

# 基于容差关系的不完备信息系统的属性约简

颜家凯, 范敏, 刘文奇, 叶荣荣

(昆明理工大学理学院, 云南昆明 650500)

**摘要:**粗糙集理论是一种处理不确定性知识的有效工具,属性约简是其核心内容之一,然而对于属性值有缺省的不完备信息系统,基于等价关系的经典粗糙集理论已经不再适用。由于容差关系下的不完备信息系统的属性约简的定义与经典粗糙集的属性约简定义相似,可以用容差关系对粗糙集理论进行扩充。文中通过定义容差关系下的可辨识矩阵,运用可辨识方法,得到了一种属性约简算法;接着分析了算法的不足之处,并且在此基础上提出了增加约简效率的改进型算法;最后通过一个数值例子,说明了该算法是合理的和有效的。

**关键词:**属性约简;不完备信息系统;容差关系;可辨识矩阵

中图分类号:TP302.1

文献标识码:A

文章编号:1673-629X(2014)01-0102-03

doi:10.3969/j.issn.1673-629X.2014.01.026

## Attributes Reduction of Incomplete Information System Based on Tolerance Relation

YAN Jia-kai, FAN Min, LIU Wen-qi, YE Rong-rong

(School of Science, Kunming University of Science and Technology, Kunming 650500, China)

**Abstract:** Rough set theory is a kind of effective tool for dealing with uncertainty knowledge. Attribute reduction is one of the most important content. Nevertheless, the classical rough set theory based on equivalence relation has not been applied for the incomplete information system which some attribute is the default value. The definition that attributes reduction of incomplete information system is similar to the classical rough set because of the tolerance relation. Can expand the classical rough set theory with tolerance relation. In this paper, by defining the discernibility matrix under the tolerance relation, obtain an attribute reduction algorithm through discernibility method. Then analyze the deficiency of the algorithm and put forward a kind of modified algorithm that can improve efficiency of the reduction. At last, prove the reasonableness and validity of the algorithm through a numerical example.

**Key words:** attribute reduction; incomplete information system; tolerance relation; discernibility matrix

## 0 引言

波兰数学家 Z. Pawlak 提出的粗糙集理论<sup>[1]</sup>是一种处理模糊和不精确问题的数学工具<sup>[2]</sup>。它以完备信息系统为研究对象,以等价关系为基础,属性约简是其核心内容之一。关于属性约简问题,前人提出了一些有效的算法,如基于属性重要性的启发式约简算法<sup>[3]</sup>、基于信息熵的约简算法<sup>[4-6]</sup>、基于可辨识矩阵和可辨识函数的约简算法<sup>[7-8]</sup>等。但是,在不完备信息系统中,一般的约简算法处理噪声数据是很难的,难以保证有较好的鲁棒性<sup>[9]</sup>。

对于不完备信息系统,传统的做法是采用

ROUSTIDA 算法<sup>[10]</sup>先进行完备化处理,然后再进行约简处理。但是,完备化处理的过程是用主观估计值来代替未知值,或多或少都会改变原始的信息系统,得到的结果不一定和客观事实相符。容差关系是对经典粗糙集理论进行扩充的一种方法,根据容差关系,不完备信息系统的属性约简的定义与传统 Rough 集的属性约简定义相似<sup>[10-11]</sup>,运用容差关系对不完备信息系统进行处理时得到的结果跟客观事实更加符合。

## 1 基本概念

定义1:设  $U$  为一个论域, $P, Q$  是定义在  $U$  上的两

收稿日期:2013-03-21

修回日期:2013-06-22

网络出版时间:2013-11-1

基金项目:科技部科技型中小企业技术创新基金项目(11C26215305906)

作者简介:颜家凯(1990-),男,硕士研究生,研究方向为粗糙集理论、直觉模糊集;范敏,副教授,硕士生导师,研究方向为数据挖掘、粗糙集理论、决策分析。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20131112.1639.033.html>

个等价关系簇,若 $P$ 中所有属性都是相对于 $Q$ 必要的,则称 $P$ 为相对于 $Q$ 独立的。

定义2:相对约简。设 $U$ 为一个论域, $P, Q$ 是定义在 $U$ 上的两个等价关系簇,若 $P$ 的 $Q$ 独立子集 $S \subset P$ 有 $\text{POS}_S(Q) = \text{POS}_P(Q)$ ,则称 $S$ 为 $P$ 的 $Q$ 约简。

定义3:不完备信息系统。给定信息系统 $S = \langle U, C, V, f \rangle$ ,其中, $U$ 是对象的非空有限集合; $C$ 是属性的非空有限集合,对于每个 $c_j \in C$ 有 $c_j: U \rightarrow V_{c_j}$ , $V_{c_j}$ 称为 $c_j$ 的值域。如果至少存在一个属性 $c_j \in C$ 使得 $V_{c_j}$ 含有空值(“\*”表示),则称 $S$ 为一个不完备信息系统,否则它是完备的。

定义4:容差关系。对于不完备信息系统 $S = \langle U, C, V, f \rangle$ ,对于具有遗漏值属性的属性子集 $B \subseteq C$ ,记遗漏值为“\*”,容差关系 $T$ 的定义如下:

$$T = \{ (x, y) \mid x \in U \wedge y \in U \wedge \forall c_j (c_j \in B \Rightarrow (c_j(x) = c_j(y) \vee c_j(x) = * \vee c_j(y) = *)) \}$$

定义5:可辨识矩阵。令决策表系统为 $S = \langle U, A, V, f \rangle$ , $A = C \cup D$ 是属性集合,子集 $C = \{a_i \mid i = 1, 2, \dots, m\}$ 和 $D = \{d\}$ 分别称为条件属性集和决策属性集, $U = \{x_1, x_2, \dots, x_n\}$ 是论域, $a_i(x_j)$ 是对象 $x_j$ 在属性 $a_i$ 上的取值。 $C_{ij}$ 表示可辨识矩阵中第 $i$ 行第 $j$ 列的元素,则可辨识矩阵 $M(S) = (C_{ij})_{n \times n}$ 可以定义为:

$$C_{ij} = \begin{cases} \{a_k \mid a_k \in C \wedge a_k(x_i) \neq a_k(x_j)\}, & d(x_i) \neq d(x_j) \\ \Phi, & d(x_i) = d(x_j) \end{cases}$$

其中, $i, j = 1, 2, \dots, n$ 。

根据定义4和5,可以定义基于容差关系下的可辨识矩阵。

定义6:设 $T = (U, C \cup D, V, f)$ 是一个不完备决策表,则基于容差关系的可辨识矩阵可定义为:

$$M(T) = (C_{ij})_{n \times n}, n = \text{card}(U)$$

$$C_{ij} = \begin{cases} \Phi, & d(x_i) = d(x_j) \\ \{a \in C \mid a(x_i) \neq a(x_j) \wedge a(x_i) \neq * \wedge a(x_j) \neq *\}, & d(x_i) \neq d(x_j), i, j = 1, 2, \dots, n \end{cases}$$

定义7:可辨识函数<sup>[12]</sup>。对于定义6给定的可辨识矩阵 $M(T)$ ,定义可辨识函数 $\Delta = \bigwedge (\bigvee C_{ij})$ 。

可辨识函数 $\Delta$ 的极小析取范式的所有合取式是属性集 $A$ 的所有约简,换句话说,约简是满足能区别由整个属性集区别的所有对象的属性的极小子集。

根据可辨识矩阵的定义,可以发现可辨识矩阵中的元素便是由能够区分两个对象的属性构成。如果某元素只有一个属性,则说明这两个对象只有靠这个属性来唯一区分,因此这个属性是不可删除的,是核属性,由此可以得到核属性的定义。

定义8:给定可辨识矩阵 $M(T)$ ,对于 $\forall C_{ij} \in M(T)$ ,如果有 $\text{card}(C_{ij}) = 1$ ,则 $C_{ij}$ 中所包含的属性就称为核属性,其中 $\text{card}(C_{ij})$ 表示 $C_{ij}$ 的基数。

定义9:设 $a, b$ 为可辨识矩阵 $M(T)$ 中的两个元素,若有 $a \subseteq b$ 则称 $b$ 为 $a$ 的重复元素。

性质1:在可辨识矩阵 $M(T)$ 中,将重复元素置空后的可辨识矩阵所生成的属性约简不变。

证明:对于给定的可辨识矩阵 $M(T)$ , $a, b \in M(T)$ ,且有 $a \subseteq b$ 。可辨识函数 $\Delta = \bigwedge (\bigvee C_{ij}) = \bigwedge (\bigvee_{C_{ij}-a-b} C_{ij}) \wedge a \wedge b$ ,由逻辑运算的基本性质可知: $a \wedge b = a$ ,故有 $\bigwedge (\bigvee_{C_{ij}-a-b} C_{ij}) \wedge a \wedge b = \bigwedge (\bigvee_{C_{ij}-a} C_{ij}) \wedge a$ ,所以令重复元素 $b = \Phi$ 不会影响可辨识函数的值,重复元素置空前或后的可辨识矩阵所生成属性约简是相同的。

## 2 属性约简

给定不完备决策表 $S = \langle U, C \cup D, V, f \rangle$ ,令 $\text{Redu}$ 是经过属性约简之后得到的条件属性集合, $C_{ij}$ 表示可辨识矩阵中的各元素, $a_k$ 表示各条件属性( $k = 1, 2, \dots, \text{card}(C)$ ), $L$ 为合取范式, $L'$ 为析取范式, $\text{Core}$ 为属性核。则属性约简算法步骤如下:

Input:  $S = \langle U, C \cup D, V, f \rangle$ ;

Output:  $\text{Redu}$ 。

Step1:  $\text{Redu} = \Phi$ , 计算决策表 $S$ 的可辨识矩阵 $M(T) = (C_{ij})_{n \times n}$ 。

Step2: 求出可辨识矩阵中的元素 $C_{ij}$ 的析取表达式 $L_{ij}$ 。

for( $i = 1$ ;  $i \leq n$ ;  $i++$ )

for( $j = 1$ ;  $j \leq n - i$ ;  $j++$ )

{

if( $C_{ij} \neq \Phi$ )

$L_{ij} = \bigvee_{a_k \in C_{ij}} a_k$ ;

}

Step3: 将所有的析取逻辑表达式 $L_{ij}$ 进行合取运算,得到一个合取范式 $L$ 。

$L = 1$

for( $i = 1$ ;  $i \leq n$ ;  $i++$ )

for( $j = 1$ ;  $j \leq n - i$ ;  $j++$ )

{

$L = L \wedge L_{ij}$ ;

}

Output:  $L$ ;

Step4: 将合取范式 $L$ 转换为析取范式的形式,得到 $L' = \bigvee_i L_i$ 。

Step5: 输出属性约简结果 $\text{Redu} = L_i$ 。

算法的不足之处:对可辨识函数进行逻辑运算时,

由于可辨识函数中有大量的重复元素,这样增加了逻辑运算中的计算量,降低了属性约简算法的效率。

针对上述算法的不足之处,基于重复元素的性质和核属性的定义提出了一种改进算法:

```
Step1: Redu = Φ, 计算决策表 S 的可辨识矩阵
M(T) = (Cij)n×n
Step2: 求出可辨识矩阵的属性核。
Core = Φ
for(i = 1; i <= n; i++)
for(j = 1; j <= n - i; j++)
{
if(Card(Cij) = 1)
Core = Core ∪ Cij;
}
Step3: 去除重复元素得到新的可辨识矩阵 M'(T)。
for(i = 1; i <= n; i++)
for(j = 1; j <= n - i; j++)
{
if(Cij ∩ Core ≠ Φ)
Cij = Φ
}
M'(T) = (Cij)n×n
Step4: 求出可辨识矩阵中的元素 Cij 的析取表达式 Lij。
for(i = 1; i <= n; i++)
for(j = 1; j <= n - i; j++)
{
if(Cij ≠ Φ)
Lij = ∪ak ∈ Cij ak
}
Step5: 将所有的析取逻辑表达式 Lij 进行合取运算, 得到一个合取范式 L。
L = 1
for(i = 1; i <= n; i++)
for(j = 1; j <= n - i; j++)
{
L = L ∧ Lij;
}
Step6: 将合取范式 L 转换为析取范式的形式, 得到 L' = ∪i Li。
Step7: 输出属性约简结果 Redu = Li ∪ Core。
```

3 实例分析

给定如表 1 所示的不完备决策表, 求其核和所有

的属性约简。

表 1 不完备决策表

U	条件属性				决策属性
	a	b	c	d	e
x <sub>1</sub>	1	*	2	0	1
x <sub>2</sub>	0	1	1	2	0
x <sub>3</sub>	0	1	0	1	1
x <sub>4</sub>	1	1	0	0	0
x <sub>5</sub>	*	0	0	1	0

根据决策表可以得到其可辨识矩阵为:

$$\begin{bmatrix} \Phi & & & & \\ acd & \Phi & & & \\ \Phi & cd & \Phi & & \\ c & \Phi & ad & \Phi & \\ cd & \Phi & b & \Phi & \Phi \end{bmatrix}$$

可辨识函数  $\Delta = (a \vee c \vee d) \wedge (c \vee d) \wedge c \wedge (a \vee d) \wedge (c \vee d) \wedge b$ 。

经过逻辑运算后可以得到  $\Delta = (a \wedge b \wedge c) \vee (d \wedge b \wedge c)$ 。

根据定义 8, 可以知道 b 和 c 是核属性。

故不完备信息系统的核  $\text{Core} = \{b, c\}$ ,  $\{a, b, c\}$  和  $\{b, c, d\}$  是它的约简。

按照改进算法可以得到可辨识矩阵为:

$$\begin{bmatrix} \Phi & & & & \\ \Phi & \Phi & & & \\ \Phi & \Phi & \Phi & & \\ \Phi & \Phi & ad & \Phi & \\ \Phi & \Phi & \Phi & \Phi & \Phi \end{bmatrix}$$

可辨识函数为  $\Delta = a \vee d$ 。

由前面的内容可以得到核  $\text{Core} = \{b, c\}$ 。

因此得到该不完备决策表的核  $\text{Core} = \{b, c\}$ ,  $\{a, b, c\}$  和  $\{b, c, d\}$  是它的约简。

4 结束语

文中运用容差关系下不完备信息系统的属性约简的定义与传统 Rough 集的属性约简定义相似的原理, 定义容差关系下的可辨识矩阵, 运用 Skowron 提出的可辨识方法对不完备信息系统进行约简, 取得了比较好的结果; 接着在原有基础上提出了增加约简效率的改进型算法, 得到的结果与原算法的结果一致, 证明了改进型算法的可行性。当然运用容差关系只是处理不完备信息系统方法中的一种, 还可以将直觉模糊集与可辨识矩阵的知识相结合, 用来对不完备信息系统进行约简, 这些都有待进一步的研究。

法优化的 LM-BP 模型”的测试误差。从表中可以看出,在该实例中多次 LM-BP 模型测试误差非常大,基本上属于失败的方法。经过遗传算法优化的 LM-BP 模型误差也比较大,但泛化能力能够得到大大提高。经过随机遗传算法优化的 LM-BP 模型比较容易找到合适的权阈值,测试效果比较理想,能够进一步提高其泛化能力。由于 LM-BP 网络的权阈值比较多,其初始权阈值都是随机产生,因此,采用不同方法测试时,每次结果都会不一样,这里只是取其中的一次测试数据。实验过程中也会出现不理想的情况,但是采用随机遗传算法的 LM-BP 网络更容易获得较理想的测试结果。

表 3 多种方法的测试误差对比 %

多次 LM-BP 模型	遗传算法优化的 LM-BP 模型	随机遗传算法优化的 LM-BP 模型
-10.238 0	3.198 9	2.236 9
25.383 0	4.066 9	-1.198 9
11.508 0	-10.191 0	0.531 1

3 结束语

LM-BP 网络具有全局收敛、速度快、拟合能力强的优点,采用 GA 优化 LM-BP 网络的初始权阈值,能克服 LM-BP 网络对初始权阈值敏感的缺点,能大大提高其泛化能力。初始权阈值的产生具有随机性,每次优化只局限在某次初始种群的范围进行,采用随机 GA 的方法能进一步提高 BP 网络的泛化能力。在实际应用中,为了取得更好的测试结果,可以适当提高优化的次数。

参考文献:

[1] 郭海如,冯 凯,邹 遵.基于 BP 网络的孝感学院未来数

(上接第 104 页)

参考文献:

[1] Pawlak Z. Rough sets[J]. International journal of information and computer science,1982,11:341-356.  
[2] Pawlak Z. Rough set:Theoretical aspects of reasoning about data [M]. Boston, London: Kluwer Academic Publishers, 1991.  
[3] Zhong N, Dong J Z, Ohsuga S. Using rough sets with heuristics for feature selection[J]. Journal of intelligent information systems,2001,16(2):199-214.  
[4] 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.  
[5] 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,1999,36(6):681-684.  
[6] Shannon C E. The mathematical theory of communication [M]. Illinois:University of Illinois Press,1963.

年招生规模预测[J].孝感学院学报,2010,30(3):60-63.  
[2] Pai T Y,Tsai Y P,Loh H M,et al. Gray and neural network prediction of suspended solids and chemical oxygen demand in hospital wastewater treatment plant effluent [J]. Computers and chemical engineering,2007,31:1272-1281.  
[3] Guo Hairu,Li Zhimin. A method of improving generalization ability for neural network based on genetic algorithm [C]//Proc of 2010 IEEE international conference on intelligent computing and intelligent systems(ICIS 2010). Beijing: Institute of Electrical and Electronics Engineers,Inc. ,2010:742-745.  
[4] 杨建刚. 人工神经网络实用教程[M]. 杭州:浙江大学出版社,2001.  
[5] Hecht-Nielsen R. Kolmogorov's mapping neural network existence theorem [C]//Proc of conf on neural networks. San Diego:IEEE,1987:11-14.  
[6] 胡 康,万金泉. 基于遗传算法的控制系统在废水处理中的应用[J]. 计算机技术与发展,2011,21(2):18-21.  
[7] 田 奕,乔俊飞. 基于遗传算法的 BOD 神经网络软测量[J]. 计算机技术与发展,2009,19(3):127-129.  
[8] 郭海如,崔雪梅,董春玲. 一种基于神经网络模型的中国能耗预测[J]. 孝感学院学报,2007,27(3):76-79.  
[9] Rumenhart D E,Hinton G E,Willians R J. Learning representation by backpropagation errors[J]. Nature,1986,323(9):533-536.  
[10] 原思聪,江祥奎. 基于 GA-BP 神经网络的双目摄像机标定[J]. 西安建筑科技大学学报(自然科学版),2011,43(4):604-608.  
[11] 雷英杰. Matlab 遗传算法工具箱及应用[M]. 西安:西安电子科技大学出版社,2005:107-118.  
[12] FECIT. 神经网络理论与 MATLAB 7 实现[M]. 北京:电子工业出版社,2005:99-107.

[7] Skowron A. The discernibility matrix and function in information system [M]//Handbook of application and advances of the rough sets theory. Dordrecht: Kluwer Academic Publishers,1992:331-362.  
[8] Guan J,Bell D. Rough computational methods for information systems[J]. Artificial intelligence,1998,105:77-103.  
[9] 蒋 瑜,张 娟,林 和,等. 基于区分矩阵的决策表相容性的判断[J]. 甘肃科学学报,2006,18(2):59-61.  
[10] 王国胤. Rough 理论与知识获取[M]. 西安:西安交通大学出版社,2001.  
[11] 张文修,吴伟志. 粗糙集理论与方法[M]. 北京:科学出版社,2001.  
[12] 杜 跃. 基于粗糙集理论的属性约简算法研究[D]. 兰州:西北师范大学,2008.

基于容差关系的不完备信息系统的属性约简

作者：[颜家凯](#)，[范敏](#)，[刘文奇](#)，[叶荣荣](#)，[YAN Jia-kai](#)，[FAN Min](#)，[LIU Wen-qi](#)，[YE Rong-rong](#)

作者单位：[昆明理工大学 理学院, 云南 昆明, 650500](#)

刊名：[计算机技术与发展](#)

ISTIC

英文刊名：[Computer Technology and Development](#)

年，卷(期)：2014(1)

本文链接：[http://d.g.wanfangdata.com.cn/Periodical\\_wjtz201401026.aspx](http://d.g.wanfangdata.com.cn/Periodical_wjtz201401026.aspx)