区间值属性决策表的数据挖掘*)

王 珏 刘三阳 张 杰
(西安电子科技大学理学院 西安710071)

Data Mining in Interval Valued Decision Table

WANG Jue LIU San-Yang ZHANG Jie
(School of Science, Xidian University. Xi'an. 710071)

Abstract Data mining in incomplete information systems is a hard problem but inevitable in uncertain decision. In this paper, an extended rough set model based on dominance relation is combined with fuzzy set theory for data mining in interval valued decision table, then decision rules can be obtained from the decision table. Simulation results show that the method is effective.

Keywords Rough set, Dominance relation, Interval number, Data mining

1. 引言

数据挖掘技术[1]是信息系统的一个重要研究内容。数据挖掘是指从大型数据库中挖掘出先前未知的、有效的、可实用的信息，并使用这些信息作出决策或丰富知识。数据挖掘与传统的数据分析的本质区别是它在没有明确假设的前提下去挖掘信息并发现知识，其目的在于从大量数据中发现令人感兴趣的规则，一般地讲，这些规则在表现形式上应比较简洁，并且具有一定程度的概括性。同时，在实际问题中，待处理的数据常有某种程度的不完备，即我们得到的表征事物行为特征的数据往往不是一些确切的数，而是一些区间数。当决策表中含有区间值属性时，传统的粗糙集理论将无能为力。Greco, Matarazzo 和 Slowinski 等[2]提出一种新的粗糙集的扩展模型，用优势关系(dominance relation)代替原来的不可分辨关系(indiscernbility relation)，在多准则决策分析中得到广泛的应用。本文利用这种扩展粗糙集模型，结合模糊数排序方法，给出了一种区间值属性决策表的数据挖掘方法，该方法能有效地挖掘出决策系统的决策规则。

2. 基于优势关系的粗糙集模型[3]

通常，一个信息表 S 可以表示为 $S=\langle U,Q,V,f\rangle$，这里 U 是论域，即有限个对象的集合，Q 是属性集，$V=\bigcup_{q\in Q}V_q$，且 V_q 是属性的值域，$f:U\times Q\to V$ 是信息函数，使得 $f(x,q)\in V_q, q\in Q, x\in U$。决策表是一类特殊而重要的知识表达系统，也是一种特殊的信息表，通常表示为 $S'=\langle U,R,V,f\rangle$，这里属性集 $R=C\cup D$ 且 $C\cap D=\Phi$。C 称为条件属性集，D 称为决策属性集。

传统粗糙集是以不可分辨关系 I_q 为基础的，$I_q=\{(x,y)\in U\times U: q(x)=q(y), q\in C\}$，不可分辨关系是一种等价关系，即满足自反性、对称性、传递性。按不可分辨关系划分论域可以得到互不相交的等价类。Greco 等对传统的粗糙集模型进行扩展，以优势关系代替原先的不可分辨关系，提出了一种新的粗糙集模型。

下面考虑区间值属性与质量型属性复合的决策表。

定义 C' 为区间值属性的集合，C'' 为质量型属性的集合。显然 $C'\cup C''=C$ 而且 $C'\cap C''=\phi$。另外，对于 C 的任意子集 $P\subseteq C$，定义 P' 为区间值属性的集合，P'' 为质量型属性的集合，即 $P'=P\cap C', P''=P\cap C''$。

令 $\geq_q$ 为论域 U 上的一个弱偏序关系。$x\geq_q y$ 是指在属性 q 上 x 至少和 y 一样好。下面定义一种优势关系 D_P：$\forall x,y\in U, xD_Py$ 如果 $\forall q\in P$，有 $x\geq_q y$。

对于既含有区间值属性又含有质量属性的复合决策表，定义一种优势-等价关系 R_P：$\forall x,y\in U, xR_Py$ 如果 $\forall q\in P'$，$x\geq_q y$ 而且 $\forall q\in P'', xI_qy$。

这个优势-等价关系是自反的、传递的，但不是对称的。所以论域按这个优势-等价关系 R_P 来划分，得到了 $|U|$ 个优势-等价类 $R_P^+(x)$ 和 $|U|$ 个逆优势-等价类 $R_P^-(x)$：

$$R_P^+(x)=\{y\in U: yR_Px\}$$

$$R_P^-(x)=\{y\in U: xR_Py\}$$

它们一般不构成对论域的划分，而是构成了对论域的覆盖。

决策属性对论域划分得到决策类 $Cl=\{Cl_t, t\in T\}$，$T=\{1,2,\cdots,n\}$，对所有的 $r,s\in T$，且 $r>s$，定义 $[x\in Cl_r, y\in Cl_s, r>s]\Rightarrow[xSy$ 而不是 $ySx]$，xSy 是指 x 至少和 y 一样优。由于决策类的这种偏序关系，我们定义 $Cl_t^{\geq}$ 为属于 Cl_t 类及优于 Cl_t 类的对象的集合，而 $Cl_t^{\leq}$ 为属于 Cl_t 类及劣于 Cl_t 类的对象的集合，即 $Cl_t^{\geq}=\bigcup_{s\geq t}Cl_s, Cl_t^{\leq}=\bigcup_{s\leq t}Cl_s, t=1,2,\cdots,n$。显然，$Cl_1^{\geq}=Cl_n^{\leq}=U, Cl_n^{\geq}=Cl_n, Cl_1^{\leq}=Cl_1$。

通常，$\forall t\in T, \forall P\subseteq C$，定义 $Cl_t^{\leq}$ 关于属性集 P 的下、上近似分别为：

$$\underline{P}Cl_t^{\leq}=\{x\in U: R_P^-(x)\subseteq Cl_t^{\leq}\}$$

$$\overline{P}Cl_t^{\leq}=\bigcup_{x\in Cl_t^{\leq}}R_P^-(x)$$

定义 $Cl_t^{\geq}$ 关于属性集 P 的下、上近似分别为：

$$\underline{P}Cl_t^{\geq}=\{x\in U: R_P^+(x)\subseteq Cl_t^{\geq}\}$$

$$\overline{P}Cl_t^{\geq}=\bigcup_{x\in Cl_t^{\geq}}R_P^+(x)$$

因此，决策类的边界为：

*)基金项目：国家自然科学基金项目(69972036)；陕西省自然科学基金项目(2000SL03)。王 珏 博士生，主要研究兴趣：数据挖掘、知识发现等。

$Bn_P(Cl_t^{\geq})=\overline{P}Cl_t^{\geq}-\underline{P}Cl_t^{\geq}$，

$Bn_P(Cl_t^{\leq})=\overline{P}Cl_t^{\leq}-\underline{P}Cl_t^{\leq}$，

对于$\forall\ t\in T,\forall\ P\subseteq C$，我们定义决策类 Cl 关于属性集 P 的近似分类质量：

$$\gamma_P(Cl)=\frac{|U-((\bigcup_{t\in T}Bn_P(Cl_t^{\leq}))\bigcup(\bigcup_{t\in T}Bn_P(Cl_t^{\geq})))|}{|U|}$$

属性集 C 的最小子集 $P\subseteq C$ 称为决策表的一个约简，当且仅当 $\gamma_P(Cl)=\gamma_C(Cl)$。需要注意的是，一个信息表可能有多个约简。

3. 模糊区间数的排序方法[4]

从模糊集的定义及其性质中知道，区间数之间的顺序关系不是通常意义下的全序关系，而是格结构下的偏序关系。

定义1 设二区间数 A=[a,b],B=[c,d]，考虑下面五种情况：

(1)当 a=c 时

(i)b=d，则 A 与 B 相等。

(ii)b>d，则 A 优于 B。

(iii)b<d，则 A 非优于 B。

(2)当 b=d 时

(i)a=c 则 A 与 B 相等。

(ii)a>c，则 A 优于 B。

(iii)a<c，则 A 非优于 B。

(3)当 a<c 且 b<d 时，B 严格优于 A。

(4)当 a>c 且 b>d 时，A 严格优于 B。

(5)当 a<c<d<b 时，

(i)b−d=c−a，则 A 与 B 相等。

(ii)b−d>c−a，则 A 优于 B。

(iii)b−d<c−a，则 A 非优于 B。

因此，我们可以把这种区间值的偏序关系作为一种优势关系，利用扩展的粗糙集模型来挖掘决策信息系统的决策规则。下面给出区间值属性决策表数据挖掘方法的具体步骤。

4. 区间值属性决策表的数据挖掘方法

这里我们只考虑区间值属性与质量型属性复合的决策表。

输入：一个决策系统 $S=\langle U,C,V,f\rangle$，其中 U 为论域，C 为属性集，C' 为区间值属性的集合，C''为质量型属性的集合。

输出：该决策表的最小决策规则。

步骤1：根据决策表属性的类型，利用优势--等价关系划分整个论域 U，得出相应的优势--等价类和逆优势--等价类。

步骤2：根据决策类并的定义计算 $Cl_t^{\leq}$、$Cl_t^{\geq}$。

步骤3：计算 $Cl_t^{\leq}$、$Cl_t^{\geq}$ 的关于属性集 C 的上、下近似和边界。

步骤4：计算 $Cl_t^{\leq}$、$Cl_t^{\geq}$ 的关于属性集 C 的近似分类质量 γ，求出决策表的约简。(可能不唯一)

步骤5：根据决策类并 $Cl_t^{\leq}$，$Cl_t^{\geq}$ 的粗糙近似，提取决策表的决策规则如下三种形式：

(1)$D_{\geq}$决策规则

如果 $f(x,q_1)\geq r_{q_1},\cdots,f(x,q_u)\geq r_{q_u}$ 而且 $f(x,q_{u+1})=r_{q_{u+1}},\cdots,f(x,q_v)=r_{q_v}$，那么 $x\in Cl_t^{\geq}$。其中 $P=\{q_1,\cdots,q_v\}\subseteq C$，$P'=\{q_1,\cdots,q_u\}$，$P''=\{q_{u+1},\cdots,q_v\}$。

$(r_{q_1},\cdots,r_{q_v})\in V_{q_1}\times\cdots,V_{q_v}$，$t\in T$。

(2)$D_{\leq}$决策规则

如果 $f(x,q_1)\leq r_{q_1},\cdots,f(x,q_u)\geq r_{q_u}$ 而且 $f(x,q_{u+1})=r_{q_{u+1}},\cdots,f(x,q_v)=r_{q_v}$，那么 $x\in Cl_t^{\leq}$

其中 $P=\{q_1,\cdots,q_v\}\subseteq C$，$P'=\{q_1,\cdots,q_u\}$，$P''=\{q_{u+1},\cdots,q_v\}$。

$(r_{q_1},\cdots,r_{q_v})\in V_{q_1}\times\cdots,V_{q_v}$，$t\in T$。

(3)$D_{\geq\leq}$决策规则

如果 $f(x,q_1)\leq r_{q_1},\cdots,f(x,q_u)\leq r_{q_u}$ 而且 $f(x,q_{u+1})\leq r_{q_{u+1}},\cdots,f(x,q_v)\leq r_{q_v}$，而且 $f(x,q_{v+1})=r_{q_{v+1}},\cdots,f(x,q_z)=r_{q_z}$，那么 $x\in Cl_s\cup Cl_{s+1}\cup\cdots,Cl_t$。

其中 $P=\{q_1,\cdots,q_z\}\subseteq C$，$P'=\{q_1,\cdots,q_u\}\cup\{q_{u+1},\cdots,q_v\}$，$P''=\{q_{v+1},\cdots,q_z\}$。

$s,t\in T,s<t$

5. 应用算例

表1是一个含有12个对象和5个属性的决策表，其中 A_1，A_2，A_3，A_5为区间值属性，A_4为质量型属性。A_1，A_2，A_3，A_4为条件属性，A_5为决策属性。

表1

	A_1	A_2	A_3	A_4	A_5
1	[3,9]	[6,9]	[1,5]	N	[1,3]
2	[3,6]	[6,9]	[1,5]	N	[1/5,1]
3	[3,6]	[6,9]	[1,5]	N	[1,3]
4	[1/5,1]	[2,4]	[1,5]	N	[1/5,1]
5	[3,6]	[2,4]	[4,9]	N	[1/5,1]
6	[3,9]	[2,4]	[4,9]	N	[1,3]
7	[3,6]	[2,4]	[1,5]	N	[1,3]
8	[3,9]	[6,9]	[1,5]	P	[1,3]
9	[3,6]	[6,9]	[1,5]	P	[1,3]
10	[1/5,1]	[2,4]	[1,5]	P	[1/5,1]
11	[3,6]	[2,4]	[4,9]	P	[1,3]
12	[3,9]	[2,4]	[4,9]	P	[1,3]

因为表1是一个复合决策表，所以利用前文介绍的优势-等价关系来划分论域 $U=\{1,2,\cdots,12\}$得到优势-等价类及逆优势-等价类如下：

R_P^+：{1}{1,2,3}{1,2,3}{1,2,3,4,5,6,7}{5,6}{6}{1,2,7}{8}{8,9}{8,9,10,11,12}{11,12}{12}。

R_P^-：{1,2,3,4,7}{2,3,4,7}{2,3,4,7}{4}{5,7}{4,5,6,7}{4,7}{8,9,10}{9,10}{10}{10,11}{11,12}。

因为只有两个决策类，所以 $Cl_1^{\leq}=Cl_1=\{2,4,5,10\}$，$Cl_2^{\geq}=Cl_2=\{1,3,6,7,8,9,11,12\}$，根据 $Cl_t^{\leq}$，$Cl_t^{\geq}$ 的上下近似定义有：

$\underline{C}Cl_1^{\leq}=\{4,10\}$ $\overline{C}Cl_1^{\leq}=\{2,3,4,5,7,10\}$ $Bn(Cl_1^{\leq})=\{2,3,5,7\}$

$\underline{C}Cl_2^{\geq}=\{1,6,8,9,11,12\}$ $\overline{C}Cl_2^{\geq}=\{1,2,3,5,6,7,8,9,11,12\}$ $Bn(Cl_2^{\geq})=\{2,3,5,7\}$

因此，根据分类质量的定义有 $\gamma=0.67$。

根据约简的定义可知，表1只有一个约简$\{A_1,A_4\}$。所以，可提取决策表1的决策规则如下：

rule1：如果 $f(x,A_1)\geq[3,9]$，那么 $X\in Cl_2^{\geq}$。

rule2：如果 $f(x,A_1)\geq[3,6]$，而且 $f(x,A_4)=P$，那么 $X\in Cl_2^{\geq}$。

rule3：如果 $f(x,A_1)\leq[1/5,1]$，那么 $X\in Cl_1^{\leq}$。

rule4：如果 $f(x,A_1)=[3,6]$，而且 $f(x,A_4)=N$，那么 $X\in Cl_1^{\leq}$或 $X\in Cl_2^{\geq}$。

结语 信息不完备是复杂决策环境中不可避免的问题[5]。由于不完备决策表中的不确定成分太多，因此从不完备决策表中挖掘出潜在的有用知识，增加了数据挖掘的难度。传统的粗糙集理论是处理不确定信息的有效方法，但它仅实用于完全决策表的情况，因此在应用中受到限制。本文利用基于优势-等价关系的扩展粗糙集模型，结合模糊集理论知识，给出了有效的区间值属性决策表的数据挖掘方法。算例证明了该方法的有效性。对于区间值属性与数量型属性复合的决策表，以及区间值属性与数量型属性、质量型属性三种情况复合的决策表，可用类似的方法来处理。

参考文献

1 李永敏，朱善君，陈湘晖，等. 基于粗糙集理论的数据挖掘模型. 清华大学学报(自然科学版)，1999，39(1)

2 Greco S，Matarazzo B，Slowinski R. Rough Approximation of a Preference relation by dominance relations. Eropean Journal of Operational Research，1997，117：63～83

3 Greco S，Mataarazzo B，Slowinski R. Rough Approximation by Dominance Relations. International Journal of Intelligent Systems，2002，17：153～171

4 张全，樊治平，潘德惠. 不确定多属性决策中区间数的一种排序方法. 系统工程理论与实践，1999(5)：129～133

5 赵卫东，李旗号，盛昭翰. 区间值属性不完全信息下的数据挖掘. 系统工程理论与方法，2001，10(2)

(上接第120页)

本文的算法运行时间和存储空间的增长分别如图7和图8所示(由于文[7]未提供对应数据，我们实现了 FP-Growth 算法。我们实现的 FP-Growth 算法运行效率与文[7]报告的结果大致相当，我们的实现稍快一点。例如，对于数据集 T100K，当最小支持度阈值为1%时，文[7]报告的运行时间大约为26秒，而我们的实现为22.18秒。图6～图8的比较涉及的 FP-Growth 算法是基于我们的实现)。可以看出，随着数据集的增大，两个算法的运行时间和存储空间都线性地增长。然而，LIFPG 算法增长的速度比较缓慢，并且时空性能始终优于 FP-Growth 算法。这表明与 FP-Growth 算法相比，本文的算法具有更好的可伸缩性。

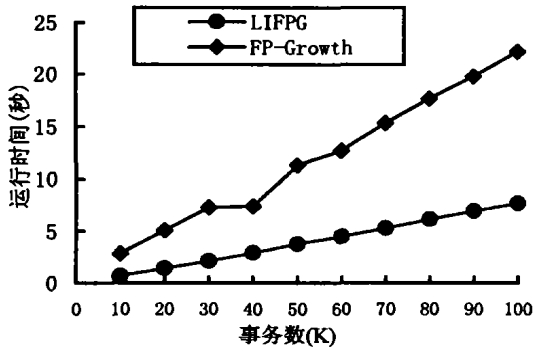

图7 算法的时间可伸缩性(最小支持度1%)

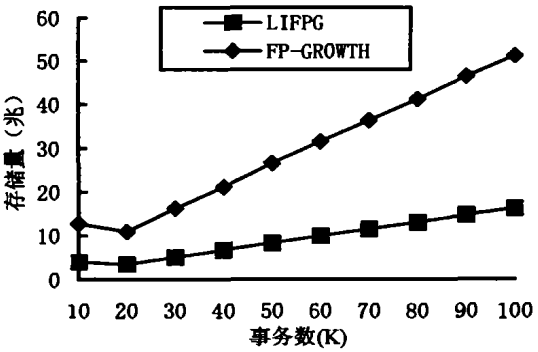

图8 算法的空间可伸缩性(最小支持度1%)

结束语 我们提出了一种直接在 Trans-树中挖掘频繁模式的有效算法 LIFPG 算法，实现了该算法，并研究了它的性能。我们的实验表明，与 FP-Growth 算法相比，LIFPG 算法的挖掘速度大约提高4倍，而存储开销节省一半。对于稠密数据，速度的提高更为显著。随着数据库容量的增长，本文的算法具有更好的可扩展性。

参考文献

1 Agrawal R，Srikant R. Fast algorithms for Mining association rules. In：Proc. 1994 Int'l Conf. on Very Large Data Bases，Sept. 1994. 487～499

2 Park J S，Chen M S. Yu P S. An effective hash-based algorithm for mining association rules. In：Proc. 1995 ACM-SIGMOD Int'l Conf. on Management of Data，May 1995. 175～186

3 Brin S，Motwani R，Silverstein C. Beyond market basket：Generalizing association rules to correlations. In：Proc. 1997 ACM-SIGMOD Int'l Conf. on Management of Data，May 1997. 265～276

4 Agrawal R，Srikant R. Mining sequential patterns，In ICDE'95，pages 3～14

5 Dong G，Li J. Efficient mining of emerging patterns：Discovering trends and differences. In：Proc. of the fifth ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining，Aug. 1999. 43～52

6 Han Jiawei，Kamber M 著，范明，孟小峰等译. 数据挖掘：概念与技术. 机械工业出版社，2001. 149～184

7 Han J，Pei J，Yin Y. Mining frequent patterns without candidate generation. In：Proc. 2000 ACM-SIGMOD Intl. Conf. on Management of Data，May 2000. 1～12

8 http：//www. ics. uci. edu/～mlearn/MLRepository. html

9 Bykowski A，Rigotti C. A Condensed Representation to Find Frequent Patterns. In：Proc. of the 20th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS 2001)，Santa Barbara，CA，USA，ACM Press，2001. 267～273