粗糙集的近似约简及其算法

申锦标,吕跃进
(广西大学 数学与信息科学学院,广西 南宁 530004)

摘 要:针对经典粗糙集中属性约简的不足,进一步拓展粗糙集属性约简的应用。提出了一种粗糙集属性近似约简的概念和一种新的粗糙集属性重要性的定义并给出和证明了属性近似约简的性质,理论证明了近似属性约简是传统属性约简的一种推广。在保持知识库分类能力基本不变的条件下,利用所给属性重要性作为启发信息给出了粗糙集属性近似约简的算法。通过一个具体的例子,说明了近似属性约简在信息系统中处理模糊和不确定性知识的可行性和有效性。

关键词:粗糙集;近似属性约简;约简算法
中图分类号:TP301 **文献标识码**:A **文章编号**:1673-629X(2009)12-0017-04

A Rough Set of Approximate Attribute Reduction and Its Algorithm

SHEN Jin-biao, LÜ Yue-jin
(School of Mathematics and Information Science, Guangxi University, Nanning 530004, China)

Abstract: In view of the deficiencies of attribute reduction in classic rough set and to further expand the application of rough set attribute reduction, Render an approximate attribute reduction of rough set and a new definition of the importance of attributes. Theory proves that approximate attribute reduction is an extension of the traditional attribute reduction. With the knowledge classification ability remaining basically unchanged, put forward a rough set of approximate attribute reduction and its methods. Finally, a concrete example demonstrates the feasibility and effectiveness of approximate attribute reduction dealing with ambiguity and uncertainty of knowledge in information systems.

Key words: rough set; approximate attribute reduction; reduction algorithm

0 引 言

粗糙集理论是1982年由波兰科学家Pawlak提出的一种新型的处理模糊和不确定知识的数学工具[1]。其主要思想就是在保持分类能力不变的前提下,通过知识约简,导出问题的决策或分类规则。经过多年的发展,该理论已经成为机器学习、知识获取、决策分析、模式识别等领域重要的基本理论[2]。

知识约简是粗糙集理论的核心内容之一。众所周知,知识库中的属性并不是同等重要的,甚至其中某些属性是冗余的。在经典的粗糙集理论中,所谓知识约简,就是在保持知识库分类能力不变的条件下,删除其中不相关或不重要的属性。很多学者投入了这方面的研究,并提出了大量有效的属性约简算法[3~8]。

收稿日期:2009-03-20;**修回日期**:2009-06-13
基金项目:广西自然科学基金(桂科自0991027);广西教育厅面上项目(200707MS061)
作者简介:申锦标(1972-),男,讲师,研究方向为数据挖掘、粗糙集;吕跃进,教授,硕士生导师,研究方向为数据挖掘、粗糙集与概念格。

如果在保持知识库分类能力完全不变的条件下,删除其中不相关或不重要的属性。在实际应用中知识约简受到很大的限制,往往无法对一些不是很重要的属性进行约简。因此,提出了一种新的属性重要性的定义,并在此基础上,在保持知识库分类能力基本不变的条件下,提出了一种近似知识约简的概念及约简算法。理论和实例表明,该方法使得知识约简的应用范围更广了,是原知识约简的一种推广。

1 粗糙集基本概念

1.1 粗糙集的定义

定义1 信息系统:信息系统 S 是一个四元组:$S=(U,A,V,f)$,其中,U 表示一组对象的非空有限集合,称为论域。A 表示属性的非空有限集合。V 是属性的值域集。f 是信息函数,$f:U\times A\to V$。设集合 $X\subseteq U$,R 是一个等价关系,称 $\underline{R}X=\bigcup\{E\in U/R\mid E\subseteq X\}$ 为集合 X 的 R 下近似集;称 $\overline{R}X=\bigcup\{E\in U/R\mid E\cap X\neq\varnothing\}$ 为集合 X 的 R 上近似集;称 $\mathrm{posr}(X)=\underline{R}X$ 为 X 的 R 正域。

定义 2 令 R 为一族等价关系，$P \in R$，如果

$\mathrm{ind}(R) = \mathrm{ind}(R - \{P\})$

则称 P 为 R 中不必要的；否则称 P 为 R 中必要的。

如果每一个 $P \in R$ 都为 R 中必要的，则称 R 为独立的；否则称 R 为依赖的。

定理 1 如果 R 是独立的，$R' \subseteq R$，则 R' 也是独立的。设 $R' \subseteq R$，如果 R' 是独立的，且 $\mathrm{ind}(R') = \mathrm{ind}(R)$，则称 R' 为 R 的一个约简。

1.2 近似约简

定义 3 令 R 为一族等价关系，$P \in R$，对于给定的 α 和 $\beta(0.5 \leqslant \alpha, \beta \leqslant 1)$，如果

$|U/\mathrm{ind}(R) \cap U/\mathrm{ind}(R - \{P\})| / |U/\mathrm{ind}(R)| \geqslant \alpha$ 且 $|\mathrm{pos}_{R-\{P\}}(R)| / |U| \geqslant \beta$

则称 P 为 R 中不必要的；否则称 P 为 R 中必要的。

说明：α 和 β 越大，表示 $\mathrm{ind}(R)$ 和 $\mathrm{ind}(R - \{P\})$ 分类能力越相近，且包含对象越多。

为方便起见，如果 $|U/\mathrm{ind}(R) \cap U/\mathrm{ind}(R - \{P\})| / |U/\mathrm{ind}(R)| \geqslant \alpha$ 且 $|\mathrm{pos}_{R-\{P\}}(R)| / |U| \geqslant \beta$，则记为 $\mathrm{ind}(R) \trianglelefteq \mathrm{ind}(R - \{P\})$，并称 $\mathrm{ind}(R)$ 和 $\mathrm{ind}(R - \{P\})$ 近似相等。

定理 2 设 $|U/\mathrm{ind}(R) \cap U/\mathrm{ind}(R - \{P\})| / |U/\mathrm{ind}(R)| = \alpha$，则 $\alpha = 1$ 充分必要条件为 $\mathrm{ind}(R) = \mathrm{ind}(R - \{P\})$。

证：由于 $\mathrm{ind}(R)$ 比 $\mathrm{ind}(R - \{P\})$ 细，$\mathrm{ind}(R) \cap \mathrm{ind}(R - \{P\}) \subseteq \mathrm{ind}(R)$，如果 $\alpha = 1$，则有 $\mathrm{ind}(R) = \mathrm{ind}(R - \{P\})$。反之，如果 $\mathrm{ind}(R) = \mathrm{ind}(R - \{P\})$，则有 $\mathrm{ind}(R) \cap \mathrm{ind}(R - \{P\}) = \mathrm{ind}(R)$，$|U/\mathrm{ind}(R) \cap U/\mathrm{ind}(R - \{P\})| / |U/\mathrm{ind}(R)| = \alpha = 1$。

定理 3 设 $|U/\mathrm{ind}(R) \cap U/\mathrm{ind}(R - \{P\})| / |U/\mathrm{ind}(R)| = \alpha$，$|\mathrm{pos}_{R-\{P\}}(R)| / |U| = \beta$。则 $\alpha = 1$ 充分必要条件为 $\beta = 1$。

证：由于 $\mathrm{ind}(R)$ 比 $\mathrm{ind}(R - \{P\})$ 细，$\mathrm{ind}(R) \cap \mathrm{ind}(R - \{P\}) \subseteq \mathrm{ind}(R)$，如果 $\alpha = 1$，则有 $\mathrm{ind}(R) = \mathrm{ind}(R - \{P\})$，$\mathrm{pos}_{R-\{P\}}(R) = U$，$|\mathrm{pos}_{R-\{P\}}(R)| / |U| = \beta = 1$。反之，如 $\beta = 1$，则有 $\mathrm{pos}_{R-\{P\}}(R) = U$，由于 $\mathrm{ind}(R) \cap \mathrm{ind}(R - \{P\}) = \mathrm{ind}(R)$，$|U/\mathrm{ind}(R) \cap U/\mathrm{ind}(R - \{P\})| / |U/\mathrm{ind}(R)| = \alpha = 1$。

由上述定理，如果给定 $\alpha = 1$ 或 $\beta = 1$，则定义 3 变为定义 2。

如果每一个 $P \in R$ 都为 R 中必要的，则称 R 为独立的；否则称 R 为依赖的。

定理 4 如果 R 是独立的，$R' \subseteq R$，则 R' 也是独立的。

证明：用反证法。假设 $R' \subseteq R$ 且 R' 是依赖的，则存在 $S \subseteq R'$，使得

$\mathrm{ind}(S) \trianglelefteq \mathrm{ind}(R')$

这意味着

$\mathrm{ind}(S \cup (R - R')) \trianglelefteq \mathrm{ind}(R)$，$S \cup (R - R') \subseteq R$

因此，R 为依赖的，与已知条件矛盾，故定理得证。

设 $R' \subseteq R$，如果 R' 是独立的，且 $|U/\mathrm{ind}(R) \cap U/\mathrm{ind}(R')| / |U/\mathrm{ind}(R)| \geqslant \alpha$，$|\mathrm{pos}_{R'}(R)| / |U| \geqslant \beta$，则称 R' 为 R 的一个约简。

1.3 近似约简例子

设 U 为如表 1 所示的知识系统。

表 1 知识系统

U	a	b	c	U	a	b	c
1	0	0	0	5	0	0	0
2	1	2	2	6	3	1	1
3	2	0	3	7	3	2	2
4	0	2	3	8	1	2	2

则有下列等价类

$U/\{a\} = \{\{x_1, x_4, x_5\}, \{x_2, x_8\}, \{x_3\}, \{x_6, x_7\}\}$

$U/\{b\} = \{\{x_1, x_3, x_5\}, \{x_6\}, \{x_2, x_4, x_7, x_8\}\}$

$U/\{c\} = \{\{x_1, x_5\}, \{x_6\}, \{x_2, x_7, x_8\}, \{x_3, x_4\}\}$

关系 $\mathrm{ind}(R)$ 有下列等价类

$U/\mathrm{ind}(R) = \{\{x_1, x_5\}, \{x_2, x_8\}, \{x_3\}, \{x_4\}, \{x_6\}, \{x_7\}\}$

因为

$U/\mathrm{ind}(R - \{a\}) = \{\{x_1, x_5\}, \{x_2, x_7, x_8\}, \{x_3\}, \{x_4\}, \{x_6\}\} \neq U/\mathrm{ind}(R)$

$U/\mathrm{ind}(R - \{b\}) = \{\{x_1, x_5\}, \{x_2, x_8\}, \{x_3\}, \{x_4\}, \{x_6\}, \{x_7\}\} = U/\mathrm{ind}(R)$

$U/\mathrm{ind}(R - \{c\}) = \{\{x_1, x_5\}, \{x_2, x_8\}, \{x_3\}, \{x_4\}, \{x_6\}, \{x_7\}\} = U/\mathrm{ind}(R)$

易知 $\{a,b\}$ 或 $\{a,c\}$ 是 $R = \{a,b,c\}$ 的约简。

对于上述知识系统，给定 $\alpha = 0.8$ 和 $\beta = 0.8$，由于 $|U/\mathrm{ind}(R) \cap U/\mathrm{ind}(R - \{a\})| / |U/\mathrm{ind}(R)| = 4/6 \leqslant 0.8$ 且 $|\mathrm{pos}_{R-\{a\}}(R)| / |U| = 5/8 \leqslant 0.8$，即 $\mathrm{ind}(R) \trianglelefteq \mathrm{ind}(R - \{a\})$ 不成立。

故 $\{a\}$ 是 R 中必要的。

对于 $\{b\}$，$|U/\mathrm{ind}(R) \cap U/\mathrm{ind}(R - \{b\})| / |U/\mathrm{ind}(R)| = 1 \geqslant 0.8$ 且 $|\mathrm{pos}_{R-\{b\}}(R)| / |U| = 1 \geqslant 0.8$，即 $\mathrm{ind}(R) \trianglelefteq \mathrm{ind}(R - \{b\})$。

故 $\{b\}$ 是 R 中不必要的。

同样地，$|U/\mathrm{ind}(R) \cap U/\mathrm{ind}(R - \{c\})| / |U/\mathrm{ind}(R)| = 1 \geqslant 0.8$ 且 $|\mathrm{pos}_{R-\{c\}}(R)| / |U| = 1$

$\geqslant 0.8$,即 $\mathrm{ind}(R) \subseteq \mathrm{ind}(R-\{c\})$。

故$\{c\}$ 也是 R 中不必要的。

因此,$\{a,b\}$ 或$\{a,c\}$ 是 $R=\{a,b,c\}$ 的近似约简。

说明:α 和β越大,表示 $\mathrm{ind}(R)$ 和 $\mathrm{ind}(R-\{P\})$ 分类能力越相近,且包含对象越多。

2 近似约简算法

2.1 近似约简算法思想

定义 4　区分矩阵:令 $S=(U,C\cup D,V,f)$ 为一个决策系统,$|U|=n$。S 的区分矩阵 $\boldsymbol{M}$ 是一个 $n\times n$ 的对称矩阵。其任一元素为

$$a(x,y)=\{a\in A \mid f(x,a)\neq f(y,a)\}$$

矩阵单元的内容是属性集,它表明了两个对象在该属性集上的值不同。对表 1 的信息表对应的区分矩阵如表 2 所示。

表 2　信息系统

U	a	b	c	d
1	0	1	1	0
2	1	1	0	1
3	1	1	1	0
4	0	1	1	1
5	1	0	0	1

性质 1:一个约简和区分矩阵的每一项的交都不能为空。

性质 2:给定决策系统 $S=(U,C\cup D,V,f)$,$\boldsymbol{M}$ 为 S 的区分矩阵,若有 $\boldsymbol{M}$ 中某项仅含一个属性,则该属性必为核成员。

基于以上的性质,文中建立了以下 3 条启发式规则:

(1) 由性质 1,2 可知,区分矩阵的某项若只有一个属性,该属性必是约简的成员,故必选;

(2) 区分矩阵某项的长度越短,该项中的属性对分类所起的作用就越大;

(3) 区分矩阵中某些属性出现的越频繁,潜在区分能力就越大,该属性就越重要;

因此可以作为一种属性选择的启发式规则,对启发知识(1)、(2) 和(3) 要建立相应的措施。为此,建立启发函数:

$$f(a)=\sum_{a\in m_{ij}}\left(\frac{n^2}{|m_{ij}|n^2-n^2+1}+1\right)$$

其中 $|m_{ij}|$ 表示该项中含属性的个数,$a\in A$,$n=|U|$。

启发函数的前半部分考虑了区分矩阵中每一项的长度,而后半部分则考虑了属性出现的频度。根据以上分析,得出近似约简的方法,首先用启发函数作为启发信息进行约简,减少比较 $\mathrm{ind}(R)\subseteq\mathrm{ind}(C)$ 的计算量,然后以约简为基础进行近似约简。

2.2 近似约简算法步骤

算法 1

输入:区分矩阵 $\boldsymbol{M}$

输出:所有约简 R

步骤 1:设置 $R=\varnothing$,删除区分矩阵中的重复项。

步骤 2:计算属性重要性,并按照属性重要性的大小进行排序。

步骤 3:根据重要性选出值最大的属性 $c_i(c_i\in C)$,$R=R\cup\{c_i\}$,并删除区分矩阵中含 c_i 的所有项。

步骤 4:如果 $\boldsymbol{M}$ 不为空,转步骤 2。

步骤 5:输出所有约简 R。

算法 2

输入:区分矩阵 $\boldsymbol{M}$ 和约简 R

输出:近似约简 R'

步骤 1:删除区分矩阵 $\boldsymbol{M}$ 的每一项中已被约简的属性,并删除区分矩阵中的重复项。

步骤 2:计算属性重要性,并按照属性重要性从小到大进行排序。

步骤 3:根据重要性选出值最小的属性 $c_i(c_i\in C)$,$R=R-\{c_i\}$,并删除区分矩阵中含 c_i 的所有项。

步骤 4:如果 $\mathrm{ind}(R)\subseteq\mathrm{ind}(C)$,转步骤 2。

步骤 5:输出近似约简 R'。

2.3 约简算法例子

设有如表 2 所示的信息系统,其区分区阵如表 3 所示。下面以此信息系统为例,说明算法的具体操作。

表 3　区分矩阵

	1	2	3	4	5
1					
2	a,c,d				
3	a	c,d			
4	d	a,c	a,d		
5	a,b,c,d	b	b,c,d	a,b,c	

以表 3 区分矩阵为例,说明约简算法的可行性和有效性。

步骤 1:设置 $R=\varnothing$,删除区分矩阵中的重复项,得

$$\boldsymbol{M}=\{(a,b,c,d),(a,b,c),(a,c,d),(a,d),(b,c,d),(a,c),(c,d)\ (a),\ (b),(d)\}$$

步骤 2:计算属性重要性,并按照属性重要性的大小进行排序:a,d,b,c。

步骤 3:根据重要性选出值最大的属性 a,$R=R\cup\{a\}$,并删除区分矩阵中含 a 的所有项,得

$R = \{a\}$ $M = \{(b,c,d),(c,d),(b),(d)\}$

步骤 4:因为 $M = \{(b,c,d),(c,d),(b),(d)\}$,重新计算属性重要性,并按照属性重要性的大小进行排序:d,b,c。

步骤 5:根据重要性选出值最大的属性 d,$R = R \cup \{d\}$,并删除区分矩阵中含 d 的所有项,得

$R = \{a,d\}$ $M = \{(b)\}$

最后选择 b,则有

$R = \{a,b,d\}$ $M = \{\}$,转步骤 6

步骤 6:输出所有约简 R。

$\{a,b,d\}$

对于约简后信息系统,给定 $\alpha = 0.8$ 和 $\beta = 0.8$,可用算法 2 进行近似约简。

步骤 1:设置 $R = \{a,b,d\}$,删除区分矩阵 M 的每一项中 c 属性,并删除区分矩阵中的重复项。得

$M = \{(a,b,d),(a,b),(a,d),(b,d),(a),(b),(d)\}$

步骤 2:计算属性重要性,并按照属性重要性从小到大进行排序:b,d,a。

步骤 3:根据重要性选出值最小的属性 b,$R = R - \{b\}$,并删除区分矩阵中含 b 的所有项,得

$R = \{a,d\}$ $M = \{(a,d),(a),(d)\}$

步骤 4:因为 $M = \{(a,d),(a),(d)\}$,重新计算属性重要性,得 a 和 d 的重要性相同。

步骤 5:由于 a 和 d 的重要性相同。

(1) 若选择 a,则有

$R = \{d\}$ $\text{ind}(R) \subseteq \text{ind}(C)$ 不成立,转步骤 6

(2) 若选择 d,则有

$R = \{a\}$ $\text{ind}(R) \subseteq \text{ind}(C)$ 不成立,转步骤 6

步骤 6:输出近似约简 R。

$\{a,d\}$

说明:不难发现,在求近似约简的步骤 3 中若选择属性 d,则可求出另一近似约简 $\{a,b\}$。

3 结束语

知识约简是粗糙集理论的核心内容之一。在经典的粗糙集理论中,知识约简是在保持知识库分类能力完全不变的条件下进行的。虽然很多学者投入了这方面的研究,并提出了大量有效的属性约简算法。但在保持知识库分类能力完全不变的条件下,删除其中不相关或不重要的属性。在实际应用中知识约简受到很大的限制,往往无法对一些不是很重要的属性进行约简。为了使粗糙集的属性约简更好的应用于实际情况,提出了一种新的属性重要性的定义,并证明了新的属性重要性是原定义的一种推广。在此基础上,在保持知识库分类能力基本不变的条件下,提出了一种近似知识约简的概念及约简算法。理论和实例表明,该方法使得知识约简的应用范围更广了,是原知识约简的一种推广。

参考文献:

[1] Pawlak Z. Rough sets[J]. International Journal of Information and Computer Science, 1982, 11(5): 341 - 356.

[2] 张文修,吴伟志. 粗糙集理论与方法[M]. 北京:科学出版社,2001.

[3] 赵士欣,彭 涛,张素娟. 一种新的求属性约简的算法[J]. 河北师范大学学报:自然科学版,2008,32(3):313 - 316.

[4] 田卫东,周创德,胡学钢,等. 基于简化分辨矩阵的粗糙集属性约简算法[J]. 计算机科学,2008,35(3):209 - 212.

[5] 陶 志,刘庆拯,李卫民. 一种基于改进区分矩阵的属性约简算法[J]. 计算机工程与应用,2007,43(32):83 - 85.

[6] 刘 飞,孔媛媛,杨习贝. 一种基于粗糙集属性频度约简算法的改进[J]. 计算机技术与发展,2008,18(12):95 - 97.

[7] 覃伟荣,秦亮曦. 基于粗糙集理论的条件属性动态约简算法[J]. 计算机技术与发展,2008,18(8):23 - 25.

[8] 徐章艳,杨炳儒,宋 威,等. 几种不同属性约简的比较研究[J]. 小型微型计算机系统,2008,29(5):848 - 853.

(上接第 16 页)

挖掘研究[J]. 计算机技术与发展,2006,16(7):102 - 104.

[6] 张素文,孟建良,庞春江. 模糊关联规则的加权挖掘算法[J]. 微机发展(现更名:计算机技术与发展),2003,13(4):64 - 70.

[7] 谢邦昌,华通人 Data Mining 团队. 商务智能与数据挖掘 Microsoft SQL Server 应用[M]. 北京:机械工业出版社,2008:97 - 98.

[8] 欧阳为民,郑 诚,蔡庆生. 数据库中加权关联规则的发现[J]. 软件学报,2001,12(4):612 - 619.

[9] Tang Zhaohui. 数据挖掘原理与应用 - SQL Server 2005 数据库[M]. 北京:清华大学出版社,2008:192 - 200.

[10] 王德兴,胡学钢,刘晓平,等. 改进购物篮分析的关联规则挖掘算法[J]. 重庆大学学报:自然科学版,2006,29(4):105 - 107.

[11] 张文献,陆建江. 加权布尔关联规则的研究[J]. 计算机工程,2003,29(9):55 - 57.

[12] Han Jiawei, Kamber M. Data Mining Concepts and Techniques[M]. Sam Francisco: Morgan Kaufmann Publishers, 2001:232 - 236.