Computer Science and Application 计算机科学与应用, 2011, 1, 24-27
http://dx.doi.org/10.12677/csa.2011.11006 Published Online June 2011 (http://www.hanspub.org/journal/csa/)

Algorithm of Relative Reduction in Incomplete Decision Table

Rui Zhang
Computer Center, institute of information technology, Yangzhou University, Yangzhou
Email: zhangrui@yzu.edu.cn
Received: Apr. 25th, 2011; revised: May 4th, 2011; accepted: May 13th, 2011.

Abstract: Many theories used for uncertainty problems are strongly complementary to each other because they focus on different aspects. This study proposes a new relative reduction algorithm by combing two theories which are frequently used to process incomplete information. It has been proved by example that the result is reliable.

Keywords: Incomplete Decision Talbe; Rough Set Theory (Rst); D-S Evidence Theory; Relative Reduction

不完备决策表的相对约简算法

张　睿
扬州大学信息工程学院计算中心，扬州
Email: zhangrui@yzu.edu.cn
收稿日期：2011 年 4 月 25 日；修回日期：2011 年 5 月 4 日；录用日期：2011 年 5 月 13 日

摘　要：诸多处理不确定性问题的理论由于各自的侧重点不同，因而相互之间具有很强的互补性。本文研究了将两种常用于处理不完备信息的理论结合来给出一种新的相对约简算法，并给出实例计算结果，证明了算法的可行性。

关键词：不完备决策表；粗糙集理论；证据理论；相对约简

1. 引言

信息系统是一个具有对象和属性关系的数据库，这种数据库隐含着知识的对象与属性之间的关系，最终表达的知识模式是用属性来表达的，具有明确的直观含义，可以被理解。当今的信息系统以电子计算机和现代通信技术为基本信息处理手段，运用数学的方法，为管理决策提供信息服务。

科学技术的进步使得信息技术的发展十分迅速，应用范围也在不断扩展。近十年来信息呈现爆炸式增长。如此海量的数据，如何找出其内在联系？如何从中提取出重要的内容，忽略错误的数据造成的影响并且缩减冗余信息？同时，现实世界中客观事物和现象往往是不确定的，或具有不完备性，而人们主观的认识领域的信息和知识大多也是不精确的，这就要求在知识的表示和处理时能够反映这种不确定性。在此背景下，数据挖掘和数据库知识发现成为了新的研究领域。

在 DM 和 KDD 的诸多理论以及方法中，模糊集、粗糙集、神经网络、遗传算法、证据理论等，都非常有效，其理论得到了不断的发展完善，应用也得到了很好的实践和推广。每种理论各自有自己的优缺点，因而可以进行互补性研究，充分利用它们的长处[1-3]。

在不完备信息系统约简方面，最常见的是利用辨识矩阵和布尔推理方法，Kryszkieewicz 给出了不完备决策表的知识约简方法，并且提出了一种获取最优规

则的方法。Leung 等提出了基于极大相容块技术的方法。国内学者也提出了许多约简算法[4-10]。

尽管不完备决策表缺少信息，但仍然蕴涵一些有用的知识，这些知识对于不完备信息下的决策，是很有意义的。本文将证据理论的概念融入粗糙集，给出了一种新的计算不完备决策表相对约简的算法，并用实例进行了验证，分析了算法的复杂度。

2. 不完备信息的相关理论

2.1. 不完备信息系统概述

定义 1[1] 设 $S = (U, AT)$是信息系统。其中 U 是对象的非空有限集合，AT 是属性的非空有限集合。对每个属性 $a \in AT$ 有 a：$U \to Va$。其中 Va 称为 a 的值域。如果至少有一个属性 $a \in AT$ 使得 Va 含有空值，则称 S 为一个不完备信息系统，并用“*”表示空值。

对于不完备信息系统中的未知信息，存在两种可能性：①该值是存在的但它被遗漏了；②该对象在其他属性值确定的情况下为空值的属性没有办法赋值。

定义 2[1] 不完备决策表(DT)是一个不完备信息系统 $DT = (U, AT \cup \{d\})$。其中属性 $d(d \notin AT, * \notin V_d)$ 称为决策属性，是完备的；AT 中的属性称为条件属性。

2.2. 不完备信息下的粗糙集理论

粗糙集(RST)[1-3]理论由波兰数学家 Z. Pawlak[2]于 1982 年提出。这是一种研究信息系统中不精确，不确定或模糊信息的理论。

与完备信息系统的不可分关系相对应，本文给出不完备信息系统的相似关系定义。

定义 3[1] 令 $A \subseteq AT$，定义相似关系如下：

$$SIM(A) = \{(x, y) \in U \times U \mid \forall a \in A, a(x) = a(y) \vee a(x) = * \vee a(y) = *\} \quad (1)$$

相似关系有如下性质：

性质 1 $SIM(A)$是一个相容关系：

$$SIM(A) = \bigcap_{a \in A} SIM(\{a\})$$

令 $S_A(x)$表示对象集 $\{y \in U \mid (x, y) \in SIM(A)\}$。对于 A 而言，$S_A(x)$是与 x 可能可区分的对象最大集合。令 $U / SIM(A) = \{S_A(x) \mid x \in U\}$ 表示分类，$U / SIM(A)$ 中的任何元素称为相容类。$U / SIM(A)$ 中的相容类一般不构成 U 的划分，它们构成 U 的覆盖。$\cup U / SIM(A) = U$。

令 $D_A(x)$表示对象集 $\{y \in U \mid (x, y) \notin SIM(A)\}$。对于 A 而言，$D_A(x)$是与 x 可能不可分的对象的最大集合。

对任意 $x \in U$，$S_A(x) \cap D_A(x) = \Phi$ 且 $S_A(x) \cup D_A(x) = U$。

定义 4[1] 不完备决策表中的广义决策函数 $\partial_A : U \to P(V_d), A \subseteq AT$ 定义为：

$$\partial_A(x) = \{i \mid i = d(y), y \in S_A(x)\} \quad (2)$$

其中，$P(V_d)$表示 V_d 的幂集。

定义 5[1] 形式上，条件属性集 AT 是不完备决策表的一个相对约简，若 $\partial_A = \partial_{AT}, A \subseteq AT$ 且 $\forall B \subset A \Rightarrow \partial_B \neq \partial_{AT}$。

2.3. 不完备信息下的粗糙集理论

证据理论[11-13]工作在封闭世界假设(CWA)的基础上。设 θ 是一有限论域，通常称为辨别框架，由一组我们需要研究的互斥且穷举的命题组成。θ 的一个子集，即 2^θ 中的元素，可以理解为一个命题。

定义 6[12] 对应于一个辨别框架 θ 的基本概率赋值函数，或称为 m 函数，定义如下：

$m : 2^\theta \to [0, 1]$

m 函数有如下性质：

(1) $m(\phi) = 0$ (2) $\sum m(X) = 1$，$X \subseteq \theta$

ϕ表示空命题，因而对其信任度为 0。反过来，对于整个论域的信任度为 1。信任度只分配到那些证据支持的命题上。对于任何 X，$m(X)$表示证据支持命题 X 发生的程度，而不支持任何 X 的真子集。

在 m 函数中，$m(\theta)$表示 m 这个概率没有赋予任何一个子集上，没有提供任何信息，因而用于对未知信息的表示。在不完备决策表中，引申为对空值的表示。

定义 7[12] 可以由决策表中的一个或多个属性对应证据理论中的一个辨别框架 θ，属性的值域对应 θ 中的元素。对于属性 a，其 m 函数定义为：

$$m(a, v) = \frac{card(X)}{card(U)} \quad (3)$$

$$m(a, *) = \frac{card(U - X)}{card(U)} \quad (4)$$

$$v \in V_a, X = \{x \in U \mid c(x) = v, v \neq *\}$$

3. 不完备决策表的相对约简算法

如果一个属性从条件属性集中去掉之后并不改变决策表的广义决策函数的分配，那么该属性对于决策分类是不重要的(冗余的)，因而可以通过逐个去除冗余属性而获得相对约简。

在决策表中，如果某一个属性缺失的值越多，那么它对决策表的决策能力的影响也就越弱，因而在约简时可以考虑先把这个属性排除。基于这样的思想，提出了一种新的关于不完备决策表的相对约简算法。

算法说明

输入：一个不完备决策表

输出：该表的一个相对约简：属性集 B

1) 为每个属性建立辨别框架，并求出相应的 *m* 函数；

2) 将条件属性集 AT 中的每个属性按其未知值(m(*))从大到小进行排序，并保存到属性数组 ATTMASS 中：

```
ATTMASS (1) = AT (1)
FOR i = 2 TO card (AT)
    FOR j = 1 TO i–1
      IF AT (I) > ATTMASS (J)
          FOR K= I – 1 TO J STEP – 1
             ATTMASS (K + 1) = ATTMASS (K)
          ENDFOR
          EXIT
      ENDIF
    ENDFOR
ATTMASS (J) = AT (I)
ENDFOR
```

3) 令 B = AT，逐个去除冗余属性：

```
FOR   I = 1 to card(AT)
IF  ∂B-ATTMASS(i) = ∂AT
B = B – ATTMASS (i)
ELSE
EXIT
ENDIF
ENDFOR
```

4) 算法结束，B 即为所求约简。

算法时间复杂度分析：设 card(AT) = n，card(U) = m，步骤 1)的复杂度为 O(n)，步骤 2)在最坏情况下的复杂度为 $O(m^2)$，步骤 3)中，计算广义决策函数的复杂度为 $O(n^2m^2)$，FOR 循环在最坏情况下的复杂度为 O(m)。故整个算法的复杂度为 $O(n^2m^2)$。

可以看出，在碰到第一个不应去除的属性时，算法就停止，因此该算法所求出的约简并不一定是最小约简；若要求最小约简，仍需遍历所有属性，当然这样需花费更长的运行时间。

4. 实例分析

表 1 中的数据利用计算机性能参数来分析计算机性价比。其中 AT = {c：CPU 主频，m：内存容量，h：硬盘容量，w：质量，p：价格}，d = {v：性价比}。表 2 为表 1 中属性的广义决策函数。

计算表 1 所示的不完备决策表的相对约简：

c 的 *m* 函数：*m*(h) = 5/6, *m*(l) = 1/6, *m*(*) = 0

m 的 *m* 函数：*m*(b) = 1/3, *m*(*) = 2/3

h 的 *m* 函数：*m*(b) = 1/3, *m*(s) = 1/3, *m* (*) = 1/3

w 的 *m* 函数：*m*(l) = 1/3, *m*(*) = 2/3

p 的 *m* 函数：*m*(h) = 1/3, *m*(l) = 1/2, *m*(*) = 1/6

U/ ind(d) = {$X_{优}$, $X_{良}$, $X_{差}$}

$X_{优}$ = {3}，$X_{良}$ = {1, 4, 5, 6}，$X_{差}$ = {2}

U/SIM(AT) = {{1}, {2}, {3, 4, 5}, {3, 4, 6}, {3, 5}, {4, 6}}

U/SIM(AT – m) = {{1}, {2}, {3, 4, 5}, {3, 4, 6}, {3, 5}, {4, 6}}

U/SIM(AT – m – w) = {{1}, {2}, {3, 4, 5}, {3, 4, 6}, {3, 5}, {4, 6}}

U/SIM(AT – m – w – h) = {{1, 4, 6}, {2}, {3, 4, 5}, {1, 3, 4, 5, 6}, {3, 4, 5}, {1, 4, 6}}

U/SIM(c) = {{1, 3, 4, 5, 6}, {2}, {1, 3, 4, 5, 6}, {1, 3, 4, 5, 6}, {1, 3, 4, 5, 6}, {1, 3, 4, 5, 6}}

ATTMASS = {m, w, h, p, c}

由此得到相对约简为 B = {c, p}。

Table 1. Incomplete decision table
表 1. 不完备决策表

U	*c*	*m*	*h*	*w*	*p*	*v*
1	高	大	大	轻	高	良
2	低	*	*	*	低	差
3	高	*	*	轻	低	优
4	高	大	小	*	*	良
5	高	*	大	*	低	良
6	高	*	小	*	高	良

Table 2. Generalized decision function
表 2. 广义决策函数

U	δ_{AT}	δ_{AT-m}	δ_{AT-m-w}	δ_{c+p}	δ_c
1	良	良	良	良	优，良
2	差	差	差	差	差
3	优，良	优，良	优，良	优，良	优，良
4	优，良	优，良	优，良	优，良	优，良
5	优，良	优，良	优，良	优，良	优，良
6	良	良	良	良	优，良

5. 致谢

在此谨向关心、指导我教学研究的领导和同事们致以诚挚的谢意！

参考文献 (References)

[1] 张文修, 吴伟志, 梁吉业等. 粗糙集理论与方法[M]. 北京: 科学出版社, 2001: 206- 213.
[2] Z. Pawlak. Rough sets: Theoretical aspects of reasoning about data. Dordrecht: Kluwer Academic publishers, 1991: 89-95.
[3] S. S. Anand, D. A. Bell, and J. G. Hughes. EDM: A general framework for data mining based on evidence theory. Data & Knowledge Engineering, 1996, 18(3): 189-223.
[4] 曾黄麟. 粗集理论及其应用[M]. 重庆: 重庆大学出版社, 1996: 55-69.
[5] 黄鲲, 陈森发, 周振国等. 基于粗集理论和证据理论的多源信息融合方法[J]. 信息与控制, 2004, 33(4): 422-425.
[6] 孙国梓, 郁鼎文, 吴志军. 个性化配置器的粗糙集方法研究[J]. 计算机集成制造系统, 2005, 11(2): 296-299.
[7] 冯朝一, 梁家荣, 黄柳萍等. 基于集合覆盖的不完备信息系统属性约简方法[J]. 计算机应用, 2006, 26(11): 2664-2666.
[8] 蔡正琦, 林和, 孔令旺等. 一种基于变精度区分矩阵的不完备信息系统属性约简[J]. 甘肃科学学报, 2006, 18(4): 71-74.
[9] J. Y. Liang, Z. B. Xu. The algorithm on knowledge reduction in incomplete information systems. International Journal of Uncertainty, Fuzziness and Knowledge-based Systems, 2002, 10(1): 952-1103.
[10] A. P. Dempster. Upper and lower probability inferences based on a sample from a finite univariate population. Biometrika, 1967, 54(3): 515-528.
[11] R. R. Yager. Minimization of regret decision making with Dempster-Shafer uncertainty. 2004 IEEE Fuzzy Systems Conference Proceedings. Budapest Hungary, 2004, 511-515.
[12] 程玉胜, 张佑生, 胡学钢. 不完备决策系统中规则提取的快速矩阵算法[J]. 系统仿真学报, 2008, 20(15): 4036-4040.
[13] 路松峰, 刘芳, 胡波. 一种基于属性依赖的属性约简算法[J]. 华中科技大学学报, 2008, 36(2): 39-41.

