Vague 决策表的知识获取*)

江 莉 刘三阳 王 珏 陆爱国
(西安电子科技大学应用数学系 西安710071)

摘 要 本文提出一种 Vague 决策表的知识获取方法。首先根据样本对于决策者需求的适合程度构造 Vague 值之间的一个序关系，将 Vague 决策表转化为二元决策表，然后利用粗糙集理论进行分析并推理出最优规则，最后再将二元决策表的决策规则转化为 Vague 决策表的有序规则。实验分析表明了该方法的有效性。

关键词 粗糙集，Vague 决策表，有序规则，规则获取

Knowledge Acquisition in Vague Decision Table

JIANG Li LIU San-Yang WANG Jue LU Ai-Guo
(Department of Applied Mathematics, Xidian University. Xi'an 710071)

Abstract A new approach rough set-based for ordering rules acquisition in vague decision table is presented. An order relation among vague values is constructed according to the degree of suitability that an alternative satisfies the decision-maker's requirement, on the basis of which we transform the vague decision table into the binary decision table. Then optimal rules can be induced using rough set theory. Finally, rules induced in binary decision table are transformed into ordering rules in vague decision table. Simulation results show that the method is effective.

Keywords Rough set, Vague set, Ordering rules, Rule acquisition

1 引言

在现实世界中，人们所面对的信息往往是不精确的、不确定的。为此，Zadeh 于1965年提出了模糊集理论[1]，它用一个隶属函数来描述每个对象属于某个集合的程度，但单个隶属函数不能同时表示支持和反对的证据，为了克服这个不足，Gau 和 Buehrer 于1993年提出了一种新的处理模糊信息的模糊理论—Vague 集[2]。作为描述不确定数据的强有力工具，Vague 集是以模糊集概念为基础的，目前在国外主要应用于模糊控制、决策、故障诊断等方面。Vague 集对模糊信息处理、知识表达等起到了重要的作用。

不确定环境下的知识获取是智能信息处理中的关键问题之一，它在实际决策中是不可避免的。如何在信息不完全、不精确或模糊的情况下，根据决策系统中已有的决策数据获取知识，一直为众多学者所关注。本文提出一种 Vague 决策表的知识获取方法。首先根据样本对于决策者需求的适合程度构造 Vague 值之间的一个序关系，将 Vague 决策表转化为二元决策表，然后利用粗糙集理论进行分析并推理出最优规则，最后再将二元决策表的决策规则转化为 Vague 决策表的有序规则。实验分析表明了该方法的有效性。

2 Vague 集理论基础[3]

定义2.1 令 X 是一个点(对象)的空间，其中的任意一个元素用 x 表示，X 中的一个 Vague 集 V 用一个真隶属函数 t_v 和一个假隶属函数 f_v 表示。$t_v(x)$是从支持 x 的证据所导出的 x 的隶属度下界，$f_v(x)$则是从反对 x 的证据所导出的 x 的否定隶属度下界，$t_v(x)$和 $f_v(x)$将区间[0,1]中的一个实数与 X 中的每一个点联系起来，即

$$t_v: X\to[0,1] \quad f_v: X\to[0,1]$$

其中 $t_v(x)+f_v(x)\leqslant 1$。

由上述定义可知，在 Vague 集中，x 的隶属度被限制在[0,1]上的一个子空间$[t_v(x), 1-f_v(x)]$内。其中 $t_v(x)$是 Vague 集 V 的真隶属函数，表示支持 $x\in X$ 的证据的必要程度，而$1-f_v(x)$则表示了支持 $x\in X$ 的证据的可能程度。这样，关于 x 的不确定性可以用差$(1-f_v(x)-t_v(x))$来表示，如果该差值小，这表明我们相当精确地知道 x；如果该差值大，则表明关于 x 我们知道得很少。

定义2.2 决策表 DT 由四元组 $DT=\langle U, A\cup\{d\}, V, f\rangle$ 组成，其中 U 为决策对象集(论域)；A 为条件属性集；d 为决策属性；$V=\bigcup_{a\in A\cup\{d\}}V_a$ 为属性值域；$f: U\times A\cup\{d\}\to V$ 为决策函数。不分明关系 $IND(A)$和 $IND(d)$的等价类分别称为条件类和决策类。

定义2.3 设 U 为论域，A，d 分别为条件属性与决策属性，$\tilde{v}_a\in\tilde{V}_a$，$\tilde{v}_d\in\tilde{V}_d$，分别为条件属性 $a\in A$ 集决策属性 d 上的取值所对应的 Vague 概念，$\tilde{V}=\bigcup_{a\in A}\tilde{V}_a\cup\tilde{V}_d$ 为决策表中所有属性的 Vague 取值，则称 $VDT=\langle U, A\cup\{d\}, \tilde{V}, \tilde{p}\rangle$为 Vague 决策表(vague decision table)。其中 Vague 决策函数 $\tilde{p}: U\times\tilde{V}\to\{S\mid S\subseteq[0,1]\}$表示将每一决策对象的 Vague 概念映射成 Vague 隶属度。

下面定义一个评分函数来估计样本对于决策者需求的适合程度。

定义2.4[4] 设 $x=[t_x, 1-f_x]$为一 Vague 值，这里 $t_x\in[0,1]$，$f_x\in[0,1]$，$t_x+f_x\leqslant 1$，定义函数 $S(x)$：$S(x)=t_x-f_x$，显然 $S(x)\in[-1,+1]$。

由上述定义可知，评分函数 $S(x)$的值越大，样本对于决

*)基金项目：教育部跨世纪优秀人才基金。江 莉 硕士生研究生，主要研究兴趣：数据挖掘，知识发现等。

策者需求的适合程度就越大，因此，根据 Vague 值的评分函数的大小，我们可以定义 Vague 值之间的一种序关系如下：

定义2.5 设有两个 Vague 值 $x=[t_x,1-f_x]$，$y=[t_y,1-f_y]$，这里 $t_x\in[0,1]$，$f_x\in[0,1]$，$t_x+f_x\leqslant 1$，$t_y\in[0,1]$，$f_y\in[0,1]$，$t_y+f_y\leqslant 1$

如果 $S(x)>S(y)$，那么 $x\succ y$。

如果 $S(x)\leqslant S(y)$，那么 $x\preccurlyeq y$。

此时，我们可以把 Vague 决策表看作有序决策表，进而利用第三部分的方法进行分析和处理，最后得出相应的决策规则。

3 有序信息系统

定义3.1[5]（有序决策表） 通常，一个有序决策表可表示为 $ODT=(U,AT,\{V_a|a\in AT\},\{I_a|a\in AT\},\{\succ_a|a\in AT\})$ 其中，U 是有限个样本的非空集合；AT 是有限个属性的非空集合；V_a 是属性 a 的值域；$I_a:U\to V_a$ 是信息函数；$\succ_a$ 是属性 a 上的一个序关系。

定义3.2 $x\succ_{(a)}y\Leftrightarrow I_a(x)\succ_a I_a(y)$。即 x 优于 y 当且仅当 x 在 a 上的属性值优于 y 在 a 上的属性值。

定义3.3 $x\succ_A y\Leftrightarrow a\in A, I_a(x)\succ_a I_a(y)\Leftrightarrow \bigwedge\limits_{a\in A} I_a(x)\succ_a I_a(y)\Leftrightarrow \bigcap\limits_{a\in A}\succ_{(a)}$

下面我们利用属性值之间的序关系来成对地比较样本，将有序决策表转化为二元决策表：

$$\forall\ a\in AT,\text{令 } I_a((x,y))=\begin{cases}1 & x\succ_{(a)}y\\ 0 & x\preccurlyeq_{(a)}y\end{cases}\quad\text{而且 } x\neq y。$$

在二元决策表中定义一个等价关系 E_A，$\forall\ A\subseteq AT$ 有

$(x,y)E_A(x',y')\Leftrightarrow\forall\ a\in A, I_a((x,y))=I_a((x',y'))$。

根据粗糙集理论的有关知识，按照这个等价关系可将论域划分为互不相交的等价类 $[(x,y)]_A$ 和决策类 $Cl_i(i=0,1)$，从而得到决策类 $Cl_i(i=0,1)$ 的上近似和下近似如下：

$$\underline{apr}(Cl_i)=\bigcup\{[(x,y)]_A|[(x,y)]_A\subseteq Cl_i\}$$

$$\overline{apr}(Cl_i)=\bigcup\{[(x,y)]_A|[(x,y)]_A\cap Cl_i\neq\phi\}$$

其中 $[(x,y)]_A$ 是 (x,y) 关于等价关系 E_A 的等价类。

为了从决策表中抽取得到适应度大的规则，我们需要对二元决策表进行约简，使得经过约简处理的决策表中的一个记录就代表一类具有相同规律特性的样本。下面给出约简的定义：

定义3.4[6] 设 U 为一个论域，P,D 为定义在 U 上的两个等价关系簇，若 P 的 D 独立子集 $S\subset P$ 有 $Pos_S(D)=Pos_P(D)$，则 S 为 P 的 D 约简。

可以记 P 的所有 D 约简关系簇为 $RED_D(P)$。

对于约简后的二元决策表，根据决策类的下近似和上近似分别可以推理出系统的必然性规则和可能性规则，然后将其转化为 Vague 决策表的有序规则，具体方法在第4部分详述。令 $s=$｛论域中满足规则条件的样本集合｝，$t=$｛论域中满足规则结论的样本的集合｝，规则的精度和覆盖度分别按照下式来计算：

$$\text{accuracy}=\frac{|s\wedge t|}{|s|}\qquad\text{coverage}=\frac{|s\wedge t|}{|t|}$$

4 Vague 决策表有序规则获取方法

输入：Vague 决策表 $VDT=(U,A\cup\{e\},\tilde{V},\tilde{p})$。

输出：Vague 决策表的有序规则。

步骤1 计算 Vague 决策表中每个 Vague 属性值的评分函数值，按照定义2.5对不同的样本在不同的属性上进行两两比较。

步骤2 将 Vague 决策表转化为二元决策表。$\forall\ a\in AT$，$u_i,u_j\in U$，如果 $S(u_i)>S(u_j)$，则 $I_a((u_i,u_j))=1$，否则为0。

步骤3 在二元决策表中，条件属性集 A 和决策属性 e 分别对应着一个等价关系 E_A,E_e，按这两个关系对论域进行划分得出等价类 $[(u_i,u_j)]_A$ 和决策类 $Cl_i(i=0,1)$。

步骤4 根据约简的定义对二元决策表进行属性约简和值约简。（一般来讲，一个决策系统的条件属性相对于决策属性的相对约简不是唯一的，这里我们利用包含属性个数最少的约简）

步骤5 求出决策类 $Cl_i(i=0,1)$ 的下近似和上近似，利用粗糙集理论知识推理出二元决策表的具有一定精度和覆盖度的必然性规则和可能性规则。

步骤6 将二元决策表中的决策规则转化为 Vague 决策表的有序规则。

如果 $I_a((x,y))=1$，那么用 $(a,\succ)$ 表示。

如果 $I_a((x,y))=0$，那么用 $(a,\prec)$ 表示。

如果 $I_e((x,y))=1$，那么用 $(e,\succ)$ 表示。

如果 $I_e((x,y))=0$，那么用 $(e,\prec)$ 表示。

5 应用实例

Vague 决策表如表1所示：论域中含有10个样本，$AT=\{a,b,c,d\}\cup\{e\}$ 均为 Vague 值属性，其中 a,b,c,d 为条件属性，e 为决策属性。

表1

\	a	b	c	d	e
1	[0.60 0.70]	[0.80 0.90]	[0.30 0.40]	[0.60 0.80]	[0.40 0.60]
2	[0.10 0.15]	[0.60 0.70]	[0.10 0.20]	[0.30 0.40]	[0.80 0.90]
3	[0.10 015]	[0.30 0.40]	[0.80 0.90]	[0.90 1.00]	[0.20 0.60]
4	[0.00 0.10]	[0.10 0.20]	[0.50 0.60]	[0.30 0.50]	[0.70 0.90]
5	[0.01 0.20]	[0.50 0.60]	[0.30 0.40]	[0.50 0.70]	[0.30 0.50]
6	[0.05 0.15]	[0.40 0.60]	[0.50 0.80]	[0.30 0.50]	[0.90 1.00]
7	[0.11 0.15]	[0.60 0.70]	[0.50 0.60]	[0.90 1.00]	[0.40 0.70]
8	[0.20 1.00]	[0.80 0.90]	[0.90 1.00]	[0.10 0.20]	[0.60 0.80]
9	[0.11 0.15]	[0.60 0.70]	[0.70 0.90]	[0.20 0.50]	[0.40 0.90]
10	[0.10 0.15]	[0.50 0.70]	[0.30 0.60]	[0.60 0.90]	[0.50 0.80]

首先计算表1中10个样本分别关于属性 a,b,c,d,e 的相应的评分函数值，根据评分函数值的大小将表1转化为二元信息系统（限于篇幅，没有给出二元决策系统）。然后根据约简的定义对二元决策系统进行属性约简，得到 $RED=\{a,b\}$。

根据粗糙集理论的有关知识，推出必然性和可能性规则如下：

rule1：$I_a(x,y)=1\wedge I_b(x,y)=1\Rightarrow I_e(x,y)=0\vee I_e(x,y)=1$

accuracy=0.58 OR 0.42 coverage=0.45 OR 0.35

rule2：$I_a(x,y)=1\wedge I_b(x,y)=0\Rightarrow I_e(x,y)=0$

accuracy =1.0 coverage =0.11

rule3：$I_a(x,y)=0\wedge I_b(x,y)=0\Rightarrow I_e(x,y)=1\vee I_e(x,y)=0$

accuracy =0.55 OR 0.45 coverage =0.56 OR 0.43

rule4：$I_a(x,y)=0\wedge I_b(x,y)=1\Rightarrow I_e(x,y)=1\vee I_e(x,y)=0$

accuracy =0.8 OR 0.2 coverage =0.09 OR 0.02

最后将二元表中的决策规则转化为 Vague 决策表的有序规则如下：

（下转第125页）

节所述，领域知识分成属性内部领域知识、分类领域知识和关联领域知识三类。

领域知识将存储在知识库中，相应的表结构和一个实例如表1所示。表中 operand1、operand2和 operator 分别对应第3节给出的领域知识描述中的 A_i、B_i 和 op，对属性内部领域知识和分类领域知识通过 num 标识出有关该属性的、该类别的不同领域知识，由字段 head 标示出其是头(结论)还是体(条件)，同时该字段还可存储 and/or。关联领域知识表达的一组关联属性通过 num 值标识。还可以将集群存储在表中并在属性上创建索引以提高查找领域知识的速度。

表1 领域知识表的一个实例

Attribute	TYPE	Num	Head	Operand1	Operator	Operand2
Salary	interfiled	1	No	Salary	>	10000
Salary	interfiled		Yes	Position	=	Manager
Salary	interfiled	2	No	Salary	≤	10000
Salary	interfiled		Yes	Position	=	Employee
Salary	category	1	No	Salary	>	10000
Salary	category		Yes	Salary	=	"High"
Salary	category	2	No	Salary	>	3000
Salary	category		and	Salary	≤	10000
Salary	category		Yes	Salary	=	"Middle"
Salary	category	3	No	Salary	≤	3000
Salary	category		Yes	Salary	=	"Low"
Salary	correlation	1				
Education	correlation	1				
Position	correlation	1				

(2)利用领域知识进行查询优化改写

当 KDDW 接收到一个知识发现查询时，根据查询中涉及的属性在领域知识表(DKT，Domain Knowledge Table)中查询相关的领域知识，将查询转换成等价的、更高效的另外一种形式。我们将对找到的每一条领域知识做如下处理：

```
select * from DKT where attribute="attribute_name"
for 每条领域知识
switch TYPE
    {"interfiled":若 head="yes"的查询条件是原知识发现查询条件
                  的一个子集，则从原查询语句中去除其余不必要
                  的、冗余的条件；
                  若 head="yes"的查询条件是对原知识发现查询条
                  件的一个有用的补充，则将其加到原查询语句的条
                  件中；
     "category":将关于该属性的所有分类信息显示出来，由用户选
                择(可放弃选择)有用的、有意义的分类，并将选中
                的分类条件加到原查询语句的条件中；
     "correlation":在原知识发现查询语句基础上，增加在与当前属
                   性具有相同 num 值的属性列表上进行投影的运
                   算；
    }
```

(3)新领域知识的积累

对在知识发现系统中新近发现的模式按(1)的方法追加到领域知识表中。

通过上述三种情况的处理，可分别以减少关系内部扫描次数、细化查询条件和削减属性的个数的方式降低数据集合的大小、缩小搜索范围，从而提高知识发现查询的时间开销。

如针对乒乓球运动员的数据库用户提出的查询是"找出那些世界排名进入前16或是世界级比赛的个人最好成绩进入前8名的运动员"，假设我们具有前面例1中给出的领域知识，则按上述处理方法可以去除其中一个查询条件或是用"技术等级标准="国际级运动健将""替换原查询。在例3中可按年龄的四种分类或收入的三种分类将"消费者的消费模式是什么?"的查询细化。在例4中对与收入有关的模式查询，通过投影运算将教育程度和职位以外的、与收入无关的属性去除。

通过以上描述和例子分析不难看出，运用我们的方法可以在查找与某一个查询相关的领域知识时节省大量的时间。

总结 本文介绍了一种在数据仓库的知识发现查询中利用领域知识进行优化处理、缩小搜索范围、使发现的数据更合乎用户要求的方法。下一步，还需要寻找某种启发式的规则以挑选最有用的领域知识，另外如何衡量每条领域知识的质量也是一个有趣的研究课题。

参考文献

1 Hummergren T. 数据仓库技术. 曹增强，王备战，岳晓奎等译. 北京：中国水利水电出版社，1998

2 Piatetsky-Shapiro G, Frawley W J. Knowledge Discovery in Databases. AAAI/MIT Press, 1991. 229～248

3 Kaufman K A, et al. Mining for Knowledge in Database: Goals and General Description of the INLEN System, Knowledge Discovery in Database. In: G. Piatetsky, W. J. Frawley, eds. AAAI/MIT Press, 1991. 449～462

4 Han J, et al. Data-Driven Discovery of Quantitative Rules in Relational Databases. IEEE Transactions on Knowledge and Data Engineering, 1993, 5(1)

5 Charkravarthy US, et al. Logic-based Approach to Semantic Query Optimization. ACM Transactions on Database Systems, 1990, 15(2): 162～207

6 Yoon S-C, et al. Using Domain Knowledge in Knowledge Discovery. In: Proc. of CIKM'99, ACM Press, 1999. 243～250

(上接第112页)

rule1': $(a,\succ)\wedge(b,\succ)\Rightarrow(e,\prec)\vee(e,\succ)$

accuracy =0.58 OR 0.42 coverage =0.45 OR 0.35

rule2': $(a,\succ)\wedge(b,\prec)\Rightarrow(e,\prec)$

accuracy =1.0 coverage =0.11

rule3': $(a,\prec)\wedge(b,\prec)\Rightarrow(e,\succ)\vee(e,\prec)$

accuracy =0.55 OR 0.45 coverage =0.56 OR 0.43

rule4': $(a,\prec)\wedge(b,\succ)\wedge(b,\prec)\Rightarrow(e,\succ)\vee(e,\prec)$

accuracy =0.8 OR 0.2 coverage =0.09 OR 0.02

结语 本文以采用 Vague 值属性描述的决策系统为例，给出了有效的 Vague 值属性决策系统的知识获取方法。首先根据样本对于决策者需求的适合程度构造 Vague 值之间的一个序关系，将 Vague 决策表转化为二元决策表，然后利用粗糙集理论进行分析并推理出最优规则，最后再将二元决策表的决策规则转化为 Vague 决策表的有序规则。实验分析表明了该方法的有效性。

参考文献

1 Zadeh L A. Fuzzy Sets. Information and Control, 1965, 8(3): 338～353

2 Gau Wen-Lung, Danied J B. Vague Sets. IEEE Transactions on Systems, Man, and Cybernetics, 1993, 23(2): 610～614

3 李凡，徐章艳，饶勇. Vague 集. 计算机科学，2000，27(9)：12～14

4 Chen S M, Tan J M. Handling Multicriteria Fuzzy Decision-making Poblems Based on Vague Set Theory. Fuzzy Sets and Systems, 1994, 67: 163～172

5 Yao Y Y, Sai Y. Mining Ordering rules using rough set theory. Bulletin of International Rough Set Society, 2001, 5: 99～106

6 王国胤. Rough 集理论与知识获取. 西安交大出版社，2001