采用布尔矩阵不完备信息系统的属性约简

周海岩
ZHOU Hai-yan

淮阴工学院 计算机工程系,江苏 淮安 223001
Department of Computer Engineering,Huaiyin Institute of Technology,Huaian,Jiangsu 223001,China
E-mail:zhy_5703@163.com

ZHOU Hai-yan.Attribute reduction based on incomplete information system.Computer Engineering and Applications, 2010,46(1):119-121.

Abstract: The popular method of completion for incomplete information system includes data complementation and extension method.In this paper,several methods solving the incomplete information system is analyzed and another method which generates boolean discernibility matrix based on information offered totally by information system is presented.And an efficient algorithm is provided for attribute reduction based on boolean discernibility matrix.

Key words: rough set;incomplete information system;boolean discernibility matrix;attribute reduction

摘 要:对于不完备信息系统完备化问题,现在常采用的方法是数据补齐法和扩充法,在研究和分析了其优劣后,提出一种完全依据信息系统提供的信息来产生布尔可辨矩阵的方法。给出一个基于对布尔可辨矩阵进行化简的求属性约简(或最小属性约简)的高效算法。

关键词:粗糙集;不完备信息系统;布尔可辨矩阵;属性约简
DOI:10.3778/j.issn.1002-8331.2010.01.037 **文章编号**:1002-8331(2010)01-0119-03 **文献标识码**:A **中图分类号**:TP301

1 引言

粗糙集理论[1]是一种新的处理模糊和不确定知识的数学工具。经过十几年的研究与发展,粗糙集理论已经在理论和实践应用上取得了长足的发展,特别是由于20世纪90年代在知识发现等领域得到了成功的应用而受到国际学术界广泛关注。目前,粗糙集理论正在被广泛应用于机器学习、决策分析、过程控制、模式识别和数据挖掘等领域。属性约简是粗糙集理论中所研究的热门问题之一[1]。粗糙集理论以一种全新的视觉审视知识,认为知识与分类能力相关,两者成正比关系,即拥有知识的多少通过分辨能力的高低来评判。所谓属性约简,就是在保持信息系统的分类能力不变的前提下, 删除其中的冗余属性,得到最佳决策规则。

遗憾的是粗糙集理论是基于完备系统这样一假设的,即信息表中所有的属性值都是已知的。完备信息系统的属性约简的研究报道很多(如文献[2-3])。而在实际应用中,由于数据的测量误差、对知识获取的限制等原因,人们往往面对的是不完备信息系统,即可能存在部分属性值未知的情况,此时,直接应用原有的粗糙集理论往往不能得到正确的结论。因此,如何有效地将粗糙集理论应用于不完备信息系统,一直是诸多学者研究的主要方向之一。

该文的主要贡献是完全依据信息系统提供的信息来产生布尔可辨矩阵,然后通过布尔可辨矩阵来求属性约简。

2 粗糙集理论中针对不完备信息系统的处理方法

对于不完备信息系统的属性约简问题,首先想到的是如何将这些属性值进行补齐,使其完备化。在此基础上,再进行属性约简。通常,信息系统完备化的方法有以下四种途径[4]。

(1)简单的将存在空缺(遗漏)属性值的记录删除,从而得到一个完备的信息系统。这种途径是用于记录数量远大于属性数目的情况。

(2)将空缺(遗漏)属性值作为一种特殊的属性值来处理,它不同于其他任何属性值,这样,不完备的信息系统便完备化了。

(3)采用统计学原理,根据信息系统中其余实例在该属性上的取值的分布情况来对一个遗漏属性值进行补充估计。如I.Kononenko等人利用贝叶斯方法确定遗失值在其取值上的概率分布,然后选择一个最可能发生的值代替遗失值。

(4)根据Rough集理论中数据不可分辨关系来对不完备的系统进行补齐,具体如ROUSTIDA算法。

当然, 上面介绍的不完备信息系统完备化方法的补齐措施,或者是删除,或者是补以主观估计值,因此会或多或少的改变原始的信息系统, 对于不完备信息系统的属性约简问题,还可以考虑的是在保持信息系统的原始信息不发生改变的情况下,对信息系统进行处理。该文正是采用该策略对不完备信息系统进行属性约简的。

基金项目:江苏省科技攻关计划(the Key Technologies R & D Program of Jiangsu Province,China under Grant No.BE2006357)。
作者简介:周海岩(1957-),男,教授,CCF高级会员,主要研究领域:知识发现(数据挖掘)、智能决策、算法设计与分析、数据库理论等。
收稿日期:2008-07-22 **修回日期**:2008-09-27

定义 1 四元组 $S=(U,A,V,f)$是一个信息系统,其中 U 表示对象的非空有限集合, 称为论域;A 表示属性的非空有限集合;$V=\bigcup_{a\in A}V_a$,V_a 是属性 a 的值域;f 表示 $U\times A\to V$ 是一个信息函数,它为每个对象的每个属性赋予一个信息值,即 $a\in A$,$x\in U$,$f(x,a)\in V_a$。

每一个属性子集 $P\subseteq A$,决定了一个二元不可辨关系(即等价关系)$IND(P)$;$IND(P)=\{(x,y)\in U\times U\mid\forall a\in P\ f(x,a)=f(y,a)\}$。

定义 2 设 $S=(U,A,V,f)$ 是一个信息系统,$a\in A$ 如果 $IND(A-\{a\})=IND(A)$,则称属性 a 在 A 中是不必要的(多余的),否则,称 a 在 A 中是必要的。

定义 3 设 $S=(U,A,V,f)$是一个信息系统,如果每个属性 $a\in A$ 在 A 中都是必要的,则称属性集 A 是独立的,否则称 A 是相依的。

定义 4 设 $S=(U,A,V,f)$ 是一个信息系统,A 中所有必要的属性组成的集合称为属性集 A 的核,记作 $Core(A)$。

定义 5 设 $S=(U,A,V,f)$是一个信息系统,$P\subseteq A$ 如果

(1)$IND(P)=IND(A)$

(2)P 是独立的。

则称 P 是 A 的一个约简(注:对决策表称为相对约简)。

可以证明核是约简的交集[5]。

定义 6 对于一个信息系统 $S=(U,A,V,f)$, 如果至少有一个属性 $a\in A$ 使得 Va 含有空值, 则称 S 为一个不完备信息系统,否则它是完备的,用"*"表示空值。

3 基于布尔可辨矩阵的属性约简

定义 7 设 $S=(U,A,V,f)$是一个完备信息系统,则相应于 S 的布尔可辨矩阵 $\boldsymbol{M}$ 构造如下:

矩阵 $\boldsymbol{M}$ 的第 i 列对应属性 a_i,共有 m 列。每一行对应论域中的一个对象对(u_p,u_q)。M 有 $n(n-1)/2$ 行,设 $M=(m_{((p,q),i)})$ 则,

$$m_{((p,q),i)}=\begin{cases}1 & \text{如果 } a_i(u_p)\neq a_i(u_q)\\ 0 & \text{否则}\end{cases}$$

在上述定义中,对于对象 u_p 与 u_q,当 $a_i(u_p)\neq a_i(u_q)$,知对象 u_p 与 u_q 可以通过属性 a_i 的值辨别开来,故用 $m_{((p,q),i)}=1$ 来标识对象 u_p 和 u_q 可以通过属性 a_i 来辨别。当 $a_i(u_p)=a_i(u_q)$时,不能通过属 a_i 的取值将对象 u_p 和 u_q 辨别开来,故命 $m_{((p,q),i)}=0$。

设 $S=(U,A,V,f)$是一个不完备信息系统,也可以采用上述处理问题的思想。对于对象 u_p 与 u_q,当 $a_i(u_p)\neq *$,$a_i(u_q)\neq *$ 且 $a_i(u_p)\neq a_i(u_q)$,在此情况下可以通过属性 a_i 将对象 u_p 与 u_q 辨别开来。当 $a_i(u_p)=*$,或 $a_i(u_q)=*$ 时根据属性 a_i 难以判断对象 u_p 与 u_q 的异同, 即根据系统给提供的信息无法通过属性值来辨别对象 u_p 与 u_q。于是给出如下不完备信息系统的布尔可辨矩阵 $\boldsymbol{M}$ 构造方法。

定义 8 设 $S=(U,A,V,f)$是一个不完备信息系统,则相应于 S 的布尔可辨矩阵 $\boldsymbol{M}$ 构造如下:

矩阵 $\boldsymbol{M}$ 的第 i 列对应属性 a_i,共有 m 列。每一行对应论域 U 中的一个对象对(u_p,u_q)。M 有 $n(n-1)/2$ 行,设 $M=(m_{((p,q),i)})$则,

$$m_{((p,q),i)}=\begin{cases}1 & \text{如果 } a_i(u_p)\neq a_i(u_q)\text{且 } a_i(u_p)\neq *\text{ 且 } a_i(u_q)\neq *\\ 0 & \text{否则}\end{cases}$$

3.1 布尔可辨矩阵的化简

布尔可辨矩阵 $\boldsymbol{M}$ 直接描述了每个属性对论域中对象的分辨情况,直接反映信息系统 S 中所蕴涵的知识,在布尔可辨矩阵 $\boldsymbol{M}$ 中,某个元素为 1 或 0 表示所在行所对应的对象 u_p,与 u_q 在属性 a_i 下可分辨与不可分辨。

根据布尔可辨矩阵的以上特性,于是有下述命题:

命题 1 若布尔可辨矩阵中某一行只有一个元素为 1,其余元素为 0,则元素 1 所在列对应某个属性,所有这样的属性构成信息系统属性集的核。若没有这样的行,则核为空。

命题 2 设 $S=(U,A,V,f)$是一个信息系统,$B\subseteq A$ 是 A 的约简,其充要条件为:(1)在由 B 中所有属性对应的各列所构成的 $\boldsymbol{M}$ 的子阵中,与 $\boldsymbol{M}$ 有相同的不全为 0 的行。(2)$\forall B'\subset B$,B'不满足条件(1)。

由命题 1,根据布尔可辨矩阵,很容易求出属性集的核或相对核,但如何求得属性集约简呢?这是一个比较复杂的问题,首先看以下事实。在可辨矩阵中,若某一行元素全为 1,说明相对应的两个对象 u_p,u_q 在任何一个属性下都可分辨, 此时将这一行删除而不影响约简。在可辨矩阵中,若某一行元素全为 0,说明相应的两个对象 u_p,u_q 在任何属性下都不可分辨, 将其删除仍不影响约简。根据以上分析,给出以下布尔可辨矩阵的约简变换的改进定义。

定义 9 布尔可辨矩阵 $\boldsymbol{M}$ 的约简变换包括以下几种形式:

(1)布尔可辨矩阵中首先将全为 0 的行删除。

(2)将布尔可辨矩阵中的行、列重新排列(注:当列重新排列时,该列所对应的属性也要随之变动)。

(3)对某两列,如 a_i 列与 a_j 列,若 $a_i+a_j=a_j$("+"表示逻辑加),则可将 a_i 列删除。

(4)对于某两行,如(u_p,u_q)行与(u_p',u_q')行,若$(u_p,u_q)+(u_p',u_q')=(u_p',u_q')$,则将$(u_p',u_q')$行删除,对于布尔可辨矩阵 $\boldsymbol{M}$ 经过约简变换之后将得到规模大为减小的矩阵, 为讨论方便,经过约简之后的矩阵记为 $\boldsymbol{M}'$,则有如下命题:

命题 3 在信息系统 $S=(U,A,V,f)$ 中,$B\subseteq A$ 是 A 的约简,其充要条件为:(1)在矩阵 $\boldsymbol{M}$ 中,由 B 中各个属性所在列的逻辑和为$(1,1,\cdots,1)^{\mathrm{T}}$;(2)$\forall B'\subset B$,$B'$不满足条件(1)。

3.2 基于布尔可辨矩阵的属性约简算法.

由于篇幅所限布尔可辨矩阵的化简算法见文献[5]。

在布尔可辨矩阵 $\boldsymbol{M}$ 中,对于那些只有一个元素为 1 其余元素均为 0 的行,元素 1 所在列的属性一定属于核,而对于那些有多个元素为 1 的行,在这些元素为 1 所在的列中,那些所含 1 的个数最多的列对应的属性虽未必是核属性,但具有很强的分辨能力,因此这样的属性在形成约简,尤其是最小约简的过程中具有重要地位。

经过前述对信息系统或决策表的布尔可辨矩阵所具有的特性的分析与论述以及给出的布尔可辨矩阵的化简算法(即算法 1)以下给出信息系统或决策表的属性约简算法。

算法(属性约简算法,AR)

1. 根据给定的信息系统 $S=(U,A,V,f)$(完备或不完备)产生布尔可辨矩阵 $\boldsymbol{M}$。

2.调用布尔可辨矩阵的化简算法对 $\boldsymbol{M}$。进行化简(结果仍记为 $\boldsymbol{M}$)。

3.置矩阵 $\boldsymbol{A}\leftarrow\boldsymbol{M}$,用 Reduction 表示属性约简,初始值为 φ。

4.将 $\boldsymbol{A}$ 的各行横向相加结果放入 $Row[i]$中,$1\leqslant i\leqslant P$。

5.求 $Row[i]$中的最小值;$Row_{\min}=\min\{Row[1],Row[2],\cdots,Row[p]\}$

(1)if $Row_{\min}=1$

①求出所有使得 $Row[r_i]=1$ 的各行$\{r_1,r_2,\cdots,r_w\}$及其元素 1 所对应的各列$\{a_1,a_2,\cdots,a_w\}$。

②$Reduction\leftarrow Reduction\cup\{a_1,a_2,\cdots,a_w\}$。

③从 $\boldsymbol{A}$ 中删除各行$\{r_1,r_2,\cdots,r_w\}$及其元素 1 所对应的各列$\{a_1,a_2,\cdots,a_w\}$,将得到的新矩阵再赋给 $\boldsymbol{A}$。

④if $\boldsymbol{A}\neq\varphi$ then go 4 else go 6。

(2)if $Row_{\min}>1$

①$h=Row_{\min}$

②求出所有使得 $Row[r_i]=h$ 的各行$\{r_1,r_2,\cdots,r_w\}$及其元素 1 所对应的各列 $L=\{a_{11},a_{12},\cdots,a_{1h},a_{21},a_{22},\cdots,a_{2h},\cdots,a_{w1},a_{w2},\cdots,a_{wh}\}$。

③将 $\boldsymbol{A}$ 的在 L 中的各个列的元素纵向相加,结果放入 $CoL[a_{ik}]$中(其中 $a_{ik}\in L$)。

④求出 $CoL[a_{ik}]$的最大值 $Col_{\max}=\max\{CoL[a_{i1}],CoL[a_{i2}],\cdots,CoL[a_{wh}]\}$并求出取得最大值 $Col_{\max}$ 的列号$\{a_{h1},a_{h2},\cdots,a_{hi}\}$若其中只有一列,则将这一列对应的属性作为 a_{choice},否则,从$\{a_{h1},a_{h2},\cdots,a_{hi}\}$中随机取一属性作为 a_{choice}。

⑤$Reduction\leftarrow Reduction\cup\{a_{choice}\}$。

⑥从 $\boldsymbol{A}$ 中消去列 a_{choice} 中所有元素 1 所对应的行及其列 a_{choice} 后,得到的新矩阵再赋给 $\boldsymbol{A}$。

⑦if $\boldsymbol{A}\neq\varphi$ then go 4 else go 6。

6.输出属性集的一个属性约简 $Reduction$。

7.算法结束。

3.3 应用实例

现以某小汽车信息表(见表 1)作为例,利用文中提出的方法进行属性约简。根据定义 8 得布尔可辨矩阵如表 2 所示。

利用文献[5]给出的布尔可辨矩阵的化简算法,得简化的矩阵如表 3 所示。

表 1 小汽车信息表

Car	Price	Mileage	Size	Max-speed
1	high	low	full	low
2	low	*	full	low
3	*	*	compact	low
4	high	*	full	high
5	*	*	full	high
6	low	high	full	*

最后得所给信息表的属性约简(也是最小约简)为{Price,Size,Max-speed}。

表 2 一个二进制可辨矩阵

$(u_p.u_q)/\boldsymbol{A}$	Price	Mileage	Size	Max-speed
(1,2)	1	0	0	0
(1,3)	0	0	1	0
(1,4)	0	0	0	1
(1,5)	0	0	0	1
(1,6)	1	1	0	0
(2,3)	0	0	1	0
(2,4)	1	0	0	1
(2,5)	0	0	0	1
(2,6)	1	1	0	0
(3,4)	0	0	1	1
(3,5)	0	0	1	1
(3,6)	0	0	1	0
(4,5)	0	0	0	0
(4,6)	1	0	0	0
(5,6)	0	0	0	0

表 3 一个二进制可辨矩阵

Price	Size	Max-speed
1	0	0
0	1	0
0	0	1

4 算法复杂度分析与结束语

给出的基于布尔可辨矩阵的不完备信息系统属性约简算法是文献[5]所给出的相应算法的拓展,算法的正确性在算法的各步中都有详细说明。给出的算法在最坏情况下,其复杂度与文献[5]相同,由于篇幅所限在此不再赘述。

在文献[5]的相应算法的基础上给出了更加高效的改进算法。完全依据信息系统提供的信息来产生布尔可辨矩阵,然后通过布尔可辨矩阵来求属性约简,与其他算法相比是高效的。但是求属性集最小约简的问题,仍然是一个未解决的问题,将继续这方面的探索与努力。

参考文献:

[1] Pawlak Z.Rough sets—Theoretical aspects of reasoning about data[M].Dordrcht:Kluwer Academic Publisher,1991.

[2] 李金海,吕跃进.一种基于关系矩阵的信息系统属性约简算法[J].计算机工程与应用,2008,44(9):147-149.

[3] 王加阳,高灿.基于分辨矩阵的快速完备约简算法[J].计算机工程与应用,2008,44(8):92-94.

[4] 肖健华.智能模式识别方法[M].广州:华南理工大学出版社,2006:58-59.

[5] 周海岩.基于二进制可辨矩阵的属性约简算法的改进[J].计算机工程与设计,2003(12):35-37.

(上接 105 页)

参考文献:

[1] Picard R W,Vyzas E,Healey J.Affective wearable[C]//Proceedings of the 1st International Symposium on Wearable Computers,Cambridge,MA,1997.

[2] Picard R W,Vyzas E,Healey J.Toward machine emotional intelligence:Analysis of affective physiological state[J].IEEE Transactions Pattern Analysis and Machine Intelligence,2001,23(10):1175-1191.

[3] Pudil P,Novovicčvá J,Kittler J.Floating search methods in feature selection[J].Pattern Recognition Letters,1994,15:1119-1125.

[4] Haag A,Goronzy S,Schaich P,et al.Emotion recognition using bio-sensors:First step towards an automatic system[M]//Affective Dialogue System,Tutorial and Research Workshop,Kloster Irsee,Germany,2004,3068:36-48.

[5] 巩敦卫,郝国生,周勇,等.分层式交互式进化计算及其应用[J].控制与决策,2004,19(10):1117-1120.

[6] 巩敦卫,郝国生,孙晓燕,等.自适应分层交互式遗传算法及其应用[J].杭州电子科技大学学报,2005,25(2):45-48.

[7] 袁健,李智勇.一种采用循环策略的改进模拟退火遗传算法[J].计算机工程与应用,2007,43(2):102-104.