

不完备信息系统下的变精度粗糙集模型及其知识约简算法^{*}

张宏宇 梁吉业

(山西大学计算机科学系 太原030006)

Variable Precision Rough Set Model and a Knowledge Reduction Algorithm for Incomplete Information System

ZHANG Hong-Yu LIANG Ji-Ye

(Department of Computer Science, Shanxi University, Taiyuan 030006)

Abstract Rough Set Theory, which has been found applicable and useful in many fields, is now a very effective method in data mining research. However, when the decision table is an incomplete one, with the original rough set theory proposed by Z. Pawlak, one can't get satisfactory results. In this paper an approach based on limited valued tolerance relation and majority inclusion relation is proposed. And furthermore a new attribute reduction method called extended discernable matrix is given. As this model is somewhat a combination of fuzzy means and majority inclusion relation, it is more effective than the previous models in practice.

Keywords Limited valued tolerance relation, Majority inclusion relation, Extended discernable matrix

1 引言

Z. Pawlak^[1]提出的粗糙集理论,提供了一种新的知识获取手段,但是,应用经典粗糙集理论的一个重要前提是它的处理对象必须是完备信息系统。现实中,由于数据采集能力不足等原因又使得大量的信息系统是不完备的,因此需要对经典粗糙集理论进行必要的扩充。目前有两种扩充方法,一是间接处理方法,即先通过一种方式将不完备信息系统转化为完备信息系统,再使用针对完备信息系统的方法对其进行处理;二是直接处理方法,即将经典粗糙集理论中的相关概念在不完备信息系统下进行扩充^[2]。然而,利用现有的直接处理模型获得的规则是一个概念层次上的,难以从多个概念级别上对原有系统进行分析处理;而且,在许多方法中,由于知识的粒度过大,使得处理结果不尽如意。本文在分析不完备信息系统的基础上,提出了一种基于不完备信息系统的变精度粗糙集模型,并给出了一种有效的知识约简算法。

2 扩充的粗糙集模型

对于不完备信息的理解,存在两种语意解释:遗漏(missing)和缺席(absent)^[3]。前者是指不完备信息系统中的未知值仅仅是被遗漏但它是存在的,后者则认为对象被不完全描述的原因是该对象不能由所有的属性描述。基于这两种理解方式,我们介绍两种直接处理方法:一种基于容差关系,另一种基于非对称相似关系^[3]。

2.1 容差关系

容差关系是为了在遗漏语意下处理不完备信息系统而定义的一种二元关系:

定义1^[3] 设不完备信息系统 $S=(U,A)$, 其中 U 是一个

对象的非空有限集合, $A=C \cup D$, $C \cap D=\emptyset$, C 是条件属性集, D 是决策属性集。对 $\forall x, y \in U, \forall B \subseteq A$, 定义容差关系 T 为:

$$T = \{(x, y) | \forall x \wedge \forall y \wedge (\forall b \in B \Rightarrow (b(x) = * \text{ 或 } b(y) = * \text{ 或 } b(x) = b(y)))\}$$

对象 $x \in U$ 的容差类定义为:

$$T(x) = \{y | y \in U \wedge (x, y) \in T\}. U \text{ 上的概念 } X \text{ 的下、上}$$

近似分别定义为:

$$\underline{T}X = \{x | x \in U \wedge T(x) \subseteq X\},$$

$$\overline{T}X = \{x | x \in U \wedge T(x) \cap X \neq \emptyset\}.$$

2.2 非对称相似关系

非对称相似关系是针对缺席语意下的不完备信息系统而定义的一种二元关系。

定义2^[3] 设不完备信息系统 $S=(U,A)$, 其中 U 是一个对象的非空有限集合, $A=C \cup D$, $C \cap D=\emptyset$. 对 $\forall x, y \in U, \forall B \subseteq A$, 定义非对称相似关系 S 为:

$$S = \{(x, y) | \forall x \wedge \forall y \wedge (\forall b \in B \Rightarrow (b(x) = * \text{ 或 } b(x) = b(y)))\}$$

非对称相似于 $x \in U$ 的对象集合及与 x 之非对称相似的对象集合定义为:

$$S(x) = \{y | y \in U \wedge S(y, x)\},$$

$$S^{-1}(x) = \{y | y \in U \wedge S(x, y)\}.$$

U 上的概念 X 的下、上近似分别定义为:

$$\underline{S}X = \{x | x \in U \wedge S^{-1}(x) \subseteq X\}$$

$$\overline{S}X = \{x | x \in U \wedge S(x) \cap X \neq \emptyset\}.$$

3 基于变精度粗糙集模型的不完备信息系统处理方法

上面介绍的容差关系和非对称相似关系都弱化了作为粗

^{*} 本文得到国家自然科学基金(No. 60275019), 国家863项目(No. 2001AA115460)和山西省自然科学基金项目的资助。张宏宇 硕士研究生, 主要研究领域为粗糙集理论、数据挖掘。梁吉业 教授, 博士, 主要研究领域为粗糙集理论、数据挖掘、人工智能。

糙集理论基础的等价关系,以此来对经典粗糙集理论的扩充。但它们有一个共同点,那就是元素之间的关系描述得不够精细,使得知识粒度过大,概念表示不精确。因此自然会想到设计一个拥有模糊性度量的模型来改善这种情况。

基于上述思想,我们认为在将不完备信息系统中的元素按某种关系分类时,应该对由此产生的每一个子集中的每个元素赋予一个其隶属于该集合的度量值。下面,我们便针对基于容差关系的粗糙集模型进行扩充。

3.1 不完备信息系统下的变精度粗糙集模型

首先,引入文[3]中容差类的概念:它是一个用关于参考元素的容差度^[2]作为成员函数的模糊集合,记为 $R(x)$ 。容差度是两个对象在所有属性上取值相同的联合概率,记为

$$\mu(x, y) = \prod_{c \in C} P(c_i(x) = c_i(y)).$$

我们定义以程度 i ($0 \leq i \leq 1$) 近似于 x 的对象集合为: $R_i(x) = \{y \in U | \mu(x, y) > i\}$,并将关系 $\{(x, y) | (x, y) \in T \wedge \mu(x, y) > i\}$ 称为限制量化容差关系,记为 $R_i(x, y)$ 。

现在,对于论域 U 中任意两个集合 X, Y ,我们引入文[6]中对多数包含关系的定义,即:

$$Y \supseteq X \Leftrightarrow c(X, Y) \leq \beta, 0 \leq \beta < 0.5$$

其中 $c(X, Y)$ 为相对分类错误率:

$$c(X, Y) = \begin{cases} 1 - |X \cap Y| / |X|, & |X| > 0 \\ 0, & |X| = 0 \end{cases}$$

在上述定义的基础上,我们给出近似空间 (U, R_i) 中概念 X 的 β 下近似、 β 上近似及 β 边界域的定义。

设 (U, R_i) 为近似空间,其中 U 是一个对象的非空有限集合, R_i 为 U 上的限制量化容差关系,对于 $X \subseteq U$, X 的基于容差度 i 的 β 下近似定义为:

$$R_{i,\beta}X = Y \{x | x \in U \wedge X \supseteq R_i(x)\}$$

或者

$$R_{i,\beta}X = Y \{x | x \in U \wedge c(R_i(x), X) \leq \beta\}$$

$R_{i,\beta}X$ 称为 X 基于容差度 i 的 β 正域,在不影响理解的情况下将其称为 X 的 β 正域。

X 的基于容差度 i 的 β 上近似定义为:

$$\bar{R}_{i,\beta}X = Y \{x \in U | c(R_i(x), X) < 1 - \beta\}$$

X 的基于容差度 i 的 β 边界域定义为:

$$bmr_{i,\beta}(X) = Y \{x \in U | \beta < c(R_i(x), X) < 1 - \beta\}$$

X 的基于容差度 i 的 β 负域定义为:

$$neg_{i,\beta}X = Y \{x \in U | c(R_i(x), X) \geq 1 - \beta\}$$

将上面的近似集定义与完备信息系统上的变精度粗糙集模型比较,我们会发现,这里定义的模型是对原模型的发展。

对于上述定义,我们可以有下面定理成立。

定理1 对于 $\forall X \subseteq U$,下列关系式成立

$$pos_{i,\beta}(\sim X) = neg_{i,\beta}(X)$$

证明:由 $neg_{i,\beta}(X)$ 的定义知,对 $\forall x \in neg_{i,\beta}(X)$ 有 $c(R_i(x), X) \geq 1 - \beta$,即 $|R_i(x) \cap X| / |R_i(x)| \leq \beta$ 。而 $|R_i(x) \cap X| = |R_i(x)| - |R_i(x) \cap \sim X|$,所以 $c(R_i(x), \sim X) \leq \beta$,也就是说对于 $\forall x \in neg_{i,\beta}(X)$ 有 $x \in pos_{i,\beta}(\sim X)$ 。同理可证,对 $\forall x \in pos_{i,\beta}(\sim X)$ 有 $x \in neg_{i,\beta}(X)$ 。命题得证。

定理2 对于任意 $X \subseteq U$,下列关系成立

$$R_{i,\beta}X \supseteq TX, TX \supseteq \bar{R}_{i,\beta}X.$$

证明:由 TX 的定义知,对 $\forall x \in TX$ 有 $T(x) \subseteq X$,又因

为 $R_i(x) \subseteq T(x)$,所以 $R_i(x) \subseteq X$ 。由上述可得,此时必有 $R_i(x) \subseteq X$,即 $x \in R_{i,\beta}X$ 。

由 $\bar{R}_{i,\beta}X$ 的定义知,对 $\forall x \in \bar{R}_{i,\beta}X$ 必有 $R_i(x) \cap X \neq \emptyset$ 。进一步推得, $T(x) \cap X \neq \emptyset$,即 $x \in TX$,命题得证。

定理3 对于每个 $0 \leq \beta < 0.5$,下列关系成立。

$$(1) R_{i,\beta}X \supseteq R_{i,\beta}X$$

$$(2) \bar{R}_{i,\beta}\emptyset = R_{i,\beta}\emptyset = \emptyset; \bar{R}_{i,\beta}U = R_{i,\beta}U = U$$

$$(3) \bar{R}_{i,\beta}(X \cap Y) \supseteq \bar{R}_{i,\beta}X \cap \bar{R}_{i,\beta}Y$$

$$(4) R_{i,\beta}X \cap R_{i,\beta}Y \supseteq R_{i,\beta}(X \cap Y)$$

$$(5) \bar{R}_{i,\beta}(X \cup Y) \supseteq \bar{R}_{i,\beta}X \cup \bar{R}_{i,\beta}Y$$

$$(6) \bar{R}_{i,\beta}X \cap \bar{R}_{i,\beta}Y \supseteq \bar{R}_{i,\beta}(X \cap Y)$$

证明:(1)对 $\forall x \in R_{i,\beta}X$ 有 $X \supseteq R_i(x)$,因此 $c(R_i(x), X) \leq \beta \leq 1 - \beta$ 。命题得证。

(2)由于对于 $\forall R_i(x)$ 有 $c(R_i(x), \emptyset) = 1$,因此对 $0 \leq \beta < 0.5$ 可得 $R_{i,\beta}\emptyset = Y\emptyset = \emptyset$ 。同样,根据 $\bar{R}_{i,\beta}X$ 的定义可得 $\bar{R}_{i,\beta}U = U$ 。由于对于 $\forall R_i(x)$ 有 $c(R_i(x), U) = 0$,因此,根据定义可以推出 $\bar{R}_{i,\beta}U = U$ 及 $\bar{R}_{i,\beta}U = U$ ($0 \leq \beta < 0.5$)。

(3)对 $\forall X, Y \subseteq U$,根据相对错误分类率的定义,可得 $c(R_i(x), X \cap Y) \leq c(R_i(x), X)$ 和 $c(R_i(x), X \cap Y) \leq c(R_i(x), Y)$ 。因此,有 $\bar{R}_{i,\beta}(X \cap Y) \supseteq \bar{R}_{i,\beta}X$ 和 $\bar{R}_{i,\beta}(X \cap Y) \supseteq \bar{R}_{i,\beta}Y$,从而得出命题。

(4)对 $\forall X, Y \subseteq U$,同(3)可得 $c(R_i(x), X \cap Y) \geq c(R_i(x), X)$ 且 $c(R_i(x), X \cap Y) \geq c(R_i(x), Y)$ 。因此 $R_{i,\beta}X \cap R_{i,\beta}Y \supseteq R_{i,\beta}(X \cap Y)$ 。

(5)证明同(3)。

(6)证明同(4)。

3.2 属性约简原则

属性约简的前提条件是保持原始决策表中条件属性和决策属性的依赖关系不发生变化。因此,我们首先给出不完备信息系统下属性间的近似依赖性定义。

设 $S = (U, A)$ 是一个不完备信息系统,其中 $A = C \cup D$, C, D 为条件属性集和决策属性集, $ind(D)$ 表示由 D 所决定的不可分辨关系, R_i 为 C 上的限制量化容差关系。决策属性集 D 与条件属性集 C 的 β 依赖性定义为:

$$\gamma(C, D, \beta) = |pos(R_i, D, \beta)| / |U|$$

其中 $pos(R_i, D, \beta) = Y_{x \in U/D} R_{i,\beta}x$

不完备信息系统下的近似依赖性是完全信息系统下的近似依赖性思想的推广。当不完备信息系统转化为完备信息系统时,上述定义的近似依赖性就变成了完备信息系统下的近似依赖性定义。

通过使用上面的定义,我们给出决策属性集 D 关于条件属性集 C 的基于近似度 i 的 β 约简 $red(C, D, \beta)$ 的公理化定义:

定义3 称 B 是不完备信息系统的一个相对约简,是指 $B \subseteq C$ 满足条件:

$$(1) \gamma(C, D, \beta) = \gamma(red(C, D, \beta), D, \beta),$$

(2)从 $red(C, D, \beta)$ 中去掉任何一个属性都将使(1)不成立。

3.3 扩展可区分矩阵

不完备信息系统下的扩展可区分矩阵定义。

定义4 设 $S = (U, A)$ 是一个不完备信息系统,其中 $A = C \cup D$ 。为讨论方便,不妨设 $D = \{d\}$ 。令 $c_i(x_j)$ 表示样本 x_j 在属性 c_i 上的取值; $m(i, j)$ 表示可区分矩阵中第 i 行 j 列的元

素,则扩展可区分矩阵为 $M = \{m(i, j)\}_{n \times n}, (1 \leq i, j \leq |U|)$, 其中:

$$m(i, j) = \begin{cases} 0, & d(x_i) = d(x_j) \\ \{c_k | c_k \in (B_v(x_i) \cap B_v(x_j)) \wedge c_k(x_i) \neq c_k(x_j)\}, & \\ d(x_i) \neq d(x_j) \wedge B_v(x_i) \cap B_v(x_j) \neq \emptyset \wedge (\exists c_k \in B_v(x_i) \cap B_v(x_j) \wedge c_k(x_i) \neq c_k(x_j)) \\ C - \{c_k | c_k \in (B_v(x_i) \cap B_v(x_j)) \wedge c_k(x_i) = c_k(x_j)\}, & \\ d(x_i) \neq d(x_j) \wedge B_v(x_i) \cap B_v(x_j) = \emptyset \wedge (\forall c_k \in B_v(x_i) \cap B_v(x_j) \wedge c_k(x_i) = c_k(x_j)) \\ \{c_k | c_k \in C\}, & d(x_i) \neq d(x_j) \wedge B_v(x_i) \cap B_v(x_j) = \emptyset \\ \emptyset, & d(x_i) \neq d(x_j) \wedge B_v(x_i) \cap B_v(x_j) = C \wedge (\forall c_k \in B_v(x_i) \cap B_v(x_j) \wedge c_k(x_i) = c_k(x_j)) \end{cases}$$

$B_v(x)$ 定义为 $B_v(x) = \{c_k \in C | c_k(x) \neq *\}$ 。

由上述定义可知,当两个样本的决策属性相同时,它们所对应的矩阵元素为0;当两个样本的决策属性不同并且可以通过某些条件属性的取值不同加以区分时,它们所对应的矩阵元素为这两个样本属性取值不同的条件属性集合;当两个样本的决策属性不同而又不能通过某些条件属性的取值不同加以区分时,它们所对应的矩阵元素为这两个样本的属性值取*的条件属性集合的并集;当两个样本发生冲突时,即所有的条件属性值都不为*且相同而决策属性不同,它们对应的矩阵元素为 $\emptyset$ 。而且,当我们的研究对象退化为完备系统时,扩展可区分矩阵也就退化为可区分矩阵了。

利用扩展可区分矩阵对原决策表进行属性约简,得到的结果是一组约简。当我们所使用两个参数 i, β 都设为0时,本文提出的不完备信息系统下的变精度粗糙集模型便退化为基于容差关系的粗糙集模型。

4 约简算法

约简算法描述如下:

输入: $S = (U, C, Y\{d\})$, k (用户要求的最低分类精度)

输出: $B \subseteq C$, B 为 C 的一个约简

步骤1: 初始化, $B = \emptyset$ 。

步骤2: 如果 $\gamma(C, d, \beta) < k$ 则转步骤6。

步骤3: 对每个 $c \in C$ 计算 U 中在属性 c 上具有遗失属性值的元素个数 $miss_c$ 。

步骤4: 获得 S 的扩展可区分矩阵 M 。

步骤5: (1) 通过 M 找到 C 的 d 核, 即 $COR D_d(C)$, 令 $B = COR D_d(C)$, 将包含核属性的元素置为 $\emptyset$, 得新的区分矩阵 M_1 。(2) 在 M_1 中选取一个出现次数最多的属性 cc , 若存在多个这样的属性, 则选取 $miss_c$ 值最小的那个属性。(3) 令 $B = B + \{cc\}$, 计算 B 在 R_i 下的 $r(B, d, \beta)$, 如果 $\gamma(B, d, \beta) = \gamma(C, d, \beta)$ 则转步骤6。(4) 置 M_1 中含有 cc 的元素为 $\emptyset$, 如果 $M_1 \neq \emptyset$ 则转至该步骤中的(2), 否则执行下一步。

步骤6: 结束, 返回 B 。

例1 给定不完备信息系统 $S = (U, A)$, 其中 $U = \{x_1, x_2, \dots, x_{12}\}$, $A = \{c_1, c_2, c_3, c_4, d\}$ 。(见文[4])

选取 $i = 0.9, \beta = 0.1$, 用上述模型处理, 有如下结果:

$R_i(a_1) = \{a_1\}, R_i(a_2) = \{a_2, a_3\},$

$R_i(a_3) = \{a_2, a_3\}, R_i(a_4) = \{a_4\},$

$R_i(a_5) = \{a_5\}, R_i(a_6) = \{a_6\},$

$R_i(a_7) = \{a_7\}, R_i(a_8) = \{a_8\},$

$R_i(a_9) = \{a_9\}, R_i(a_{10}) = \{a_{10}\},$

$R_i(a_{11}) = \{a_{11}\}, R_i(a_{12}) = \{a_{12}\}.$

$R_{i,\beta}\Phi = \{a_1, a_4, a_7, a_{10}, a_{12}\},$

$R_{i,\beta}\Psi = \{a_1, a_2, a_3, a_4, a_7, a_{10}, a_{12}\}.$

$R_{i,\beta}\Psi = \{a_5, a_6, a_8, a_9, a_{11}\},$

$R_{i,\beta}\Psi = \{a_2, a_3, a_5, a_6, a_8, a_9, a_{11}\}.$

对象 \ 属性	c_1	c_2	c_3	c_4	d
x_1	3	2	1	0	Φ
x_2	2	3	2	0	Φ
x_3	2	3	2	0	Ψ
x_4	*	2	*	1	Φ
x_5	*	2	*	1	Ψ
x_6	2	3	2	1	Ψ
x_7	3	*	*	3	Φ
x_8	*	0	0	*	Ψ
x_9	3	2	1	3	Ψ
x_{10}	1	*	*	*	Φ
x_{11}	*	2	*	*	Ψ
x_{12}	3	2	1	*	Φ

运用我们提出的约简算法进行属性约简可得: 原决策表中条件属性 C 相对于决策属性 d 的约简为 $\{a_1, a_2, a_4\}$ 。

结束语 本文提出了一种基于不完备信息系统的变精度粗糙集模型, 在此基础上引入了扩展可区分矩阵的概念, 给出了一种有效的知识约简算法。它的最大特点是通过调节 i , 可以得到不同概念层次上的规则; 通过引入 β 增强了方法的鲁棒性。

参考文献

- 1 Pawlak Z. Rough Set-Theoretical Aspect of Reasoning about Data. Kluwer Academic Publishers, Dordrecht, Boston, London, 1991
- 2 余英泽, 王国胤, 吴渝. 一种基于 Rough 集理论的不完备信息系统处理方法. 计算机科学, 2001, 28(5. 专辑)
- 3 王国胤. Rough 集理论与知识获取. 西安交通大学出版社, 2001
- 4 Stefanowski J, Tsoukisa A. On the Extension of Rough Sets under Incomplete Information. In: 7th Intl. workshop, RSFDGRC'99 Yamaguchi, Japan, Proc. New Directions in Rough Sets, Data Mining, and Granular-soft Computing, 1999. 73~81
- 5 梁吉业, 王江, 苗夺谦. 推广粗糙集模型下粗糙集与粗糙分类的模糊性度量. 计算机科学, 2001, 28(5. 专辑)
- 6 张文修, 吴伟志, 梁吉业, 李德玉. 粗糙集理论与方法. 科学出版社, 2001
- 7 Liang Jiye, Xu Zongben. The algorithm on knowledge reduction in incomplete information systems. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(1): 95~103