

基于边界域的不完备信息系统属性约简方法

刘芳¹ 李天瑞²

(内江师范学院数学与信息科学学院 内江 641101)¹ (西南交通大学信息科学与技术学院 成都 611756)²

摘要 研究了用矩阵来计算不完备信息系统的属性约简方法,引入了容差关系矩阵等概念来计算决策表的上下近似集;然后给出了基于容差关系矩阵的决策表边界域的计算方法,再利用边界域的基数相等作为评价属性约简的准则,提出了基于边界域的启发式约简方法;最后,举例说明了属性约简的操作方法和所提算法的可行性。

关键词 粗糙集,属性约简,边界域,容差关系,不完备信息系统

中图法分类号 TP311 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.3.044

Method for Attribute Reduction Based on Rough Sets Boundary Regions

LIU Fang¹ LI Tian-rui²

(Department of Mathematics and Information Science, Neijiang Normal University, Neijiang 641101, China)¹

(School of Information Science and Technology, Southwest Jiaotong University, Chengdu 611756, China)²

Abstract The method of attribute reduction for incomplete information systems was studied by using matrix. The concept of tolerance relation matrix was introduced to calculate the upper and lower approximations in the decision table. A method for calculating the boundary region of decision table based on the tolerance relation matrix was presented. The criterion for evaluating the attribute reduction was based on the equal of cardinal number of the boundary region. A heuristic reduction method based on boundary region was proposed in this paper. At last, the feasibility of the operation method and the algorithm of attribute reduction were illustrated by examples.

Keywords Rough sets, Attribute reduction, Boundary region, Tolerance relation, Incomplete information system

1 引言

粗糙集理论是一种处理不确定性问题的数学工具,已成功应用到模式识别、数据挖掘和决策支持等领域。在数据分析任务中,并不是所有的属性都是必要的,实际上,一些属性是多余的,这些多余的属性会占用很多存储空间和耗费更多的计算时间。因此,期望能找到一个属性子集,它与原始属性集具有相同的数据分类能力,这个处理过程就叫做属性约简。通过属性约简可以减少数据冗余,从而简化决策规则^[1,2]。

经典的粗糙集理论是基于完备信息系统的,文献[3,4]提出了用不可区分矩阵进行属性约简的技术,文献[5,6,16]提出了基于正域的属性约简,文献[7,20]利用信息熵进行属性约简,这些技术为属性约简的研究开拓了新的视野,并且已经成功地应用到了完备决策信息系统中。但是现实生活中,很多情况下信息都是不完备的。专家学者对经典粗糙集理论在不完备信息系统下进行扩充,提出了容差关系、非对称相似关系、限制容差关系等。文献[8-10]分别用容差粗糙集模型、非对称容差粗糙集模型、限制容差粗糙集模型对不完备信息系统进行了研究。在这些模型中,容差粗糙集模型已被广泛应用到不完备信息系统的属性约简研究中^[11,12]。

基于矩阵的计算方法是很有实用价值的数学方法,在经典的粗糙集模型中已经有很多基于矩阵的属性约简算法,例如文献[13]运用区分矩阵研究信息系统的属性约简,文献[14]对信息系统中属性形成的不可分辨关系及等价关系矩阵进行了研究,文献[15]提出了一种利用矩阵来计算粗糙集上下近似集的方法,但以上算法并没有将矩阵的计算应用到属性约简的过程中。

目前,几乎所有研究容差粗糙集模型的属性约简算法都是基于下近似集的,很少一部分研究是基于边界域的属性约简,实际上,基于边界域的属性约简具有更高的约简效率^[16-19]。本文在文献[15,19]的基础上,给出了在不完备信息系统中利用矩阵来计算基于边界域的属性约简的启发式算法,并从理论上证明了该方法的正确性。

2 容差粗糙集相关知识

本节简单介绍容差粗糙集的相关概念^[8-10]。

决策表 $S = (U, C \cup D, V, f)$, 其中 U 是非空有限对象集, C 是条件属性集, D 是一个决策属性集, $\forall a \in C \cup D$, 映射 $f: U \rightarrow V_a$ 是信息函数, 其中 V_a 是 a 的值域。如果在条件属性 C 的值域中有缺失的数值, 这样的决策信息系统就是不完备的, 通常用“*”表示缺失的数值。

到稿日期:2015-10-08 返修日期:2015-11-12 本文受国家自然科学基金项目(61175047)资助。

刘芳(1984—),女,硕士,讲师,主要研究方向为粗糙集与粒计算等,E-mail:lytb@163.com;李天瑞(1969—),男,教授,博士生导师,主要研究方向为数据挖掘与知识发现、粗糙集与粒计算等,E-mail:trli@swjtu.edu.cn(通信作者)。

令决策表 $S=(U, C \cup D, V, f)$ 为一个不完备决策信息系统, 对任意的子集 $P \subseteq C$, P 决定的 U 上的容差关系可定义为 $SIM(P)$ 。

$$SIM(P) = \{(u, v) \in U \times U \mid \forall a \in P, f(u, a) = f(v, a) \vee f(u, a) = * \vee f(v, a) = *\}$$

容易得出 $SIM(P)$ 满足自反性、对称性, 不满足传递性。对于 $\forall u \in U$, $SIM(P)$ 决定的容差类表示为 $S_P(u)$, $S_P(u)$ 描述了与对象 u 在属性集合 P 下的最大不可区分对象子集。 $S_P(u) = \{v \in U \mid (u, v) \in SIM(P)\}$, 显然 $S_P(u)$ 是 U 上的一个覆盖, $\bigcup_{u \in U} S_P(u) = U$ 。

$\pi_D = \{D_i \mid i=1, 2, \dots, m\}$ 是由决策属性集 D 决定的在 U 上的划分, 其中 D_i 为决策类, 容差决策粗糙集模型的上、下近似集分别定义为 $\bar{P}(D_i) = \{u \in U \mid S_P(u) \cap D_i \neq \emptyset\}$, $\underline{P}(D_i) = \{u \in U \mid S_P(u) \subseteq D_i\}$ 。

通过上、下近似集, 决策系统 S 的对象集合 U 被划分到两个互不相交的决策域: 正域 $POS_P(\pi_D)$ 和边界域 $BND_P(\pi_D)$ 。

$$POS_P(\pi_D) = \bigcup_{i=1}^m \underline{P}(D_i)$$

$$BND_P(\pi_D) = \bigcup_{i=1}^m (\bar{P}(D_i) - \underline{P}(D_i))$$

因为 $\pi_D = \{D_i \mid i=1, 2, \dots, m\}$ 是由决策属性集 D 决定的在 U 上的划分, 所以 $\bigcup_{i=1}^m D_i = U$; 又因为对于 $\forall u \in U$, 总是存在一个 $D_i \in \pi_D$ 使得 $S_P(u) \cap D_i \neq \emptyset$, 所以 $\bigcup_{i=1}^m \bar{P}(D_i) = U$, 即对象集合 U 负域为空, 所以 $POS_P(\pi_D) \cup BND_P(\pi_D) = U$, 并且 $POS_P(\pi_D) \cap BND_P(\pi_D) = \emptyset$ 。当且仅当 $BND_P(\pi_D) = \emptyset$ 或者 $POS_P(\pi_D) = U$ 时, 决策表 S 是一致的, 否则决策表 S 是不一致的。

3 基于矩阵的边界域计算方法

本节提出了在容差粗糙集模型中边界域的基于矩阵计算方法的理论分析。

3.1 理论分析

定义 1 给定一个决策表 $S=(U, C \cup D, V, f)$, $P \subseteq C$, 令 $U = \{u_1, u_2, \dots, u_n\}$, 并且 $SIM(P)$ 是 U 上的容差关系, $S_P(u)$ 是容差类, $M_{n \times n}^{S_P} = (m_{ij})_{n \times n}$ 是 $SIM(P)$ 对应的容差关系矩阵。

$$m_{ij} = \begin{cases} 1, & (u_i, u_j) \in S_P(u) \\ 0, & (u_i, u_j) \notin S_P(u) \end{cases}$$

定义 2^[15] 令 X 为 U 的子集, 特征函数 $G(X) = (g_1, g_2, \dots, g_n)^T$ (T 表示转置操作), 其中

$$g_i = \begin{cases} 1, & u_i \in X \\ 0, & u_i \notin X \end{cases}$$

定义 3^[15] 给定一个决策表 $S=(U, C \cup D, V, f)$, $P \subseteq C$, 决策属性 D 导出的划分为 $\pi_D = \{D_i \mid i=1, 2, \dots, m\}$, $G(D_i)$ 是基于 D_i 的 n 列的布尔向量, 令决策矩阵 $GD_{n \times m} = (G(D_1), G(D_2), \dots, G(D_m))$, 它是一个 $n \times m$ 的布尔矩阵。

定义 4 令 $P \subseteq C$, 并且 $SIM(P)$ 是 U 上的容差关系, $\Lambda_{n \times n}^{S_P}$ 是 $M_{n \times n}^{S_P} = (m_{ij})_{n \times n}$ 的诱导对角矩阵, $\Lambda_{n \times n}^{S_P} = \text{diag}(\frac{1}{\lambda_1}, \frac{1}{\lambda_2}, \dots, \frac{1}{\lambda_n})$, 其中 $\lambda_i = \sum_{j=1}^n m_{ij}$, $1 \leq i \leq n$ 。

定义 5^[15] 给定决策表 $S=(U, C \cup D, V, f)$, $P \subseteq C$,

$\pi_D = \{D_i \mid i=1, 2, \dots, m\}$ 是决策 D 上的一个划分, $\forall D_i \in \pi_D$, $H(D_i)$ 是一个基本向量, 令 $HD = (H(D_1), H(D_2), \dots, H(D_m))$ 为一个 $n \times m$ 的基本矩阵, 则 $HD = \Lambda_{n \times n}^{S_P} \cdot (M_{n \times n}^{S_P} \cdot GD_{n \times m}) = \Lambda_{n \times n}^{S_P} \cdot \Omega_{n \times m}^{S_P}$, 其中 $\cdot$ 表示矩阵的数量积, $\Omega_{n \times m}^{S_P}$ 叫做中间矩阵。

定义 6 基本矩阵 HD 的 λ -截矩阵 $(HD_\lambda)_{n \times m}$ ($0 \leq \lambda \leq 1$) 可定义为:

$$(u_\lambda)_{ij} = \begin{cases} 1, & u_{ij} \geq \lambda \\ 0, & u_{ij} < \lambda \end{cases}$$

定理 1 给定决策表 $S=(U, C \cup D, V, f)$, $P \subseteq C$, $\pi_D = \{D_i \mid i=1, 2, \dots, m\}$ 是决策 D 上的一个划分, 基本向量 $H(D_i)$ 的 1-截矩阵为 $H(D_i)_1$, 则子集 D_i 的下近似集的布尔列矩阵为 $G(\underline{P}(D_i)) = H(D_i)_1$ 。

证明: 如果 $u_i \in \underline{P}(D_i)$, 那么 $[u_i]_P \subseteq D_i$, 即 $\frac{|[u_i]_P \cap D_i|}{|[u_i]_P|} = 1$, 所以 $|[u_i]_P \cap D_i| = |[u_i]_P|$ 。

由此可以得到 $(M_{n \times n}^{S_P} \cdot GD_{n \times m})_{i1} = \sum_{j=1}^m m_{ij} g_{j1} = |[u_i]_P|$, 从而有 $(\Lambda_{n \times n}^{S_P} \cdot (M_{n \times n}^{S_P} \cdot GD_{n \times m}))_{i1} = \frac{1}{|[u_i]_P|} \cdot \sum_{j=1}^m m_{ij} g_{j1} = 1$, 所以 $(\Lambda_{n \times n}^{S_P} \cdot (M_{n \times n}^{S_P} \cdot GD_{n \times m}))_{i1}$ 的 1-截矩阵等于 1, 故有 $G(\underline{P}(D_i)) = H(D_i)_1$ 。

类似地, 可以证明 $G(\underline{P}(D_i)) = H(D_i)_1$, $i = \{1, 2, \dots, m\}$ 。证毕。

定理 2^[16] 决策表 $S=(U, C \cup D, V, f)$, $P \subseteq C$, $\pi_D = \{D_i \mid i=1, 2, \dots, m\}$ 是决策 D 上的一个划分, D 关于 P 的正域为 $POS_P(\pi_D) = \bigcup_{i=1}^m \underline{P}(D_i)$, 则 $G(\bigcup_{i=1}^m \underline{P}(D_i)) = \sum_{i=1}^m G(\underline{P}(D_i))$ 。

定理 3 决策表 $S=(U, C \cup D, V, f)$, $P \subseteq C$, $\pi_D = \{D_i \mid i=1, 2, \dots, m\}$ 是决策 D 上的一个划分, D 关于 P 的边界域为 $BND_P(\pi_D) = \bigcup_{i=1}^m (\bar{P}(D_i) - \underline{P}(D_i))$, 则 $G(\bigcup_{i=1}^m (\bar{P}(D_i) - \underline{P}(D_i))) = J - \sum_{i=1}^m G(\underline{P}(D_i))$, 其中 $J = (1, 1, \dots, 1)^T$ 为对象集合 U 的布尔列向量。

证明: 因为 $\pi_D = \{D_i \mid i=1, 2, \dots, m\}$ 是由决策属性集 D 决定的在 U 上的划分, 所以 $\bigcup_{i=1}^m D_i = U$; 又因为对于 $\forall u \in U$, 总是存在一个 $D_i \in \pi_D$ 使得 $S_P(u) \cap D_i \neq \emptyset$, 所以 $\bigcup_{i=1}^m \bar{P}(D_i) = U$, 所以 $POS_P(\pi_D) \cup BND_P(\pi_D) = U$, 而且 $POS_P(\pi_D) \cap BND_P(\pi_D) = \emptyset$, 可以得到 $BND_P(\pi_D) = U - POS_P(\pi_D)$ 。因为对象集合 U 的布尔列向量 $G(U) = (1, 1, \dots, 1)^T$, 则 $G(BND_P(\pi_D)) = G(U) - G(POS_P(\pi_D))$ 。

因为 $BND_P(\pi_D) = \bigcup_{i=1}^m (\bar{P}(D_i) - \underline{P}(D_i))$, 令 $J = G(U)$, 则根据定理 2 的结论可以得到 $G(\bigcup_{i=1}^m (\bar{P}(D_i) - \underline{P}(D_i))) = J - \sum_{i=1}^m G(\underline{P}(D_i))$ 。

证毕。

3.2 算法描述

算法 1 计算容差粗糙集模型中决策类的边界域的算法输入: 一个决策表 $S=(U, C \cup D, V, f)$ 和条件属性子集 $P \subseteq C$ 输出: 边界域集合 $BND_P(\pi_D)$

Step 1 构建容差关系矩阵 $M_{n \times n}^{S_P}$;

Step 2 计算诱导矩阵 $\Lambda_{n \times n}^{S_P}$;

Step 3 构建决策矩阵 $GD_{n \times m}$;

Step 4 计算中间矩阵 $\Omega_{n \times m}^S$;

Step 5 计算基本矩阵 HD ;

Step 6 计算基本矩阵 HD 的 1-截矩阵 $(HD_1)_{n \times m}$;

Step 7 计算决策属性 D 的划分 π_D 的任意子集 D_i 的下近似集合矩阵 $G(\underline{P}(D_i)) = H(D_i)_1, i = \{1, 2, \dots, m\}$;

Step 8 计算决策属性 D 关于条件属性 P 的边界域 $G(BNG_P(D)) = G(\bigcup_{i=1}^m (\bar{P}(D_i) - \underline{P}(D_i))) = J - \sum_{i=1}^m G(\underline{P}(D_i))$;

Step 9 得出决策类的边界域。

3.3 算例

例 1 给定一个不完备决策系统 $S = (U, C \cup D, V, f)$, 如表 1 所列。其中 $U = \{u_1, u_2, \dots, u_6\}$, $C = \{c_1, c_2, c_3, c_4\}$, $D = \{d\}$ 为决策属性。属性集 c_1, c_2, c_3, c_4 和 d 分别表示 Price、Mileage、Size、Max-Speed 和 Acceleration。令 $P = \{c_3\}$, 则边界域的计算过程如下。

表 1 一个不完备决策系统

Car	C ₁	C ₂	C ₃	C ₄	d
u ₁	high	high	full	low	good
u ₂	*	*	full	low	good
u ₃	*	*	compact	low	poor
u ₄	low	*	full	high	good
u ₅	*	*	full	high	excellent
u ₆	high	high	full	*	good

根据定义 1 可以得到不完备决策表 1 的容差关系矩阵

$$M_{6 \times 6}^{S_P} = (m_{ij})_{6 \times 6}:$$

$$M_{6 \times 6}^{S_P} = \begin{bmatrix} 1 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 \end{bmatrix}$$

根据定义 2 和定义 3 可以计算出决策矩阵 $GD_{6 \times 3} = (G(D_1), G(D_2), G(D_3))$ 。其中 $G(D_1) = (1, 1, 0, 1, 0, 1)^T$, $G(D_2) = (0, 0, 1, 0, 0, 0)^T$, $G(D_3) = (0, 0, 0, 0, 1, 0)^T$ 。

根据定义 4 可以计算出诱导矩阵 $\Lambda_{6 \times 6}^{S_P}$:

$$\Lambda_{6 \times 6}^{S_P} = \text{diag}(\frac{1}{5}, \frac{1}{5}, 1, \frac{1}{5}, \frac{1}{5}, \frac{1}{5})$$

由以上计算结果可以得到中间矩阵 $\Omega_{6 \times 3}^{S_P} = M_{6 \times 6}^{S_P} \cdot GD_{6 \times 3}$:

$$\Omega_{6 \times 3}^{S_P} = \begin{bmatrix} 1 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 4 & 0 & 1 \\ 4 & 0 & 1 \\ 0 & 1 & 0 \\ 4 & 0 & 1 \\ 4 & 0 & 1 \\ 4 & 0 & 1 \end{bmatrix}$$

根据诱导矩阵和中间矩阵可以计算出基本矩阵 $HD = \Lambda_{6 \times 6}^{S_P} \cdot \Omega_{6 \times 3}^{S_P}$:

$$HD = \text{diag}(\frac{1}{5}, \frac{1}{5}, 1, \frac{1}{5}, \frac{1}{5}) \cdot \begin{bmatrix} 4 & 0 & 1 \\ 4 & 0 & 1 \\ 0 & 1 & 0 \\ 4 & 0 & 1 \\ 4 & 0 & 1 \\ 4 & 0 & 1 \end{bmatrix} = \begin{bmatrix} \frac{4}{5} & 0 & \frac{1}{5} \\ \frac{4}{5} & 0 & \frac{1}{5} \\ 0 & 1 & 0 \\ \frac{4}{5} & 0 & \frac{1}{5} \\ \frac{4}{5} & 0 & \frac{1}{5} \\ \frac{4}{5} & 0 & \frac{1}{5} \end{bmatrix}$$

所以, 基本矩阵 HD 的 1-截矩阵 $(HD_1)_{6 \times 3}$ 为:

$$(HD_1)_{6 \times 3} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

根据定理 1—定理 3 以及 $(HD_1)_{6 \times 3}$ 的计算结果, 可以得到不完备决策表 1 在容差粗糙集模型下的下近似集和边界域:

$$G(\underline{P}(D_1)) = H(D_1)_1 = (0, 0, 0, 0, 0, 0)^T$$

$$G(\underline{P}(D_2)) = H(D_2)_1 = (0, 0, 1, 0, 0, 0)^T$$

$$G(\underline{P}(D_3)) = H(D_3)_1 = (0, 0, 0, 0, 0, 0)^T$$

然后, 可以计算出

$$G(BNG_P(D)) = J - \sum_{i=1}^m G(\underline{P}(D_i)) = (1, 1, 0, 1, 1, 1)^T$$

所以, 该容差粗糙集模型 D 关于 P 的边界域为: $BND_P(D) = \{u_1, u_2, u_4, u_5, u_6\}$ 。

4 基于边界域的启发式约简

4.1 属性递增计算边界域的方法

定义 7^[19] 给定决策表 $S = (U, C \cup D, V, f)$, $C = \{c_1, c_2, \dots, c_n\}$, 并且 $P_k = \{c_1, c_2, \dots, c_k\}, k \leq n$ 。则属性个数的递增序列为 $P = \{P_1, P_2, \dots, P_n\}$, $\forall P_i, P_j \in P$, 当 $P_i \subseteq P_j$ 时, 由 P_i 决定的容差关系比 P_j 决定的容差关系更细; 换言之, P_i 决定的容差关系比 P_j 决定的容差关系更粗。

定义 7 说明了当属性个数单调增加时, 正域的个数单调增加, 边界域的个数单调减少。该定义描述了通过属性的逐步增加计算不完备决策表的边界域的方法。该方法如算法 2 所示。

算法 2 不完备决策信息系统中属性个数逐步增加求解边界域的计算方法

输入: $S = (U, C \cup D, V, f)$, $D = \{d\}$, $C = \{c_1, c_2, \dots, c_n\}$

输出: 不完备信息系统的边界域 $BND_C(D)$

Step 1 通过算法 1 计算 $BND_{c_1}(D) = X, X \subseteq U$ 为对象集合 U 的子集。

Step 2 令 U 等于 Step1 计算得到的边界集合 X , 同时在新的对象集合下计算决策类集合 ND , 通过算法 1 计算出 $BND_{c_2}(ND)$ 。同样, 令 $BND_{c_2}(D) = X$, 循环 Step2, 直至计算出 $BND_{c_n}(ND)$ 。

Step 3 输出 $BND_C(D) = BND_{c_n}(ND)$ 。

例 2(继续例 1) 给定一个如表 1 所列的不完备信息系统, 令 $C = \{c_3, c_4, c_1, c_2\}$ 为代入算法 2 的属性计算顺序。通过例 1 计算出 $BND_{c_3}(D) = X = \{u_1, u_2, u_4, u_5, u_6\}$, 所以令 $U =$

X, 则此时在属性 c_4 决定的容差关系矩阵为:

$$M_{5 \times 5}^{c_4} = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \end{bmatrix}$$

决策矩阵 $GD_{5 \times 3} = (G(D_1), G(D_2), G(D_3))$ 。其中 $G(D_1) = (1, 1, 1, 0, 1)^T$, $G(D_2) = (0, 0, 0, 0, 0)^T$, $G(D_3) = (0, 0, 0, 1, 0)^T$ 。

诱导矩阵 $\Lambda_{5 \times 5}^{c_4} = \text{diag}(\frac{1}{3}, \frac{1}{3}, \frac{1}{2}, \frac{1}{3}, \frac{1}{4})$ 。

中间矩阵 $\Omega_{5 \times 3}^{c_4} = M_{5 \times 5}^{c_4} \cdot GD_{5 \times 3}$ 。

$$\Omega_{5 \times 3}^{c_4} = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 3 & 0 & 0 \\ 3 & 0 & 0 \\ 1 & 0 & 1 \\ 2 & 0 & 1 \\ 3 & 0 & 1 \end{bmatrix}$$

基本矩阵

$$HD = \text{diag}(\frac{1}{3}, \frac{1}{3}, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}) \cdot \begin{bmatrix} 3 & 0 & 0 \\ 3 & 0 & 0 \\ 1 & 0 & 1 \\ 2 & 0 & 1 \\ 3 & 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{2}{3} & 0 & \frac{1}{3} \\ \frac{3}{4} & 0 & \frac{1}{4} \end{bmatrix}$$

所以, 基本矩阵 HD 的 1-截矩阵 $(HD)_1$ 为:

$$(HD)_1 = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

下近似集的布尔列向量分别为:

$$G(\underline{P}(D_1)) = H(D_1)_1 = (1, 1, 0, 0, 0)^T$$

$$G(\underline{P}(D_2)) = H(D_2)_1 = (0, 0, 0, 0, 0)^T$$

$$G(\underline{P}(D_3)) = H(D_3)_1 = (0, 0, 0, 0, 0)^T$$

然后, 可以计算出边界域的布尔列向量:

$$G(BNG_{c_4}(D)) = J - \sum_{i=1}^m G(\underline{P}(D_i)) = (0, 0, 1, 1, 1)^T$$

所以, 该容差粗糙集模型 D 关于 $P = \{c_3, c_4\}$ 的边界域为: $BND_P(D) = \{u_4, u_5, u_6\}$ 。

然后, 令 $U = \{u_4, u_5, u_6\}$, 同时在决策类集合 D 中也删除不在边界域中的对象, 与新的对象集合 U 保持一致, 则在新的对象集合下决策矩阵为 $GD_{3 \times 3} = (G(D_1), G(D_2), G(D_3))$ 。其中 $G(D_1) = (1, 0, 1)^T$, $G(D_2) = (0, 0, 0)^T$, $G(D_3) = (0, 1, 0)^T$ 。循环使用算法 1 继续计算出 $BND_{c_1}(D) = \{u_4, u_5, u_6\}$, $BND_{c_2}(D) = \{u_4, u_5, u_6\}$ 。这时对所有的属性都进行了边界域的计算, 所以该不完备信息系统的边界域为 $BND_C(D) = \{u_4, u_5, u_6\}$ 。

4.2 基于边界域的属性约简

定义 7 说明了当属性个数单调增加时, 边界域中的对象

个数呈单调递减的趋势, 因此确定一个最优的属性顺序对于属性约简的计算起至关重要的作用。

在决策表 $S = (U, C \cup D, V, f)$ 中, 粗糙集理论通过一个变量的依赖度来描述属性 a 的重要性, $Sig(a, C, D) = \gamma_C(D) - \gamma_{C-\{a\}}(D)$, 其中 $Sig(a, C, D)$ 称为属性 a 的重要度, $\gamma_C(D)$ 是 D 对 C 的依赖度, 而且 $\gamma_C(D) = |POS_C(\pi_D)| / |U|$ 。 $Sig(a, C, D)$ 的值越大, 代表属性 a 越重要; 反之, 则代表属性 a 越不重要。有了这种关系, $Sig(a, C, D)$ 经常用来评估一个属性的重要程度。本节将该重要度方法进行变换, 通过边界域的分类质量来得出属性的重要度, 从而构建出一个计算属性约简的最优的属性选择顺序。

定理 4^[19] $Sig(a, C, D) = (|BND_{C-\{a\}}(\pi_D)| - |BND_C(\pi_D)|) / |U|$ 。

定理 4 描述了一种基于边界域的属性重要性的计算方法。如果 $Sig(a, C, D) = 0$, 则属性 a 是多余的, 否则属性 a 是一个核属性。因此可以将各个属性按照它们的属性重要度的值由大到小进行排序, 具有最大重要度值的一个属性被优先选择去引导搜索属性约简的处理过程。

定理 5^[19] 一个子集 P 是一个约简, 当且仅当 $|BND_P(\pi_D)| = |BND_C(\pi_D)|$ 和 $|BND_{P'}(\pi_D)| \neq |BND_C(\pi_D)|$ 同时成立, 其中 $\forall P' \subset P$ 。

定理 5 描述了一个基于边界域的属性约简评价标准, 比较子集 P 与初始集合 C 的边界域基数, 如果相等, 而且没有任何子集 $P' \subset P$ 满足该相等的条件, 则子集 P 即为一个属性约简。基于边界域的启发式属性约简算法如算法 3 所示。

算法 3 基于边界域的启发式属性约简算法

输入: $S = (U, C \cup D, V, f)$, $D = \{d\}$, $C = \{c_1, c_2, \dots, c_n\}$

输出: 一个约简 R

Step 1 通过算法 2 计算 $BND_C(\pi_D)$;

Step 2 循环 n 次通过算法 2 计算 $BND_{C-\{c_i\}}(\pi_D)$, 同时计算出各个属性 c_i 的重要度 $Sig(c_i, C, D) = (|BND_{C-\{c_i\}}(\pi_D)| - |BND_C(\pi_D)|) / |U|$, 其中 $i = \{1, 2, \dots, n\}$;

Step 3 将属性 $c_1, c_2, \dots, c_n$ 按计算出的属性重要度降序排列为属性序列 $a_1, a_2, \dots, a_n$;

Step 4 通过算法 2 按 Step 3 得到的降序序列计算 $BND_{a_i}(\pi_D)$, 直到 $BND_{a_i}(\pi_D) = BND_C(\pi_D)$ 为止;

Step 5 令 $R = \{a_1, a_2, \dots, a_i\}$ 即是要得到的一个属性约简, 输出 R 。

结束语 本文基于矩阵的基本运算和容差关系矩阵等概念, 提出了使用矩阵方法计算粗糙集中不完备信息系统的上、下近似集, 并在此基础上研究了不完备决策表边界域的计算方法, 然后利用边界域的基数大小计算出各个属性的重要度, 实现了基于矩阵的不完备信息系统的启发式约简算法。理论分析和实例表明, 本文提出的基于边界域的不完备信息系统属性约简的矩阵计算方法是可行的。

参考文献

- [1] Pawlak Z. Rough Sets-Theoretical Aspects of Reasoning about Data[D]. Dordrecht: Kluwer Academic, 1991
- [2] Yao Y Y, Zhao Y. Attribute reduction in decision-theoretic rough set models[J]. Information Sciences, 2008, 178(17): 3356-3373

- [14] Wu S T, Li Y, Xu Y. Deploying Approaches for Pattern Refinement in Text Mining[C]//Proc. IEEE Sixth Int'l Conf. Data Mining (ICDM '06). 2006;1157-1161
- [15] Liao Jia-bin, Lin Shi-ping. Cluster-based Shallow Semantic Pattern Recognition In Chinese [C]//Fifteenth Annual Academic of Information Theory and First National Annual Academic on Network Coding. Qingdao, China, 2008;1214-1219(in Chinese)
廖加彬, 林世平. 基于聚类的中文浅层语义模式识别[C]//中国电子学会第十五届信息论学术年会暨第一届全国网络编码学术年会. 青岛, 2008;1214-1219
- [16] Zhong Ning, Li Yue-feng, Wu Sheng-tang. Effective Pattern Discovery for Text Mining[J]. IEEE Transactions on Knowledge and Data Engineering, 2012, 24(1): 30-44
- [17] Wen Hao, Wen You-kui. Research on Discovery Method in Literature Implicit Knowledge Using Semantic Complementary Reasoning[J]. Computer Science, 2014, 41(6): 171-175(in Chinese)
温浩, 温有奎. 基于语义互补推理的文献隐含知识的发现方法研究[J]. 计算机科学, 2014, 46(6): 171-175
- [18] Huang Cheng-hui, Yin Jian, Hou Fang. Improved Text Retrieve Algorithm Based on Subject-verb-object Structure[J]. Computer Science, 2010, 37(9): 173-176(in Chinese)
黄承慧, 印鉴, 侯防. 一种基于主谓宾结构的文本检索算法[J]. 计算机科学, 2010, 37(9): 173-176

(上接第 245 页)

- [3] Ma X, Zhang J F, Yang H F. Heuristic algorithm of attribute reduction based on discernibility matrix[J]. Journal of Computer Applications, 2010, 30(8): 1999-2002(in Chinese)
马翔, 张继福, 杨海峰. 基于区分矩阵的启发式属性约简算法[J]. 计算机应用, 2010, 30(8): 1999-2002
- [4] Yao Y Y, Zhao Y. Discernibility matrix simplification for constructing attribute reducts[J]. Information Sciences, 2009, 179(7): 867-882
- [5] Liu S H, Sheng Q J, Wu B, et al. Research on efficient algorithms for rough set methods[J]. Chinese Journal of Computers, 2003, 26(5): 524-529(in Chinese)
刘少辉, 盛秋戩, 吴斌, 等. Rough 集高效算法的研究[J]. 计算机学报, 2003, 26(5): 524-529
- [6] Ge H, Li L S, Yang C J. Improvement to quick attribution reduction algorithm[J]. Journal of Chinese Computer Systems, 2009, 30(2): 308-312(in Chinese)
葛浩, 李龙澍, 杨传健. 改进的快速属性约简算法[J]. 小型微型计算机系统, 2009, 30(2): 308-312
- [7] Wang G Y, Yu H, Yang D C. Decision table reduction based on conditional information entropy[J]. Chinese Journal of Computers, 2002, 25(7): 759-766(in Chinese)
王国胤, 杨大春. 基于条件信息熵的决策表约简[J]. 计算机学报, 2002, 25(7): 759-766
- [8] Kryszkiewicz M. Properties of incomplete information systems in the framework of rough sets[C]//Polkowski L, Skowron A, eds. Rough Sets in Data Mining and Knowledge Discovery Physica. Heidelberg, 1998, 1: 422-450
- [9] Stefanowski J, Tsoukias A. On the extension of rough sets under incomplete information[C]//Proceedings of the Seventh International Workshop on New Directions in Rough Sets, Data Mining and Granular-Soft Computing. Physica Verlag, Yamaguchi, 1999; 73-81
- [10] Wang G Y. Extension of rough set under incomplete information systems[J]. Computer Research and Development, 2002, 39(10): 1238-1243(in Chinese)
王国胤. Rough 集理论在不完备信息系统中的扩充[J]. 计算机研究与发展, 2002, 39(10): 1238-1243
- [11] Wang W, Xu Z Y, Li X Y. Attribute Reduction Algorithm Based on Object Matrix in Incomplete Decision Table[J]. Computer Science, 2012, 39(4): 201-204(in Chinese)
王炜, 徐章艳, 李晓瑜. 不完备决策表中基于对象矩阵属性约简算法[J]. 计算机科学, 2012, 39(4): 201-204
- [12] Sun L, Xu J C, Tian Y. Feature selection using rough entropy-based uncertainty measures in incomplete decision systems[J]. Knowledge-Based Systems, 2012, 36: 206-216
- [13] Skowron A, Swinarski R, Synak P. Approximation Spaces and Information Granulation[C]//Proc of the International Conference on Rough Sets and Current Trends in Computing. Uppsala, Sweden, 2004; 116-126
- [14] Guan J W, Bell D A, Guan Z. Matrix Computation for Information System[J]. Information Sciences, 2001, 131(1-4): 129-156
- [15] Wang L, Li T R. Matrix-Based Computational Method for Upper and Lower Approximations of Rough Sets[J]. Pattern Recognition and Artificial Intelligence, 2011, 24(6): 757-762(in Chinese)
王磊, 李天瑞. 基于矩阵的粗糙集上下近似的计算方法[J]. 模式识别与人工智能, 2011, 24(6): 757-762
- [16] Jing Y G, Li T R. Reduction Algorithm of Positive Domain for Decision Table Based on Relationship Matrix[J]. Computer Science, 2013, 40(11): 261-264(in Chinese)
景运革, 李天瑞. 一种基于关系矩阵的决策表正域约简算法[J]. 计算机科学, 2013, 40(11): 261-264
- [17] Inuiguchi M, Tsurumi M. Measures based on upper approximations of rough sets for analysis of attribute importance and interaction[J]. International Journal of Innovative Computing, Information and Control, 2006, 2(1): 1-12
- [18] Parthala N M, Shen Q. Exploring the boundary region of tolerance rough sets for feature selection[J]. Pattern Recognition, 2009, 42: 655-667
- [19] Lu Z C, Zheng Q, Zhang Y Q, et al. A fast feature selection approach based on rough set boundary regions[J]. Pattern Recognition Letters, 2014, 36: 81-88
- [20] Chen Y Y, Chen Y M. Attribute Reduction Algorithm Based on Information Entropy and Ant Colony Optimization[J]. Journal of Chinese Computer Systems, 2015, 36(3): 586-590(in Chinese)
陈颖悦, 陈玉明. 基于信息熵与蚁群优化的属性约简算法[J]. 小型微型计算机系统, 2015, 36(3): 586-590