

Jelonek 属性约简算法的一个改进

叶东毅

(福州大学计算机系,福州 350002)

摘 要: 受 Jelonek 粗糙集属性约简算法思想的启发,本文利用单属性的近似精度和 Hu 的差别矩阵方法,得到一个改进的属性约简算法.理论分析表明,该算法比 Jelonek 算法具有更低阶的计算复杂性.

关键词: 粗糙集;近似精度;属性约简;差别矩阵

中图分类号: TP182 **文献标识码:** A **文章编号:** 0372-2112 (2000) 12-0081-02

An Improvement to Jelonek's Attribute Reduction Algorithm

YE Dong-yi

(Department of Computer, Fuzhou University, Fuzhou 350002, China)

Abstract: In this paper, inspired by Jelonek's algorithm, an improved attributes reduction algorithm is given based on the approximation quality of a single attribute and the modified discernibility matrix due to Hu. It is shown that the presented algorithm is of lower computational complexity than Jelonek's algorithm.

Key words: rough set; approximation quality; reduction of attribute; discernibility matrix

1 引言

属性约简是粗糙集理论中的一个核心部分^[1,2],至今,人们已提出了若干个属性约简算法^[3~6].Jelonek^[6]等人提出的算法是其中比较典型的一个(它与 HU 的算法^[5]本质上是-一样的),取得了较好的效果,但也存在一些不足,它必须计算很多不同属性子集的近似精度才能决定如何扩展候选属性约简,因此,需要较多的计算量.本文对此进行改进,利用单属性的近似精度决定如何扩展候选属性约简,并利用改进的差别矩阵进行修正,得到一个新的属性约简算法,明显降低了计算复杂性,提高了计算速度.

2 改进的约简算法

有关粗糙集的一些基本概念,读者可参阅有关文献^[1,2].这里,仅介绍一下属性约简、核等概念.

给定一个信息系统^[2]:

$$L = (U, Q, V_q, F_q), q \in Q$$

其中 $U = \{x_1, \dots, x_n\}$ 是论域, Q 是属性集合, V_q 为属性取值的集合, F_q 是 $U \times Q \rightarrow V_q$ 的映射.在决策分析中,属性集合 Q 通常分为条件属性集 C 与决策属性集 D .

以下,设条件属性集合 C 中有 m 个属性: C_1, C_2, \dots, C_m , 其值域为有限离散集合.不失一般性,假设仅有一个决策属性 D ,其取值范围是 $1, 2, \dots, k$.由 D 决定的等价类构成 U 的一个划分: $\{Y_1, Y_2, \dots, Y_k\}$, 其中, $Y_i = \{x \in U : F_q(x, D) = i\}$, $i = 1, \dots, k$.

定义 1 设 $P \subseteq C$, 对划分 $\{Y_1, Y_2, \dots, Y_k\}$ 的 P -近似精度 (approximation quality) 为^[2]:

$$p = \frac{\sum_{i=1}^k \text{card}(\underline{P}Y_i)}{\text{card}(U)} \quad (1)$$

其中 $\underline{P}Y_i$ 为 Y_i 的 P -下近似^[2] (lower approximation), $\text{card}(\cdot)$ 表示集合的基数.

定义 2 设 $R \subseteq C$, 若 $R = C$, 且不存在 $R \subset R'$, 使得 $R' = R$, 则称 R 为 C 的一个属性约简^[2].所有 C 的属性约简的交称为 C 的核, 记为 $\text{CORE}(C)$.

从核的定义可知,“不存在 $R \subset R'$, 使得 $R' = R$ ”等价于“ $R = \text{CORE}(R)$ ”, 因此, 下列性质是显然的.

性质 1 如果 R 满足 $R = C$, 且 $R = \text{CORE}(R)$, 则 R 为 C 的一个属性约简.

文献^[5]给出一个利用改进的差别矩阵直接计算核的方法, 其中, 改进的差别矩阵 $M = \{m_{ij}\}$ 定义为:

$$m_{ij} = \begin{cases} \{a \in C : F_q(x_i, a) \neq F_q(x_j, a)\}, & \text{当 } F_q(x_i, D) \neq F_q(x_j, D) \text{ 时} \\ (\text{空集}), & \text{其他情况时} \end{cases} \quad (2)$$

当且仅当某个 m_{ij} 为单个属性时, 该属性属于核 $\text{CORE}(C)$. 容易验证, 这种求核方法的计算复杂性为 $O(n^2 m)$.

对于求属性约简的问题, Jelonek 等人提出了逐步扩展型算法^[6], 其基本思想是从 $R = \text{核}(\text{CORE}(C))$ 开始, 如果 $R = C$, 则 R 为属性约简; 否则, 对所有的属性 $a \in C \setminus R$ 计算 $\text{gain} = p_{R \cup \{a\}} - p_R$, 若 a 是使相应的 gain 达到最大的属性, 则

收稿日期: 1999-10-12; 修回日期: 2000-05-15

基金项目: 福建省自然科学基金 (No. A0010009); 福建省优秀留学回国人员科研基金 (No. JA00144); 国家 973 计划 (No. G1998030600)

置 $R = R \setminus \{a\}$ 并判断它是否为约简 ($R = C$ 否, 若是, 则停机; 否则, 继续上述过程. 在该算法中, 从核 (C 可以是空集) 开始, 每扩展一次, 都要对所有的属性 $a \in C \setminus R$, 计算新的近似精度 $\rho_R \setminus \{a\}$, 计算量是比较大的. 容易验证, ρ_R 的计算复杂性为 $O(n^2 \text{card}(R))$, 完成一次扩展的计算复杂性为 $O(n^2 \text{card}(R) \text{card}(C \setminus R))$. 因此, 在最坏情况下, Jelonek 算法的计算复杂性为 $O(n^2 m^3)$. 本文利用单属性的近似精度进行扩展, 并利用改进的差别矩阵进行修正, 得到一个新的属性约简算法 (简称算法 A), 其基本过程如下:

步骤 1 对每个条件属性 $C_j, j = 1, \dots, m$, 计算 c_j ; 令 $Q = \text{CORE}(C)$, $P = Q$, $D = C \setminus P$; 如果 P 满足 $\rho_P = \rho_C$, 则停止, P 为约简.

步骤 2 计算 $c_g = \max(C_i : C_i \subseteq D)$, 置 $P = P \cup \{C_g\}$, $D = D \setminus \{C_g\}$, 转下一步继续执行;

步骤 3 如果 P 满足 $\rho_P = \rho_C$, 转步骤 4; 否则, $\rho_P < \rho_C$, 转步骤 2.

步骤 4 如果 $P = \text{CORE}(P)$ (采用前述的差别矩阵方法计算^[5]), 则停止, P 为一个属性约简; 否则, 转步骤 5.

步骤 5 任取 $C_d \in P \setminus \text{CORE}(P)$, 令 $P = P \setminus \{C_d\}$ 转步骤 4.

下面对算法 A 的正确性和计算复杂性进行分析.

命题一: 算法 A 得到的 P 为一个属性约简.

证明: 根据算法 A, 所求得的 P 满足 $\rho_P = \rho_C$, 且 $P = \text{CORE}(P)$. 由性质 1 即知, P 为一个属性约简. 证毕.

命题二: 在最坏情况下, 算法 A 的计算复杂性为 $O(n^2 m^2)$.

证明: ρ_R 的计算复杂性为 $O(n^2 \text{card}(R))$. 所以, m 个单属性近似精度 c_j 的计算复杂性为 $O(n^2 m)$, 对 m 个 c_j 进行快速排序和求核的计算复杂性分别为 $O(m \log(m))$ 和 $O(n^2 m)$. 因此, 当算法进入到步骤 4 之前的计算复杂性为 $O(n^2 m)$; 以后, 每执行一次步骤 4, 要对逐步缩小的 P 计算 $\text{CORE}(P)$ (采用差别矩阵的方法), 由前述可知, 其计算量为 $O(n^2 \text{card}(P))$, 因此, 求出约简至多需要 $\text{card}(P) O(n^2 \text{card}(P)) =$

$O(n^2 (\text{card}(P))^2)$ 的计算量. 因此, 在最坏情况下, 算法 A 的计算复杂性为 $O(n^2 m^2)$. 证毕.

从上述两个命题可知, 就计算复杂性而言, 算法 A 比 Jelonek 算法 ($O(n^2 m^3)$) 减少了一个数量级, 因此, 从理论上改进了 Jelonek 算法. 另外, 应用算法 A 对 UCI 机器学习数据库中的具有离散属性值的一部分数据库表进行计算, 虽然不能保证每次得到最小属性约简 (Jelonek 算法也无法保证做到这点), 但大部分情况下, 能很快得到属性个数较少的属性约简, 且平均计算时间明显少于 Jelonek 算法. 因此, 算法是实用的.

参考文献:

- [1] 曾黄麟. 粗集理论及其应用 (修订版) [M]. 重庆大学出版社, 1998.
- [2] Z. Pawlak and R. Slowinski. Rough set approach to multiattribute decision analysis [J]. Invited Review. European Journal of Operational Research, 1994, 72: 443 - 459.
- [3] 王珏等. 基于 Rough Sets 理论的“数据浓缩” [J]. 计算机学报, 1998, 21(5): 393 - 400.
- [4] 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, 6(36): 681 - 684.
- [5] X H Hu, N Cercone. Learning in relational databases: a rough set approach [J]. Computational Intelligence, 1995, 11(2): 323 - 338.
- [6] J. Jelonek et al. Rough set reduction of attributes and their domains for neural networks [J]. Computational Intelligence, 1995, 11(2): 339 - 347.

作者简介:



叶东毅 1964 年出生, 教授, 博士, 主要从事神经网络、最优化理论和粗糙集等方面的研究, 已发表有关论文 30 余篇.

(上接第 80 页)

参考文献:

- [1] Jack H. Winters. Smart antennas for wireless systems [J]. IEEE Pers. Commun., Feb. 1998: 23 - 27.
- [2] Ryuji Kohno. Spatial and temporal communication theory using adaptive antenna array [J]. IEEE Pers. Commun., Feb. 1998: 28 - 35.
- [3] L. C. Godara. Application of antenna arrays to mobile communications, Part I: Performance Improvement, Feasibility, and System Considerations [J]. Proc. IEEE, July 1997.
- [4] Richard B. Ertel et al. Overview of spatial channel models for antenna array communication systems [J]. IEEE Pers. Commun., Feb. 1998: 10

- 22.

- [5] Arogyaswami J. Paulraj and Boon Chong Ng. Space-time models for wireless personal communications [J]. IEEE Pers. Commun., Feb. 1998: 36 - 48.
- [6] N. Ishii and R. Kohno. Joint optimization of spatial and temporal multiuser equalization in both transmitter and receiver using an adaptive array antennas for DS/CDMA [J]. IEEE GLOBECOM '96, Commun. Theory Mini-Conf., Nov. 1996: 137 - 141.
- [7] N. Ishii and R. Kohno. Spatial and temporal joint transmitter-receiver using an adaptive array antenna [J]. IEICE Trans, Commun., Mar. 1996, E79-B(3): 361 - 367.