一种基于粗糙集理论的规则获取算法[1]

安利平[1]，吴育华[1]，仝凌云[2]
（1. 天津大学管理学院，天津 300072；2. 河北工业大学管理学院，天津 300130）

摘要：获取规则是数据挖掘中的一项重要技术，根据粗糙集理论及决策值归纳函数的概念，可以把不相容的决策系统转化为相容的决策系统，并提出规则参数的合并方法. 在此基础上，利用决策矩阵和决策函数，提出了一种在决策系统中获取规则的算法. 同传统的算法相比，该算法得出的规则集没有信息丢失的现象发生. 最后以例子作了说明.

关键词：粗糙集；规则获取；数据挖掘
中图分类号：TP311　　**文献标识码**：A　　**文章编号**：1007-9807(2001)05-0074-05

0 引　言

在智能决策中，通常的知识库表示成决策规则 if…then… 形式的集合. 粗糙集理论[2,3]的提出和应用为获取规则提供了重要方法，其导出的规则精炼且便于存储和使用. 但是，目前多数文献使用单个最小决策规则集作决策，即用户根据不同的原则来选择一个认为最好的约简，比如属性个数最少的约简或具有最少费用的约简[4,5]，通过此约简计算最小决策规则集. 许多获取规则的算法也致力于寻找单个最小决策规则集. 根据粗糙集理论，一个决策系统的属性集可能存在多个约简，每个约简对应的属性和规则互不相同. 另一方面，在多数情况下事先很难预料利用哪些约简得出的规则，有时某些属性的信息因为某种原因不能得到或存在错误，对这种信息不完全的情况，有的文献建议找出所有的约简以便得出相应的规则，通过避开不完全的或有错误的信息而达到正确决策的目的[6]，所以，不少算法将求约简作为获取规则的一个必不可少的步骤. 由于约简只是区分不同对象属于哪个类所必须的属性集，而提取规则是指利用哪些必须信息可以得出结论，两者是有区别的. 所以，即使找出所有约简而得到规则，也会导致有用信息的丢失而影响决策.

利用决策矩阵和决策函数得出所有决策规则，可以避免信息丢失现象的发生. 决策者可针对不同的问题需要，选取不同的规则获取方法，如利用某些启发式算法求约简从而获取规则. 这种算法复杂性低，运行速度快，但缺点是得到的规则集不完善，对有特殊需要的问题，利用本文所述的方法可以得到完备的规则集，有利于在信息不完全的情况下作决策，但付出的时间将有所增加.

1 粗糙集理论的有关概念

定义 1　称 $S=(U,A)$ 为信息系统，其中 U 为论域，是一非空有限集，即 $U=\{x_1,x_2,\cdots,x_n\}$；A 是非空有限的属性集合. 对任一 $a\in A$，可视为一个函数，即 $a:U\to V_a$，集合 V_a 是 a 的值域.

称 $S=(U,A\cup\{d\})$ 为决策系统，其中 A 为条件属性集，$d\notin A$ 为决策属性. 决策系统是一种特殊的信息系统.

定义 2　令 $S=(U,A\cup\{d\})$ 为一决策系统，$B\subseteq A$，定义 B 不可分辨关系 $\mathrm{IND}(B)$ 为 $\mathrm{IND}(B)=\{(x,x')\in U^2\mid\forall a\in B, a(x)=a(x')\}$，$a(x)$ 为元素 x 在属性 a 上的值.

如果 $(x,x')\in\mathrm{IND}(B)$，说明根据已有的信息不能将 x 和 x' 区分开.

[1] 收稿日期：2000-05-17；修订日期：2001-08-28.
作者简介：安利平(1971-)，男，河北张家口人，博士生.

定义 3　令集合 V_d 表示决策属性 d 的值域，不失一般性，设 $V_d=\{1,2,\cdots,r(d)\}$，则决策属性 d 将论域 U 划分成 $\{Y_1,Y_2,\cdots,Y_{r(d)}\}$，其中 $Y_k=\{x\in U|d(x)=k\}$，$1\leqslant k\leqslant r(d)$，$d(x)$ 为决策属性 d 在对象 x 上的值，集合 Y_k 称为决策系统 S 的第 k 个决策类.

定义 4　设 $S=(U,A\cup\{d\})$ 为一决策系统，$B\subseteq A$，定义 A 中的 B- 决策值归纳函数为

$\partial_B:U\to 2^{V_d}$

$\partial_B(x)=\{v\in V_d|\exists\ x'\in U\ \text{s.t.}\ x\ \text{IND}(B)x,\text{且}\ d(x')=v\}$

定义 5　设 $S=(U,A\cup\{d\})$ 为一决策系统，如果对任意 $x\in U$，有 $|\partial_A(x)|=1$，则 S 称为相容决策系统，否则 S 称为不相容决策系统.

不相容意味着决策系统中存在着一些对象，它们的条件属性值均相同，却属于不同的决策类. 对于这样的决策系统，通过将决策属性 d 替换为 ∂_A，可以把不相容决策系统转化为相容的决策系统，以便利于规则获取.

定义 6　令 $S=(U,A\cup\{d\})$ 为一决策系统，$U=\{x_1,x_2,\cdots,x_n\}$，定义决策矩阵为

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

其中 $m_{ij}=\{a\in A|a(x_i)\neq a(x_j)\ \text{且}\ d(x_i)\neq d(x_j)\}$

从定义可以看出，决策矩阵的元素 m_{ij} 是对象 x_i 的某些条件属性的集合，用于区分对象 x_i 和 x_j. 决策矩阵保留了原系统的所有分类信息，提供了获取决策规则的有效方法. 决策矩阵是对称矩阵，且属于同一决策类的对象不必相互比较，这意味着在实际操作中只需列出少于一半的矩阵元素，以避免对象之间的重复比较，最大限度地利用原有的计算信息.

2　规则参数的计算

设 $S=(U,C\cup\{d\})$ 为一决策系统，$U/\text{IND}(C)=\{X_1,X_2,\cdots,X_n\}$，$U/\text{IND}(d)=\{Y_1,Y_2,\cdots,Y_m\}$. 决策系统 S 描述了 $U/\text{IND}(C)$ 与 $U/\text{IND}(d)$ 之间的关系 $X_i\to Y_j(i=1,2,\cdots,n;\ j=1,2,\cdots,m)$. 利用四个参数来描述 $X_i\to Y_j$，即(1) 等价类 X_i 包含的元素数量 $|X_i|$；(2) 决策类 Y_j 包含的元素数量 $|Y_j|$；(3)$SI(X_i,Y_j)=\dfrac{|X_i\cap Y_j|}{|X_i|}$，简记为 SI_{ij}；(4)$CI(X_i,Y_j)=\dfrac{|X_i\cap Y_j|}{|Y_j|}$，简记为 CI_{ij}.

由决策矩阵及决策函数，可以得出每个等价类 X_i 的最简规则，$X_i\to Y_j$ 的参数很自然地传递给每一条规则 r，得出与规则 r 相关的三个参数为

1) 规则适用的对象数 $\text{support}(r)=|X_i\cap Y_j|$；

2) 规则的精度 $\text{accuracy}(r)=SI_{ij}$；

3) 规则的适用度 $\text{coverage}(r)=CI_{ij}$.

有时，同一决策类中的不同等价类得出形式相同的规则，则需要将这几条规则的参数合并，作为一条规则使用. 设规则 r 分别由 $X_i\to Y_j$ 和 $X_{i'}\to Y_j$ 得出，它们的参数分别为 $(|X_i|,SI_{ij},CI_{ij})$ 和 $(|X_{i'}|,SI_{i'j},CI_{i'j})$，则

$$\text{support}(r)=|(X_i\cup X_{i'})\cap Y_j|$$

$$\text{accuracy}(r)=\frac{|X_i\cap Y_j|+|X_{i'}\cap Y_j|}{|X_i|+|X_{i'}|}=\frac{SI_{ij}\cdot|X_i|+SI_{i'j}\cdot|X_{i'}|}{|X_i|+|X_{i'}|}$$

$$\text{coverage}(r)=\frac{|X_i\cap Y_j|+|X_{i'}\cap Y_j|}{|Y_j|}=CI_{ij}-CI_{i'j}$$

上述参数合并过程不难推广到由多个等价类得出相同规则的情况.

3　获取规则的算法

规则获取算法如下：

1) 对研究的决策系统 S 检查其是否相容，若不相容，则利用决策值归纳函数将其转化为相容的决策系统 S'；

2) 计算所有 $X_i\to Y_j$ 的参数；

3) 构建 S' 的决策矩阵 $M(S')=(m_{ij})_{n\times n}$，由于决策矩阵是对称矩阵，故设 $M(S')$ 为上三角矩阵；

4) 由决策矩阵建立每个条件等价类 $X_i(i=1,2,\cdots,n)$ 的决策函数 B_i：

$$B_i=\left(\bigwedge_{\substack{k=1\\ a\in m_{kl}}}^{i-1}\bigvee a\right)\wedge\left(\bigwedge_{\substack{l=i+1\\ a\in m_{il}}}^{n}\bigvee a\right),k=1,2,\cdots,i-1;\ l=i-1,\cdots,n,$$

$\wedge$ 和 $\vee$ 分别表示合取和析取运算.

5) 利用幂等率($a \wedge a = a$)和吸收率($a \wedge (a \vee b) = a$)化简决策函数为析取范式，其中每个合取子式对应一条规则的前件，由于合取子式中只有属性名，所以将各属性匹配以 X_i 的属性值，并明确其决策值，得到 X_i 的规则；

6) 重复过程(4)、(5)，得到 S 中所有等价类的规则，进行规则参数的合并，根据问题的需要，利用一定的规则选取策略选择规则进行决策.

设利用上述算法得到的规则集为 R_1，通过计算所有约简得出的规则集为 R_2，则有如下关系存在：$R_2 \subseteq R_1$，原因就是此算法保留了原系统的所有有用信息. 另外，本算法把不相容的决策系统转化为相容的决策系统，从而利用了原系统的统计信息.

4 举例

某决策系统如表 1 所示[11]，其中不相容数据 8、9 的决策值归纳函数值为 {1,2}. 决策类 $Y_1 = \{1,2,3\}$，$Y_2 = \{4,5,9\}$，$Y_3 = \{6,7,8\}$.

利用上述算法，首先根据原系统构建决策矩阵如表 2(1) 所示. 其次，对每个类计算决策函数并化为析取范式，以第一个等价类为例，化简其决策函数为

$B_1^1 = (b \vee c \vee d) \wedge (b \vee d) \wedge (a \vee b \vee c) \wedge (a \vee c) \wedge (a \vee b \vee c \vee d) = (a \wedge b) \vee (a \wedge d) \vee (b \wedge c) \vee (c \wedge d)$，将各属性匹配以相应的属性值，得出的决策规则为

$(a=0) \wedge (b=0) \rightarrow (e=0)$，$(a=0) \wedge (d=1) \rightarrow (e=0)$，$(b=0) \wedge (c=1) \rightarrow (e=0)$，$(c=1) \wedge (d=1) \rightarrow (e=0)$. 同理可以得出所有的确定性和可能性的规则集如表 2(2) 所示.

表 1 决策系统

序号	样本个数	条件属性 a	b	c	d	决策属性 e
1	15	0	0	1	1	0
2	10	1	0	1	2	0
3	5	1	1	0	1	0
4	60	0	1	0	2	1
5	10	0	2	1	2	1
6	30	1	2	0	1	2
7	25	2	0	0	1	2
8	40	2	1	2	0	2
9	5	2	1	2	0	1

表 2 决策矩阵和规则集

		(1)						(2)
		类 1		类 2		类 {1,2}		规 则 前 件
		4	5	6	7	8	9	
类 0	1	bcd	bd	abc	ac	$abcd$		$(a,0) \wedge (b,0)$，$(a,0) \wedge (d,1)$，$(b,0) \wedge (c,1)$，$(c,1) \wedge (d,1)$
	2	abc	ab	bcd	acd	$abcd$		$(a,1) \wedge (b,0)$，$(a,1) \wedge (c,1)$，$(a,1) \wedge (d,2)$，$(b,0) \wedge (c,1)$，$(b,0) \wedge (d,2)$
	3	ad	$abcd$	b	ab	acd		$(a,1) \wedge (b,1)$，$(b,1) \wedge (d,1)$
类 1	4			abd	abd	acd		$(a,0) \wedge (b,1)$，$(a,0) \wedge (c,0)$，$(a,0) \wedge (d,2)$，$(b,1) \wedge (d,2)$，$(c,0) \wedge (d,2)$
	5			acd	$abcd$	$abcd$		$(a,0) \wedge (b,2)$，$(a,0) \wedge (d,2)$，$(b,2) \wedge (c,1)$，$(b,2) \wedge (d,2)$
类 2	6					$abcd$		$(a,1) \wedge (b,2)$，$(b,2) \wedge (c,0)$，$(b,2) \wedge (d,1)$
	7					bcd		$(a,2) \wedge (b,0)$，$(a,2) \wedge (c,0)$，$(a,2) \wedge (d,1)$，$(b,0) \wedge (c,0)$
类 {1,2}	8					$(a,2) \wedge (b,1)$，$(c,2)$，$(d,0)$		
	9							

进行规则的参数合并，例如表 2 中条件等价类 X_1 和 X_2 均得出规则 $(b,0) \wedge (c,1) \rightarrow (e,0)$，$X_1 \rightarrow Y_1$ 和 $X_2 \rightarrow Y_2$ 的参数分别为：$|X_1| = 15$，$|X_2| = 10$，$|Y_2| = 30$，$SI(X_1,Y_1) = SI(X_2,Y_2) = 1$，$CI(X_1,Y_1) = 0.5$，$CI(X_2,Y_2) = 0.333$.

设 r 表示规则 $(b,0) \wedge (c,1) \rightarrow (e,0)$，其参数合并如下：

$$\text{support}(r) = |(X_1 \cup X_2) \cap Y_1| = 25$$

$$\text{accuracy}(r) = \frac{SI_{11} \times |X_1| + SI_{21} \times |X_2|}{|X_1| + |X_2|} = \frac{1 \times 15 + 1 \times 10}{15 + 10} = 1$$

$$\text{coverage}(r) = CI_{11} + CI_{12} =$$

$$0.5 - 0.333 = 0.833$$

如果利用传统的方法,首先计算系统的所有约简,该系统的约简集为 $RED(S,e) = \{\{ab\},\{bcd\}\}$,再进行属性值约简,得到的决策规则如表 3 中(1)所示,表 3 中(2)列出了传统算法得不出而利用本文算法可以得出的规则.

表 3　规则集的比较

		(1)	(2)
类 0	1	$(a,0) \wedge (b,0),(b,0) \wedge (c,1),(c,1) \wedge (d,1)$	$(a,0) \wedge (d,1)$
	2	$(a,1) \wedge (b,0),(b,0) \wedge (c,1),(b,0) \wedge (d,2)$	$(a,1) \wedge (c,1),(a,1) \wedge (d,2)$
	3	$(a,1) \wedge (b,1),(b,1) \wedge (d,1)$	
类 1	4	$(a,0) \wedge (b,1),(b,1) \wedge (d,2),(c,0) \wedge (d,2)$	$(a,0) \wedge (c,0),(a,0) \wedge (d,2)$
	5	$(a,0) \wedge (b,2),(b,2) \wedge (c,1),(b,2) \wedge (d,2)$	$(a,0) \wedge (d,2)$
类 2	6	$(a,1) \wedge (b,2),(b,2) \wedge (c,0),(b,2) \wedge (d,1)$	
	7	$(a,2) \wedge (b,0),(b,0) \wedge (c,0)$	$(a,2) \wedge (c,0),(a,2) \wedge (d,1)$
类{1,2}	8 9	$(a,2) \wedge (b,1),(c,2),(d,0)$	

5　结论

根据粗糙集理论,利用决策矩阵提出了一种从决策系统中获取规则的算法,该算法克服了一般算法导致的有用信息丢失这一不足,这对于有效地获取规则和在信息不完备的情况下作决策十分有意义.

参 考 文 献:

[1] Pawlak Z. Rough set[J]. Inter J. of Computer and Information Sciences, 1982,11:341-356

[2] Pawlak Z, Grzymala-Busse J, Slowinski R, et al. Rough sets[J]. Communications of the ACM, 1995,38(11):89-95

[3] A. Skowron. Extracting laws from decision table: a rough set approach[J]. Computational Intelligence, 1995, 11(2):371-388

[4] 赛 英,陈文伟. 从数据库中发现知识的方法研究与应用[J]. 管理科学学报,1999,2(3):92-96

[5] Hu Xiao-hua, Nick Cercone. Learning in relational databases: a rough set approach[J]. Computational Intelligence, 1995,11(2):323-338

[6] Wojciech Ziarko. Discovery though rough set theory[J]. Communications of the ACM, 1999, 42(11): 55-57

[7] 张 琦,韩祯祥,文福拴. 一种基于粗糙集理论的电力系统故障诊断和警报处理新方法[J]. 中国电力, 1998, 31(4): 32-38

[8] Ning Shan, Wojciech Ziarko. Data-based acquisition and incremental modification of classification rules[J]. Computational Intelligence, 1995,11(2):357-370

[9] Øhrn A. Discernibility and rough sets in medicine: tools and applications[D]. Norwegian University of Science and Technology, Department of Computer and Information Science, Dec. 1999

[10] 常犁云,王国胤,吴 渝. 一种基于 Rough Set 理论的属性约简及规则提取方法[J]. 软件学报, 1999, 10(11): 1206-1211

[11] 李永敏,朱善君,陈湘晖等. 基于粗糙集理论的数据挖掘模型[J]. 清华大学学报, 1999, 39(1): 110-113

[12] Tsumoto S, Tanaka H. Primerose: probabilistic rule induction method based on rough sets and resampling methods [J]. Computational Intelligence, 1995,11(2): 389-405

Rule generation approach based on rough set theory

AN Li-ping[1], *WU Yu-hua*[1], *TONG Ling-yun*[2]

1. School of Management, Tjanjin University, Tianjin 300072, China;
2. School of Management, Hebei University of Technology, Tianjin, 300130, China

Abstract: Rule generation is an important technology in data mining. In this paper, based on rough set theory, using the concept of generalized decision attribute function or uncertain class, an inconsistent decision system is transformed to one that is consistent as an initial preprocessing step. On the basis of this, by means of discernibility matrix and decision function, a rule generation algorithm is presented. The algorithm can generate all decision rules directly, instead of computing all the reducts of decision system. Furthermore, the obtained rule set using the algorithm keeps all useful information, which is different from that using some other algorithm. In the end, an example illustrates that the algorithm is reasonable and effective.

Key words: rough sets; rule generation; data mining

(上接第12页)

Granger causality analysis of stock markets in China

ZHU Hong-quan, *LU Zu-di*, *WANG Shou-yang*

Institute of Systems Science, Academy of Mathematics and Systems Sciences, Chinese Academy of Sciences, Beijing 100080, China

Abstract: This paper analyzes the auto- and cross- correlations of Hong Kong, Shanghai and Shenzhen stock markets. Based on some econometric techniques, the Granger causality on the cross-correlation and co-movement of the three stock markets is explored. The results show that there are strong auto-correlation, long memory and persistence for the volatility of each of the three stock markets. Hong Kong stock market has little influence on Shanghai and Shenzhen stock markets. Shenzhen stock market has strong Granger causality on Shanghai stock market, but the reversal does not exist.

Key words: return; volatility; granger causality