计算机科学 2005Vol. 32№. 11

基于等价类矩阵的属性约简*)

闫德勤
(辽宁师范大学计算机系 大连 116029)

摘 要 由于不完备信息系统不能完全适用于粗糙集等价类模型,其合理的属性约简方法的研究在当前是一个备受关注的研究热点。文章给出不完备信息系统等价关系的矩阵表示,同时给出了关于等价类矩阵以及核属性的相关定理,给出了应用等价类矩阵进行属性约简的方法和应用举例,为不完备信息系统的属性约简提供了一种新的方法。

关键词 粗糙集,属性约简,等价类矩阵

Matrix Method for Attribute Reduction for Incomplete Information Systems

YAN De-Qin
(Department of Computer Science, Liaoning Normal University, Dalian 116029)

Abstract In this paper, matrix expression for equivalence relation is proposed, some theorems related to equivalence matrix and the property of core are presented. And, a new method of attribute reduction for incomplete information systems is given. The validity of the method is verified by an example.

Keywords Rough sets, Attribute reduction, Equivalent matrix

1 引言

不完备信息系统中由于数据的丢失或不确定,其等价关系不能完全适用于粗糙集等价类模型。为对不完备信息系统进行属性约简,一些学者对信息系统中丢失属性或不确定数据从不同的角度提出了一些处理方法[2],同时也有一些对传统粗糙集扩展模型的提出。所有这些方法与模型都是在不同意义下对不完备信息系统进行属性约简的一种处理方式。由于不同的信息系统中缺省数据(或不确定数据)关联的意义不同,使得建立有效的属性约简的统一模型和方法变困难。这样也使得研究不完备信息系统属性约简的理论方法变得更加重要。Guan 等人在文[3]中对于完备信息系统提出了等价类的矩阵表示,并给出了相关的属性约简方法。该文的方法从一个新的角度发展了属性约简的算法和理论,但该方法不适用于不完备信息系统的属性约简。为此,本文提出了一种针对不完备信息系统的等价类矩阵表示方法,给出了关于等价类矩阵以及核属性的相关定理。同时,给出了应用等价类矩阵进行属性约简的方法和应用举例。

2 基本概念

设 $S=(U,Q,V,F)$为一信息系统,其中 $U=\{x_1,x_2,\cdots,x_n\}$是论域,Q 是属性集合,V 是属性取值集合,F 是 $U\times Q\to V$ 的映射。设属性集合包含 m 个条件属性 $C=\{C_1,C_2,\cdots,C_m\}$和一个决策属性 D。若 D 的取值有 s 个,如 $D_1,D_2,\cdots,D_s$,则由 D 导出的等价类构成 U 的一个划分:$\{Y_1,Y_2,\cdots,Y_s\}$。其中,$Y_i=\{x\in U|F(x,D)=D_i\}$,$i=1,2,\cdots,s$。

信息系统中每一 x_i 及其所对应的属性值称为一个(信息表示的)规则。

若信息系统中的每个属性值都是已知的,则称为完备的信息系统。在有些情况下,由于种种原因信息系统中的某些属性值不能得到或确定,则称信息系统为不完备信息系统。

3 等价类矩阵

设 $S=(U,Q,V,F)$为一完备信息系统,对于 $P\subseteq C$ 在文[3]中 Guan 给出了属性集合 P 的等价类矩阵 M_p 元素的定义:对于 $x_i,x_j\in U$ 如果 $F(x_i,P)\cong F(x_j,P)$,则 $M_p(i,j)=1$;否则 $M_p(i,j)=0$。其中,符号"$\cong$"表示对于每一 $C_k\in P$ 都有 $F(x_i,C_k)=F(x_j,C_k)$。

可见等价类矩阵为一个 $n\times n$ 阶(0,1)矩阵,其元素 $M_p(i,j)=1$ 表示在属性集合 P 的划分下,x_i 和 x_j 为同一等价类。即等价类矩阵表示出了在等价关系 P 下论域 U 中等价类的划分。以此为基础,Guan 给出了一种属性约简方法。

对不完备信息系统难于直接用这样的方法产生等价矩阵。下面提出一种关于不完备信息系统等价矩阵构造方法。

设 $S=(U,Q,V,F)$为一个不完备信息系统,V 是属性取值集合,信息系统中的不确定属性值用 * 表示,系统中属性取值的种类用 v 表示。

定义 1 给定不完备信息系统 S,对于 $C_k\subseteq C$,其等价矩阵 M_{C_k} 的元素 $M_{C_k}(i,j)$定义为:

当 $F(x_i,C)$与 $F(x_j,C_k)$都是确定值时:

若 $F(x_i,C_k)=F(x_j,C_k)$,$M_{C_k}(i,j)=1$;否则 $M_{C_k}(i,j)=0$。

当 $F(x_i,C)$与 $F(x_j,C_k)$有一个是 * 时:

$$M_{C_k}(i,j)=\frac{1}{v};$$

当 $F(x_i,C)$与 $F(x_j,C_k)$两个都是 * 时:

*)国家自然科学基金(60372071)资助;辽宁省教育厅高等学校科学研究基金(2004C031)资助;辽宁师范大学校基金资助。闫德勤 博士,教授,主要研究领域为模式识别、数据挖掘等。

$M_{C_k}(i,j)=\frac{1}{v^2}$。

表1 不完备信息系统1

U	a	b	d
x_1	1	*	1
x_2	*	1	2
x_3	0	0	3
x_4	*	1	4

例1 表1为一不完备信息系统表，a,b 为条件属性，由定义1分别得到关于 a 和 b 的等价类矩阵为：

$$M_a=\begin{pmatrix}1 & \lambda & 0 & \lambda\\ & 1 & \lambda & \lambda^2\\ & & 1 & \lambda\\ & & & 1\end{pmatrix}\quad M_b=\begin{pmatrix}1 & \lambda & \lambda & \lambda\\ & 1 & 0 & 1\\ & & 1 & 0\\ & & & 1\end{pmatrix}$$

其中，$\lambda=\frac{1}{v}$。由于 M_a 和 M_b 都是对称矩阵，上面只列出了它们的上三角部分。在本例中由于已知确定的属性值有0和1两种形式，因此 $v=2$。

由等价类矩阵的定义可知，属性所对应的等价类矩阵实际上刻画出了等价类的划分，即给出了论域 $U=\{x_1,x_2,\cdots,x_n\}$ 中任两个元素 x_i,x_j 的等价程度，是一种软划分。

定义2 设 M_P 和 M_Q 为两个等价类矩阵，其交运算 $M_P\bigcap M_Q$ 定义为：$M=M_P\bigcap M_Q$ 的元素 $M(i,j)=\min\{M_P(i,j),M_Q(i,j)\}$。

由定义2可知，若 $P\subseteq C$ 是包含多个条件属性 $A_1,A_2,\cdots,A_s$ 的集合，其等价矩阵 M_P 的元素 $M_P(i,j)$ 定义为：

$M_P(i,j)=\min\limits_{1\leqslant k\leqslant s}\{M_{A_k}(i,j)\}$。

定理1 给定不完备信息系统 S，对于 $P,Q\subseteq C$，有 $M_P\bigcap M_Q=M_{P\cup Q}$。

证明：根据等价类矩阵的定义可得到：

$$\begin{aligned}M_P\bigcap M_Q&=(\bigcap_{a\in P}M_a)\bigcap(\bigcap_{a\in Q}M_a)\\&=((\bigcap_{a\in P-P\cap Q}M_a)\bigcap(\bigcap_{a\in P\cap Q}M_a))\bigcap((\bigcap_{a\in Q-P\cap Q}M_a)\bigcap(\bigcap_{a\in P\cap Q}M_a))\\&=((\bigcap_{a\in P-P\cap Q}M_a)\bigcap(\bigcap_{a\in Q-P\cap Q}M_a))\bigcap(\bigcap_{a\in P\cap Q}M_a)\\&=\bigcap_{a\in P\cup Q}M_a。\end{aligned}$$

证毕。

根据等价类矩阵的定义以及定理1可得到以下定理2和3(这里证明略去)。

定理2 给定不完备信息系统 S，对于 $P,Q\subseteq C$，有 $M_P\bigcap M_Q\leqslant M_P,M_Q$。

定理3 给定不完备信息系统 S，对于 $P,Q\subseteq C$，若 $P\supseteq Q$ 则有 $M_P\leqslant M_Q$。

需要说明的是，对于含有决策属性的不完备信息系统，由于对应同一决策属性的规则属于同一决策类，条件等价类矩阵相应的元素值要置1。如两个规则 x_i,x_j 如果对应同一决策属性，则对于任何 $P\subseteq C$ 其等价类矩阵 M_P 中的元素 $M_P(i,j)=1$。事实上，这个说明的结论完全可以根据定义1和2得出。这样的等价类矩阵称为具有决策属性影响的等价类矩阵。本文以下所述的等价类矩阵都是指的这种情况。

定义3 给定不完备信息系统 S，对于 $a\in P$ 若 $M_P<M_{P-\{a\}}$，则称 a 是重要的，否则称 a 是不重要的。a 在 p 中的重要性定义为：$sig_P(a)=|M_{P-\{a\}}-M_P|$。

其中，$||$ 表示矩阵元素的总和。

定义4 给定不完备信息系统 S，对于 $a\in C$ 若 $M_C<M_{C-\{a\}}$，则称 a 是核属性。

定理4 给定不完备信息系统 S，对于对应不同的决策属性的两个规则 x_i,x_j，若存在唯一的属性 $C_k\in C$ 有 $F(x_i,C_k)=F(x_j,C_k)=*$（$*$ 表示属性值为不确定），而两个规则中具有确定的属性值相等，即 $F(x_i,a)=F(x_j,a)(a\in C$，对应两规则的属性值为确定值），则 C_k 为核属性。

证明：考察每个条件属性 $C_s(s=1,2,\cdots,m)$ 对应的等价类矩阵中第 i 行第 j 列的元素 $M_{C_s}(i,j)$。由定理的条件可知，$M_{C_k}(i,j)=\lambda^2$（这里 $\lambda=\frac{1}{v}$），而其它的元素 $M_{C_s}(i,j)(s\neq k)$ 则大于 λ^2。因此，$M_C(i,j)=\min\limits_{1\leqslant s\leqslant m}\{M_{C_s}(i,j)\}=M_{C_k}(i,j)$。即 $M_C(i,j)<M_{C-\{C_k\}}(i,j)$。由定义4知，$C_k$ 为核属性。

证毕。

根据以上关于等价类矩阵的定义和定理即可以对不完备信息系统进行属性约简。具体方法是：对于条件属性 C 计算出相应的等价类矩阵 M_C，然后对于每个 $C_s(s=1,2,\cdots,m)$ 依次计算 $M_{C-\{C_s\}}$，若存在 C_k 使得 $M_C<M_{C-\{C_k\}}$，则删去 C_k。最后不被删去的属性即是约简结果。

对不完备信息系统进行属性约简的另一个方法是：首先找到核属性，根据属性重要性的定义不断把重要性大的属性加到核属性所在的集合中，直到集合中属性所对应的等价矩阵的交运算的结果等于 M_C 为止。此时，集合中的元素即是属性约简结果。

下面给出一个简单例子说明利用等价类矩阵对不完备信息系统属性约简的方法。

例1 表2为一不完备信息系统。其中，a,b,c 和 d 为条件属性，e 为决策属性。由于在表中条件属性确定值为0和1两种，因此 $v=2$。

根据等价类矩阵的定义，可以得到以下关于属性 a,b,c 和 d 的等价类矩阵（其中 $\lambda=\frac{1}{2}$，由于等价类矩阵是对称矩阵，这里只写出上三角部分）：

$$M_a=\begin{pmatrix}1 & 1 & 1 & \lambda & 1 & 0\\ & 1 & 1 & 1 & 1 & 0\\ & & 1 & 1 & 1 & 0\\ & & & 1 & \lambda & \lambda\\ & & & & 1 & 1\\ & & & & & 1\end{pmatrix}\quad M_b=\begin{pmatrix}1 & 0 & 1 & 0 & 0 & 1\\ & 1 & 1 & 1 & 1 & 0\\ & & 1 & 1 & 0 & 1\\ & & & 1 & 1 & 0\\ & & & & 1 & 1\\ & & & & & 1\end{pmatrix}$$

$$M_c=\begin{pmatrix}1 & \lambda & \lambda & \lambda^2 & \lambda^2 & \lambda\\ & 1 & 1 & 1 & \lambda & 0\\ & & 1 & 1 & \lambda & 1\\ & & & 1 & \lambda^2 & \lambda\\ & & & & 1 & 1\\ & & & & & 1\end{pmatrix}\quad M_d=\begin{pmatrix}1 & 1 & 0 & 0 & 0 & 0\\ & 1 & 1 & 1 & 0 & 0\\ & & 1 & 1 & 1 & 1\\ & & & 1 & 1 & 1\\ & & & & 1 & 1\\ & & & & & 1\end{pmatrix}$$

（下转第174页）

将字符图像分成 4×4 个网格，则每个网格白色像素最多为 256 个。则构成的决策表条件属性 $C=\{C1,C2,\cdots,C16\}$，且属性值值域为 $V_c=\{0,1,\cdots,256\}$，决策属性 $D=\{d\}$ 且属性值域 $V_d=\{0,1,2\cdots,9\}$。

建立决策表后，需先对其进行离散化处理。在这里我们采用了简单的等距离离散化方法，经过实验，距离间隔为 20 的时候效果最佳。则离散化后属性值的值域为 $V_d=\{0,1\cdots,16\}$。

我们使用基于差别矩阵的约简算法，可以得出全部的约简。从约简结果选择区别比较大的几个约简，分别用来构造分类器。

3.3 BP 神经网络的构造

在进行 BP 网络的设计时，一般应从网络的层数、每层中神经元的个数、激活函数、初始权值以及学习效率几个方面进行考虑[6]。经过实验，选择如下：

(1)网络的层数：我们选择了三层网络，即输入层、隐含层和输出层。

(2)各层神经元个数：输入层为约简结果中属性的个数，即约简特征向量元素格数。隐含层神经元个数为 30 时，网络收敛效果比较好。输出层神经元个数为 10，即对应于 10 个数字。

(3)初始权值的选取：取在(−1,1)之间的随机数。

(4)学习速率：学习速率的范围是 0.01～0.8，我们选择的是 0.1。

3.4 字符识别算法

使用训练后的网络识别待识样本：

(1)取一个待识样本，由某个约简结果属性组成向量。$X=[x_1,x_2,\cdots,x_n]$

(2)将向量输入相应的 BP 网络，得到某个输出结果，如 $D1=[0.1,0.8,0.1,\cdots,0.0]$。

(3)同样的可以在其他的 BP 网络上，得到另两个输出结果，如 $D2=[0,0.7,0,0.1,0.2,\cdots,0]$，$D3=[0,0.9,\cdots,0.1]$。

(4)由神经网络组合的投票法的到识别结果，比如是字符“1”。

4 实验结果分析

实验的编程环境为 Visual C++ 6.0。实验的对象为车牌数字字符集(因为汉字和字母样本数量太少)。样本是拍摄的车牌图像经字符自动分割和归一化算法得到的[7]。训练样本 200 个(每个字符 20 个)，测试样本为 200 个(每个字符 20 个)，对测试样本进行 10 次试验，取平均值作为最终结果。

我们采用了三种方法对样本进行测试。

(1)基于粗糙集规则匹配：由粗糙集属性约简和值约简得到匹配规则，输入样本特征进行匹配。

(2)基于 BP 网络：属性没有经过粗糙集离散化和属性约简，直接送入 BP 网络进行训练。

(3)基于粗糙集和 BP 网络：即本文中论述的方法。

三种方法实验结果比较如下表所示。

方法	匹配时间(s)	平均正确识别样本个数	平均拒识样本个数	平均误识样本个数	识别率(%)
1	5.2	145	30	25	72.5
2	4.3	162	15	23	81
3	2.8	183	8	9	91.5

结论 用粗糙集理论对属性约简后，神经网络的神经元数减少，连接也随之减少。本文从理论上得出从网络上移除一个连接对网络输出影响的极限值。实验论证，经过粗糙集离散化和约简处理后，神经网络的结构简单，学习效率高。基于多个约简结果的分类器的组合比单个分类器效果好。总之，基于粗糙集和神经网络结合的分类系统识别率高，识别需要的时间短。

参 考 文 献

1 王国胤. Rough 集理论于知识获取[M]. 西安：西安交通大学出版社，2001

2 边肇琪，张学工，等. 模式识别(第二版)[M]. 北京：清华大学出版社，1999

3 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

4 Setiono R. Extracting rules from Pruned neural networkd for breast cancer diagnosis. Artificial Intelligence in Medicine，1996，8(1)：37～51

5 Swiniarski R W，Skowron A. Rough set methods in feature selection and recognition[J]. Pattern recognition letters[J]，2003，24：833～849

6 Pandya A S，Macy R B，著，徐勇，等译. 神经网络模式识别及其实现[M]. 电子工业出版社，1999

7 刘智勇，刘迎建. 车牌识别(LPR)中的图像提取及分割[J]. 中文信息学报，2003，14(4)

(上接第 171 页)

表 2 不完备信息系统 2

U	a	b	c	d	e
x_1	1	1	*	0	1
x_2	1	0	1	0	0
x_3	1	1	0	1	0
x_4	*	0	*	1	0
x_5	1	0	*	1	2
x_6	0	1	0	1	2

由于 $M_c(4,5)=\lambda^2$ 是 $M_a(4,5)$、$M_b(4,5)$、$M_c(4,5)$、$M_d(4,5)$中唯一最小的，因此属性 c 是核属性。因为 $M_c \cap M_d = M_c$，所以属性 d 是可约的。由于 $M_c \cap M_a < M_c$，可知属性 a 不可约，同样，由于 $(M_c \cap M_a) \cap M_b < (M_c \cap M_a)$ 可知属性 b 不可约。因此，最后得到的约简结果为 $\{a,b,c\}$。

结论 本文提出了一种针对不完备信息系统的等价类矩阵表示方法，同时给出了相关的理论研究结果以及属性约简方法。由于采用矩阵表示，使得等价类的描述简洁明确，给属性约简带来方便。更重要的是为进一步研究属性约简开辟新的思路。

参 考 文 献

1 Pawlak Z. Rough set approach to multi-attribute decision analysis. European Journal of Operational Research，1994，72：443～459

2 Grzymala-Busse J W，Hu M. A comparison of several approaches to missing attribute values in data mining. In：Proc. of the 2nd Int'l Conf. on Rough Sets and Current Trends in Computering. Berlin：Springer Verlag，2000. 378～385

3 Guan J W，Bell D Z，Guan Z. Matrix computation for information systems. Information Sciences，2001，131：129～156

4 曾黄麟. 粗集理论及其应用(修订版). 重庆：重庆大学出版社，1998