 负域：　$NEG_\beta^{\gamma^t}(X)=\{x\in U\mid c(\gamma^t(x),X)\geqslant1-\beta\}$。

由上面的定义可以证明有以下性质和定理。

性质 2　$\overline{\gamma}_\beta{}^t(X)=POS_\beta^{\gamma^t}(X)\cup BN_\beta^{\gamma^t}(X)$。

性质 3　$POS_\beta^{\gamma^t}(\sim X)=NEG_\beta^{\gamma^t}(X)$。

定理 1　若 $0<t_1\leqslant t_2<1$，则有：1) $\forall x\in U,\gamma^{t_2}(x)\subseteq\gamma^{t_1}(x)$；2) $\forall X\subseteq U,\underline{\gamma}_\beta^{t_1}(X)\subseteq\underline{\gamma}_\beta^{t_2}(X)$；3) $\forall X\subseteq U,\overline{\gamma}_\beta^{t_2}(X)\subseteq\overline{\gamma}_\beta{}^{t_1}(X)$。

3 基于重要度的约简算法

本节利用 β 下近似集构造了 β 正域，进一步定义了 β 近似依赖度和相对重要度，最后以定理 2 为基础提出了基于重要度的约简算法并给出了算法步骤。

定义 6　设信息系统 $S=(U,C\cup d,V,f)$，其中 C 为条件属性集，d 为决策属性；$\forall B\subseteq C,\gamma^t$ 为灰色关联关系；U/d 表示决策属性 d 的等价类集合。

条件属性集 B 相对于决策属性 d 的 β 正域 $POS_\beta^{\gamma^t}(B,d)$ 定义为 $POS_\beta^{\gamma^t}(B,d)=\bigcup\limits_{X\in U/d-}\gamma^t{}_\beta(X)$

决策属性 d 与条件属性集 B 的 β 近似依赖度 $\lambda(B,d,\beta)$ 定义为 $\lambda(B,d,\beta)=\dfrac{|POS_\beta^{\gamma^t}(B,d)|}{|U|}$

记条件属性集 C 关于决策属性 d 的 β 近似约简为 $RED(C,d,\beta)$，$RED(C,d,\beta)\subseteq C$。$RED(C,d,\beta)$ 满足下面两个条件：1) $\lambda(C,d,\beta)=\lambda(RED(C,d,\beta),d,\beta)$；2) 从 $RED(C,d,\beta)$ 中去掉任何一个属性都将使 1) 式不成立。

定义 7[6]　设 $B\subseteq C$，定义属性 $c\in C$ 相对于属性集 B 的重要度 $SIG_B(c)$ 为 $SIG_B(c)=\lambda(B\cup\{c\},d,\beta)-\lambda(B-\{c\},d,\beta)$。$SIG_B(c)$ 越大，说明在条件属性集 B 中属性 c 相对于决策属性 d 越重要。

定理 2[7]　在条件属性集 B 中所有重要度不为 0 的属性构成 B 的核 $CORE(C)$。

约简的算法步骤如下：1) 计算核 $CORE(C)$：$\forall c\in C$，计算重要度 $SIG_C(c)$，所有重要度大于 0 的属性构成核 $CORE(C)$；2) 令 $RED(C)\leftarrow CORE(C)$；3) 计算 $\lambda(C,d,\beta)$ 及 $\lambda(RED(C),d,\beta)$，并判断 $\lambda(C,d,\beta)=\lambda(RED(C),d,\beta)$ 是否成立。若成立，则转 6)，否则转 4)；4) 对所有 $c\in C-RED(C)$ 计算 $SIG_{RED(C)}(c)$ 并计算其中最大值，即

$SIG_{RED(C)}(c_{\max})=\max\limits_{c\in C-RED(C)}\{SIG_{RED(C)}(c)\}$；5)令 $RED(C)\leftarrow CORE(C)\cup\{c_{\max}\}$，转3)；6)输出最小约简 $RED(C)$。

4 实例

例 表1是一个决策表，其中 $C=\{a_1,a_2,a_3,a_4\}$ 为连续值条件属性集，d 为离散值决策属性。

表1 决策表

U	a_1	a_2	a_3	a_4	d
x_1	5.4	3	4.5	1.5	1
x_2	5.5	2.4	3.8	1.1	1
x_3	5.7	4.4	1.5	0.4	0
x_4	5.8	2.7	3.9	1.2	1
x_5	5.8	4.1	2.0	0.2	0
x_6	5.1	3.8	1.6	0.2	0
x_7	5.0	3.8	1.9	0.4	0
x_8	5.7	2.8	4.1	1.3	1

根据灰色关联度计算公式可以计算得到15个相似度矩阵，例如：

$$H_{\{a_2,a_3,a_4\}}=\begin{bmatrix}1&0.74&0.91&0.69&0.99&0.91&0.86&0.77\\0&1&0.70&0.83&0.66&0.71&0.76&0.62\\0&0&1&0.71&0.91&0.95&0.92&0.73\\0&0&0&1&0.63&0.68&0.72&0.70\\0&0&0&0&1&0.91&0.86&0.77\\0&0&0&0&0&1&0.92&0.71\\0&0&0&0&0&0&1&0.68\\0&0&0&0&0&0&0&1\end{bmatrix}$$

设 $t=0.95$，则可以得到关联类集合：$U/d=\{\{x_1,x_2,x_4,x_8\},\{x_3,x_5,x_6,x_7\}\}$

$U/\gamma_C=\{\{x_1,x_5,x_7\},\{x_2,x_4\},\{x_3\},\{x_6\},\{x_8\}\}$，$U/\gamma_{\{a_1,a_2,a_3\}}=\{\{x_1\},\{x_3\},\{x_2,x_4\},\{x_5,x_7\},\{x_6\},\{x_8\}\}$

$U/\gamma_{\{a_1,a_2,a_4\}}=\{\{x_1\},\{x_3\},\{x_2\},\{x_4\},\{x_5,x_7\},\{x_6\},\{x_8\}\}$，$U/\gamma_{\{a_1,a_3,a_4\}}=\{\{x_1\},\{x_2\},\{x_3,x_6\},\{x_4\},\{x_7,x_5\},\{x_8\}\}$

$U/\gamma_{\{a_2,a_3,a_4\}}=\{\{x_1,x_5\},\{x_2\},\{x_3,x_6\},\{x_4\},\{x_7\},\{x_8\}\}$，$U/\gamma_{\{a_1,a_2\}}=\{\{x_1,x_7\},\{x_2,x_3,x_5\},\{x_4,x_8\},\{x_6\}\}$

$U/\gamma_{\{a_1,a_3\}}=\{\{x_1,x_2,x_3,x_4,x_6\},\{x_5,x_7\},\{x_8\}\}$，$U/\gamma_{\{a_1,a_4\}}=\{\{x_1,x_2,x_3,x_4,x_5,x_6,x_8\},\{x_7\}\}$

设 $\beta=0.4$，则条件属性集相对于决策属性 d 的 β 正域为：

$$POS_\beta^{\gamma'}(C,d)=\{x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8\}$$

$$POS_\beta^{\gamma'}(\{a_1,a_2,a_3\},d)=\{x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8\}$$

$$POS_\beta^{\gamma'}(\{a_1,a_2,a_4\},d)=\{x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8\}$$

$$POS_\beta^{\gamma'}(\{a_1,a_3,a_4\},d)=\{x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8\}$$

$$POS_\beta^{\gamma'}(\{a_2,a_3,a_4\},d)=\{x_2,x_3,x_4,x_6,x_7\}$$

依赖度为：$\lambda(C,d.\beta)=1$，$\lambda(\{a_1,a_2,a_3\},d,\beta)=1$，$\lambda(\{a_1,a_2,a_4\},d,\beta)=1$，$\lambda(\{a_1,a_3,a_4\},d,\beta)=1$，$\lambda(\{a_2,a_3,a_4\},d,\beta)=\frac{3}{4}$。重要度为：$SIG_C(a_1)=\frac{1}{8}$，$SIG_C(a_2)=SIG_C(a_3)=SIG_C(a_4)=0$。

因此，$CORE(C)=\{a_1\}$。$SIG_{\{a_1\}}(a_3)=\max\limits_{c\in C-\{a_1\}}\{SIG_{\{a_1\}}(c)\}$ 并且 $\lambda(C,d,\beta)=\lambda(\{a_1,a_3\},d,\beta)=1$，所以 $\{a_1,a_3\}$ 就是所求的决策表的约简。

5 结束语

本文建立了基于灰色关联度的变精度粗糙集模型，避免了属性离散化造成的信息损失同时允许一定程度上的错误分类。进一步研究了模型的相关性质、定理并给出了基于重要度的约简算法。最后，通过一个实例说明了模型建立、约简的整个过程。模型中 t 和 β 的取值对模型的影响比较大，但是目前取值具有较强的主观性，因此如何选择合适的 t 和 β 取值是需要进一步研究的问题。

参考文献：

[1] Pawlak Z. Rough set[J]. International Journal of Computer and Information Sciences,1982,11:341-356.

[2] 孟科. 粗糙集连续属性离散化通用模型及GASA方法[J]. 兰州理工大学学报,2011,37(1):91-94.

Meng K. Universal model and GASA method for discretization of continuous attribute in rough sets[J]. Journal of Lanzhou University of Technology,2011,37(1):91-94.

[3] Ziarko W. Variable precision rough set model[J]. Journal

of Computer and System Science,1993,46(1):39-59.

[4] 徐怡,李龙澍.基于(α,λ)联系度容差关系的变精度粗糙集模型[J].自动化学报,2011,37(3):303-308.
Xu Y,Li L S. Variable precision rough set model based on (α,λ) connection degree tolerance relation[J]. Acta Automatica Sinica,2011,37(3):303-308.

[5] 孙林凯,金家善,耿俊豹. 基于修正邓氏灰色关联度的设备费用影响因素分析[J]. 数学的实践与认识,2012,42(8):140-144.
Sun L K,Jin J S,Geng J B. Research on the influence factors of the equipment's expense based on the amend grey correlation[J]. Mathematics in Practice and Theory,2012,42(8):140-144.

[6] 赵晓雨,周润珍.基于等价关系的信息熵及概率分配函数[J].重庆师范大学学报:自然科学版,2009,26(3):75-78.
Zhao X Y,Zhou R Z. The information entropy and probability assignment based on equivalent relations[J]. Journal of Chongqing Normal University:Natural Science,2009,26(3):75-78.

[7] 付昂,王国胤,胡军.基于信息熵的不完备信息系统属性约简算法[J].重庆邮电大学学报:自然科学版,2008,20(5):586-592.
Fu A,Wang G Y,Hu J. Information entropy based attribute reduction algorithm in incomplete information systems [J]. Journal of Chongqing University of Posts and Telecommunications:Natural Science Edition,2008,20(5):586-592.

[8] 李健,常太华,杨婷婷.变精度粗糙集模型属性约简分析[J].计算机工程与应用,2012,48(13):130-132.
Li J,Chang T H,Yang T T. Analysis of attribute reduction in variable precision rough set model[J]. Computer Engineering and Applications,2012,48(13):130-132.

[9] 李晓瑜,徐章艳,王炜,等.不完备信息系统中一种新的求核算法[J]. 计算机工程,2011,37(11):56-58.
Li X Y,Xu Z Y,Wang W,et al. New core computing algorithm in incomplete information system[J]. Computer Engineering,2011,37(11):56-58.

Variable Precision Rough Set Model Based on Grey Correlation Degree

WANG Jin-shan, WANG Lei

(Artillery Academy, P. L. A, Mathematic Teaching & Researching Section, Hefei 230031, China)

Abstract: Classical rough set theory deals with the problems of continuous valued attribute decision table, discrediting continuous valued attributes will result in loss of information, while the classical rough set model can not deal with imprecise classification. To solve the above problem, variable precision rough set model based on the gray absolute correlation degree($\underline{\gamma'}_{\beta}(X)$,$\overline{\gamma'}_{\beta}(X)$) has been established. In the model, firstly, similarity class and rough sets are defined by gray absolute correlation degree and some properties of the model are studied(when $0<t_1\leqslant t_2<1$, then $\gamma'^{t_2}(x)\subseteq\gamma'^{t_1}(x)$、$\underline{\gamma}_{\beta}^{t_1}(X)\subseteq\underline{\gamma}_{\beta}^{t_2}(X)$and $\overline{\gamma}_{\beta}^{t_2}(X)\subseteq\overline{\gamma}_{\beta}^{t_1}(X)$). And then, a reduction algorithm based on the significant degree($SIG_B(c)$) is designed to calculate the minimum reduction. Finally, an example is used to illustrate the effectiveness and feasibility of the model.

Key words: continuous valued attribute decision table; variable precision rough set model; grey correlation degree; indicator reduction; significant degree

(责任编辑 游中胜)