# 基于粗糙集理论的面向个性化知识发现算法

周 军[1,2],张庆灵[2]
ZHOU Jun[1,2],ZHANG Qing-ling[2]

1.辽宁工学院 计算机科学与工程学院,辽宁 锦州 121001
2.东北大学 理学院,沈阳 110004
1.Department of Computer Science,Liaoning Institute of Technology,Jinzhou,Liaoning 121001,China
2.College of Science Northeastern University,Shenyang 110004,China
E-mail:lnjunzhou@163.com

**ZHOU Jun,ZHANG Qing-ling.Algorithm for personalized knowledge discovery based on rough set theory.Computer Engineering and Applications,2007,43(16):172-174.**

**Abstract**: A heuristic algorithm for personalized knowledge discovery based on the rough set theory and descriptive knowledge discovery is presented.The correctness of the algorithm is proved in theory,the algorithms of extracting personalized knowledge are described and the performance of the algorithms is analyzed.The main parts of the algorithm are how to compose rule,how to calculate certainty factor,coverage factor and the strength of the decision rule.At last,the efficiency and practicability of the algorithm is illustrated by the example in this paper.

**Key words**: personalized knowledge discovery;rough set;decision rule

**摘 要**:在基于粗糙集理论的数据处理与决策分析的基础上,从实际应用的角度出发,提出了面向个性化知识发现的启发式算法。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,分析了算法的性能。算法的关键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。通过例子说明了算法的有效性和实用性。

**关键词**:个性化知识发现;粗糙集;决策规则
**文章编号**:1002-8331(2007)16-0172-03 **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、有趣的模式的过程。这种模式即为知识,它揭示了数据内部的一种本质和客观的联系和规律。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[2-4]。然而,一个拥有海量数据的信息系统的所有模式也是海量的。海量的数据转换成了海量的知识, 对于用户来说从海量的知识中提取所需的知识,无疑是困难的。由此出现了各种属性的约简方法,目的是为了消除冗余的信息,当然是有意义。但是,已经证明了求取系统的所有约简是 NP 难问题。同时,约简的信息有可能是用户最感兴趣的。也就是说,在同一决策系统中,可提取满足不同应用的知识,但对一既定用户,不是所有的知识都是必要的,如何发现对用户有价值的知识(称为个性化知识),是一个挑战性的研究课题。同时,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。

本文在基于粗糙集理论的数据处理、决策分析[5-7]的基础上,从实际应用的角度出发,给出了面向个性化知识发现的启发式算法。首先,从理论上证明了算法的正确性。其次,给出了面向个性化的知识发现算法的描述。最后,分析了算法的性能,其时间和空间的复杂度都不超过 O(|$AT$|$n$)(其中|$AT$|为系统中属性的个数)。算法的关键在于规则的合成和可信度、覆盖度和规则强度计算的方法。通过例子说明了算法的有效性和实用性。

## 2 一些基本概念

决策信息系统(又称决策表)是指一个系统 $DT=(U,AT=CU\cup\{d\},V,f)$,其中,$U$ 为一个非空有限对象集(称为论域);$AT$ 是有限的非空属性集,$AT$ 分为两部分: 其中,$C$ 是条件属性集合,$d$ 是一个决策属性,且 $CU\cap\{d\}=\Phi$。$V=\bigcup_{a\in AT}V_a$ 是 $AT$ 中的属性的值域,其中 $V_a$ 是属性 $a$ 的值域。$f$ 是映射,满足 $f_a:U\to V_a$,$\forall a\in AT$。$Inf(x)=\{(a,f_a(x))|a\in AT\}$称为 $x$ 的信息向量集。设 $X\subseteq U$,$A\subseteq AT$,定义集合:$\underline{A}(X)=\{x\in U:[x]_A\subseteq X\}$,$\overline{A}(X)=\{x\in U:[x]_A\cap X\neq\Phi\}$,$Bn(X)=\underline{A}(X)-\overline{A}(X)$,则 $\underline{A}(X)$、$\overline{A}(X)$、$Bn(X)$分别称为 $X$ 的相对于 $A$ 的下近似、上近似和边界,其中,$[x]_A$ 是元素 $x$ 关

**基金项目**:辽宁省教育厅资助科研课题(the Research Project of Department of Education of Liaoning Province,China under Grant No.20031066);辽宁省优秀青年骨干教师基金资助。

**作者简介**:周军(1966-),女,教授,博士研究生,主要从事数据挖掘与知识发现研究;张庆灵(1956-),男,博士,教授,博士生导师,主要从事智能控制、数据处理研究。

于属性 $A$ 所在的等价类。

一般地,属性值序对 $(a,v)$,$a\in AT$,$v\in V_a$,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,属性集 $A\subseteq AT$ 中的原子属性的描述的合取被称为 $A$–描述,具有原子属性 $(a,v)$ 的所有元素的集合记为 $\|(a,v)\|$,即 $\|(a,v)\|=\{x\in U|f_a(x)=v\}$。满足描述 $t$ 的所有元素的集合记为 $\|t\|$。如果 $t$ 和 $s$ 是两个描述,则有 $\|t\wedge s\|=\|t\|\cap\|s\|$ 和 $\|t\vee s\|=\|t\|\cup\|s\|$,文中 $\wedge$(或 $\vee$)表示逻辑合取(析取)运算符。

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

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

其中,$C$ 是规则的条件部分的所有属性的集合,$C'\subseteq C$,$w\in V_d$。$t$ 与 $s$ 分别称为规则的条件部分和决策部分。带有单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 $AT$ 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 $\|t\|\cap\|s\|\neq\phi$,并且 $\|t\|\not\subset\|s\|$ 时,称规则 $t\to s$ 是不确定的(或可能的)规则。

对于决策表 $DT=(U,AT=C\cup\{d\},V,f)$ 中的每一规则 $t\to s$

$$Cer(t\to s)=\pi(s/t)=pu(\|s\|/\|t\|)$$

$$Cov(t\to s)=\pi(s/t)=pu(\|t\|/\|s\|)$$

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

分别称为规则 $t\to s$ 的可信度因子、覆盖度因子和规则强度,其中,$pu(X)$ 表示 $X$ 的概率分布,$pu(X/Y)$ 表示当 $Y$ 满足时 $X$ 的条件概率。

规则 $t=\bigwedge_{a\in C'\subseteq C}(a,v)\to s\bigvee_{i=1,2,\cdots,m}(d,w_i)$,可以改写为如下的 $m$ 个明确的规则:

$$r_1:t=\bigwedge_{a\in C'\subseteq C}(a,v)\to s=(d,w_1)$$

$$r_2:t=\bigwedge_{a\in C'\subseteq C}(a,v)\to s=(d,w_2)$$

$$\cdots$$

$$r_m:t=\bigwedge_{a\in C'\subseteq C}(a,v)\to s=(d,w_m)$$

上面的规则形式与规则演绎推理的 $B$ 规则形式相同,因此,可以直接用于规则演绎推理。因而,在本文中使用的规则的一般形式如下:

$$t\to s(Cer(t\to s),Cov(t\to s),\sigma(t\to s))$$

其中 $t=\bigwedge_{a\in C'\subseteq C}(a,v)$,$s=(d,w)$,$Cer(t\to s)$、$Cov(t\to s)$ 和 $\sigma(t\to s)$ 分别为规则的可信度、覆盖度和规则强度。

## 3　算法的理论依据

设决策表 $DT=(U,AT=C\cup\{d\},V,f)$,其中 $C$ 为条件属性集,$d$ 为决策属性。令 $U/C=\{X_1,X_2,\cdots,X_n\}$,$U/d=\{Y_1,Y_2,\cdots,Y_m\}$,$des(X_i)$(或 $des(Y_i)$)表示等价类 $X_i$(或 $Y_i$)的描述,即等价类 $X_i$(或 $Y_i$)的对于各个条件属性的特定取值。如果 $\|X_i\|\cap\|Y_j\|\neq\phi$,则称规则 $des(X_i)\to des(Y_j)$ 为 $DT$ 的一个基本决策规则(以下简称基本规则)。由 $DT$ 中的所有基本规则组成的集合,称为 $DT$ 的基本规则集,记为 $DRS(DT)$。设 $\underline{C}(Y_i)=\{X_{i_1},X_{i_2},\cdots,X_{i_n}\}$,规则 $des(X_{i_j})\to des(Y_i)$,$j=1,2,\cdots,p_i$,即为所有结论为 $des(Y_i)$ 的确定性基本规则。设 $\overline{C}(Y_i)=\underline{C}(Y_i)+Bn(Y_i)=\{X_{i_1},X_{i_2},\cdots,X_{i_{p_i}}\}\cup\{X_{i_{p_i+1}},\cdots,X_{i_{q_i}}\}$,$Bn(Y_i)=\{X_{i_{p_i+1}},\cdots,X_{i_{q_i}}\}$,规则 $des(X_{i_j})\to des(Y_i)$,$j=p_i+1,\cdots,q_i$,为所有结论为 $des(Y_i)$ 的不确定性(或称可能性)基本规则。

显然,决策表的基本规则集中只有确定性和不确定性两类基本规则。

**定理 1** 设 $DT=(U,AT=C\cup\{d\},V,f)$ 是一个决策表,$t_i\to s_i$,$i=1,2,\cdots,n$ 是 $n$ 个互不相同的基本规则,如果存在描述 $t$、$s$ 满足:

(1) $\|t\|\bigcup_{i=1,2,\cdots,n}\|t_i\|$;

(2)至少存在一个 $k$,使得 $s=s_k$,$1\leqslant k\leqslant n$。

则规则 $t\to s$ 可由 $t_i\to s_i$,$i=1,2,\cdots,n$,生成,其中:

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

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

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

**证明**　显然规则 $t\to s$ 是明确的,由于 $t_i\to s_i$,$i=1,2,\cdots,n$,中至少存在一个 $k$,使得 $s=s_k$,有 $\|t_k\|\cap\|s\|\neq\phi$,则有

$$\|t\|\cap\|s\|=(\bigcup_{i=1,2,\cdots,n}\|t_i\|)\cap\|s\|=\bigcup_{i=1,2,\cdots,n}(\|t_i\|\cap\|s\|)\neq\phi \tag{1}$$

又因

$$\|t\wedge s\|=\|t\|\cap\|s\|=(\bigcup_{i=1,2,\cdots,n}\|t_i\|)\cap\|s\|=\bigcup_{i=1,2,\cdots,n}(\|t_i\|\cap\|s\|)=\bigcup_{i=1,2,\cdots,n}(\|t_i\wedge s\|)$$

进而,有

$$card(\|t\wedge s\|)=card(\bigcup_{i=1,2,\cdots,n}\|t_i\wedge s\|)=\sum_i card(\|t_i\wedge s\|) \tag{2}$$

由式(1)、(2)和规则 $t\to s$ 的可信度、覆盖度和规则强度的定义,可得:

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

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

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

显然,当规则 $t\rightarrow s$ 是确定性规则时,$Cer(t\rightarrow s)=1$。

**定理 2** 决策表中的任一性规则都可由若干个基本规则生成。

**证明** 分为两种情况证明:确定性规则和不确定性规则。

设 $t\rightarrow s$ 是决策表 $DT=(U,AT=C\cup\{d\},V,f)$ 中的任一确定性规则,则有 $\|t\|\cap\|s\|\neq\phi$。假设 $t=\bigwedge_{a\in C'\subseteq C}(a,v)$。由于 $C'\subseteq C$,有 $\forall A\in U/C'$,$\exists B_1,B_2,\cdots,B_n\in U/C$,$A$ 和 $B_i$ 可分别描述为 $\|\bigwedge_{a\in C'\subseteq C}(a,v)\|$ 和 $\|\bigwedge_{a\in C}(a,v_i)\|$,使得 $A=\bigcup_{i=1,2,\cdots,n}B_i$,且有 $B_i=\|\bigwedge_{a\in C}(a,v_i)\|\subseteq\|\bigwedge_{a\in C'\subseteq C}(a,v)\|=A\subseteq\|S\|$,设 $t_i=\bigwedge_{a\in C}(a,v_i)$,有 $\|t_i\|=B_i\subseteq\|s\|$,故有 $t_i\rightarrow s(i=1,2,\cdots,n)$为确定性基本规则,且有 $\|t\|=\bigcup_{i=1,2,\cdots,n}\|t_i\|$,由定理 1 可得

$$Cer(t\rightarrow s)=\frac{\sum_i card(\|t_i\wedge s\|)}{\sum_i card(\|t_i\|)}=1$$

$$Cov(t\rightarrow s)=\frac{\sum_i card(\|t_i\wedge s\|)}{\sum_i card(\|s\|)}$$

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

规则 $t\rightarrow s$ 可由确定性基本规则 $t_i\rightarrow s,i=1,2,\cdots,n$,生成。

当 $t\rightarrow s$ 是决策表 $DT=(U,AT=C\cup\{d\},V,f)$ 中的任一不确定性规则时的证明类似确定性规则的证明,略。证毕。

## 4 决策规则合成的启发式算法

用户决策规则合成的总体思路是根据用户提供的信息,利用基本知识库,生成最终的规则。主要分为用户的启发性信息中有决策属性值和无决策属性值两种情况进行算法设计。

**算法** ExtractionRule

输入:基本规则集(基本知识库)$KB$,其中含有 $N$ 条规则:

$r_i:t_i\rightarrow s_i,i=1,2,\cdots,N$;

用户感兴趣的一组属性和属性值:

$UA=\{(a_{u_l},v_{u_l})|a_{u_l}\in AT,v_{u_l}\in V_{a_{u_l}},u_l\in\{1,2,\cdots,N\},l=1,2,\cdots,M\}$

其中 $M=card(UA)$;

输出:用户感兴趣的决策规则集(用户知识库)。

步骤:

step 1. 初始化,$URB=\phi$;$K=0$($k$ 表示入库 $URB$ 中的规则的个数);

step 2. if 用户输入属性中有决策属性值 $s=(d,w)$,即$(a_M,v_M)=(d,w)$;

step 3. then 启发性信息有决策属性的规则生成,入库 $URB$;

step 4. else 启发性信息无决策属性的规则生成,入库 $URB$;

step 5. 显示结果;

step 6. return.

其中,stept 3、stept 4 需要进一步精化,具体步骤如下:

step 3.1 for ($i$=1 to $N$)

step 3.2 {

step 3.2.1 比较当前的基本规则是否是用户感兴趣的;

step 3.2.2 如果是用户感兴趣的规则,将基本规则合成,包括 $card(t)$,$card(s)$,$card(t\wedge s)$的合成;

$card(\|t\|)=card(\|t\|)+card(\|t_i\|)$

$card(\|t\wedge s\|)=card(\|t\wedge s\|)+card(\|t_i\wedge s\|)$

$card(\|s\|)=card(\|s\|)$

step 3.3 }

step 4.1 for ($i$=1 to $N$)

step 4.2 {

step 4.2.1 比较当前的基本规则是否是用户感兴趣的;

step 4.2.2 如果该规则是用户感兴趣,求取该规则的 $card(t)$,$card(s_j)$,$card(t\wedge s_j)$,入 $URB$,并记数;

step 4.3 }

其中 stept 3.2.1、stept 4.2.1 可以进一步求精为;

$mn=0$;

for ($j$=1 to $M$)

{if $(a_{u_j},v_{u_j})\in r_i$ then $mn=mn+1$;} //* 这里$(a_{u_j},v_{u_j})\in UA$*//

注:当 $mn=M$($M$ 为用户给出的启发性信息的属性个数)时,说明当前的规则是用户感兴趣的。另外,在显示用户的规则集(用户知识库)$URB$ 时,需要利用 $card(t)$,$card(s_i)$,$card(t\wedge s_i)$和公式

$$Cer(t\rightarrow s_i)=\frac{card(\|t\wedge s_i\|)}{card(\|t\|)}$$

$$Cov(t\rightarrow s_i)=\frac{card(\|t\wedge s_i\|)}{card(\|s_i\|)}$$

$$\partial(t\rightarrow s_i)=\frac{card(\|t\wedge s_i\|)}{card(U)}$$

计算规则的可信度、覆盖度和规则强度。

算法的时间复杂度为 $MAX\{O(|N||M|),O(|N|^2)\}$,其中,$N$ 表示基本知识库中的规则数,$M$ 表示用户提供的启发性信息的属性值的个数。

## 5 例子

表 1 表示了一个简单的信息系统的例子。9 个商店的特征通过 4 个属性来表征,即 $E$—销售人员的授权情况、$Q$—产品质量的认识情况、$L$—高交易位置情况、$P$—商店的赢利情况。其中属性 $E$、$Q$ 和 $L$ 为条件属性,$P$ 为决策属性。

**表 1 信息系统的例子**

| Store | $E$ | $Q$ | $L$ | $P$ |
|---|---|---|---|---|
| 1 | high | good | no | profit |
| 2 | high | good | no | profit |
| 3 | med | good | no | loss |
| 4 | med | good | no | profit |
| 5 | no | avg. | no | loss |
| 6 | no | avg. | no | loss |
| 7 | med | avg. | yes | loss |
| 8 | high | avg. | yes | profit |
| 9 | high | avg. | yes | profit |

如果用户输入的启发性信息为$\{(E,high),(Q,good),(L,no),(P,profit)\}$由启发式算法生成如下规则:

$ur:(E,high)\wedge(Q,good)\wedge(L,no)\wedge(P,profit)(1.00,0.40,0.22)$

如果用户输入的启发性信息为$\{(Q,good),(L,no)\}$由启发式算法生成如下规则:

$ur1:(Q,good)\wedge(L,no)\rightarrow(P,profit)(0.75,0.60,0.33)$

$ur2:(Q,good)\wedge(L,no)\rightarrow(P,loss)(0.50,0.25,0.11)$

## 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。本文给出了面向个性化的知识获取算法,从理论上论证了算法的正确性。算法设计的出发点是在现有的数据库基础上,在不进

(下转 212 页)