

一种基于关系矩阵维度增量式约简算法

景运革^{1,2}, 黄峻福²

(1 山西运城学院 公共计算机教学部, 山西 运城 044000;

2 西南交通大学 信息科学与技术学院, 四川 成都 610031)

摘 要: 目前决策表的数据信息都是动态变化的, 为了有效地对这些数据集进行属性约简处理, 在改进的等价关系矩阵的基础上, 提出一种增量式属性约简算法, 决策表增加了一些属性, 利用更新前决策表的等价关系矩阵和属性约简等信息, 便能快速求解出更新后的决策表属性约简. 实验结果进一步验证了该方法的有效性和正确性.

关键词: 粗糙集; 增量式更新; 正域约简; 关系矩阵

中图分类号: TP 311

文献标识码: A

文章编号: 1000-7180(2015)04-0155-04

An Incremental Reduction Algorithm for Decision Table
Based on Relationship MatrixJING Yun-ge^{1,2}, HUANG Jun-fu²

(1 Department of Public Computer Teaching, Yuncheng College, Yuncheng 044000, China;

2 School of Information Science and Technology, Southwest Jiaotong University, Chengdu 610031, China)

Abstract: Many real data increase dynamically in size. As an effective and efficient mechanism to deal with such data, Incremental algorithms for attribute reduction based on modified equivalence relation matrix are proposed. When a group of attributes are added to a decision table, the algorithm aims to find the new minimal attribute reduction in a much shorter time. Our example illustrate that this algorithm is effective.

Key words: rough set; incremental update; positive domain reduction; relation matrix

1 引言

由于计算机硬件和网络的发展, 目前大多数的信息系统不仅积累了海量的数据而且信息数据大多是动态变化的, 每时每刻都有大量的实时数据增加到决策表中. 信息系统变化包含对象集、属性集以及属性值的变化. 针对静态的信息系统, 许多学者提出了大量的约简算法, 但动态属性约简算法研究相对较少, 由于信息系统中的属性集会随时间的变化而增加, 属性集的增加导致了信息系统的正域发生改变.

传统的知识获取方法对信息系统中不断变化的数据进行挖掘时, 不仅在时间上耗费巨大, 而且很难从动态数据中高效、及时地提取出有用的知识, 针对

传统只是获取方法的缺陷, 很多科学研究提出了增量式学习方法, 该方法是利用以前已有的知识, 不需要重新开始对所有知识进行操作, 仅仅通过更新变化部分数据, 所以能提高数据挖掘效率, 满足实时需求, 目前许多研究者对增量式属性约简, 已经做了大量的研究工作^[1-6].

2 粗糙集的相关理论

2.1 粗糙集与矩阵的相关理论

定义 1 设 $S = (U, A = C \cup D, V, f)$ 是一个四元组的信息系统, $U = \{x_1, x_2, \dots, x_n\}$ 表示有限的对象集合且是非空的, 称为论域; 其中 C 和 D 分别为条件属性集与决策属性集, $V = \bigcup_{a \in C \cup D} V_a$, 其中 V_a

收稿日期: 2014-05-09; 修回日期: 2014-07-02

基金项目: 国家自然科学基金联合项目(U1230117); 国家自然科学基金联合项目(61175047)

是属性 a 的值域, $f:U \times C \cup D \rightarrow V$ 是一个信息函数, 即 $a \in C \cup D, x \in U$ 有 $f(x, a) \in V_a^{[9]}$

定义 2 信息系统 $S = (U, A = C \cup D, V, f)$, U 为论域, $C, D \subseteq A$ 为条件属性集和决策属性集, 决策属性集 D 关于条件属性集 C 的依赖度定义为 $\gamma_C(D) = |POS_C(D)|/|U|$.

定义 3 (矩阵乘法): 如果 $A = (a_{ij})_{m \times n}, B = (b_{ij})_{n \times p}$, 则 $A \cdot B = (c_{ij})_{m \times p}$, 其中 $(c_{ij})_{m \times p} = \sum_{k=1}^n a_{ik} \cdot b_{kj}$.

定义 4 信息系统 $S = (U, A = C \cup D, V, f)$, T_C 是论域 U 上的等价关系, 则其等价关系矩阵 $M_{n \times n}^{T_C} = (m_{i,j})_{n \times n}$ 的元素为

$$(m_{i,j})_{n \times n} = \begin{cases} 1, & (u_i, u_j) \in T_C \\ 0, & (u_i, u_j) \notin T_C \end{cases}$$

定义 5 信息系统 $S = (U, A = C \cup D, V, f)$, 决策属性 D 的等价类 $U/D = \{D_1, D_2, \dots, D_m\}$, D_i 是 U/D 的任意子集, 则决策属性的矩阵 $M_{n \times m}^{D}$ 的元素为

$$M_{n \times m}^{D} = (m_{i,j})_{n \times m} = \begin{cases} 1, & u_i \in D_j \\ 0, & u_i \notin D_j \end{cases}$$

2.2 非增量式约简矩阵计算方法

文献[7] 中针对静态信息系统给出了等价关系矩阵的正域约简算法, 本文在文献[7] 的基础上, 当决策表增加一些属性时, 利用增量式更新技术对等价关系进行更新, 非增量约简算法 1^[7] 如下:

设 T_C 为决策表 S 的等价关系, 且其关系矩阵为 $M_{n \times n}^{T_C} = (m_{i,j})_{n \times n}$, 矩阵 $M_{n \times n}^{T_C}$ 的诱导矩阵 $\Lambda_{n \times n}^{T_C}$ 表示为 $\Lambda_{n \times n}^{T_C} = \text{diag}(1/\sum_{k=1}^n m_{1k}, 1/\sum_{k=1}^n m_{2k}, \dots, 1/\sum_{k=1}^n m_{nk})$.

定理 1 设 $S = (U, A = C \cup D, V, f)$ 是一个信息系统, U/C 和 U/D 是信息系统中的等价类, 其中 $U/D = \{D_1, D_2, \dots, D_m\}$, D_i 是 U/D 的任意子集, D 关于 C 的正域记为 $POS_C(D)$, 正域的矩阵计算方法: $POS_C(D) = (\Lambda_{n \times n}^{T_C} \cdot (M_{n \times n}^{T_C} \cdot M_{n \times m}^D))_1 = (\Lambda_{n \times n}^{T_C} \cdot C_{n \times m})_1$ 且用 $\cdot$ 来表示矩阵的代数乘法运算 $\Lambda_{n \times n}^{T_C} \cdot (M_{n \times n}^{T_C} \cdot M_{n \times m}^D)$ 的 1 - 截矩阵表示为 $(\Lambda_{n \times n}^{T_C} \cdot (M_{n \times n}^{T_C} \cdot M_{n \times m}^D))_1$.

2.3 决策信息系统属性约简算法的描述

算法 1: 非增量式矩阵约简的算法^[11]

输入: 信息系统 $S = (U, A = C \cup D, V, f)$.

输出: 信息系统的属性约简 RED .

Setp1 根据定理 1 计算 D 关于 C 的正域 POS_C^D ;

Step2 计算条件属性 C 的依赖度并降序排列,

设序列为 $\gamma_{C_1}(D) > \gamma_{C_2}(D) > \dots > \gamma_{C_m}(D)$;

Setp3 $RED = C_1$;

Setp4 for $i = 1:m$

4.1 根据文献[11] 中的定理 3 计算更新后的 $M_{RED}^{T_{RED \cup C_i}}$ 等价关系矩阵.

4.2 根据文献[11] 中的定理 4 计算更新后的对

角矩阵 $\Lambda_{n \times n}^{RED+} = \text{diag}(\lambda_1^+, \lambda_2^+, \dots, \lambda_n^+) (\lambda_i^+ = \sum_{j=1}^n m_{ij})$

4.3 根据定理 1 计算更新后 D 关于 $M_{RED}^{T_{RED \cup C_i}}$ 的正域 $POS_{RED}(D)$

if $POS_C(D) = POS_{RED}(D)$

Then

$RED = RED \cup C_i$

End

For ($a' \in RED$)

If ($POS_{RED-\{a'\}}(D) = POS_C(D)$) then

$RED = RED - a'$

End if

End

Setp5 输出信息系统的属性约简 RED .

3 增量式矩阵的约简算法

3.1 属性集增加时, 增量式更新的矩阵方法

定理 2 设 $S = (U, A = C \cup D, V, f)$ 是一个决策信息系统, $C = \{a_1, a_2, \dots, a_n\}$ 是条件属性集, $B \subseteq C$ 且 $B \neq \emptyset, a_i \in C - B$, 属性 B, a_i 的等价关系矩阵分别为 $M_{n \times n}^{T_B} = (m_{ij})_{n \times n}, M_{n \times n}^{T_{a_i}} = (m_{ij}')_{n \times n}$, 则把属性 $\{a_i\}$ 添加到属性集 B 后, B 的等价关系矩阵更新为 $M_{n \times n}^{T_{B \cup \{a_i\}}}$ 的元素为 $(m_{ij})^+ = (\min(m_{ij}, m_{ij}'))$.

注: 把属性 $\{a_i\}$ 增加到属性集 C 时, 如果 $m_{i,j} = 0$ 时, 更新后的 $m_{i,j}$ 是不变的^[5].

定理 3 设 $S = (U, A = C \cup D, V, f)$ 是一个信息系统, $B \subseteq A, B \neq \emptyset, a_i \in A - B, \Lambda_{n \times n}^{T_B} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n) (\lambda_i = \sum_{j=1}^n m_{ij})$.

设 $\Lambda_{n \times n}^{T_{B \cup \{a_i\}}} = \text{diag}(\lambda_1^+, \lambda_2^+, \dots, \lambda_n^+) (\lambda_i^+ = \sum_{j=1}^n m_{ij}')$,

则有: $\lambda_i^+ = \lambda_i - \sum_{j=1}^n (1 - \omega_{i,j}) \cdot m_{ij}$.

3.2 增加属性的增量式约简算法

当把一些属性增加到决策表时, 通过增量式更新技术更新等价关系矩阵, 我们提出了粗糙集增量式矩阵的约简法

算法 2: 增量式矩阵的约简算法

输入:增加属性前的等价关系 $M_{n \times n}^{T_{C_n}}$ 和约简的等价关系矩阵 $M_{n \times n}^{T_{RED}}$ 以及属性的约简 RED , 增加新的属性 C_X .

输出:增加属性后的属性约简.

Setp1: 根据定理 2 增量式更新等价关系矩阵 $M_{n \times n}^{T_{C_n}}$;

Setp2: 根据定理 3 增量式更新诱导矩阵 $\Lambda_{n \times n}^{T_{C_n}}$;

Step3: $RED^{\uparrow} = RED$;

Step4: 根据定理 1 计算更新后的 $POS_C(D)^{\uparrow}$ 和 $POS_{RED}(D)^{\uparrow}$, 如果 $POS_C(D)^{\uparrow} = POS_{RED}(D)^{\uparrow}$ 转到 Setp6, 否则执行 Setp5;

Setp5: for $i = 1: |C - RED|$

5.1 按照属性依赖度的大小来增加属性 C_i ;

5.2 根据定理 2 计算更新后的等价关系矩阵 $M_{n \times n}^{T_{RED}^{\uparrow}}$;

5.3 根据定理 3 计算更新后的对角矩阵 $\Lambda_{n \times n}^{T_{RED}^{\uparrow}}$

$= diag(\lambda_1^+, \lambda_2^+, \dots, \lambda_n^+) (\lambda_i^+ = \sum_{j=1}^n m_{ij}^+)$;

5.4 根据定理 1 计算更新后 D 关于 $RED^{\uparrow}$ 的正域 $POS_{RED}(D)^{\uparrow}$;

if $POS_C(D)^{\uparrow} = POS_{RED}(D)^{\uparrow}$

Then

$RED^{\uparrow} = RED \cup C_i$

End if

End

For ($a' \in RED^{\uparrow}$)

If ($POS_{RED^{\uparrow} - \{a'\}}(D) = POS_C^{\uparrow}(D)$) then

$RED^{\uparrow} = RED^{\uparrow} - a'$

End if

End

Setp6: 输出增加属性后的约简 $RED^{\uparrow}$

4 算例

表 1 是一个决策表: $U = \{x_1, x_2, x_3, x_4\}$ 是论域, $C = \{a, b, c\}$ 是条件属性, $D = \{d\}$ 是决策属性. 假设 $\{c_1, c_2\}$ 是决策表的一个约简, 如果把属性 $c_5 = \{1, 0, 1, 1\}$ 增加到决策表, 求属性增加后的约简.

表 1 决策信息系统

U	c_1	c_2	c_3	c_4	c_5	d
x_1	1	0	1	0	1	0
x_2	1	0	1	0	0	1
x_3	1	1	0	0	1	0
x_4	0	1	0	0	1	1

4.1 $M_{n \times m}^L$ 属性

根据文献[11], 可知决策表矩阵 $M_{n \times m}^D$ 属性为

$$M_{n \times m}^D = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

4.2 按照定理 2 计算等价关系矩阵 $M_{n \times n}^{T_{C_n}}$

新增属性 $\{c_5\}$ 后, 按照定义 4, 计算新增属性的等价关系矩阵 $M_{n \times n}^{T_{C_n}}$:

$$M_{n \times n}^{T_{C_n}} = \begin{bmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 \end{bmatrix},$$

$$M_{n \times n}^{T_{C_n}} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}.$$

按照定理 3, 计算更新诱导矩阵 $\Lambda_{n \times n}^{T_{C_n}}$ 为

$\Lambda_{n \times n}^{T_{C_n}} = \text{diag}(1/1, 1/1, 1/1, 1/1)$.

4.3 计算增加属性后的约简

按照定理 1 计算增加属性后的条件属性 C 的正域

$POS_C(D)^{\uparrow} =$

$$(\Lambda_{n \times n}^{T_{C_n}} \cdot (M_{n \times n}^{T_{C_n}} \cdot M_{n \times m}^D))_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix},$$

则 D 关于 C 的正域为 $POS_C(D)^{\uparrow} = \{x_1, x_2, x_3, x_4\}$.

根据文献[11], 可知 D 关于 RED 的正域为

$POS_{RED}(D)^{\uparrow} = \{x_3, x_4\}$.

因为

$POS_C(D)^{\uparrow} \neq POS_{RED}(D)^{\uparrow}$,

计算增加属性后的依赖度并降序排列 $\gamma_{c_5}(D) > \gamma_{c_4}(D) = \gamma_{c_3}(D)$, 所以增加属性 c_5 到 RED 中, 由等价关系矩阵 $M_{n \times n}^{T_{C_n}}$ 和 $M_{n \times n}^{T_{RED}}$ 增量式求等价关系矩阵 $M_{n \times n}^{T_{(RED) \cup \{c\}}}$:

$$M_{n \times n}^{T_{(RED) \cup \{c\}}} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

按照定理 3 计算变化后的诱导矩阵 $\Lambda_{n \times n}^{T_{(RED) \cup \{c\}}}$:

$\Lambda_{n \times n}^{T_{(RED) \cup \{c\}}} = \text{diag}(1/1, 1/1, 1/1, 1/1)$.

根据定理 1 计算变化后的正域:

$$POS_{RED}(D)^{\uparrow} = (\mathbf{A}_{n \times H}^{T_{RED} \cup \{c\}} \cdot (\mathbf{M}_{n \times H}^{T_{RED} \cup \{c\}} \cdot \mathbf{M}_{n \times m}^D))_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$POS_{RED}(D)^{\uparrow} = \{x_1, x_2, x_3, x_4\}$$

故 $POS_C(D)^{\uparrow} = POS_{RED}(D)^{\uparrow}$ 所以 $\{c_1, c_2, c_4\}$ 是表 1 的一个属性约简.

5 实验测试与分析

为了验证增量式约简的矩阵算法比非增量式约简的矩阵算法的有效性,我们从 UCI 数据集下载了 3 个数据集分别为 Cancer, Balance, 数据集描述见表 2,分别用两种方法对 3 个数据集进行了测试,并对所消耗的时间进行比较,实验测试的软硬件环境: CPU Intel Core™ 双核 2GHz, 内存 1.0GB, Windows7.0 操作系统, C++ 开发平台.

表 2 数据集描述

序号	数据集	属性个数	对象个数
1	Wine	13	178
2	Cancer	9	683
3	Balance	5	625

在测试实验中,把每个数据集属性的 60% 作为基本数据集,剩余的属性作为增加得属性集,当增加属性集时,我们分别用增量式矩阵约简和非增量式矩阵约简得方法测试,比较两类矩阵算法所消耗的时间,实验测试结果如表 3 所示.

表 3 增量式与非增量式时间

序号	数据集	非增量式/s	增量式/s
1	Wine	0.51	0.22
2	Cancer	0.89	0.4
3	Balance	0.78	0.3

从表 3 可得增量式的矩阵约简算法花费的时间明显比非增量式的矩阵约简算法所用时间少,且增量式的矩阵约简算法能在较短时间找到一个有效的约简.

6 结束语

当有新属性添加到决策表时,常常会引起决策表的正域发生变化,如果重新计算会花费很多时间,且造成资源浪费,为此提出了增加属性条件下的增量式矩阵的约简算法.该算法利用增加属性前的关系矩阵和属性约简知识,通过增量式学习方法,能够在较短时间内找到一个约简,在理论分析的基础上,通过 UCI 数据实验表明,本文所给出的增量式矩阵的约简算法是可行的,下一步工作将探讨属性值变化时信息系统增量式矩阵约简算法.

参考文献:

- [1] Qian Y, Liang J, Pedrycz W, et al. Positive approximation: an accelerator for attribute reduction in rough set theory[J]. Artificial Intelligence, 2010, 174(9/10): 597-618.
- [2] Wang F, Liang J, Dang C, et al. Attribute reduction: a dimension incremental strategy [J]. Knowledge-Based Systems, 2013(39): 95-108.
- [3] 王磊,李天瑞. 一种基于矩阵的知识粒度计算方法[J]. 模式识别与人工智能, 2013, 26(5): 448-453.
- [4] Shaoyong Li, Tianrui Li, Dun Liu. Incremental updating approximations in dominance-based rough sets approach under the variation of the attribute sets[J]. Knowledge-Based Systems, 2013, 40(1): 17-26.
- [5] Chuan Luo, Tianrui Li, Hongmei Chen. Dynamic maintenance of approximations in set-valued ordered decision systems under the attribute generalization[J]. Information Sciences, 2014, 257(1): 210-228.
- [6] 李抒音,姚崇怀. 基于增量式更新带权差别集合的属性约简算法[J]. 微电子学与计算机, 2008, 25(11): 68-71.
- [7] 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, 286.

作者简介:

景运革 男, (1970-), 博士研究生, 副教授. 研究方向为粗糙集理论. E-mail: jyq701022@163.com.

黄峻福 男, (1990-), 硕士. 研究方向为云计算与知识发现.