不完备信息系统规则获取的矩阵算法*)

瞿彬彬　卢炎生
(华中科技大学计算机学院　武汉 430074)

摘　要　对象信息的不完备性是从实例中归纳学习的最大障碍。本文定义了限制非对称相似关系，并将经典的可辨识关系矩阵加以扩充，定义了限制非对称相似关系下的可辨识关系矩阵，采用布尔推理方法，直接从不完备决策系统中提取规则而无需改变初始不完备信息系统的结构。实验结果表明，所获得的决策规则简洁、高效，与缺省值无关。

关键词　粗糙集，不完备信息系统，限制非对称相似关系，规则获取

Matrix Computation for Rule Extraction in Incomplete Information Systems

QU Bin-Bin　LU Yan-Sheng
(College of Computer Science & Technology, Huazhong University of Science & Technology, Wuhan 430074)

Abstract　The incompleteness of information about objects may be the greatest obstruct to performing induction learning from example. In this paper, the concept of limited non-symmetric similarity relation is defined and then classical discernibility matrix is extended based on limited non-symmetric similarity relation. By taking the method of Boolean reasoning, rules are extracted directly from the incomplete decision systems without changing the size of original incomplete systems. The experiment shows that the algorithm provides precise and simple decision rules and does not affected by the missing values.

Keywords　Rough sets, Incomplete information systems, Limited non-symmetric similarity relation, Rule extraction

1　引言

不完备信息系统大量存在于现实生活中，如数据库、数据集市等。对象信息的不完备性是从实例中归纳学习的最大障碍。因此，从不完备信息系统中获取规则是人工智能、数据挖掘研究的热点之一。目前，对不完备信息系统的处理通常有两种策略：其一，通过对不完备对象进行处理使不完备信息系统完备化。主要方法有：1)删除法[1]。删除含有不完备属性的对象。2)默认值法[2]。当信息系统中的不完备属性为数值型时，用该属性的所有平均值作为默认值替代；当信息系统中的不完备属性为非数值型时，用该属性的最常见值作为默认值替代。3)扩展法[3]。对信息系统中的不完备属性用该属性的所有可能取值分别替代。例如，对象 x 有不完备属性 a 和 b，$|V_a|=3$，$|V_b|=4$，应用扩展法将对象 x 按 a 和 b 可能取值组成 12 种情况。方法 1、2 实现简单，但由于改变了原不完备信息系统的结构，使得在此基础上获得的规则不能真实地反映属性间的关系。方法 3 虽然克服了上述缺点，但在实际应用中，对于不完备的大数据集而言，计算代价太大。其二，基于扩展的粗糙集理论，直接从不完备决策表中提取规则。如 Stefanowski. J 提出了基于量化容差关系模型的规则获取算法[4]，从上、下近似集中获取超过设定阈值的规则。但阈值的设置需要预先知道决策表中属性值的概率分布情况，这对于一个新的不完备决策表而言是很困难的，而且规则数的多少还与阈值的高低相关。Kryszkiewicz M. 在相容关系模型基础上，利用广义决策函数，从不完备决策表及其所有扩展形式中获取确定规则[5]。上述粗糙集扩展模型对于不完备信息的理解，只考虑了“遗漏”语意而没有考虑“缺席”语意。而在现实生活中，很多未知值是无法再得到的，因此不能与任一值相比较。

本文试图从限制非对称相似关系模型出发，利用扩充的可辨识矩阵和布尔推理方法，直接从不完备决策表中提取规则。

2　基本概念与原理

本节简述与后续工作相关的粗糙集理论的主要概念。

定义 1[1,6]　信息系统(Information System)：一个信息系统 IS 是四元组：$IS=\langle U,A,V,f\rangle$，其中 U 是对象的非空有限集合；A 是属性的非空有限集合；对任何 $a\in A$，V_a 表示属性 a 的值域，即 $V=\bigcup V_a$，$f:U\times A\to V$ 称为信息函数，它为每个对象赋予一个信息值，$\forall a\in A, x\in U$，有 $f(x,a)\in V_a$。

对于 IS，若 $A=C\cup D$，$C\cap D=\varnothing$，则称 IS 为一个决策表(Decision Table, DT)，其中 C 中的属性为条件属性，D 中的属性为决策属性。

定义 2[6]　不可分辨(Indiscernibility)关系：给定 $IS=\langle U,A,V,f\rangle$，$B\subseteq A$，定义 B 在 U 上的不可分辨关系 $IND(B)=\{(x,y)\in U\times U: f(x,a)\}$。

定义 3[5,6]　给定 $IS=\langle U,A,V,f\rangle$，对于每个子集 $X\subseteq U$ 和不可分辨关系 $IND(B)$，定义两个子集：下近似集 $B_-(X)=\bigcup\{Y_i\in U/IND(B)\mid Y_i\subseteq X\}$ 和上近似集 $B^-(X)=\bigcup\{Y_i\in U/IND(B)\mid Y_i\cap X\neq\varnothing\}$。

定义 4[3]　设相容决策表 $DT=\langle U,A,V,f\rangle$，$A=C\cup D$，$C=\{a_i\mid i=1,\cdots,m\}$ 和 $D=\{d\}$ 分别为条件属性集和决策属性集，$U=\{x_1,\cdots,x_n\}$ 是论域，$a_i(x_j)$ 是样本 x_j 在属性 a_i 上的取值。$M_D(i,j)$ 表示可辨识矩阵中第 i 行 j 列的对象，则可辨识矩阵 M_D 定义为：

$M_D(i,j)=$

*)基金项目：作“十五”国家科技攻关计划资助项目(2001 BA102 A04-04-03)。瞿彬彬　博士，研究方向为数据挖掘、人工智能；卢炎生　教授，博士生导师，研究方向为数据挖掘、软件构件化、软件测试、人工智能等。

$$\begin{cases}\{a_k \mid a_k \in C \wedge a_k(x_i) \neq a_k(x_j)\} & , d(x_i) \neq d(x_j) \\ 0 & , d(x_i) = d(x_j)\end{cases}$$

其中 $i,j=1,2,\cdots,n$。

3 限制非对称相似关系模型

定义 5 不完备决策表 $IDT=\langle U,C\cup D,V,f\rangle$，$B\subseteq C$，限制非对称相似关系 LS 定义为：

$$\forall_{x,y\in U}(LS_B(x,y)\Leftrightarrow\forall_{c_j\in B}(c_j(x)=*\vee c_j(x)=c_j(y)\wedge c_j(x)=c_j(y)\neq *))$$

定义 6 限制非对称相似于 x 的对象集合 $LR_B(x)$，x 与之限制非对称相似的对象集合 $LR_B^{-1}(x)$ 定义为：

$LR_B(x)=\{y\mid y\in U\wedge LS_B(y,x)\}$，

$LR_B^{-1}(x)=\{y\mid y\in\in U\wedge LS_B(x,y)\}$。

定义 7 不完备决策表 $IDT=(U,C\cup D,V,f)$ 的对象集合 X 关于属性 $B\subseteq C$ 的下近似(X_B^{LS})和上近似(X_{LS}^{B})定义为：

$X_B^{LS}=\{x\mid x\in U\wedge LR_B^{-1}(x)\subseteq X\}$，

$X_{LS}^{B}=\{x\mid x\in U\wedge LR_B(x)\cap X\neq\varnothing\}$。

定理 1 给定信息表 $IS=\langle U,C,V,f\rangle$，集合 $B\subseteq C$，个体对象集合 $X\subseteq U$，在限制非对称相似关系下 X 的上近似集和下近似集是对非对称相似关系下 X 的上近似集和下近似集的改进。

证明：即要证：$X_B^S\subseteq X_B^{LS}$，$X_{LS}^{B}\subseteq X_S^B$ 成立。

根据非对称相似关系 $S_B(x,y)$ 和限制非对称相似关系 $LS_B(x,y)$ 定义知：

对于 $\forall x,y\in U$，有

$LS_B(x,y)\rightarrow S_B(x,y)$，$LS_B(y,x)\rightarrow S_B(y,x)$ 成立。

1)证 $X_B^S\subseteq X_B^{LS}$。根据定义

$R_B^{-1}(x)=\{y\mid y\in U\wedge S_B(x,y)\}$，

$LR_B^{-1}(x)=\{y\mid y\in U\wedge LS_B(x,y)\}$，

$\therefore\forall y\in U$，

$y\in LR_B^{-1}(x)\Rightarrow y\in LS_B(x,y)\Rightarrow y\in S_B(x,y)$，

即 $LR_B^{-1}(x)\subseteq R_B^{-1}(x)$，

$\because X_B^{LS}=\{x\mid x\in U\wedge LR_B^{-1}(x)\subseteq X\}$，

$X_B^S=\{x\mid x\in U\wedge R_B^{-1}(x)\subseteq X\}$，

$\therefore\forall x\in U, x\in X_B^S\Rightarrow R_B^{-1}(x)\subseteq X\Rightarrow LR_B^{-1}(x)\subseteq X\Rightarrow x\in X_B^{LS}$

但反过来不一定成立。

$\therefore X_B^S\subseteq X_B^{LS}$ 成立。

2)证 $X_{LS}^{B}\subseteq X_S^B$。根据定义 $LR_B(x)=\{y\mid y\in U\wedge LS_B(y,x)\}$，$R_B(x)=\{y\mid y\in U\wedge S_B(y,x)\}$，

$\therefore\forall y\in U, y\in LS_B(y,x)\Rightarrow y\in S_B(y,x)$，

即 $LR_B(x)\subseteq R_B(x)$，

$\because X_{LS}^{B}=\bigcup LR_B(x)\mid x\in X, X_S^B=\bigcup R_B(x)\mid x\in X$，

$\therefore X_{LS}^{B}\subseteq X_S^B$ 成立。 证毕。

4 基于限制非对称相似关系模型的可辨识矩阵

定义 8 不完备决策表 $IDT=\langle U,C\cup D,V,f\rangle$，条件属性集 $C=\{a_i\mid i=1,\cdots,m\}$，决策属性集 $D=\{d\}$，$U=\{x_1,\cdots,x_n\}$ 是论域，定义限制非对称相似关系的下近似可辨识矩阵如下：

$$\underline{M}=(m_{ij})_{|U|\times|U|}$$

$$=\begin{cases}\{a: f(x_i,a)\neq f(x_j,a)\vee(f(x_i,a)\neq * \\ \wedge f(x_j,a)=*\}, x_j\notin LR_C^{-1}(x_i), f(x_i,d)\neq f(x_j,d) \\ \varnothing, \text{else}\end{cases}$$

定义 9 不完备决策表 $IDT=\langle U,C\cup D,V,f\rangle$，条件属性集 $C=\{a_i\mid i=1,\cdots,m\}$，决策属性集 $D=\{d\}$，$U=\{x_1,\cdots,x_n\}$ 是论域，定义限制非对称相似关系的上近似可辨识矩阵如下：

$$\overline{M}=(m_{ij})_{|U|\times|U|}$$

$$=\begin{cases}\{a: f(x_i,a)\neq f(x_j,a)\vee(f(x_i,a)\neq * \\ \wedge f(x_j,a)\neq *\}, x_j\notin LR_C(x_i), f(x_i,d)\neq f(x_j,d) \\ \varnothing, \text{else}\end{cases}$$

与完备决策表的可辨识矩阵不同，限制非对称相似关系的下近似可辨识矩阵和上近似可辨识矩阵都不是一个依主对角线对称的矩阵。显然，$\overline{M}=[\underline{M}]^T$。

5 规则获取矩阵算法

算法：规则获取矩阵算法

输入：不完备决策表 $IDT=\langle U,C\cup D,V,f\rangle$；

输出：决策规则集；

Step1：计算不完备决策表 IDT 的限制非对称相似关系的下近似分辨矩阵；

Step 2：计算相应的逻辑表达式

$L_i=\bigwedge_j(\bigvee_k a_k)$，$a_k\in C_{ij}$，$C_{ij}\neq 0$，$C_{ij}\neq\varnothing$；

Step 3：生成基于 L_i 的决策规则，规则表示为：

$R_i: L_i(x_i)\rightarrow d(x_i)$。

6 实验

为验证基于限制非对称相似关系模型的规则获取矩阵算法效果，实验选用 4 个含不完备信息的 UCI 数据集，分别为：

Breast cancer 数据集，由原南斯拉夫卢布尔维那肿瘤研究所提供，有 286 个对象，9 个属性，含 0.2%的未知条件属性值。

Hepatitis 数据集，由 Carnegie-Mello 大学提供，有 155 个对象，19 个属性，含 5.7%的未知条件属性值。

Congressional Votes 数据集记录了美国国会 1984 年 435 个国会议员的投票情况，16 个属性，含 5.7%的未知条件属性值。

Credit Approval 申请信用卡数据集，有 690 个对象，15 个属性，含 0.6%的未知条件属性值。

对于数据集中连续属性，采用连续属性值，用 ROSETTA 软件中的 Entropy/MDL 算法进行离散化处理，同时采用 ROSETTA 软件中的 Split 功能随机地将数据集分成两半，一半用作训练集，对各训练集分别提取规则，另一半用作测试集，用所提取的规则分别对测试样本进行识别。实验结果包括规则数目和规则匹配率。规则匹配率定义为能够与规则集中规则条件匹配的测试样本占总样本的比例。实验结果如表 1 所示。

表 1 矩阵规则获取算法结果

数据集	对象数	属性数	缺省属性值	规则数	规则匹配率
Breast cancer	286	9	0.2%	73	86.4%
Hepatitis	155	19	5.7%	60	78.5%
Congressional Votes	435	16	5.7%	186	76.3%
Credit Approval	690	15	0.6%	305	85.7%

实验结果显示，对于含缺省属性值少的数据集，矩阵规则提取算法能获得更高的规则匹配率。对于含缺省属性值多的数据集，将来要考虑采用合适的规则推理策略，来提高规则的适应能力。

结束语 本文提出了一种新的基于粗糙集的从不完备信息系统获取规则的矩阵算法。该算法具有以下优点：1)不改变初始不完备信息系统结构；2)获取的规则不受缺省值的影响。实验结果表明，所获得的规则简洁，规则集的规模小，具有较好的可理解性和较强的泛化能力。进一步的工作，研究限制非对称相似关系模型下规则的增量获取以及如何定义规则的可信度、覆盖度来解决规则冲突消解问题。

参 考 文 献

1 Komorowski J. Øhrn A. Skowron A. The ROSETTA Rough Set Software System: In: Handbook of Data Mining and Knowledge Discovery, Lendon: Oxford University Press, 2002

2 Clark P, Niblett T. The CN2 induction algorithm. Machine learning, 1989(3): 261～283

3 Grzymala-Busse J W. On the unknown attribute values in learning from examples, proc of the ISMIS-91. In: 6th Internetional symposium on Methodologies for Intelligent Systems, 1991, Lecture Notes in Artifical Intelligence, vol 542, Springer-Verlag, Berlin Heidelberg New York, 1991. 368～377

4 Stefanowski J, Tsoukias A. Valued Tolerance and Decision Rules. In: W. Ziarko, Y. Yao, eds, Rough Sets and Current Trends in Computing, Berlin: Springer, 2002. 271～278

5 Kryszkiewicz M. Rough set approach to incomplete information systems. Information Sciences, 1998, 112(1-4): 39～49

6 王国胤著. Rough 集理论与知识获取. 西安：西安交通大学出版社，2001

(上接第 144 页)

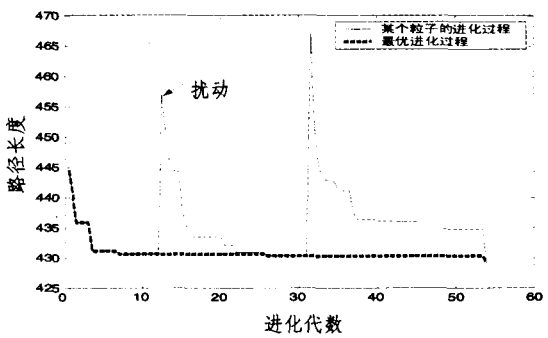

图 2 一个完整的搜索过程

表 1 算法 ACS+2-OPT 和 SEHDPSO 的比较
(各运行 10 次)

实例	ACS+2-OPT			SEHDPSO		
	平均误差 %	最小误差%	CPU 时间(s)	平差均误%	最小误差%	CPU 时间(s)
Eil51	0.00	000	0.00	1	0.00	
Berlin52	0.00	0.00	1	0.00	0.00	3.2
St70	0.40	0.00	26	0.00	0.00	6.4
Eil76	0.00	0.00	3	0.00	0.00	9.8
Pr76	0.00	0.00	3	0.05	0.02	8.7
Kroc100	0.00	0.00	41	0.01	0.00	37.5
Krod100	0.06	0.00	19	0.05	0.12	27.2
Krob100	0.23	0.00	55	0.09	0.00	41.8
Rd100	0.77	0.45	9	0.25	0.30	54.1
Eil101	0.64	0.08	28	0.02	0.00	16.3
Lin105	0.00	0.00	9	0.03	0.00	12.4
Pr107	0.40	0.30	15	0.18	0.00	16.4
Pr124	0.05	0.00	9	0.23	0.30	47.5
Bier127	0.48	0.10	136	0.10	0.00	101
Ch130	0.45	0.21	236	0.29	0.11	239
Kroa150	0.01	0.00	661	0.05	0.00	66
Krob150	0.05	0.05	93	0.18	0.10	45.7
U159	0.47	0.19	223	0.00	0.00	153
Krob200	0.35	0.03	1459	0.25	0.05	254
D198	0.61	0.55	229	0.13	0.03	198
Tsp225	0.18	0.10	34	0.08	0.02	168
Tsp225	0.56	0.40	673	0.23	0.00	95
A280	0.86	0.38	1514	0.15	0.03	561
Rd400	1.34	1.18	4581	0.25	0.16	1250
P654	1.10	0.72	13582	0.93	0.41	7004
U724	2.26	1.85	24340	0.64	0.22	2381
平均值	0.42	0.24	1845	0.16	0.07	492

为了检验 SEHDPSO 的有效性，本文的算法和文[9]中的蚁群混合算法(ACS+2-OPT)进行了比较，仿真结果见表 1。

从表 1 中的数据对比来看，在实例规模 51 个城市到 124 个城市之间，算法 SEHDPSO 和算法 ACS+2-OPT 的效果相当。但随着规模的增加，SEHDPSO 显示出明显的优势，尤其在实例 U724 上，它不仅搜索精度远高于 ACS+ 2-OPT，而且收敛的时间不到它的十分之一。表 1 的最后一行列出了 26 个实例的平均值，从平均误差、最小误差、运行时间来看，SEHDPSO 分别是 ACS+2-OPT 的 35%，29%，26.6%。

结论 本文提出的 SEHDPSO 利用自逃逸思想很好地保持了群体的多样性，利用局部搜索算法能加快收敛速度。和混合蚁群算法的比较表明，SEHDPSO 算法是有效的。传统理论分析和实验数据显示，速度公式(2)的系数变化对算法的结果有很大的影响。所以，进一步的工作将放在以下两个方面：(1)调整各系数，协调算法局部搜索和全局搜索。(2)利用种群信息，从候选边集合中选择更好的边，找出更有效的逃逸行为，增强算法的全局搜索能力。

参 考 文 献

1 Eberhart R, Kennedy J. A New Optimization Using Particles Swarm Theory. Proc Sixth International Symposium on Micro Machine and Human Science Nagiya [C]. Japan: IEEE Service Center, Piscataway, 1995. 39～43

2 Kennedy J, Eberhart R. A discrete binary version of the particle swarm optimization [C]. In: Proc. IEEE Int Conf. on Neural Networks. Perth, Austrilia, 1997. 4104～4108

3 WANG KANG-PING. Particle swarm optimization for traveling salesman problem [J]. In: Proceedings of the Second International on Machine Learning and Cybernetics, Xi'an, November 2003. 2～5

4 Glover F. Ejection chains, reference structures and alternating path methods for traveling salesman problems [J]. Discrete Applied Mathematics, 1996, 65:223～253

5 Helsgaun K. An effective implementation of the Lin-Kernighan traveling salesman heuristic [J]. European Journal of Operational Research, 2000, 126:106～130

6 Ozcan E, Mohan CK. Particle swarm optimization: Surfing the waves [J]. In: Proc. of the IEEE Int'l Conf. on Evolutionary Computation. Washington: IEEE Inc, 1999. 1939～1944

7 www. iwr. uniheidelberg. de/groups/comopt/software/TSPLIB95/tsp/

8 Rego C. Relaxed tours and path ejections for the traveling salesman problem [J]. European journal of Operational Research, 1998, 106: 522～538

9 LE Louarn F, Gendreau M. Geni Ants for the Traveling Salesman Problem [J]. Annals of Operations Research, 2004, 121: 187～201

10 Croes G A. A method for solving traveling salesman problem [J]. Operations Research, 1958, 6:791～812