◎数据库、信号与信息处理◎

个性化决策规则获取算法及规则表示

申锦标,吕跃进
SHEN Jin-biao,LV Yue-jin

广西大学 数学与信息科学学院,南宁 530004
School of Mathematics and Information Science,Guangxi University,Nanning 530004,China
E-mail:sjinbiao@163.com

SHEN Jin-biao,LV Yue-jin.Decision rules acquisition algorithm for personalized knowledge based on rough set and rule presentation.Computer Engineering and Applications,2009,45(14):122-124.

Abstract: This paper presents a new decision rules acquisition algorithm for personalized knowledge based on rough set and rule presentation.The correctness of the algorithm is proved in theory,and the algorithm of extracting personalized knowledge is described.The main parts of the algorithm are how to compose rules,and how to calculate certainty factor,coverage factor and the strength of decision rules.At last,the efficiency and practicability of the algorithm is illustrated by one example.
Key words: rough set;personalized knowledge discovery;decision rules

摘 要:提出了一种基于粗糙集理论的面向个性化知识的决策规则获取算法。从理论上证明了算法的正确性,给出了面向个性化的知识获取算法的描述。算法的重点在于规则合成的方法和可信度、覆盖度和规则强度计算的方法。最后通过例子说明了算法的有效性和实用性。
关键词:粗糙集;个性化知识发现;决策规则
DOI:10.3778/j.issn.1002-8331.2009.14.037 **文章编号**:1002-8331(2009)14-0122-03 **文献标识码**:A **中图分类号**:TP301

1 引言

知识发现是指从海量数据中抽取新颖的、有趣的知识的过程。它揭示了数据内部的一种本质和客观的联系和规律。但是这种揭示可能是全方位的,多方面的。对于用户来说,如何从海量的知识中提取真正感兴趣的、符合其实际需要的知识(称为个性化知识)是一个挑战性的研究课题。

Pawlak 提出的粗糙集理论[1]是智能数据分析和数据挖掘的一种新的数学方法,它在知识生成和规则提取等方面有着很强的优势。它已经成为机器学习、知识获取、决策分析、模式识别等领域重要的基本理论[2-3]。一般来说,从决策系统中发掘知识的过程可分为两步:首先是对属性进行约简,然后找出决策规则(知识)。属性的约简方法已有许多学者进行了研究,而关于个性化知识发现,虽已经有了一些初步的探讨[4-6],但在面向个性化知识提取算法方面的研究还没有引起足够的重视。

在文献[7]基础上进一步研究,给出了一种新的面向个性化知识发现的规则提取算法和规则表示方法。在理论上证明了在多决策属性下算法的正确性,给出并证明了合成规则的可信度、覆盖度和规则强度的计算公式。其次,给出了面向个性化的知识发现算法的描述和规则的表示方法。最后,通过例子说明了算法的有效性和实用性。

2 一些基本概念

在粗糙集中,决策表可以表示为 $S=(U,A,V,f)$,其中:U 是非空有限对象集(称为论域)。A 为有限非空属性集,$A=C\cup D$,$C\cap D\neq\Phi$,其中 C 为条件属性集,D 为决策属性集。$V=\bigcup_{a\in A}V_a$ 是 A 中属性的值域,V_a 是属性 a 的值域。f 是映射,满足 $f_a:U\to V_a$,$\forall a\in A$。可以根据条件属性和决策属性对论域进行划分,论域中的对象根据条件属性的不同。被划分到具有不同决策属性的决策类中。

一般地,属性值序对 (a,v),$a\in A$,$v\in V_a$,被称为原子属性。任何原子属性或者它们的逻辑联合体被称为描述,满足描述 t 的所有对象的集合记为 $\|t\|$。如果 t 和 s 是两个描述,则有 $\|t\wedge s\|=\|t\|\cap\|s\|$ 和 $\|t\vee s\|=\|t\|\cup\|s\|$,其中 $\wedge$(或 $\vee$)表示逻辑合取(析取)运算符。

在决策表 $S=(U,A,V,f)$ 中,决策规则的广义表示形式为 $t\to s$,即:

$$t=\bigwedge_{a\in C'\subseteq C}(a,v)\to s=\bigvee\left(\bigwedge_{d\in D}(d,w)\right)$$

基金项目:广西自然科学基金(the Natural Science Foundation of Guangxi of China under Grant No.桂科自 0991027);广西教育厅面上项目(No.200707MS061)。
作者简介:申锦标(1972-),男,硕士,主研方向:数据挖掘、粗糙集;吕跃进(1958-),男,教授,硕士生导师,主研方向:数据挖掘、粗糙集与概念格。
收稿日期:2008-03-17 **修回日期**:2008-06-06

其中,C 是规则的条件部分的所有属性的集合,$C'\subseteq C$,$w\in V_d$。t 和 s 分别称为规则的条件部分和决策部分。带有唯一决策值的规则被称为明确的,否则,称为不明确的。并且,一个元素 $x\in U$ 支撑规则 $t\to s$ 被定义为 x 在 A 中同时满足描述 t 和 s,即 $x\in \|t\wedge s\|$。

如果一个决策规则 $t\to s$ 是明确的,并且满足 $\|t\|\subseteq\|s\|$ 时,称规则 $t\to s$ 是确定的;如果一个决策规则 $t\to s$ 是明确的,同时满足 $\|t\|\cap\|s\|\neq\Phi$ 时,并且 $\|t\|\not\subset\|s\|$ 时,称规则$t\to s$ 是不确定的(或可能的)规则。显然,决策表的基本规则只有确定性和不确定性两类。

对于决策 $S=(U,A,V,f)$表中的每一条规则 $t\to s$

$$Cer(t\to s)=\frac{card(\|t\wedge s\|)}{card(\|t\|)}$$

$$Cov(t\to s)=\frac{card(\|t\wedge s\|)}{card(\|s\|)}$$

$$\sigma(t,s)=\frac{card(\|t\wedge s\|)}{card(U)}$$

分别称为规则 $t\to s$ 的可信度、覆盖度和规则强度。

在本文中使用的规则的一般形式为:$t\to s(Cer(t\to s),Cov(t\to s),\sigma(t,s))$。

3　算法的理论依据

定义 1　设 $S=(U,A=C\cup D,V,f)$是一个决策表。令 X_i 和 Y_j 分别代表 U/C 与 U/D 中的各个等价类,$des(X_i)$表示对等价类 X_i 的描述,$des(Y_j)$表示对等价类 Y_j 的描述,如果 $\|des(X_i)\|\cap\|des(Y_j)\|\neq\Phi$,则称 $r_{ij}:des(X_i)\to des(Y_j)$为决策表 S 的一个基本决策规则(以下简称基本规则)。为了表述方便,记 $t_i=des(X_i)$,$s_j=des(Y_j)$。

定义 2　设 $r_{ij}:t_i\to s_j$ 是 n 个互不相同的基本规则,如果存在描述 t、s 满足:

$$(1)\ \|t\|=\bigcup_{i=i,1,\cdots,n}\|t_i\|$$

$$(2)\ \|s\|=\bigcup_{j=i,1,\cdots,n}\|s_j\|$$

则称规则 $t\to s$ 可由 $r_{ij}:t_i\to s_j$ 生成。

定理 1　决策表中的任一规则都可由若干个基本规则生成。

证明　由文献[7]知,如果规则是明确的,结论成立。若规则不是明确的,设规则为

$$t=\bigwedge_{a\in C'\subseteq C}(a,v)\to s=\vee(\bigwedge_{d\in D}(d,w))$$

上式可以表示为形如 $t=\bigwedge_{a\in C'\subseteq C}(a,v)\to s=\bigwedge_{d\in D}(d,w)$规则的析取,易知 $t=\bigwedge_{a\in C'\subseteq C}(a,v)\to s=\bigwedge_{d\in D}(d,w)$是明确的,故结论成立。证毕。

定理 2　设 $S=(U,A,V,f)$是一个决策表,$A=C\cup D$,$C\cap D\neq\Phi$,其中 C 为条件属性集,D 为决策属性集,规则 $t\to s$ 由 $t_i\to s_i$,$i=1,2,\cdots,n$ 是 n 个互不相同的基本规则生成,则

$$Cer(t\to s)=\frac{\sum_i\sum_j card(\|t_i\wedge s_j\|)}{card(\cup\|t_i\|)}$$

$$Cov(t\to s)=\frac{\sum_i\sum_j card(\|t_i\wedge s_j\|)}{card(\cup\|s_j\|)}$$

$$\sigma(t,s)=\frac{\sum_i\sum_j card(\|t_i\wedge s_j\|)}{card(U)}$$

证明　由规则 $t\to s$ 是 n 个互不相同的基本规则 $t_i\to s_j$ 生成,则有

$$\|t\|\cap\|s\|=(\bigcup_i\|t_i\|)\cap\|s\|=\bigcup_i(\|t_i\|\cap\|s\|)=$$

$$\bigcup_i(\|t_i\|\cap(\bigcup_j\|s_j\|))=$$

$$\bigcup_i(\bigcup_j(\|t_i\|\cap\|s_j\|))=$$

$$\bigcup_i\bigcup_j(\|t_i\|\cap\|s_j\|)$$

又因为

$$\|t\wedge s\|=\|t\|\cap\|s\|==\bigcup_i\bigcup_j(\|t_i\|\cap\|s_j\|)=$$

$$\bigcup_i\bigcup_j(\|t_i\wedge s_j\|)$$

进而,有

$$card(\|t\wedge s\|)=card(\bigcup_i\bigcup_j(\|t_i\|\cap\|s_j\|))=$$

$$\sum_i\sum_j card(\|t_i\wedge s_j\|)$$

由上式及规则 $t\to s$ 的可信度、覆盖度和规则强度的定义,可得:

$$Cer(t\to s)=\frac{card(\|t\wedge s\|)}{card(\|t\|)}=\frac{\sum_i\sum_j card(\|t_i\wedge s_j\|)}{card(\cup\|t_i\|)}$$

$$Cov(t\to s)=\frac{card(\|t\wedge s\|)}{card(\|s\|)}=\frac{\sum_i\sum_j card(\|t_i\wedge s_j\|)}{card(\cup\|s_j\|)}$$

$$\sigma(t,s)=\frac{card(\|t\wedge s\|)}{card(U)}=\frac{\sum_i\sum_j card(\|t_i\wedge s_j\|)}{card(\cup\|U\|)}$$

定理表明,任一规则可由基本规则合成,合成规则的可信度、覆盖度、规则强度可通过基本规则中的 $\|t_i\|$,$\|s_j\|$,$t\|t_i\wedge s_j\|$ 来计算。

4　决策规则合成算法描述

4.1　决策规则的表示方法

为了利用数据库管理系统的强大功能,用关系数据库的方法表示决策表的基本规则[8]。由定理 2,为了减少数据的冗余,结合基本规则的数据特点,用三个相关的数据表表示基本规则,分别称为条件分类表、决策分类表和规则表。表结构分别为:条件分类表(条件号,条件属性 1,条件属性 1,…,$\|t_i\|$ 的值)、决策分类表(决策号,决策属性 1,决策属性 1,…,$\|s_j\|$ 的值)、规则表(条件号,决策号,$\|t_i\|\cap\|s_j\|$ 的值)。

每一个 U/C 等价类对应条件分类表中一条记录,每一个 U/D 等价类对应决策分类表中一条记录,每一基本规则对应规则表中一条记录。

4.2　算法描述

为了实现个性化知识规则提取,用户决策规则合成的总体思路是根据用户提供的信息,利用基本规则集(基本知识库),生成相关的 3 个数据表。然后使用 SQL 语言检索符合条件的

基本决策规则:根据输入的条件属性,在条件分类表中检索所有符合条件的描述s_i、求$\sum card(\|s_i\|)$并保存条件号的值,根据输入的决策属性,在决策分类表中检索所有符合条件的描述s_j、求$\sum card(\|s_j\|)$并保存决策号的值;在决策规则表中检索条件号和决策号为上述保存的值的所有记录,求$\sum card(\|t_i\| \cap \|s_j\|)$,最后根据规则表生成规则并由定理 2 中的公式计算合成规则的可信度、覆盖度、规则强度。

由上述可知,根据用户给定的条件。使用 SQL 语言检索符合条件的决策规则,进行相应的规则生成,可以从前提(给定的设计条件)推导出目标(决策),这是正向推理;也可以从目标推导出前提,这是反向推理。

算法

输入:决策表和基本规则集(基本知识库)KB;用户感兴趣的一组属性和属性值;

输出:用户感兴趣的决策规则集(用户知识库)。

步骤 1 根据基本规则集生成 3 个相关的基本规则数据表,详情参见例子。

步骤 2 使用 SQL 语言检索符合条件的基本决策规则并计算:$\sum card(\|s_i\|)$、$\sum card(\|s_j\|)$、$\sum card(\|t_i\| \cap \|s_j\|)$。

步骤 3 对符合条件的基本决策规则进行相应的规则生成。

步骤 3.1 合成规则:根据用户输入的条件对基本规则进行合取或析取;

步骤 3.1 由定理 2 中的公式计算合成规则的可信度、覆盖度、规则强度。

步骤 4 显示结果。

5 例子

以决策表表 1 为例,说明算法的具体操作。

表 1 决策表

	C				D	
U	a	b	c	d	f	g
1	1	2	1	1	1	1
2	1	2	1	1	1	2
3	1	2	1	1	1	2
4	1	2	1	1	1	1
5	1	1	2	1	2	1
6	2	1	2	1	2	2
7	1	1	2	1	2	2
8	1	1	2	2	1	2

(1)对决策表 1 进行属性约简,生成基本规则$t_i \to s_i$,并计算:$nt=card(\|t_i\|)$,$ns=card(\|s_j\|)$,$nst=card(\|t_i\| \cap \|s_j\|)$

对每个规则中的t_i和nt生成表 2 中的一条记录,对每个规则中的s_j和ns生成表 3 中的一条记录,对每个规则$t_i \to s_i$和nst生成表 4 中的一条记录,结果为表 2、表 3 和表 4。

表 2 条件类

序	a	b	c	d	nt
1	1	2	1	1	4
2	1	1	2	1	2
3	2	1	2	1	1
4	1	1	2	2	1

表 3 决策类

序	f	g	ns
1	1	1	2
2	1	2	3
3	2	1	1
4	2	2	2

表 4 规则类

序	t	s	nst
1	1	1	2
2	1	2	2
3	2	3	1
4	2	4	1
5	3	4	1
6	4	3	1

(2)根据用户输入的信息,用 SQL 语言找出所有满足条件t_i和s_j,并根据表 4 合成规则。如果用户输入的个性化信息为b=1 and d=1 and f=2 and g=2,由表 2 知第 2、3 条记录满足条件信息,$\sum card(\|s_i\|)$=2+1=3;由表 3 知第 4 条记录满足决策信息,$\sum card(\|s_j\|)$=2,由表 4 知第 4、5 条记录为满足条件基本规则,$\sum card(\|t_i\| \cap \|s_j\|)$=1+1=2。由算法可生成如下规则:

ur:$(b,1) \wedge (d,1) \to (f,2) \wedge (g,2)(0.67,1,0.25)$

如果用户输入的个性化信息为f=2 and g=1,由算法可生成如下规则:

$ur1$:$(a,1) \wedge (b,1)(c,2) \wedge (d,1) \to (f,2) \wedge (g,1)(0.5,1,0.13)$

$ur2$:$(a,1) \wedge (b,1)(c,2) \wedge (d,2) \to (f,2) \wedge (g,1)(1,1,0.13)$

6 总结

个性化知识发现的研究是一个具有挑战性的研究课题。给出了一种新的面向个性化知识发现的规则提取算法和规则表示方法。在理论上证明了在多决策属性下算法的正确性,给出并证明了合成规则的可信度、覆盖度和规则强度的计算公式。算法设计的出发点是利用数据库和规则的特点设计数据库和算法下,进行规则的提取与合成。提高了规则合成的效率,这对于基于数据库系统的知识获取系统的设计具有实际的指导意义。

参考文献:

[1] Pawlak Z.Rough sets[J].International Journal of Computer and Information Science,1982,11:341-356.

[2] Swinizrski R W,Skowron A.Rough set methods in feature selection and recognition[J].Pattern Recognition Letters,2003,24:833-849.

[3] 张文修,吴伟志.粗糙集理论与方法[M].北京:科学出版社,2001.

[4] 刘文军.基于粗糙集的最简规则提取算法[J].华东理工大学学报,2007,33(6):10-12.

[5] 时希杰,沈睿芳.基于粗糙集的两阶段规则提取算法与有效性度量[J].计算机工程,2006,32(3):60-61.

[6] 王一萍,陈波,吴坚.基于数据库操作的粗糙集的规则生成[J].计算机应用研究,2005,22(5):55-57.

[7] 周军,张庆灵.基于粗糙集理论的面向个性化知识发现算法[J].计算机工程与应用,2007,43(16):172-174.

[8] 周春来,李志刚,孟跃进,等.决策规则获取算法及规则表示[J].计算机工程与应用,2007,43(4):102-105.