文章编号:1001 - 9081(2008)04 - 0872 - 02

一种基于零值原则的属性约简方法

罗来鹏,刘二根
(华东交通大学 基础科学学院,南昌 330013)
(loulp@ tom. com)

摘　要:根据 Guan 等提出的完备信息系统下矩阵约简算法,提出一种改进的属性约简计算方法。该方法根据矩阵的运算特点,通过引入唯一零值概念,使得计算过程更为简易。证明了它与区分矩阵下属性约简的等价性,最后将该方法运用到协调决策表中,并用实例对此进行了说明。

关键词:粗糙集;属性矩阵;属性约简
中图分类号: TP311.32　　**文献标志码**:A

Attribute reduction method based on zero value principle

LUO Lai-peng, LIU Er-gen
(*School of Basic Sciences, East China Jiaotong University, Nanchang Jiangxi* 330013, *China*)

Abstract: Based on matrix reduction algorithm for information system introduced by Guan, a new improved computational method of attribute reduction was presented. According to the characteristics of operation between matrixes, the computation was simplified by introducing only the concept of zero value. This method was proved to be equal to attribute reduction by discernable matrix. At last, the method was applied to complete decision table. Its correctness and effectiveness are shown in an example.

Key words: rough set; attribute matrix; attribute reduction

粗集理论自 1982 年由波兰 Z. Pawlak 教授[1,2]等提出以来,在人工智能、机器学习与知识发现、模型识别、分类、故障诊断等方面得到了较成功的应用。随着应用发展需要,Pawlak 模型得到拓广,近些年提出了许多新的粗集模型与属性约简方法[3-6],它们分别从不同角度对知识表示和获取进行了描述。J. W. Guan 等在文献[7]提出了完备信息系统下矩阵约简算法,该方法通过矩阵方式分别就基本知识和属性重要性等进行了描述并给出了相应的属性约简的定义。该方法给出了一种新的认识粗糙集的工具,它的主要特点:一是每次属性约简都要重新构造关系矩阵;二是构造的新关系矩阵与最初的关系矩阵要进行所有元素之间的比较,可以很容易证明它与 Pawlak 方法是等价的,因此对于一个信息系统求其所有属性约简将是很困难的。本文基于对矩阵之间所定义的运算分析,通过引入唯一零值属性概念和关系矩阵的一种表示方法对该方法的计算进行改进。它不仅能减少原来约简过程当中一些不必要的计算,而且能很方便求出所有约简,它是一个先求核属性再求属性约简的过程。

1　信息表的矩阵约简算法[7]

定义 1　一个知识表示系统 S 是一个四元组 $S = (U, R, V, f)$,其中 U 是对象的集合,$R = C \cup D$ 是有限个属性的非空集合,V 是属性值的集合,f 为信息函数,$f: U \times R \rightarrow V$。当子集 C 和 D 分别为条件属性和决策属性,知识表达系统又称为决策表。

定义 2　信息系统 $S = (U, R, V, f)$,对于任何一个 $A \subseteq R$,都可以决定一个等价关系,所有等价类集合表示为 U/A 或者 $U/IND(A)$,$U/IND(R) = U/R = \{X_1, X_2, \cdots, X_m\}$,其中 $X_i \cap X_j = \varnothing$, $i \neq j$, $i, j = 1, \cdots, m$ 并且 $\bigcup_{i=1}^{m} X_i = U$。

定义 3　设 A 是论域 U 上的等价关系,则 A 对应一个方阵为:$\boldsymbol{M}_A = [m_{ij}]_{n \times n}$。其中 $n = |U|$, $m_{ij} = \begin{cases} 1, & x_i A x_j \\ 0, & \text{否则} \end{cases}$, 矩阵 $\boldsymbol{M}_A$ 称为 A 的关系矩阵或者属性 A 矩阵。

定义 4　设两个关系矩阵 $\boldsymbol{M} = (m_{ij})_{n \times n}$, $\boldsymbol{N} = (n_{ij})_{n \times n}$,则 $\boldsymbol{M} \cap \boldsymbol{N} = [r_{ij}]_{n \times n}$, 其中 $r_{ij} = \min\{m_{ij}, n_{ij}\}$。

定理 1　信息系统 $S = (U, A, V, f)$,其中 $A = \{a_1, a_2, \cdots, a_m\}$,则 S 的属性矩阵 $\boldsymbol{M}_A = \bigcap_{i=1}^{m} M_{a_i}$, 矩阵元素 $(A)_{ij} = (a_1)_{ij} \wedge \cdots \wedge (a_m)_{ij}$, 其中 $(a)_{ij}$ 表示为属性 a 的矩阵 $[(a)_{ij}]$ 第 i 行和第 j 列元素, $\wedge$ 为取小运算。因此 $\boldsymbol{M}_A = (A)_{ij} = [(a_1)_{ij} \wedge \cdots \wedge (a_m)_{ij}]$,这种表示更能突出它们之间关系,下文都采用这种表示。

定义 5　设 $S = (U, A, V, f)$ 为一个信息系统, $a \in A$, 且 $\boldsymbol{M}_A = \boldsymbol{M}_{A-\{a\}}$, 则称 a 为不重要的。所有重要属性集称为独立集。

定义 6　信息系统 $S = (U, A, V, f)$,称 $C \subseteq A$ 为 A 的一个约简, 如果:

1) $\boldsymbol{M}_C = \boldsymbol{M}_A$;

2) 对任意 $C' \subset C$, $\boldsymbol{M}_{C'} \neq \boldsymbol{M}_C$, 或者说 C 是独立的。

2　属性矩阵性质与进一步结论

根据上述定义可以得到属性矩阵具有如下一些结论:

性质 1　属性矩阵是对称的,即 $\boldsymbol{M}_A = \boldsymbol{M}_A^{\mathrm{T}}$(转置)。

性质 2　若属性集 $A \subset B$,则 $\boldsymbol{M}_B \leqslant \boldsymbol{M}_A$。

性质 3　设属性集合 $A = \{a_1, a_2, \cdots, a_m\}$, 若存在 $a \in A$ 且 $(a)_{ij} = 0$, 则 $(A)_{ij} = 0$。

定义 7　设属性集合 $A = \{a_1, a_2, \cdots, a_m\}$, 若 $a_k \in A$ 且 $(a_k)_{ij} = 0$, 对于任何 $p \neq k$ 且 $a_p \in A$,都有 $(a_p)_{ij} \neq 0$, 称

收稿日期:2007 - 10 - 12;**修回日期**:2007 - 12 - 20。　**基金项目**:国家自然科学基金资助项目(10661007);江西省自然科学基金资助项目(0611009);华东交通大学校立科研基金资助项目(07JC05)。
作者简介:罗来鹏(1973 -),男,江西吉水人,讲师,硕士,主要研究方向:数据挖掘、智能信息处理;　刘二根(1965 -),男,江西吉水人,教授,主要研究方向:图论与优化。

$(a_k)_{ij}$ 为 $(A)_{ij}$ 唯一零值，a_k 为含有唯一零值属性，否则为零值属性。

定理 2　设信息系统 $S = (U, A, V, f)$，$A = \{a_1, a_2, \cdots, a_m\}$，$a \in A$ 是核属性的充分必要条件是 a 为含有唯一零值属性。

证明　设 $a \in A$ 为核属性，若在 $\boldsymbol{M}_A$ 中任何 $(A)_{ij} = \min\{(a_1)_{ij}, (a_2)_{ij}, \cdots, (a_m)_{ij}\} = 0$ 中至少还有其他某个属性所对应属性矩阵值为 0，这样在 $\boldsymbol{M}_A$ 中删除属性 a 后 $\boldsymbol{M}_A$ 值不变，即 $\boldsymbol{M}_a = \boldsymbol{M}_{\{A-a\}}$，因而存在 $C \subseteq \{A-a\}$ 是信息系统的一个约简，而 $a \notin C$，这与 a 是属性核矛盾。

反之，若 a 为含有唯一零值属性，则 a 必为核属性。为此只要证明 $\boldsymbol{M}_A \neq \boldsymbol{M}_{\{A-a\}}$。事实上，由 a 是含有唯一零值属性，那么 a 在 $\boldsymbol{M}_A$ 的元素 $(A)_{ij}$ 中至少有一个是唯一零值，不妨设这个元素就是 $(A)_{ij}$，这样，若将 a 删除，值 $(A)_{ij}$ 将发生改变，从而整体矩阵值也发生改变，即有 $\boldsymbol{M}_A \neq \boldsymbol{M}_{\{A-a\}}$。

定理 3　设信息系统 $S = (U, A, V, f)$ 的辨识矩阵 $\boldsymbol{D} = [(C_i, C_j)]$，其中 C_i，C_j 为等价类的描述，则具有如下性质：$a \in A$ 是核属性当且仅当存在 C_i，$C_j (i \neq j)$，使 $D(C_i, C_j) = \{a\}$。

事实上，若存在 C_i，$C_j (i \neq j)$，使 $D(C_i, C_j) = \{a\}$，也就是说在论域 U 中至少有两个对象可以用唯一属性来区分，那么从属性矩阵 $\boldsymbol{M}_A$ 来考虑，就是在所有 $(A)_{ij} = 0$ 中至少有一个是由属性 a 唯一零值所决定，因此上述两个定理实际上是等价的。

定理 4　设信息系统 $S = (U, A, V, f)$，$C \subseteq A$ 为 A 的一个属性约简当且仅当满足：对于 $\boldsymbol{M}_A$ 中任何 $(A)_{ij} = 0$，

1) $(C)_{ij} = 0$；

2) 不存在 $C' \subset C$，使 $(C')_{ij} = 0$。

证明　根据定义 6，只需要证明根据属性矩阵进行属性约简判定时不需要考虑 $(A)_{ij} = 1$ 的元素对应关系。事实上，根据定理 1，若 $(A)_{ij} = 1$，则对于任何 $a \in A$，$(a)_{ij} = 1$，所以无论怎样约简 $(A)_{ij} = 1$ 的值始终不变，因此上述结论成立。这个定理与定义 6 的最大区别是在进行等式判定时减少了比较元素个数。

同时定理 4 也说明在用矩阵方法判定属性是否可约简时，只需要考虑约简后 $\boldsymbol{M}_A$ 中零值元素是否发生了改变。此外由于对于任意 $a \in A$，$(a)_{ij} = 0$，无论怎样约简 $(A)_{ij} = 0$ 值不变，因此在进行值判定时也可以不用考虑。

性质 4　若 $a \in A$ 为核属性且 $(a)_{ij} = 0$，则无论怎样对其他属性约简，值 $(A)_{ij} = 0$ 不变。

因此，一旦确定了核属性，在式 $(A)_{ij} = (a_1)_{ij} \wedge \cdots \wedge (a_m)_{ij} = 0$ 中有核属性值为 0，则在求属性约简时属性矩阵 $\boldsymbol{M}_A$ 中 $(A)_{ij}$ 也可以不考虑。设信息系统 $S = (U, A, V, f)$，$C \subset A$ 为核属性，将属性矩阵 $\boldsymbol{M}_A$ 进行判定属性约简前可以将由 $(a_1)_{ij} = \cdots = (a_m)_{ij} = 1$(或0) 和对于任何 $c \in C$，$(c)_{ij} = 0$ 所对应的元素在矩阵 $\boldsymbol{M}_A$ 中删除，并记为：$\boldsymbol{M}_A''$。

这样定理 4 又可以改为：

推论 1　设信息系统 $S = (U, A, V, f)$，$C \subseteq A$ 为 A 的一个属性约简当且仅当满足：对于 M_A'' 中任何 $(A)_{ij} = 0$，

1) $(C)_{ij} = 0$；

2) 不存在 $C' \subset C$，使 $(C')_{ij} = 0$。

定理 5　设信息系统 $S = (U, A, V, f)$，$C \subset A$ 为核属性，对于任何 $(A)_{ij} = (a_1)_{ij} \wedge \cdots \wedge (a_m)_{ij} = 0$，若 $a_k \in A - C$ 在 $\boldsymbol{M}_A''$ 任何元素中都有 $(a_k)_{ij} = 0$，那么属性集 $C \cup \{a_k\}$ 为信息系统的一个约简。

由推论 1，这个结论显然成立。

综合起来就是，可以根据零值原则对矩阵先约简，然后求核属性和属性约简。

3　属性约简算法描述

3.1　信息系统属性约简算法描述

设信息系统 $S = (U, A, V, f)$，其中 $A = \{a_1, a_2, \cdots, a_m\}$。

1) 求属性 A 中各属性的关系矩阵，并按照定理 1 形式书写 $\boldsymbol{M}_A$。

2) 将 $\boldsymbol{M}_A$ 中所有为 1 或为 0 的表达式删除(或者记为 $\varnothing$)得矩阵 $\boldsymbol{M}_A'$。

3) 在 $\boldsymbol{M}_A'$ 中找出所有唯一零值及其所对应的属性，这些属性集即为核属性集。

4) 在 $\boldsymbol{M}_A'$ 中删除核属性集所对应值为 0 的元素，得到 $\boldsymbol{M}_A''$。

5) 根据定理 5，在 $\boldsymbol{M}_A''$ 中求出所有的属性约简。

6) 将所有属性约简输出。

3.2　决策表属性约简算法

3.2.1　几个概念

设决策系统 $S = (U, C \cup D, V, f)$，C 为条件属性集，D 为决策属性集。

定义 8　决策系统 S 为协调的充要条件为 $\boldsymbol{M}_C \leqslant \boldsymbol{M}_D$。

定义 9　协调决策系统 $S = (U, C \cup D, V, f)$，属性集 $C' \subset C$ 为一个约简，当且仅当满足：

1) $\boldsymbol{M}_{C'} \leqslant \boldsymbol{M}_D$；

2) 不存在 $C'' \subset C'$，且 $\boldsymbol{M}_{C''} \leqslant \boldsymbol{M}_D$。

根据定义 2，在 $\boldsymbol{M}_D$ 中，对于 $(D)_{ij} = 1$，无论怎样约简始终成立，因此使用上述方法判定就只归结为 $(D)_{ij} = 0$ 的对应元素之间关系，这个过程实际上跟信息系统的属性约简相似并且比它还要简化。

3.2.2　算法描述

1) 分别求出各条件属性与决策属性 $\boldsymbol{M}_D$ 的属性矩阵，按定理 1 写 $\boldsymbol{M}_C$。

2) 删除 $\boldsymbol{M}_C$ 中有关项，包括全部为 1 和 0 的元素，决策属性矩阵中值为 1 在条件矩阵中所对应的元素。

3) 根据唯一零值确定核属性。

4) 根据定理 5 求所有属性约简。

4　实例说明

表 1 为一决策系统，条件属性集为 $C = \{a, b, c\}$，决策属性集为 $D = \{d\}$。

1) 条件属性矩阵和决策属性矩阵为：

$$
\boldsymbol{M}_C = \begin{bmatrix}
1 & & & & \\
0 \wedge 0 \wedge 0 & 1 & & & \\
0 \wedge 0 \wedge 1 & 1 \wedge 1 \wedge 0 & 1 & & \\
1 \wedge 0 \wedge 1 & 0 \wedge 0 \wedge 0 & 0 \wedge 0 \wedge 1 & 1 & \\
1 \wedge 0 \wedge 0 & 0 \wedge 0 \wedge 1 & 0 \wedge 0 \wedge 0 & 1 \wedge 1 \wedge 0 & 1
\end{bmatrix}
$$

表 1　决策系统

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

$$
\boldsymbol{M}_D = \begin{bmatrix}
1 & & & & \\
0 & 1 & & & \\
0 & 0 & 1 & & \\
1 & 0 & 0 & 1 & \\
0 & 0 & 1 & 0 & 1
\end{bmatrix}
$$

2) 删除条件关系矩阵相关元素：

$$
\boldsymbol{M}_{C'} = \begin{bmatrix}
 & & & & \\
 & & & & \\
0 \wedge 0 \wedge 1 & 1 \wedge 1 \wedge 0 & & & \\
 & & 0 \wedge 0 \wedge 1 & & \\
1 \wedge 0 \wedge 0 & 0 \wedge 0 \wedge 1 & & 1 \wedge 1 \wedge 0 &
\end{bmatrix}
$$

(下转第 876 页)

对多盘算法改进前根据式(1)得到：$P_1 = 0.125$，$P_2 = 0.044$，$P_3 = 0.047$，$P_{总} = P_1 + P_2 + P_3 = 0.216$；改进后根据式(4)得：$P'_1 = 0.1425$，$P'_2 = 0.052$，$P'_3 = 0.028$，$P'_{总} = P'_1 + P'_2 + P'_3 = 0.2225$。

从所得到的数据分析，当数据项的长度固定为 1 个时间单位时，改进前后盘 B_i 内所有数据项访问概率的平均值 P_i 没有变化；当数据项不固定长度（单位时间的整数倍）时，改进后数据项访问概率的平均值 P'_i 要优于改进前的 P_i，因此 $P'_{总}$ 比 $P_{总}$ 的值更大。

在经典多盘调度算法中，当给定盘数 K 以及按照访问概率把数据项分配到盘中，各盘的相对广播频率 f_i 正比于该盘平均访问概率的平方根 $\mathrm{sprt}(p_i)$ 即 $f_i \propto \mathrm{sqrt}(p_i)$ 时，该广播调度的平均访问时间取得最优值[4]。通过线性规划对可变数据项分配改进后，那么广播调度的平均访问时间取得最优值的必要条件是 $f_i \propto \mathrm{sqrt}\left(\frac{1}{c_i}\sum_{j \in B_i} q_j\right)$。

改进前的调度队列为：d_1，d_2，d_3，d_5，d_1，d_2，d_4，d_6，d_1，d_2，d_3，d_7，其中广播周期 L 为 25 个时间单位。改进后的调度队列为：d_1，d_3，d_6，d_2，d_4，d_1，d_3，d_6，d_5，d_7，其中广播周期 L 为 19 个时间单位。

文献[7]证明了基于 Zipf 多盘调度的平均访问时间为：

$$S = \sum_{i=1}^{k} \frac{c_i p_i L}{2 f_i} = \frac{L}{2} \sum_{i=1}^{k} \frac{c_i p_i}{f_i} \tag{5}$$

平坦广播调度和多盘广播调度下访问时间的比较如表 3。

表 3 多种调度方式下访问时间的比较

调度方式	周期	平均访问时间/s
平坦广播调度	15	15/2 =7.5000
多盘广播调度	25	6.9875
优化后的多盘广播调度	19	6.7735

从表 3 的数据中可以得出当广播数据项为变长时，多盘广播调度的平均访问时间比平坦广播调度小，但广播周期有所增长；优化后的多盘广播调度在平均访问时间上比改进前的多盘广播调度小，而且广播周期基本接近平坦广播调度。

5 结语

本文介绍了移动计算的系统模型，描述了移动数据库广播关键技术以及数据调度策略。在总结非平坦调度策略中经典多盘调度算法的基础上，运用统筹学的线性规划对所要广播的数据进行最优规划并通过该调度算法优化调度。通过实验，减少了广播数据的平均访问时间，改进后的多盘调度算法对数据广播调度的访问时间有较好的优化效果，能更好地进行数据广播。

参考文献：

[1] IMIELINSKI T, BADRINATH B R. Mobile wireless computing: challenges in data management[J]. Communications of the ACM, 1994, 37(10): 18 – 28.

[2] LIU CHUAN - MING, LIN KUN - FENG. Disseminating dependent data in wireless broadcast environments[J]. Distributed and Parallel Databases, 2006, 22(1): 1 – 25.

[3] 刘天禄. 统筹学概论[M]. 北京: 中国商业出版社, 2004.

[4] HUNG H - P, HUANG J - W, HUANG J - L, *et al*. Scheduling dependent items in data broadcasting environment[C]// Proceedings of the 2006 ACM Symposium on Applied Computing. NewYork: ACM, 2006: 1177 – 1181.

[5] 潘海琴. 移动环境中数据广播相关技术的研究[D]. 杭州: 浙江大学, 2004.

[6] 何新贵. 特种数据库技术[M]. 北京: 科学出版社, 2000: 50 – 58.

[7] 胡虚怀. 移动计算环境中数据广播调度算法的研究[J]. 湖南理工学院学报, 2005, 18(2): 79 – 82.

（上接第 873 页）

3）含唯一零值属性为核属性。只有 $\{c\}$。

4）属性约简。删除含 $c = 0$ 的元素。

$$M_{C''} = \begin{bmatrix} & & \\ & & \\ & & \\ 0 \wedge 0 \wedge 1 & & \\ & & 0 \wedge 0 \wedge 1 \\ & 0 \wedge 0 \wedge 1 & \end{bmatrix}$$

只剩下 $0 \wedge 0 \wedge 1$，整体值为 0，而 c 逻辑值为 1，为了保证整体值不变，必要有一个 0 和 c 的组合，即得到属性约简为 $\{a, c\}$，$\{b, c\}$。

这个结果与用其他方法的结果完全相同。

5 算法特点

相对根据定义 6、定义 9 得到文献[3]中信息系统的属性约简和完备决策表的属性相对约简而言，本文根据推论 1 与定理 5 得到的属性约简与属性相对约简具有以下特点：一是属性约简后的关系矩阵与最初关系矩阵之间元素比较由原来每次属性约简都要进行一次比较变成只需要比较一次；其次所关心的矩阵元素由原来的全部变成只关心整体为 0 的元素；此外约简过程不需要每次重新构造新的关系矩阵，而是通过一种新的关系矩阵表示方法，根据零值不可变的条件直接求出所有的属性约简和属性相对约简，过程简单，在这个步骤中又充分利用了运算特点。

6 结语

本文主要就文献[3]所提的属性约简的矩阵方法的两个主要问题：一是每次属性约简都要构造新的关系矩阵，二是判定一个属性是否可约简需要进行关系矩阵所有元素之间进行比较，在理论上证明了关系矩阵元素随属性变化而变化之间的关系，由此通过给出一种关系矩阵的新表示和引入唯一零值属性概念使得信息系统的属性约简和属性相对约简得到简化。同时还就该方法与区分矩阵方法进行了对比分析。

参考文献：

[1] PAWLAK Z. Rough Sets[J]. International Journal of Information and Computer Science, 1982, 11(5): 341 – 356.

[2] PAWLAK Z. Rough sets and intelligent data analysis[J]. Information Sciences, 2002, 147(1/4): 1 – 12.

[3] GUAN J W, BELL D Z, GUAN Z. Matrix computation for information systems[J]. Information Sciences, 2001, 131: 129 – 156.

[4] 张文修, 吴伟志, 梁吉业, 等. 粗糙集理论和方法[M]. 北京: 科学出版社, 2001.

[5] 刘清. Rough 集及 Rough 推理[M]. 北京: 科学出版社, 2001.

[6] 王国胤. Rough 集理论与知识获取[M]. 西安: 西安交通大学出版社, 2001.

[7] 苗夺谦, 胡桂荣. 知识约简的一种启发式算法[J]. 计算机研究与发展, 1999, 36(6): 681 – 684.