不协调决策表几种约简标准及其关系分析

杜卫锋[1],秦克云[2]
(1.嘉兴学院数理与信息工程学院,浙江嘉兴 314001;2.西南交通大学数学学院,四川成都 610031)

摘 要: 叶东毅教授在文中指出 Hu Xiaohua 等学者提出的约简方法在某些情况下会给出错误的结果,本文通过分析得出叶的方法实际上就是正域约简,而 Hu 的方法实质上保持边界域划分不变,与叶东毅教授提出的约简方法只是标准不同,而无所谓对错问题.在此基础上,阐明了决策表各种约简标准的关系,并给出了当两个约简标准存在强弱关系时,其约简结果之间的关系.

关键词: 粗糙集;决策表;约简标准;区分函数
中图分类号: TP182 **文献标识码:** A **文章编号:** 0372-2112 (2011) 06-1336-05

Analysis of Several Reduction Standards and Their Relationship for Inconsistent Decision Table

DU Wei-feng[1], QIN Ke-yun[2]
(1. *School of Mathematics, Physics & Information Engineering, Jiaxing University, Jiaxing, Zhejiang* 314001, *China*;
2. *School of Mathematics, Southwest Jiaotong University, Chengdu, Sichuan* 610031, *China*)

Abstract: Prof. Ye Dongyi has pointed out in his paper that the reduction approach proposed by Hu Xiaohua et al. will lead to wrong results under some circumstances. In this paper, we find out through analysis that Ye's approach is actually positive region reduction, whereas Hu's approach is to make sure that the boundary regions are kept unchanged. The main difference between the two approaches is that each use a different standard, thus there is no ground to judge which one is correct or wrong. What is more, we clarify the relationship among various reduction standards for decision table, and we also give the relationship between reduction results when there is a strong-week correlation between two reduction standards.

Key words: rough set; decision table; reduction standard; discernibility function

1 引言

粗糙集理论是一种新的处理不确定和不完备信息的数学工具[1,2].自 1982 年波兰数学家 Pawlak[1] 首次提出以来,经过二十多年的研究与发展,在理论和应用上均取得了长足的发展,特别是由于八十年代末九十年代初在知识发现等领域的成功应用而受到了国际上广泛关注.目前,它已在人工智能、机器学习、模式识别、数据库知识发现、决策分析与故障检测等领域获得了较为成功的应用.

粗糙集理论的创立基于独特的哲学与认知学思考:粗糙集理论认为知识是区分事物的能力[2],知识是有粒度的.粗糙集理论独特的哲学与认知学基础决定了它不同于概率论、模糊集理论及证据理论方法.在粗糙集理论中,借助近似算子,待认识的概念被两个精确概念从外延的角度近似逼近.这两个精确概念可通过数据本身被确定地计算出来,而这一过程无需提供所处理的数据集合之外的任何额外的先验信息.近年来,粗糙集与其它软计算方法如模糊数学、神经网络、证据理论等的结合,使得粗糙集的数据挖掘能力得到进一步提高.

粗糙集理论的基本思想是利用定义在数据集合上的等价关系作为知识,而对知识不确定程度的度量则是对被分析数据整体处理之后自然获得,这样,粗糙集理论无需对数据的局部给予主观评价,从这个角度上讲,粗糙集理论对不确定性的描述相对客观.

知识约简是粗糙集理论的核心内容之一.众所周知,知识库中的知识(属性)并不是同等重要的,甚至其中某些知识是冗余的.所谓知识约简,就是在保持知识库分类能力不变的前提下,删除其中不相关或不重要的知识[3].我们只需保留能形成约简的属性子集,则新信

收稿日期:2009-04-19;修回日期:2011-02-15
基金项目:国家自然科学基金(No.60875034,No.61070213);浙江省自然科学基金(No.Y1080901);嘉兴学院高层次人才科研启动基金

息系统和原信息系统具有同样的分类能力.求解决策表所有的约简是一个 NP 问题,已有相当的文献[4~6]讨论了求解决策表约简的算法或近似算法.

对于协调决策表,各种约简标准给出的结果是一致的,而在不协调决策表中,各种约简标准给出的结果不一定一致,文献[7]讨论了某些约简标准之间的关系.本文在此基础上,论述了不协调决策表几种约简标准的性质,并探讨了目前所见的所有约简标准之间的关系.

2 基本概念

粗糙集理论中,原始的知识表示以决策表的形式给出.

定义 1[3]　决策表是一个五元组 $S=(U,A,d,V,f)$

其中,U 为对象的非空有限集合,称为论域;A 为条件属性的非空有限集合;d 为决策属性;$V=U_{a\in A}V_a$,V_a 是属性 a 的值域;$f: U\times A\to V$ 称为信息函数,它为每个对象的每个属性赋予一个信息值,即 $f(x,a)\in V_a$,$\forall x\in U, a\in A$,信息函数有时也可记为 $a(x)$.

在不引起混淆的情况下,决策表通常可简写为 $S=(U,A,d)$.

容易看出,决策表中每个属性对应一个等价关系.对于 $a\in A$,a 对应的等价关系 R_a 定义为:

$$\forall x,y\in U,(x,y)\in R_a\Leftrightarrow a(x)=a(y)$$

定义 2　$S=(U,A,d)$是原始决策表,如果 $B\subseteq A$ 能保持决策表的某些性质,则称 B 是 S 的某种协调集[8],约简后的决策表由 $S'=(U,B,d)$表出.

定义 3　决策表的区分矩阵是一个 n 阶方阵($|U|=n$),其元素为:

$$\alpha(x,y)$$

决策表的区分函数定义为:

$$\Delta=\bigwedge_{x,y\in U}\bigvee\alpha(x,y)$$

区分函数是一个布尔公式,其中元素看成布尔变量.区分函数的极小析取范式中的所有合取式恰为 S 的所有约简.

不同的约简标准对 $\alpha(x,y)$的规定不同,Hu Xiaohua 等学者[9]提出

$$\alpha(x,y)=\{a\in A\mid a(x)\neq a(y)\wedge d(x)\neq d(y)\}$$

叶东毅[10]在论文中指出,该约简方法在某些情况是错误的,王国胤[11]在论文中也对该方法以及叶东毅的方法作了论述,并探讨了粗糙集理论的代数观和信息观及其关系,指出在协调决策表中,各种约简标准给出的结果是一致的,而在不协调决策表中,各种约简标准给出的结果不一定一致.

3 叶方法实为正域约简

文献[3,12]给出了正域约简关于对象的区分条件,其区分矩阵的元素为:

$$\alpha(x,y)=\{a\in A\mid a(x)\neq a(y)\wedge\omega(x,y)\}$$

$\forall x,y\in U,\omega(x,y)$ 满足:

$$x\in pos_A(d)\wedge y\notin pos_A(d)\vee x\notin pos_A(d)\wedge y\in pos_A(d)\vee x,y\in pos_A(d)\wedge d(x)\neq d(y)$$

在文献[13,14]中,已经证明了条件 $\omega(x,y)$可以转换为 $v(x,y)$而不改变正域约简的实质,其中 $v(x,y)$满足:

$$d(x)\neq d(y)\wedge(x\in pos_A(d)\vee y\in pos_A(d))$$

该结论正好与叶东毅在文献[10]中提出的条件完全等价,并且形式上也十分一致,现将叶在文中给出的条件用本文的符号写为:

$$d(x)\neq d(y)\wedge\min\{|d([x]_A)|,|d([y]_A)|\}=1$$

现在只需证明

引理 1　$x\in pos_A(d)\vee y\in pos_A(d)$ 与 $\min\{|d([x]_A)|,|d([y]_A)|\}=1$ 等价.

证明　$x\in pos_A(d)$,有$[x]_A\subseteq[x]_d$,则有

$$|d([x]_A)|=1,$$

同理,或有

$$|d([y]_A)|=1,$$

因此有

$$\min\{|d([x]_A)|,|d([y]_A)|\}=1.$$

事实上,决策表的约简标准尚有许多,在王国胤的论文中提出的代数观其实是最传统的正域约简标准,这种标准保证约简前后的决策表产生等价的确定性规则,但一般说来,产生的不确定性决策规则未必相同.粗糙集代数观还有其它一些约简标准如分配约简、近似约简、分布约简、最大分布约简.

由此,Hu 的方法也可看作为某种约简标准,而无所谓对错问题,约简结果不一致只是探讨的约简标准不同而已,为了明确 Hu 方法的含义,不妨分析一下 Hu 方法的性质.

4 Hu 方法的性质

引理 2　$[x]_A\neq[x']_A$,如果 $\forall z\in[x]_A, z'\in[x']_A$ 均有 $d(z)=d(z')$,则 $|d([x]_A)|=|d([x']_A)|=1$,$d([x]_A)=d([x']_A)$.

证明　如果 $|d([x]_A)|\neq1$ 或 $|d([x']_A)|\neq1$,则必然 $\exists z\in[x]_A, z'\in[x']_A, d(z)\neq d(z')$,与题设矛盾,因此有 $|d([x]_A)|=|d([x']_A)|=1$,并可得 $d([x]_A)=d(x)$,$d([x']_A)=d(x')$,由 $\forall z\in[x]_A, z'\in[x']_A$ 均有 $d(z)=d(z')$,可得 $d(x)=d(x')$,因此 $d([x]_A)=d$

$([x']_A)$.

定理 1 $B\subseteq A$ 是决策表 $S=(U,A,d,V,f)$的一个 Hu 约简,如果$[x]_A\neq[x']_A,[x]_B\neq[x']_B$ 则 $d([x]_A)=d([x']_A)=d([x]_B)$且$|d([x]_B)|=1$.

证明 $[x]_A\neq[x']_A,[x]_B=[x']_B$,由 Hu 约简的定义,$\forall z=[x]_A,z'=[x']_A$ 均有 $d(z)=d(z')$,由引理 2得 $|d([x]_A)|=|d([x']_A)|=1$, $d([x]_A)=d([x']_A)$.

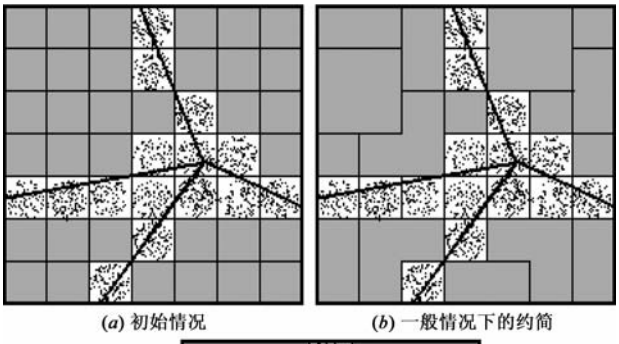
定理 2 $B\subseteq A$ 是决策表 $S=(U,A,d,V,f)$的一个 Hu 约简,若$|d([x]_A)|\neq 1$,则$[x]_B=[x]_A$.

证明 显然有$[x]_A\subseteq[x]_B$.

现需证明$[x]_B\subseteq[x]_A$,即 $\forall y\in[x]_B$ 有 $y\in[x]_A$,下面用反证法加以证明,反设 $\exists y\in[x]_B$ 有 $y\notin[x]_A$ 得到$[y]_A\neq[x]_A$,由$|d([x]_A)|\neq 1$,$\exists z\in[x]_A,d(z)\neq d(y)$,由 Hu 约简的定义,必然 $\exists a\in B,a(z)\neq a(y)$,因此 $y\notin[x]_B$,与反设矛盾! 结论得证,有$[x]_B=[x]_A$.

由定理 1 和定理 2 可知,Hu 约简使正域中的某些等价类得到了合并,但是边界域的划分保持不变,因此 Hu 约简可称为边界域划分约简.

下面就是边界域划分约简的示意图,图 1(a)给出了一个 4 类决策表约简前的情况,从理论上讲,在极端情况下,边界域划分约简后能将决策类的下近似合并为一类,(如图 1(c)所示).对 n 类决策表,在极端情况下,边界域划分约简后则最少可将决策类的下近似约简成 n 类.当然,取决于具体决策表各属性的实际情况,不太可能出现这种极端情况下的约简.一般情况下的约简如图 1(b)所示.



(a) 初始情况 (b) 一般情况下的约简

(c) 可能出现的极端情况下约简

图1 边界域划约简

5 决策表几种约简标准之间的关系

5.1 边界域划分协调集与正域协调集的关系

不同约简标准对 $\alpha(x,y)$的规定不同,Hu Xiaohua 等学者[9]提出的边界域划分约简,其区分矩阵的元素为:

$$\alpha_B(x,y)=\{a\in A\mid a(x)\neq a(y)\wedge d(x)\neq d(y)\}$$

文献[13,14]给出了正域约简关于对象的区分条件,其区分矩阵的元素为:

$$\alpha_P(x,y)=\{a\in A\mid a(x)\neq a(y)\wedge v(x,y)\}$$

其中 $v(x,y)$满足:

$$d(x)\neq d(y)\wedge(x\in pos_A(d)\vee y\in pos_A(d)$$

可见,在正域约简中用于区分元素对的属性需满足的条件要比边界域划分约简中的强.因此,在正域约简中需要区分的元素对在边界域划分约简中必定会加以区分,因此,边界域划分协调集一定是正域协调集,其关系如图 2 所示.

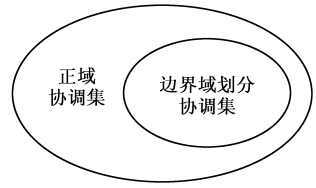


图2 不协调决策表正域协调集与边界域划分协调集的关系

5.2 边界域划分协调集与分布协调集的关系

定义 4[8] 设$(X,\preccurlyeq)$是偏序集,若 $\forall x,y\in X$,都有 $D\left(\frac{y}{x}\right)$与之对应,且满足:

(1)$0\preccurlyeq D\left(\frac{y}{x}\right)\preccurlyeq 1$;

(2)$x\preccurlyeq y\Rightarrow D\left(\frac{y}{x}\right)=1$;

(3)$x\preccurlyeq y\preccurlyeq z\Rightarrow D\left(\frac{x}{z}\right)\preccurlyeq D\left(\frac{x}{y}\right)$.

定义 5 设 $S=(U,A,d)$是决策表,R_A 与 R_d 分别是由条件属性集 A 和决策属性 d 生成的 U 上等价关系,对于 $B\subseteq A$,记

$$\frac{U}{R_B}=\{[x]_B\mid x\in U\}$$

$$\frac{U}{R_d}=\{D_1,D_2,\cdots,D_r\}$$

记

$$D\left(\frac{D_j}{[x]_B}\right)=\frac{|D_j\bigcap[x]_B|}{|[x]_B|},j=1,2,\cdots,r$$

则 D 是 2^U 上的包含度.若记

$$\mu_B(x)=\left(D\left(\frac{D_1}{[x]_B}\right),D\left(\frac{D_2}{[x]_B}\right),\cdots,D\left(\frac{D_r}{[x]_B}\right)\right)$$

则称 $\mu_B(x)$是对象 x 关于属性集 B 在决策表中的广义决策分布函数.

定义 6 设 $S=(U,A,d)$是决策表,$B\subseteq A$,若 $\forall x\in U$ 有 $\mu_B(x)=\mu_A(x)$,则称 B 是分布协调集.若 B 是

分布协调集,且 B 的任何真子集不是分布协调集,则称 B 为分布约简.

定理 3　设 $S=(U,A,d)$是决策表,则边界域划分协调集必为分布协调集.

证明　分如下两种情况分别证明:

(1)如果约简后的等价类在正域中,即 $|d([x]_B)|=1$,则必有 $\mu_B(x)=\mu_A(x)$;

(2)如果约简后的等价类在边界域中,即 $|d([x]_B)|\neq 1$,则由边界域划分协调集的定义知,边界域划分保持不变,即$[x]_B=[x]_A$,故 $\mu_B(x)=\mu_A(x)$.

因此,边界域划分协调集必为分布协调集.

在文献[14,15]中,已经得到了其它各种约简标准之间的关系,再附上本文讨论的边界域划分标准与其它约简标准之间的关系,由此,所有讨论的协调集之间存在图 3 所示的关系.

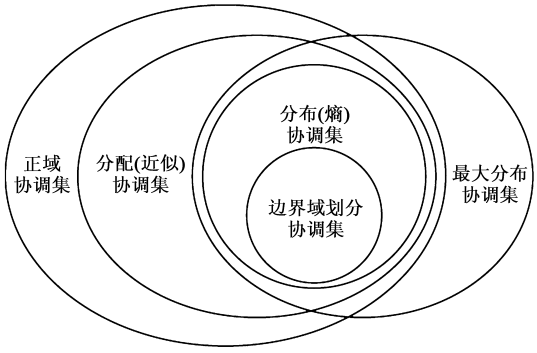


图3　不协调决策表各协调集之间的关系

6　决策表约简结果的关系

假设对某不协调决策表,不同标准产生的协调集有如图 4 所示的关系,则有如下定义.

定义 7　某不协调决策表有两个约简标准 R 和 R',若对应的 R 协调集一定是 R'协调集,则称约简标准 R 为强约简标准,约简标准 R'为弱约简标准.

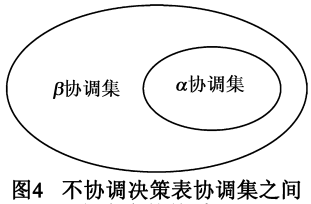


图4　不协调决策表协调集之间可能存在的关系

强约简标准产生的协调集称为强协调集,产生的约简集称为强约简集,强约简集中的约简称为强约简;弱约简标准产生的协调集称为弱协调集,产生的约简集称为弱约简集,弱约简集中的约简称为弱约简.

若有强约简集 A,弱约简集 B,则有如下定理:

定理 4　$\forall \alpha\in A$,必 $\exists \beta\in B$,使得 $\beta\subseteq\alpha$.

证明　由于 α 是强约简,则 α 必是强协调集(约简是最小的协调集),则 α 必定是弱协调集(强协调集必为弱协调集).

(1)若 $\nexists\beta\subset\alpha$,使得 β 是弱约简,则 α 本身就是弱约简,故 $\alpha\in B$;

(2)若 $\exists\beta\subset\alpha$,$\beta$ 是弱约简,而 α 仅是一般弱协调集,$\beta\in B$.

因此总有 $\beta\in B$,使得 $\beta\subseteq\alpha$,定理得证.

推论　若某单元集是强约简,则必是弱约简.

而对于 $\beta\in B$,却未必 $\exists\alpha\in A$,使得 $\beta\subseteq\alpha$.如表 1 所示的不协调决策表,若以正域约简为弱约简标准,分布约简为强约简标准.可得正域约简集 $R_P=\{\{a_1,a_2\},\{a_3\}\}$,分布约简集 $R_D=\{\{a_2,a_3\}\}$.

$\{a_2,a_3\}$是分布约简(强约简),其子集$\{a_3\}$构成正域约简(弱约简);$\{a_1,a_2\}$是正域约简(弱约简),但$\{a_1,a_2\}$的超集不构成分布约简(强约简),原因很简单,$\{a_1,a_2\}$的超集$\{a_1,a_2,a_3\}$仅能构成分布协调集,因为其真子集$\{a_2,a_3\}$是分布约简.

表 1　不协调决策表

U	a_1	a_2	a_3	d
x_1	1	1	2	0
x_2	0	0	1	0
x_3	0	0	1	1
x_4	1	0	1	0
x_5	1	0	1	1
x_6	0	0	0	0
x_7	0	0	0	1
x_8	0	0	0	1
x_9	0	1	1	0
x_{10}	0	1	1	0
x_{11}	0	1	1	1

7　结论

本文通过分析得出叶东毅教授给出的约简方法实质上是正域约简,而 Hu Xiaohua 等学者提出的约简方法保持边界域划分约简,这样这两种约简方法只是标准不同,而无所谓对错问题;在此基础上,讨论了不协调决策表各种约简标准之间的关系,并给出了当某两种约简标准存在强弱关系时,其约简结果之间存在的关系.

参考文献

[1] Z Pawlak. Rough sets[J]. International Journal of Computer and Information Science, 1982, 11: 341 – 356.

[2] Z Pawlak. Rough Sets: Theoretical Aspects of Reasoning about Data[M]. Boston: Kluwer Academic Publishers, 1991

[3] 张文修,吴伟志,梁吉业,李德玉. 粗糙集理论与方法[M]. 北京:科学出版社,2001:158 – 186.
Zhang Wenxiu, Wu Weizhi, Liang Jiye, Li Deyv. Theory and Approach of Rough Set[M]. Beijing: Science Press, 2001. 158 – 186(in Chinese)

[4] 叶东毅. Jelonek 属性约简算法的一个改进[J]. 电子学报,

2000,28(12):81 – 82.
Ye Dongyi. An improvement to Jelonek's attribute reduction algorithm[J]. Acta Electronica Sinica, 2000, 28(12): 81 – 82. (in Chinese)

[5] 杨明.决策表中基于条件信息熵的近似约简[J].电子学报,2007,35(11):2156 – 2160.
Yang Ming. Approximate reduction based on conditional information entropy in decision tables[J]. Acta Electronica Sinica, 2007, 35(11): 2156 – 2160. (in Chinese)

[6] 张腾飞,肖健梅,王锡淮.粗糙集理论中属性相对约简算法[J].电子学报,2005,33(11):2080 – 2083.
Zhang Tengfei, Xiao Jianmei, Wang Xihuai. Algorithms of Attribute Relative Reduction in Rough Set Theory[J]. Acta Electronica Sinica, 2005, 33(11): 2080 – 2083. (in Chinese)

[7] 邓大勇,黄厚宽,李向军.不一致决策系统中约简之间的比较[J].电子学报,2007,35(2):252 – 255.
D Y Deng, H K Huang, X J Li. Comparison of various types of reductions in inconsistent systems[J]. Acta Electronica Sinica, 2007, 35(2): 252 – 255. (in Chinese)

[8] 张文修,梁怡,吴伟志.信息系统与知识发现[M].北京:科学出版社,2003.48.
Zhang Wenxiu, Liang Yi, Wu Weizhi. Information System and Knowledge Discovery[M]. Beijing: Science Press, 2003. 48. (in Chinese)

[9] Hu Xiaohua, Cercone N. Learning in relational databases: A rough set approach[J]. Computational Intelligence, 1995, 11(2): 323 – 337.

[10] 叶东毅,陈昭炯.一个新的差别矩阵及其求核方法[J].电子学报,2002,30(7):1086 – 1088.
Ye Dongyi, Chen Zhaojiong. A new discernibility matrix and the computation of a core[J]. Acta Electronica Sinica, 2002, 30(7): 1086 – 1088. (in Chinese)

[11] 王国胤.决策表核属性的计算方法[J].计算机学报,2003,26(5):611 – 615.
Wang Guoyin. Calculation methods for core attributes of decision table[J]. Chinese Journal of Computers, 2003, 26(5): 611 – 615. (in Chinese)

[12] A. Skowron, C. Rauszer. The discernibility matrices and functions in information systems[A]. R. Slowinski. Intelligent Decision Support-Handbook of Applications and Advances of the Rough Sets Theory[C]. Dordrecht: Kluwer Academic Publishers, 1992. 331 – 362.

[13] 杜卫锋,秦克云.决策表正域约简区分函数条件的改进[J].计算机工程与应用,2006,42(20):16 – 18.
Du Weifeng, Qin Keyun. The improvement to condition in discernibility function of positive reduct of decision table[J]. Computer Engineering and Applications, 2006, 42(20): 16 – 18.

[14] 杜卫锋.粗糙集理论在中文文本分类中的应用[D].四川成都:西南交通大学,2006.
Du Weifeng. Application of Rough Set Theory in Chinese Text Categorization[D]. Chengdu: Southwest Jiaotong University, 2006. (in Chinese)

[15] 杜卫锋,秦克云.不协调决策表正域约简与其它约简的关系[J].海南师范学院学报,2005,18(1):8 – 11.
Du Weifeng, Qin Keyun. The relationship of positive domain reduction to other reductions of inconsistent decision tables[J]. Jouranl of Hainan Normal University, 2005, 18(1): 8 – 11.

[16] Keyun Qin, Zheng Pei, Weifeng Du. The relationship among several knowledge reduction approaches[A]. Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science)[C]. vol. 3613, n PART I. Fuzzy Systems and Knowledge Discovery Berlin Heidelberg: Springer-Verlag, 2005. 1232 – 1241.

作者简介

杜卫锋 男,1977 年生于江苏省常熟市,博士,主要从事粗糙集理论、智能信息处理等方面的研究,已发表论文 10 余篇.
E-mail: woodmud@tom.com

秦克云 男,1962 年生于河南省新乡市,博士生导师,教授,主要研究领域为代数逻辑与不确定性推理,已发表论文 30 余篇.
E-mail: qinkeyun@home.swjtu.edu.cn