

计算机技术

基于扩展可辨识矩阵的混合决策系统属性约简

赵焱¹ 杨静² 刘海峰¹ 石瀚洋¹

(太原理工大学信息工程学院¹, 信息化管理与建设中心², 太原 030024)

摘要 经典粗糙集理论的研究对象只能是完备的离散决策系统。为了直接对不完备混合决策系统进行属性约简,因此通过引入邻域关系和限制容差关系的概念对可辨识矩阵的定义进行了扩展,提出了一种基于扩展可辨识矩阵的属性约简算法;该算法可以兼容处理完备与不完备混合决策系统。通过 UCI 数据集的仿真实验证明了该算法的有效性,最后讨论了扩展可辨识矩阵中的邻域阈值选择对属性约简结果的影响。

关键词 完备与不完备混合决策系统 邻域 限制容差关系 扩展可辨识矩阵 属性约简

中图法分类号 TP183; **文献标志码** A

粗糙集理论是波兰华沙理工大学 Pawlak 教授在 1982 年提出的^[1],研究处理不精确、不一致、不完整等信息的有效工具,目前被广泛应用于机器学习、人工智能、数据挖掘等领域。粗糙集理论中的重要概念之一就是约简^[2]。实现决策系统信息处理的关键内容是知识的约简及其算法。由于经典粗糙集只能处理完备的离散型数据,为了使粗糙集理论能适应于不完备系统,用它时应该先预处理:离散化数值型属性。但是信息丢失在离散化的过程中是无法避免的,并且或多或少地改变原系统的信息成分。但很大程度上,计算处理的结果都取决于离散化的效果。当利用经典粗糙集处理不完备决策系统时,也要进行数据的预处理,不同的数据预处理方法将产生不同的特征选择和属性约简结果。

本文利用扩展可辨识矩阵对混合决策系统提出了一种属性约简算法。该算法的第一步采用欧氏距离,通过邻域信息粒子逼近论域中任一对象,

构造出一个邻域粒子族,此邻域粒子族覆盖论域。第二步依照限制容差关系的定义,通过扩展可辨识矩阵来找到能区分任意两个邻域粒子的属性组合。最后通过理论分析以及 UCI (University of California Irvine) 机器学习数据库数据集的仿真实验共同例证,本文算法在约简不完备混合决策系统的冗余属性方面有着先进之处,并且兼容适用于完备混合决策系统,此外还讨论了邻域阈值的大小对属性约简结果的影响。

1 混合决策系统下的粗糙集扩展模型

1.1 决策系统

粗糙集理论中的决策系统 (decision system) 为 $S_{DS} = (U, A, V, f)$, 其特点是 V 中既含有离散值集合,又含有连续值集合甚至模糊值集合。其中 U 为论域,是对象的非空有限集合, $A = C \cup D$ 为属性集合,且 $C \cap D = \emptyset$, C 和 D 分别表示条件属性集和决策属性集, V 为属性值域, $f: U \times A \rightarrow V$ 为映射函数。假如在所有属性中至少存在一个 $a \in B \subseteq A$ 使得 V_a 含有空值(用“*”表示), V_a 为属性集和 A 的值域,那么我们就把 S_{DS} 叫做是不完备混合决策系统,表示为 IDS (incomplete decision system); 否则称为完

2013年8月1日收到,8月30日修改

第一作者简介:赵焱(1986—),女,硕士研究生。研究方向:数据挖掘。E-mail: jzyhxf2011@sina.com。

备混合决策系统。

决策属性是完备的,并且未知属性值只是在对象的条件属性中出现,即 $* \in V_C, * \notin V_D$,其中 V_C, V_D 分别为对应条件属性集和决策属性集的值域,这样混合决策系统无论是否含有未知的属性值,都能够做出正确的决策^[3]。

1.2 基于限制容差关系的扩充粗糙集模型

当 IDS 中的所有未知属性值都是遗漏型时, M. Kryszkiewicz 提出了基于容差关系的扩充粗糙集模型^[4]其具有自反性和对称性,但不一定满足传递性。Stefanowski 提出了基于相似关系的扩充粗糙集模型^[5],其不满足对称性,但满足自反性和传递性。在 M. Kryszkiewicz 的基于容差关系的扩充粗糙集模型中,两个体很容易被误判定在同一个容差类中:如极少相同已知属性信息时,以及没有明确相同的已知属性信息时;在 Stefanowski 的基于相似关系模型中,一些信息不可以在同一个相似类中被划分,这些信息具有很多相同已知属性。针对这些扩充关系所遇到的问题,一种限制容差关系被学者们提出。

定义 1^[6]: 令 S 为一 IDS, 对于 $\forall B \subseteq A$, 令 $P_B(x) = \{a \in B: f(x, a) \neq *\}$, 则由属性集合 A 的任意子集 B 决定的限制容差关系记为 $L_B(x, y)$ 且 $L_B(x, y) = \{(x, y) \in U^2: \forall a \in B, f(x, a) = f(y, a) = * \vee ((P_B(x) \cap P_B(y) \neq \emptyset) \wedge \forall_{a \in B} (f(x, a) \neq *) \wedge (f(y, a) \neq *) \rightarrow (f(x, a) = f(y, a)))\}$ (1)

显然,自反性和对称性是限制容差关系所具备的,可是传递性在这里没有体现。

性质 1^[7] $\underline{B}^T(X) \subseteq \underline{B}^L(X) \subseteq \underline{B}^S(X), \bar{B}^S(X) \subseteq \bar{B}^L(X) \subseteq \bar{B}^T(X)$ 。

限制容差关系不走非对称相似关系和容差关系的极端,正好在两者之间。其所划分的粒度比非对称相似关系所划分的大,比容差关系所划分的小。同时在容差关系中易形成不一致性规则,以及非对称相似关系中每条规则的对象数太少的短处也一并被解决了。

2 扩展可辨识矩阵

2.1 可辨识矩阵

定义 2^[8]: 令 $S = (U, A, V, f)$ 是一个决策表系统, $A = C \cup D$ 是属性集合,其中子集 C 和 D 分别表示条件属性集和决策属性集, U 为论域且 $U = \{x_1, x_2, \dots, x_n\}$, $V = \bigcup_{r \in R} V_r$ 是属性值集合, $f: U \times R \rightarrow V$ 为信息函数,其以函数形式指定 U 中每一个对象 x 的属性值, V_r 为属性 $r \in R$ 的属性值范围,也就是属性 r 的值域。 $C_D(i, j)$ 为可辨识矩阵内第 i 行,第 j 列元素, $a_i(x_j)$ 为样本 x_j 在属性 a_i 上的取值。故可辨识矩阵可记:

$$C_D(i, j) = \begin{cases} \{a_k \mid a_k \in C \wedge a_k(x_i) \neq a_k(x_j)\} & d(x_i) \neq d(x_j) \\ 0, & d(x_i) = d(x_j) \end{cases} \quad (2)$$

式(2)中 $i, j = 1, 2, \dots, n$ 。

2.2 邻域粒子

定义 3^[9]: 给定实数空间上的非空有限集合 $U = \{x_1, x_2, \dots, x_n\}$, 对于 U 上的任意对象 x_i , 定义其 δ 邻域为 $\delta(x_i) = \{x \mid x \in U, \Delta(x, x_i) \leq \delta\}$ 其中 $\delta \geq 0$ 。 $\delta(x_i)$ 称为由 x_i 生成的 δ 邻域信息粒子, 简称为 x_i 的邻域粒子。

由邻域性质可知,邻域关系满足自反性、对称性和传递性。对于 N 个属性的样本集,距离常用 P 范数表示为 $\Delta_p(x_1, x_2) = [\sum_{i=1}^N |f(x_1, a_i) - f(x_2, a_i)|^p]^{1/p}$, 其中 $f(x, a_i)$ 为样本 x 在属性 a_i 上的取值。 $\Delta_p(x_1, x_2)$ 的定义是对于数值型属性而言的,但邻域模型很容易将距离计算扩展到含有名义型属性和数值型属性的混合数据上来。对于名义型属性 a_i , 可定义

(1) $|f(x_1, a_i) - f(x_2, a_i)| = 0$, 若 x_1, x_2 在 a_i 上的取值相同;

(2) $|f(x_1, a_i) - f(x_2, a_i)| = 1$, 若 x_1, x_2 在 a_i 上的取值不同^[10,11]。

2.3 扩展可辨识矩阵

将之前提到的邻域等价关系和限制容差关系引入到上述定义的可辨识矩阵中,可得到扩展可辨

识矩阵。

给定一个混合决策系统 $S_{\text{IDS}} = \langle U, A, D \rangle$, 定义扩展可辨识矩阵 $M = (m_{ij})_{|U| \times |U|}$, 其元素定义为:

$$m_{ij} = \begin{cases} \{a \mid a \in B \wedge (|f(x_i, a) - f(x_j, a)| > \delta) \\ \quad \wedge f(x_i, a) \neq * \wedge f(x_j, a) \neq * \}, \\ x_j \notin L_B(x_i), f(x_i, d) \neq f(x_j, d) \\ 0, \quad \text{else} \end{cases} \quad (3)$$

式(3)中 $\exists x_i, x_j \in U, i, j = 1, 2, \dots, n; \forall B \subseteq A, L_B(x)$ 表示样本 x 的限制容差类, δ 表示邻域大小。

3 基于扩展可辨识矩阵的属性约简算法

设 $M = (m_{ij})_{|U| \times |U|}$ 为混合决策系统 $S_{\text{IDS}} = \langle U, A, D \rangle$ 的扩展可辨识矩阵, 如果 B 满足

充分条件: $M_B = M_A$;

必要条件: $\forall a \in B, M_{B-\{a\}} \neq M_B$, 称 $B \subseteq A$ 是 A 的一个约简。

在前者约束下, 条件属性 B 能形成与全部条件属性 A 大小一致的可辨识矩阵, 这样充分的分类信息可以被保留下来。而在后者约束下, 多余的属性不会在约简过程中产生, 也就是说条件属性 B 内的每一个属性都是必要的。从条件属性集中滤除冗余属性, 从而获得满足定义中条件的子集就是混合数据的属性约简过程。

设混合决策系统 $S_{\text{IDS}} = \langle U, A, D \rangle, B \subseteq A$, 设 $red_B(D)$ 为决策属性的所有条件属性约简关系簇, $core_B(D)$ 为决策属性的条件属性核, 则 $core_B(D) = \cap red_B(D)$ [3]。

冗余属性不会引发分类边界变化, 因为它与系统的分类问题无关, 但分类器复杂、学习速度慢、过拟合等问题会出现。在海量数据的处理过程中, 运算时间过长会使算法的优势减弱, 从而去除冗余属性被看作是一个重要的问题。

由此, 基于扩展可辨识矩阵的混合决策系统属性约简算法:

输入: 不完备或者完备混合决策系统 $S_{\text{IDS}} = \langle U, A, D \rangle$;

输出: 约简 red 和核属性 $core$ 。

Step1: 计算 $S_{\text{IDS}} = \langle U, A, D \rangle$ 的扩展可辨识矩阵

M , 这里的混合决策系统不但可以是不完备而且可以是完备的;

Step2: 在扩展可辨识矩阵内, 对每一个非空集合元素 $m_{ij} (m_{ij} \neq 0, m_{ij} \neq \emptyset)$, 都与之对应的构建一个析取逻辑表达式 $L_{ij}, L_{ij} = \bigvee_{a_i \in m_{ij}} a_i$;

Step3: 对每一个析取逻辑表达式 L_{ij} 都应用合取运算进行运算, 最后合取为合取范式 L , 即 $L = \bigwedge_{m_{ij} \neq 0, m_{ij} \neq \emptyset} L_{ij}$;

Step4: 对合取范式 L 等价变换, 获得析取范式 $L' = \bigvee_i L_i$;

Step5: 属性约简的结果就与 L' 中的每个合取项相对应, 得到约简后的条件属性集合 red , 且 $core = \cap red$;

Step6: Return red 和 $core$ 。

4 实验分析

为验证所提算法求出的约简的有效性, 本文分别选用来自 UCI 机器学习数据库数据集^[12] 的完备和不完备数据集进行属性约简实验, 其中包括符号型数据集、数值型数据集和混合型数据集, 数据集的描述见表 1。在不完备混合决策系统中, 样本的分辨函数是一个合取范式 L , 当将其等价地转化为它的析取范式 L' 时, 析取范式的子式确定了样本的所有约简, 数目最小的子式便是样本的最小相对约简。实验利用本文的约简算法进行属性约简, 采用 matlab 进行编程, 实验结果包括核属性数、规则数和最小相对约简属性数, 见表 2。

表 1 数据集描述

序号	数据集	类型	样本数	属性数	是否完备	类别
1	Wisconsin prognostic breast cancer	实型	198	34	不完备	2
2	Hepatitis	符号型、 整型、实型	155	19	不完备	2
3	Tic-Tac-Toe Endgame	符号型	958	9	完备	2
4	Wisconsin diagnostic breast cancer	实型	569	31	完备	2
5	Wine recognition	整型、实型	178	13	完备	3
6	Wisconsin original breast cancer	整型	699	10	不完备	2

在表示每个样本的扩展邻域关系过程中,对每一个数值型属性进行标准化,本文把它们标准化到 $[0,1]$ 区间,这样做的好处是减轻量纲不同对最终结果的巨大影响。同时,本次实验中设置 $\delta = 0.2$, $P = 2$,即邻域半径为0.2,度量函数采用2范数即欧氏距离函数。

表2 基于扩展可辨识矩阵的属性约简
($\delta = 0.2, P = 2$)

数据集	缺省属性值	核属性数	规则数	最小相对约简属性数
Wpbc	0.059%	16	65	19
Hepatitis	5.671%	10	15	13
Tic-Tac-Toe	0	0	9	8
Wdbc	0	25	1	25
Wine	0	6	8	9
Wdbc	0.228 9%	7	2	8

通过实验结果验证了本文提出的约简算法的正确性和有效性。数据集 Tic-Tac-Toe Endgame 所得到的结果可知,核属性定义为决策系统所有条件属性约简关系簇的交集,核属性不一定存在;数据集 Wdbc 的约简结果可知,当规则数只有一条时,该规则中的所有约简属性都是核属性,同时也是最小相对约简属性。此外实验结果可能受到邻域大小 δ 的影响,并且结果数据具有一定的随机性。

在扩展可辨识矩阵中采用2范数距离函数,如果 δ 较小,粒度就小,区分决策只用很少的特征,所以得到的属性约简结果较多,但是如果 δ 小至0时,便退化成 Pawlak 粗糙集,这种经典粗糙集不能对混合型数据分类。如果 δ 较大,所需特征则较多,规则就少,如果这时粒度超过一个限度,就会没有特征能够区分样本,可以看出如果规则数减小至0,则得不到属性约简的结果。

图1给出了不完备数据集“Hepatitis”和完备数据集“Wine”的约简规则数随邻域大小 δ 变化的曲线,其中 δ 的取值以0.05为步长从0到1逐步变化。从图中可以直观地发现,邻域大小 δ 的取值和具体分类问题有关,相对而言,参数 δ 在 $[0.1, 0.2]$ 之间取值时较为理想。

5 结束语

本文讨论了混合决策系统下的粗糙集扩展模

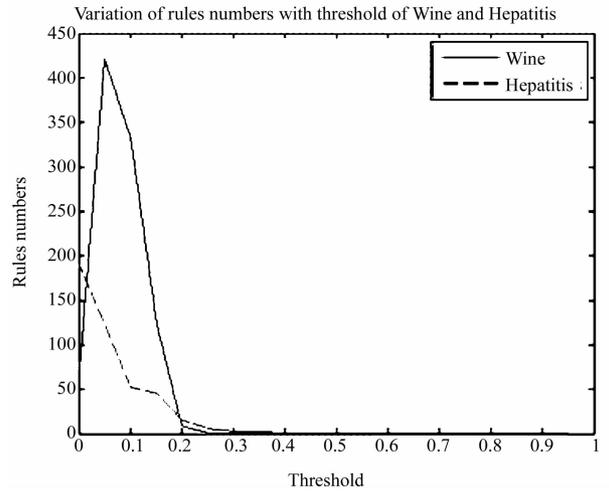


图1 Wine 和 Hepatitis 约简规则数随邻域大小 δ 的变化曲线

型,并通过引入邻域关系和限制容差关系的概念对可辨识矩阵的定义进行扩展,最后给出一种基于扩展可辨识矩阵的混合决策系统的属性约简算法的具体实现。仿真实验采用UCI的六组数据进行了属性约简结果、核属性和约简规则数的分析,并讨论了邻域大小对属性约简结果和规则数的影响。实验结果证明了扩展可辨识矩阵的合理性以及约简算法的有效性,得到的结果具有较好的可理解性和较强的泛化能力,本文算法不但可以应用于传统的完备的决策系统,还可以应用于对不完备的混合值域的决策系统,在不改变初始决策系统结构的基础上,获取到不受缺省值和混合值影响的属性约简结果。

参 考 文 献

- 1 Pawlak Z. Rough sets. International Journal of Parallel Programming, 1982;11(5):341—356
- 2 周献中,黄兵. 基于粗糙集的不完备信息系统属性约简. 南京理工大学学报(自然科学版),2003;27(5):630—635
- 3 Grzymala-Busse J W. Data with missing attribute values: generalization of indiscernibility relation and rule induction. http://sci2s.ugr.es/keel/pdf/specific/cong-reso/grzymala_busse04.pdf. 2012—06—05
- 4 Kryszkiewicz M. Rough set approach to incomplete information systems. Information Sciences,1998;112:39—49
- 5 Lin T Y. Granular computing: practices, theories, and future directions. Encyclopedia of C-complexity and Systems Science,2008;770:

- 1—17
- 6 胡清华,于达仁,谢宗霞. 基于邻域粒化和粗糙逼近的数值属性约简. 软件学报,2008;19(3):640—649
- 7 胡清华,赵 辉,于达仁. 基于邻域粗糙集的符号与数值属性快速约简算法. 模式识别与人工智能,2008;21(6):732—738
- 8 徐久成,张灵均,孙 林,等. 广义邻域关系下不完备混合决策系统的约简. 计算机科学,2013;40(4):244—248
- 9 Skowron A,Rauszer C. The discernibility matrix-based functions in information systems. *Slo-winski: Fundamenta Informaticae*, 1991; 331—362
- 10 王 超,罗 可. 不完备信息系统中基于限制容差关系的属性约简方法. 计算机应用,2011;31(12):3236—3239
- 11 霍忠诚,曾 玲,范 婷,等. 混合值不完备信息系统一种新的数据分析方法. 计算机应用研究,2011;28(9):3321—3323
- 12 Frank A, Asuncion A. UCI machine learning repository. (2010). <http://archive.ics.uc-i.edu/ml>. 2012—06—05

Attribute Reduction of Hybrid Decision System Based on Expanded Discernibility Matrix

ZHAO Yan¹, YANG Jing², LIU Hai-feng¹, SHI Han-yang¹

(School of Information Engineering¹, Center of Information Management and Development², Taiyuan University of Technology, Taiyuan 030024, P. R. China)

[**Abstract**] Classical Rough Set Theory, introduced by Professor Pawlak in 1982, its research object could only be the complete discrete decision system. In order to reduce the attribute of the incomplete hybrid decision system directly, expanded the discernibility matrix through the introduction of the concept of neighbourhood relation and limited tolerance relation, a attribute reduction algorithm based on expanded discernibility matrix is proposed. The attribute reduction can either deal with incomplete or complete hybrid decision system. Demonstrated the effectiveness of the algorithm by simulating experiments of UCI data sets, concluded with a discussion on the influence of selection for neighbouring threshold of expanded discernibility matrix on attribute reduction results.

[**Key words**] complete and incomplete hybrid decision system neighbourhood relation limited tolerance relation expanded discernibility matrix attribute reduction