

动态知识系统的规则约简方法

Rules Reduction for Dynamic Knowledge Systems

闫德勤¹ 迟忠先²

(辽宁师范大学计算机系 大连116029)¹ (大连理工大学计算机系 大连116026)²

Abstract In recent years, due to its practical use in knowledge discovery and data mining, the research for rules reduction is becoming more important aspect of computer science. Based on rough set theory, a method of rules reduction for Dynamic Knowledge Systems is proposed in this paper. The experiments show the method is effective.

Keywords Rough set, Dynamic knowledge system, Rules reduction

1 引言

知识系统中的规则约简是知识发现、数据挖掘的重要内容。近年来粗集(Rough sets)理论在知识约简的应用已成为计算机科学的热点问题^[1~4]。对静态数据的规则约简已有多种算法^[5,6]。但知识系统往往是动态的,本文给出了一种针对动态知识系统的基于粗集理论的规则约简方法。对知识系统而言,动态一般指记录或论域的元素不断增加。因此,有关处理动态系统的一些方法又称为增量式算法。文章给出了所提出方法的有效性的结论以及实验的结果。

2 相关的基本概念与定义

设 $U = \{x_1, x_2, \dots, x_n\}$ 是一有限集,称为论域, R 是 U 上的一个等价关系, U/R 表示在 U 上导出的所有等价类; $[x]_R$ 表示包含元素 x 的 R 的等价类, $x \in U$ 。对任集合 $X \subseteq U$

$$R_-(X) = \{x \in U | [x]_R \subseteq X\}$$

$$R_+(X) = \{x \in U | [x]_R \cap X \neq \emptyset\} (\emptyset \text{ 为空集})$$

分别称 $R_-(X)$ 与 $R_+(X)$ 为 X 的 R 下近似和 X 的 R 上近似。 $BN_R(X) = R_+(X) - R_-(X)$ 称为 X 的 R 边界。

令 P 和 Q 为 U 中的等价关系, Q 的 P 正域(Positive Region)记为 $\text{Pos}_P(Q)$,

$$\text{Pos}_P(Q) = \bigcup_{(X \subseteq U/Q)} P_-(X)$$

指对于 U/P 的分类, U/Q 的正域是论域中所有通过分类 U/P 表达的知识能够确定地划入 U/Q 类的对象的集合。

一个知识系统(或称为信息系统)就是一个关系系统 $K = (U, R)$, 其中 U 为论域, R 是 U 上的一个等价关系族。 $\cap R$ (所有等价关系的交)也是一个等价关系, 记作 $\text{IND}(R)$ 。

定义1 令 R 为一等价关系族, 且 $r \in R$, 当 $\text{IND}(R) = \text{IND}(R - \{r\})$ 时, 称 r 为 R 中可省略的, 否则 r 为 R 中不可省略的。当对于任一 $r \in R$, 若 r 为 R 中不可省略的, 则 R 为独立的。

定义2 设 R 为一等价关系族, 且 $r \in R$, 如果 $P = R - \{r\}$ 是独立的, 则 P 是 R 中的一个约简。

知识系统 $K = (U, R)$ 以属性表示可写为 $S = (U, A)$, 又设 $C, D \subseteq A$ 是属性集 A 的两个子集, C 和 D 分别为 A 的条件属性和决策属性, V 为 A 的值域, $f: U \times A \rightarrow V$ 是一个信息函数, 则 S 可写为 $T = (U, C, D, V, f)$ 。对应于 U 中每一个元素 x , 都有相应的 f 中的一个 f_i 与之对应。我们称 f_i 为规

则。规则约简即是约去知识系统中的冗余规则。

设条件属性 C 有 N 个变量, 则 U 中元素 x_i 的条件属性值构成一 N 元向量, 称为条件属性向量, 记为 v_i 。

定义3 条件属性向量 v_i 与 v_j 的异或运算(记做 $v_i \oplus v_j$)是指两向量经运算后得一 N 元 $(0, 1)$ 向量, 两向量中对应位的元若相同, 则运算结果在相应位的值为0, 否则为1。

定义4 两条件属性向量 u 与 v 的与运算(记做 $u \cap v$)是指两向量经运算后得一 N 元 $(0, 1)$ 向量, 两向量中对应位的元若相同, 则运算结果在相应位的值为1, 否则为0。

定义5 两 N 元 $(0, 1)$ 向量 u 与 v 的与运算(记做 $u \cap v$)是指两向量经运算后得一 N 元 $(0, 1)$ 向量, 两向量中对应的元若都是1, 则运算结果在相应位的值为1, 否则为0。

3 规则约简

在知识系统中, 对于同一条件属性若有不同的决策属性, 则称该系统为不一致或不相容的。本文只考虑具有一致性的知识系统中的规则约简。

首先, 对于静态的知识系统根据粗集理论我们给出一个规则约简算法(算法A):

设信息系统 T 中条件属性有 N 列; 决策属性有 M 个等价类 $D_k (k=1, 2, \dots, M)$, 每个等价类含有 S_k 个元素。

步骤(一)

For $L=1$ to $M-1$

$$k = \sum_{i=1}^L S_{i-1} + 1 \quad (S_0 = 0)$$

For $i=k+1$ to $k+S_L$

For $j=1$ to $N-1$

在 $L+1$ 到 M 子块中寻找与 $N-1$ 阶子元 U_{ij} 相同的 $N-1$ 阶子元。若有相同的 $N-1$ 阶子元如 U_{ij} , 则同时标记属性中的元素 U_{ij} 与 U_{ij} 为 T 。否则, 检查 U_{ij} 是否已被标记为 T , 如果是则不变, 否则此时标记为 F 。

Next j

Next i

Next L

标记矩阵为 B , 其元素为相应条件属性 C 的标记: 当 U_{ij} 为 T 时, $B_{ij} = 1$, 当 U_{ij} 为 F 时, $B_{ij} = 0$ 。

步骤(二)

For $L=1$ to $M-1$

$$k = \sum_{i=1}^L S_{i-1} + 1 \quad (S_0 = 0)$$

For $i=k+1$ to $k+S_L-1$

如果第 i 行全为0则 $i=i+1$

For $j=i+1$ to $k+S_L$

若

$$\sum_{k=1}^N (B_{ik} + B_{jk}) \times (U_{ik} - U_{jk})^2 = 0$$

闫德勤 博士, 副教授, 主要研究领域为模式识别、数据挖掘。迟忠先 教授, 博导, 主要研究领域为知识发现、数据仓库、数据挖掘等。

则量 B 的第 j 行元素全为 0。

Next j
Next i
Next L

步骤(三)

与 B 矩阵中非 0 行对应的 C 属性中的行为不可约简行, 相应的规则为不可约规则。每个等价类 $D_k (k=1, 2, \dots, M)$ 对应的不可约规则得出不可约规则向量集合 B_k 。

对动态知识系统而言, 新增加记录(规则)后不仅已得到的不可约规则受到影响而且已经约去的规则也可能成为不可约规则。记新增加记录在论域 U 中的元素为 x , 规则为 f , 则相应的条件属性向量为 v 。对动态知识系统的规则约简我们提出如下算法(算法 B):

步骤(一)

若规则 f 的决策属性不属于已有决策属性 $D_k (k=1, 2, \dots, M)$ 中任何一个, 则把向量 v 列为一新向量集合 B_{M+1} , 程序结束。否则:

假定 f 的决策属性属于已有决策属性 D_i ,

赋值 w 为 N 元 0 向量

For $j=1$ to $M (j \neq i)$

For $s=1$ to S_j

$T = v \oplus jv_s$ (jv_s 表示 D_j 属性块中的条件属性向量)

If $T \cap (jv_s \cap B_j) = \{0\}$

($\{0\}$ 表示 0 向量) Then

把 jv_s 加入到 B_j 中

Next s

Next j

步骤(二)

If $w \cap (v \cap B_i) \supseteq w$ Then

把 v 加入到 B_i 中

程序结束。

定理 1 由算法 B 得到的约简规则构成知识系统 T 的最

小约简。

证明: 由算法 B 知, 所得到的约简规则构成动态知识系统 T 的等价系统中的核集。

例 给定一知识系统如表 1, 其中 a, b, c, d 为条件属性, e 为决策属性。使用算法 A 得到约简结果为 $B_1 = \{f_1\}, B_2 = \{f_3, f_4\}, B_3 = \{f_5\}$ 。若新增加一规则 $x_8 = \{1, 0, 0, 0, 1\}$, 使用算法 B 得到的新的约简结果同原来的结果。再新增加一规则 $x_9 = \{1, 1, 0, 0, 0\}$, 使用算法 B 得到的新的约简结果为原来的结果。再新增加一规则 $x_{10} = \{2, 0, 0, 1, 2\}$, 使用算法 B 得到的新的约简结果为 $B_1 = \{f_1\}, B_2 = \{f_3, f_4\}, B_3 = \{f_5, f_{10}\}$ 。

假设原知识系统如表 1 中没有信息, $x_1, x_2, \dots, x_{10}$ 为不断加入的新记录, 使用算法 B 进行规则约简后得到的结果为 $B_1 = \{f_1\}, B_2 = \{f_3, f_4\}, B_3 = \{f_5, f_{10}\}$

表 1 知识系统表

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

(下转第 78 页)

2003 年全国理论计算机科学学术年会 征文通知

由中国计算机学会理论计算机科学专业委员会主办, 青岛大学信息工程学院、青岛大学海尔软件学院承办的“2003 年全国理论计算机科学学术年会”将于 2003 年 8 月在山东青岛召开。会议录用论文将收录在正式出版的论文集中, 欢迎大家积极投稿。现将有关征文要求通知如下:

1. 应征论文应未在其他刊物或学术会议上正式发表过。特别欢迎有创见的论文和有应用前景的论文。

2. 稿件要求用计算机打印, 格式为 38 行 $\times$ 38 字, 字体为 5 号宋体。稿件中的图形要求画得工整、清晰、紧凑, 尺寸要尽量小; 图中字体要求为 6 号宋体。稿件正文不超过六千字。标题、作者姓名、作者单位、摘要、关键词采用中英文间隔行文。稿件各部分依次为: 一、引言; 二、...; 最后是结束语。附录放在参考文献之后; 参考文献限已公开发表的, 文中最好不要出现文献序号。参考文献的格式为:

序号 作者·书名·出版社所在地: 出版社名, 出版年代

序号 作者·论文名·出处, 年代·卷号(期号): 起迄页码

务必附上第一作者简历(姓名、性别、出生年月、职称、学位、研究方向等)、通信地址和联系电话, 并注明论文所属领域。请提供打印稿和电子稿各一份, 来稿一律不退, 请自留底稿。

3. 征文范围

程序理论(程序逻辑、程序正确性验证、形式开发方法等); 计算理论(算法设计与分析、复杂性理论、可计算性理论等); 语言理论(形式语言理论、自动机理论、形式语义学、计算语言学等); 人工智能(知识工程、机器学习、模式识别、机器人等); 逻辑基础(数理逻辑、多值逻辑、模糊逻辑、模态逻辑、直觉主义逻辑、组合逻辑等); 数据理论(演绎数据库、关系数据库、面向对象数据库等); 计算机数学(符号计算、数学定理证明、计算几何等); 并行算法(分布式并行算法、大规模并行算法、演化算法等)。

4. 征文截止日期: 2003 年 4 月 1 日

5. 论文投寄地址: (266071) 山东青岛大学信息工程学院 郭振波收 联系电话: 0532-5953151(郭振波) 0532-5952834(王彬、李涛) 电子信箱: gzb@qdu.edu.cn 或 gzb@qingdaonews.com

表8 项目数量及相应阈值表

	数量	通过阈值	输出阈值
海尔台式机	2	1	2
纯平彩电	1	2	2
海信纯平彩电	1	1	2
佳能激光打印机	2	1	2
台式机	2	1	2
配件	3	2	3
南岸	3	2	2
江北	2	2	2
W1	3	2	2
W2	2	2	2

五、实验和性能评价

所有实验都是在800MHz, Intel Celeron256M 内存的台式机上进行, 程序用 Borland Dephi5.0 编写。事务总数据为1980条(随机产生), 平均事务长度为27, 平均最大频繁序列长度为12。实验结果如图3所示。E-FP 方法与 ML-T2L1 相比较有以下优点:

·只需要扫描数据库两次, 而 ML-T2L1 和 Apriori 却需要根据序列的长度确定。如果序列长度为 n , 则需要扫描 n 次数据库, 从而大大地节约了运行时间。

·E-FP 方法能够产生跨层的频繁序列, 而 ML-T2L1 算法不能产生。

·E-FP 方法可以处理多层高维的序列, 而 ML-T2L1 只能处理多层序列。

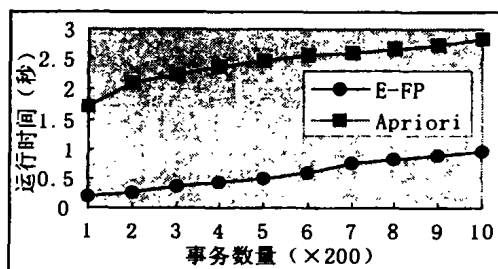


图3 实验结果比较

(上接第49页)

结论 本文以粗集理论为基础给出了一种有效可行的对动态知识信息系统进行规则约简的算法。实验结果证明该算法是灵活可行有效的。由于该算法易于编程实现, 适于处理大型知识系统。进一步的工作将是把该方法用于知识发现、数据挖掘的研究与应用方面。

参考文献

- 1 Pawlk Z. Rough Sets: Theoretical Aspects of Reasoning about

小结 本文针对多层高维频繁序列挖掘给出了一种新方法。由于该方法采用了 FP-growth 的结构, 只需扫描三次数据库, 从而大大节约了运行时间。同时我们又对 FP-growth 的统一支持度进行了改进, 从而避免了丢失一些有用信息。实验结果表明, 此方法优于以往的方法。下一步我们将考虑 FP 树的更新问题。

参考文献

- 1 Agarwal R, et al. A tree projection algorithm for generation of frequent itemsets. In: Journal of Parallel and Distributed Computing, 2000
- 2 Agarwal R, Srikant S. Mining sequential patterns. In: Proc. 1995 Int. Conf. Data Engineering (ICDE'95), Taipei, Taiwan, March 1995. 3~14
- 3 Han J, Fu Y. Discovery of multi-level association rules from large database. In: Proc. 1995 Int. Conf. Very large DataBases (VLDB'95), Zurich, Switzerland, Sep. 1995. 420~431
- 4 Han J, Pei J, Yin Y. Mining frequents without Candidate Generation. In: Proc. 2000 ACM-SIGMOD Int. Conf. On Mangement of Data (SIGMOD'00), dallas, TX, May 2000. 1~12
- 5 Liu B, Hsu W, Ma Y. Mining association rules with multiple minimum supports. In: Proc. 1999 Int. Conf. Knowledge Discovery and Data Mining (KDD'99), San diego, CA, Aug. 1999. 337~341
- 6 Pei J, Han J, Mao R. CLOSET: An Efficient Algorithm for Mining Frequent Closed Itemsets. In: Proc. 2000 ACM-SIGMOD Int. Workshop Data Mining and Knowledge Discovery (DMKD'00), Dallas, TX, May 2000. 11~20
- 7 Pei J, Han J, et al. PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern BibliographyGrowth. In: Proc. 2001 Int. Conf. Data Engineering (ICDE'01), Heidelberg Germany, April 2001
- 8 Srikant R, Agrawal R. Mining generalized association rules. In: Proc. 1995 Int. Conf. Very Large Data Bases (VLDB'95), Zurich, Switzerland, Sept. 1995. 407~419
- 9 Wang K, He Y, Han J. Mining Frequent Itemsets Using Support Constraints. In: Proc. 2000 Int. Conf. Very Largo Data Bases (VLDB'00), Cairo, Egypt, Sep. 2000. 43~52

Data. Dordrecht: Kluwer Publishers, 1991

- 2 Pawlk Z, et al. Rough Sets. Communications of the ACM, 1995, 38(11):89~95
- 3 Pawlk Z. Rough set theory and its application to data analysis. Cybernetics and Systems, 1998, 29(9):661~668
- 4 Ziarko W. Introduction to the special issue on rough sets and knowledge discovery. International Journal of Computational Intelligence, 1995, 11(2):223~226
- 5 曾黄麟. 粗集理论及其应用. 重庆: 重庆大学出版社, 1996
- 6 刘清. Rough 集及 Rough 推理. 北京: 科学出版社, 2001