

基于差别矩阵的增量式属性约简完备算法

刘洋, 冯博琴, 周江卫

(西安交通大学计算机科学与技术系, 710049, 西安)

摘要: 为了解决基于差别矩阵的属性约简完备算法得不到最小约简的问题, 提出了一种改进的属性约简方法. 该方法将信息论定义的属性重要性作为启发式信息, 并通过构造一个条件信息熵算子对差别集合进行运算, 同时利用算子来计算候选属性的剔除次序, 采用宽度优先搜索策略使约简集合中含有最重要的属性, 这样就解决了完备算法约简率低的问题. 结合该方法并在分析对象集增量与差别矩阵关系的基础上, 证明了增量约简定理, 由此提出了一种增量式约简完备算法(CAIR), 当新数据加入决策表时, 算法可增量构造差别集合. 实验结果表明, 所提 CAIR 在大大缩短计算差别集合时间的同时, 约简率比非完备算法提高了 20.3%, 是同条件下完备算法执行效率的 13.2 倍.

关键词: 差别矩阵; 差别集合; 属性约简; 完备算法

中图分类号: TP18 **文献标识码:** A **文章编号:** 0253-987X(2007)02-0158-04

Complete Algorithm of Increment for Attribute Reduction Based on Discernibility Matrix

Liu Yang, Feng Boqin, Zhou Jiangwei

(Department of Computer Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China)

Abstract: Aiming at the problem that complete algorithm of attribute reduction based on discernibility matrix can not find approximately minimal reduction, an improved attribute reduction method is proposed based on the original algorithm, and the attribute importance defined from the viewpoint of information theory is regarded as heuristic information. The discriminated set is operated by constructing an operator of conditional information entropy, and the excluding order of candidate attributes is calculated as the algorithm is iterated. Meanwhile, the breadth-first search strategy is utilized to make minimal reduction set contain the most important attribute. Thus, the method can resolve the problem that complete algorithm is of low reduction rate. On the basis of the analysis of the relation between the objects increase and discernibility matrix, a theorem of the increment reduction is proved, and then a complete algorithm for increment reduction (CAIR) is presented. When new data are added into the decision table, the discernibility set can be constructed incrementally. The experimental results show that the computation time of the discernibility set is significantly reduced by CAIR, the reduction rate is 20.3% higher than that of incomplete algorithm, and the execution efficiency is enhanced by 13.2 times compared to the complete algorithm under same conditions.

Keywords: discernibility matrix; discernibility set; attribute reduction; complete algorithm

粗糙集理论^[1]是一种处理不精确、不相容和不 完全数据的新的数学工具, 已经在人工智能、数据

挖掘、模式识别等领域得到了较为成功的应用^[2]. 属性约简是粗糙集理论的核心内容之一,研究者已从计算复杂性的角度证明了找出决策表的最小约简是 NP 难问题^[3]. 所以,一般采用启发式算法可以找出近似最优约简以提高约简效率,而这些算法的共同特点是利用属性的重要性作为启发式信息,但并不完备^[4-5],不能保证一定能够得到约简^[6].

文献[6]给出了一种以差别矩阵为基础的基于属性序的完备算法,约简质量取决于预先定义的性质顺序. 文献[7-8]有所改进,它使约简算法在缺少先验知识的前提下同样有效,约简的质量也有所提高,但约简效率远低于非完备算法.

本文在完备算法的基础上引入了属性的重要性算子,用条件熵作为完备算法的启发式信息,并采用宽度优先搜索策略选择剔除的属性,该方法也是一种启发式算法. 为了提高完备算法在处理大规模应用时的效率,本文又建立了对象集增量与差别矩阵的关系,得到了基于差别矩阵的增量约简定理和增量式属性约简完备算法,它比非完备算法有更好的约简效果.

1 属性约简完备算法的原理

1.1 相关概念和定义

一个决策表可以表示为一个五元组 $S=(U, C, D, V, f)$, 其中 U 为论域, C 为条件属性集合, D 为决策属性集合, V 为对象属性的值域, f 为一个映射函数 $U \times A \rightarrow V$.

决策表 S 的差别矩阵 M_S 是一个 $n \times n$ 矩阵, 其中 $n = |U|$, $M_S = (m_{ij}) = \{a \in C \mid f(x_i, a) \neq f(x_j, a) \wedge t(x_i, x_j) = 1\}$, $U_1 = P_C(D)$, $U_2 = U - U_1$, 且有

$$t(x_i, x_j) = \begin{cases} 1 & \begin{matrix} x_i \in U_1, x_j \in U_2 \\ x_j \in U_1, x_i \in U_2 \\ x_i, x_j \in U_1, (x_i, x_j) \notin \text{IND}(D) \end{matrix} \\ 0 & \text{其他} \end{cases}$$

决策表 S 的差别集合 $A_S = \{m_{ij} \mid m_{ij} \in M_S \wedge m_{ij} \neq \emptyset\}$ ^[8], A_S 的核元素集合 $\text{Atom}(A_S) = \{a \mid \{a\} \in A_S\}$, A_S 的属性集 b 剔除 $A_S^{\text{exclude}}(b)$ 以及属性集 β 选择 $A_S^{\text{select}}(\beta)$ 分别定义为

$$A_S^{\text{exclude}}(b) = \{\alpha' \mid \alpha' = \alpha - \{b\}, \alpha' \neq \emptyset, \alpha \in A_S\}$$

$$A_S^{\text{select}}(\beta) = \{\alpha \mid \alpha \cap \beta = \emptyset, \alpha \in A_S\}$$

文献[8]提出了一种基于差别集合 A_S 的属性约简完备算法,其主要思想是:在每次迭代过程中选择必要的条件属性加入到属性约简结果集 R 中,如果在某次迭代过程中找不到这样的条件属性,

则任意剔除一个条件属性. 这种算法需要计算 $\text{Atom}(A_S)$, 如果 $\text{Atom}(A_S)$ 为空集,说明剩余属性集合中存在冗余属性,此时需要从 A_S 的所有元素中剔除一个冗余属性. 当进行属性剔除时,如果任意选择属性,在进行剔除时有可能丢失对决策表起关键作用的属性,这会导致算法得出的约简结果不是最小的.

1.2 改进的属性约简完备算法

本文引入属性重要性算子来描述算法,它在已有约简子集的条件下对剔除的属性进行优先排序,并采用宽度优先搜索策略使约简集合中含有最重要的属性,再利用基于信息熵相对重要性的度量标准^[5]来度量属性重要性.

定义 1 在决策表 S 中,对于任意的候选剔除属性 $a \in C - R$,在完备算法已选择约简属性子集为 $R \subset C$ 的条件下,属性 a 的重要性为 $\sigma_{R,D}(a) = H(D|R) - H(D|R \cup \{a\})$,其中 $H(D|R)$ 表示已知 R 时, D 的条件信息熵.

算法 1 计算条件属性的某个完备约简.

输入: 决策表 S 的差别集合 A_S .

输出: 决策表 S 的最小属性约简集合.

步骤 1: $R = \emptyset, T = \emptyset; A = A_S$.

步骤 2: 如果 A 为空集,则输出 R 并退出,否则执行步骤 3.

步骤 3: $P = \text{Atom}(A)$, 如果 P 为空集,则转步骤 4, 否则执行步骤 5.

步骤 4: 选择一个属性 $a \in C - (R \cup T)$, 使得 $\forall b \in C - (R \cup T), \sigma_{R,D}(a) < \sigma_{R,D}(b)$ 成立, 令 $A = A^{\text{exclude}}(a), T = T \cup \{a\}$, 并上转步骤 2.

步骤 5: $R = R \cup P, A = A^{\text{select}}(R)$, 上转步骤 2.

2 增量式属性约简完备算法

当数据以增量方式输入时,计算增量数据产生的差别集合(简称增量差别集合). 下面将证明增量约简定理,并基于该定理提出增量式属性约简完备算法(CAIR),当增量差别集合存在非原始差别集合中的元素时,它可进行启发式约简,从而避免重复计算差别集合.

2.1 增量约简定理和增量式属性约简完备算法

定义 2 设有 2 个结构完全相同的决策表 $S_1 = (U_1, A, V, f_1)$ 和 $S_2 = (U_2, A, V, f_2)$, 则 S_1 与 S_2 的交互差别集合为 $A_{S_1 \oplus S_2} = \{m_{ij} \mid x_i \in U_1 \wedge x_j \in U_2, x_i, x_j \text{ 为 } m_{ij} \text{ 对应的实例}\}$.

由定义 2 可知, $A_{S_1 \oplus S_2} = A_{S_2 \oplus S_1}$.

引理 1 有 2 个结构相同的决策表 $S_1=(U_1, A, V, f_1)$ 和 $S_2=(U_2, A, V, f_2)$, 设合并后的决策表为 $S=(U, A, V, f)$, 其中 $U=U_1 \cup U_2, f=f_1 \cup f_2$, 则 $A_S=A_{S_1} \cup A_{S_2} \cup A_{S_1 \oplus S_2}$ 成立.

证明 对于任意的 $m_{ij} \in U_1 \cup U_2, x_i, x_j$ 为 m_{ij} 对应的对象实例, 则有: ①如果 $x_i, x_j \in U_1$, 于是 $m_{ij} \in A_{S_1}$; ②如果 $x_i, x_j \in U_2$, 于是 $m_{ij} \in A_{S_2}$; ③如果 $x_i \in U_1, x_j \in U_2$, 于是 $m_{ij} \in A_{S_1 \oplus S_2}$; ④如果 $x_i \in U_2, x_j \in U_1$, 于是 $m_{ij} \in A_{S_2 \oplus S_1} = A_{S_1 \oplus S_2}$. 综合以上 4 种情况得到: $A_S \subseteq A_{S_1} \cup A_{S_2} \cup A_{S_1 \oplus S_2}$. 同理可以证明 $A_{S_1} \cup A_{S_2} \cup A_{S_1 \oplus S_2} \subseteq A_S$.

定理 1 设决策表 $S_1=(U_1, A, V, f_1)$ 的差别集合为 $A_{S_1}, S_2=(U_2, A, V, f_2)$ 的差别集合为 A_{S_2} , 若 $A_{S_1} \cup A_{S_2} \cup A_{S_1 \oplus S_2} = A_S$ 成立, 则决策表 S_1 的任意一个约简也是决策表 $S=(U, A, V, f)$ 的一个约简, 其中 $U=U_1 \cup U_2, f=f_1 \cup f_2$.

证明 根据引理 1, 决策表 S 的差别集合 $A_S=A_{S_1} \cup A_{S_2} \cup A_{S_1 \oplus S_2}$, 由于 $A_S \subseteq A_{S_1}$ 并且 $A_{S_1} \subseteq A_S$, 因此 $A_{S_1}=A_S$, 将 A_S 和 A_{S_1} 分别代入文献[8]的完备约简算法中, 可以得到相同的约简结果.

由定理 1 可知, 增量式约简算法首先计算增量差别集合 $\Delta A_S=A_{\Delta S} \cup A_{\Delta S \oplus S}$ 并判断 ΔA_S 和 A_S 的关系, 如果 $\Delta A_S \subseteq A_S$ 成立, 增量数据在加入原决策表之后, 便可获得决策表的相对约简. 对于数据以增量形式输入的情况, 该算法只需要计算增量差别集合.

算法 2 增量式属性约简完备算法(Complete Algorithm of Increment Reduction, CAIR).

输入: 增量数据.

输出: 决策表 S 近似最小属性约简集合.

步骤 1: 选取初始数据集 d_0 , 对 d_0 进行预处理, 包括数据清理、离散化和补足缺失属性, 得到 d_c , 计算 d_c 的差别集合 A_S .

步骤 2: 将 A_S 作为输入调用算法 1, 得到最小属性约简集合.

步骤 3: 如果不更新数据集, 下转步骤 7, 否则执行步骤 4.

步骤 4: 读取增量数据并进行预处理, 得到 Δd_c .

步骤 5: 计算 Δd_c 的增量差别集合 ΔA_S , 如果 $\Delta A_S \subseteq A_S$, 说明约简结果达到了预期的效果, 将 d_c 与 Δd_c 合并, 即 $d_c = d_c \cup \Delta d_c$, 并返回步骤 4, 否则合并差别集合 $A_S = A_S \cup \Delta A_S$, 上转步骤 2.

步骤 6: 将 A_S 作为算法 1 的输入计算最小属性约简, 并返回步骤 4.

步骤 7: 输出当前决策表的最小属性约简.

系统首先从数据库接收一批数据作为初始训练集, 并调用基于差别集合的启发式约简算法, 得到初始训练集的最小约简, 之后接收增量数据并计算增量差别集合 ΔA_S . 由定理 1 可知, 如果满足 $\Delta A_S \subseteq A_S$ 的条件, 不必调用算法 1, 按照算法 2 不断对一次接受到的增量数据计算增量差别集合 ΔA_S 并进行属性约简. 利用 CAIR 可避免重复计算决策表的差别矩阵所付的代价, 当处理大数据集时, 它比完备算法有更好的约简效率.

2.2 算法复杂性分析

在每次迭代过程中, CAIR 需选择必要的条件属性加入到约简结果集 R 中. 如果在某次迭代过程中找不到必要的属性, 则通过宽度优先搜索策略对候选剔除属性集合进行扫描, 删除属性重要性算子 $\sigma_{R,D}(a)$ 最小的属性, 从而确保 R 中每个属性都是不可省的, 而采用宽度优先搜索最后得到的 R 一定是完备约简. 计算差别集合 A 的核元素集合 $\text{Atom}(A)$ 的时间复杂度是 $O(|A|)$, 计算 $A^{\text{exclude}}(b)$ 和 $A^{\text{select}}(\beta)$ 的时间复杂度分别为 $O(|A|)$ 和 $O(|A| \cdot |C|)$, 剔除非核心属性需要计算候选属性的启发式信息, 其属性重要性算子的时间复杂度为 $O(|A|)$. 由于算法 1 最多重复执行 $|C|$ 次, 因此算法的时间复杂度为 $O(|A||C| + |A| + |A||C|) = O(|A| \cdot |C|^2) = O(|U|^2 |C|^2)$.

3 实验仿真及结果分析

3.1 实验设计

用 Microsoft Visual Studio C++ 6.0 分别实现了本文算法、文献[8]中的完备约简算法(简称为算法 A)和文献[5]中的非完备约简算法(简称为算法 B), 并选用了 UCI 机器学习(<http://www.ics.uci.edu/~mlearn/MLRepository.html>)中 8 个典型数据库实例. 用 n 表示数据集的实例数, m 表示原始条件属性数, r 表示约简后的条件属性数, $\eta = (m - r)/m$ 表示约简率, t 表示约简时间(算法的平均执行时间). η 和 t 可以反映算法对决策表的约简效果, 即当 t 基本相同时, r 就越多, 而当 r 相同时, t 就越少, 算法的约简效果也就越好.

3.2 实验结果及分析

图 1 给出了 CAIR 的性能测试结果, 从中可以看出执行增量式完备约简所需的时间比执行基本完备约简的要少, CAIR 的效率随增量实例数量的增加而下降. 表 1 记录了 3 种算法的 η 和 t .

表1 在 UCI 数据集上的约简算法实验结果

数据集	n	m	算法 CAIR		算法 A		算法 B	
			r	t/s	r	t/s	r	t/s
Echocardiogram	87	7	4	0.348	4	7.544	4	0.029
Parity5+5	100	10	5	0.467	7	8.160	5	0.043
Wine	118	13	4	0.712	5	10.274	4	0.274
Audiology	150	69	14	0.955	15	10.520	20	0.522
Heart	180	13	9	1.463	11	12.738	10	1.015
Lymphography	148	18	7	1.632	8	15.312	8	1.129
Chess End-Game	3 196	36	30	3.681	34	42.671	31	3.334
Mushroom Database	8 124	22	5	5.144	8	68.174	7	4.971

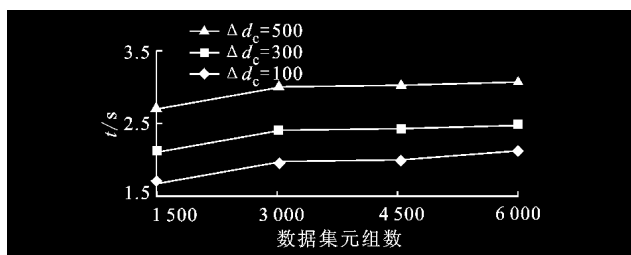
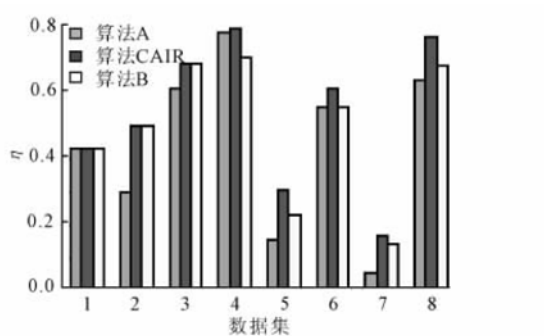


图1 算法 CAIR 的性能测试

从表1可以看出,CAIR在 t 和 η 上都优于算法A.对于大数据集,CAIR和算法B的 t 基本一致,但CAIR的约简质量比较高.从实验结果又知,在大多数数据集上,CAIR的约简效果比算法B更好,且通常可以找到最小或近似最小的约简.

图2给出了3种约简算法的 η ,结果表明,CAIR在大大缩短 t 的同时,约简质量比非完备算法提高了20.3%,其效率是同条件下其他完备算法的13.2倍.随着实例数的增加,CAIR的约简效果更为显著.



1:Echocardiogram; 2:Parity5+5; 3:Wine; 4:Audiology;
5:Heart; 6:Lymphography; 7:Chess End-Game;
8:Mushroom Database

图2 3种算法的约简率对比

4 结束语

现有的很多启发式属性约简算法是不完备的,

本文在属性约简完备算法的基础上引入了属性的重要性算子,用条件熵作为启发式信息设计了启发式属性约简完备算法.在仔细分析对象集增量与差别集合关系的基础上,提出了一种基于差别集合的增量式属性约简完备算法,使完备算法能够更好地面对大数据集的挑战.实验结果表明,增量式属性约简算法在较多数据集上的约简质量均高于非完备约简算法.下一步的工作将探讨更为有效的差别集合更新方法,并在不同的测试集上对算法的效果做进一步的实验.

参考文献:

- [1] Pawlak Z. Rough sets [J]. International Journal of Computer and Information Science, 1982, 11(5):341-356.
- [2] Pawlak Z. Rough sets: theoretical aspects of reasoning about data [M]. Boston: Kluwer Academic Publishers, 1991.
- [3] Wong S K M, Ziarko W. On optimal decision rules in decision tables [J]. Bulletin of Polish Academy of Sciences, 1985, 33(11/12):693-696.
- [4] Hu X H, Cercone N. Learning in relational databases: a rough set approach [J]. Computational Intelligence, 1995, 11(2):333-338.
- [5] 王国胤,于洪,杨大春.基于条件信息熵的决策表约简[J].计算机学报,2002,25(7):759-766.
Wang Guoyin, Yu Hong, Yang Dachun. Decision table reduction based on conditional information entropy [J]. Chinese Journal of Computers, 2002, 25(7):759-766.
- [6] Wang Jue, Wang Ju. Reduction algorithms based on discernibility matrix: the ordered attributes method [J]. Journal of Computer Science and Technology, 2001, 16(6):489-504.

- [7] Zhang Wenxiu, Ling Wei, Qi Jianjun. Attribute reduction in concept lattice based on discernibility matrix [C]//Proceedings of Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing. Berlin: Springer-Verlag, 2005:157-165.
- [8] Wang Bing, Chen Shanben. A complete algorithm for attribute reduction [C]//Proceedings of Robotic Welding, Intelligence and Automation. Berlin: Springer-Verlag, 2004:345-352.

(编辑 苗凌)