一种不确定性条件下的自主式知识学习模型*

王国胤+, 何 晓

(重庆邮电学院 计算机科学与技术研究所,重庆 400065)

A Self-Learning Model under Uncertain Condition

WANG Guo-Yin+, HE Xiao

(Institute of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China)

+ Corresponding author: Phn: 86-23-62460066, Fax: 86-23-62461882, E-mail: wanggy@cqupt.edu.cn

http://www.cqupt.edu.cn

Received 2002-06-03; Accepted 2002-11-06

Wang GY, He X. A self-learning model under uncertain condition. *Journal of Software*, 2003,14(6): 1096~1102.

http://www.jos.org.cn/1000-9825/14/1096.htm

Abstract: It is a very difficult problem in machine learning to learn uncertain knowledge automatically without prior domain knowledge. In this paper, a theory is developed to express, measure and process uncertain information and uncertain knowledge according to uncertainty measure of decision table and decision rule. Based on the Skowron's default rule generation algorithm, a self-learning model and the method is developed to solve this problem. Simulation results illustrate the efficiency of this self-learning method.

Key words: uncertainty; rough set; self-learning; knowledge acquisition; machine learning

摘 要: 在没有领域先验知识条件下的不确定知识主动式学习是机器学习领域中的一个难题.通过研究决策表和决策规则的不确定性,建立基于粗集表示、度量和处理不确定性信息和知识的理论,并且结合 Skowron 的缺省规则获取算法,提出一种不确定性条件下的数据自主式学习模型和方法,以解决这一问题.通过仿真实验,验证了该自主式学习方法的有效性.

关键词: 不确定性;粗集;自主式学习;知识获取;机器学习

中图法分类号: TP18 **文献标识码**: A

* Supported by the National Natural Science Foundation of China under Grant No.69803014 (国家自然科学基金); the National Climb Program of the Ministry of Science and Technology of China (国家科技部攀登特别支持经费); the Foundation for University Key Teacher by the State Education Ministry of China under Grant No.GG-520-10617-1001 (高等学校骨干教师资助计划); the Scientific Research Foundation for the Returned Overseas Chinese Scholars by the State Education Ministry of China (教育部留学回国人员科研启动基金); the Application Science Foundation of Chongqing of China (重庆市应用基础研究基金); the Science and Technology Research Program of the Municipal Education Committee of Chongqing of China under Grant No.02050 (重庆市教育委员会科学技术研究项目)

第一作者简介: 王国胤(1970－),男,重庆人,博士,教授,主要研究领域为粗集理论,神经网络,智能信息系统,网络安全,多媒体数据处理.

© 中国科学院软件研究所 http://www.jos.org.cn

粗集(rough set,简称 RS)理论[1]由波兰逻辑学家 Pawlak 教授于 1982 年提出,由于它能有效地分析和处理不精确、不一致、不完整等各种不完备信息,并从中发现隐含的知识,揭示潜在的规律,近年来在机器学习、数据挖掘、人工神经网络等多个领域得到了广泛应用[2,3].

在粗集理论的研究中,关于不确定性问题的研究是最重要的内容之一.众多学者纷纷提出各种研究方法:Skowron 提出一种通过投影得到缺省决策规则的算法[4],能够在不确定性条件下获取规则;王国胤等人给出了一种决策表信息系统的不确定性度量方法[5],有助于实现知识获取过程不依赖于领域先验知识.

在不确定性条件下的数据自主式机器学习方法,或者称为主动式学习方法,是人工智能知识获取研究中的一个难题.如果能够摆脱学习过程中对先验知识的依赖,由数据自主地完成知识的获取过程,无疑将对机器学习理论的发展和应用的推广起到重要的推动作用.在传统的机器学习研究中,人们都借助于部分领域先验知识.概率论根据人们的概率模型假定这一先验知识处理不确定性问题,模糊集理论根据人们对隶属函数的假设这一先验知识处理不确定性问题.这些方法都将人类的先验知识用于处理不确定性,不是完全根据原始数据来进行分析,而这些先验知识(假设)往往不能很好地满足实际情况,这样就在很多问题上受到限制.如果人类对有待研究的问题还没有很好的认识,这些方法就难以适用.而且,多专家决策中权重的分配也需要不同专家权重的先验知识,这往往也是难以很好设定的,但它恰恰又是决定系统性能的关键之一,而粗集理论却摆脱了这一局限,这也是粗集理论脱颖而出的关键.但是,事实上问题并没有这么简单,在基于粗集理论进行智能数据分析研究时,研究人员还是将自己对问题的假设引入了问题之中.例如,在产生规则知识的时候对于不确定规则的取舍往往根据产生规则的可信度来决定,这个取舍界限的标准就是一个先验知识.所以,对于不确定性问题的处理,现在还没有一个能够摆脱先验知识,完全由原始数据自主决策的理论方法.这是不确定智能数据分析系统研究中最为困难的问题之一,也是主动式机器学习理论研究的核心内容之一.

本文提出一种度量决策表和决策规则不确定性的方法,对二者不确定性度量的关系进行研究,将决策表的局部最小确定性作为控制规则生成过程中的阈值来控制规则生成.这样就得到了一种在不确定性条件下,完全由数据自主控制规则生成的机器学习方法,建立了一种不确定性条件下的自主式知识学习模型.

1 相关基本概念

为了方便叙述,我们首先简单介绍一些粗集理论中的有关基本概念.

定义 1(决策表信息系统). 一个决策表信息系统(简称决策表)$S=\langle U,R,V,f\rangle$,其中,U 是对象的集合,也称为论域,$R=C\cup D$ 是属性集合,子集 C 和 D 分别称为条件属性集和决策属性集,$D\neq\varnothing$,$V=\cup_{r\in R}V_r$ 是属性值的集合,V_r 表示属性 $r\in R$ 的属性值范围,即属性 r 的值域,$f:U\times R\to V$ 是一个信息函数,它指定 U 中每一个对象 x 的属性值.

定义 2(条件分类和决策分类). 给定决策表 S,C 和 D 分别为决策表的条件属性和决策属性,$U|IND(C)$和 $U|IND(D)$分别为论域 U 在属性集 C 和 D 上形成的划分,条件分类定义为 $E_i\in U|IND(C)(i=1,\ldots,m,m$ 为条件分类的个数);决策分类定义为 $X_j\in U|IND(D)(j=1,\ldots,n,n$ 为决策分类的个数).

定义 3(条件分类的一致性和不一致性). 给定决策表 S,C 为条件属性集合,对任意一个条件分类 $E_i\in U|IND(C)$,如果属于它的所有记录都有相同的决策值,则称 E_i 为一致的,否则称 E_i 为不一致的.

定义 4(确定决策表和不确定决策表). 对任意一个决策表 S,如果它所有的条件分类都是一致的,则称 S 为确定决策表,否则称 S 为不确定决策表.

例 1:见表 1,决策表有条件属性 a,b,c,决策属性 d;条件分类为 E_1,E_2,E_3,E_4,E_5 这 5 个等价类;决策分类为 X_1,X_2,X_3,X_4 这 4 个等价类.E_5 分为 $E_{5,1}$ 和 $E_{5,2}$ 两部分,分别对应不同的决策值,因此是不一致的,该决策表为不确定决策表.

Table 1 An uncertain decision table

表 1 一个不确定决策表

U	a	b	c	d	X_j
E_1	1	2	3	1(50x)	X_1
E_2	1	2	1	2(5x)	X_2
E_3	2	2	3	2(30x)	X_2
E_4	2	3	3	2(10x)	X_2
$E_{5,1}$	3	5	1	3(4x)	X_3
$E_{5,2}$	3	5	1	4(1x)	X_4

© 中国科学院软件研究所 http://www.jos.org.cn

2 不确定性度量方法

2.1 决策表不确定性度量

我们用决策表整体不确定性和决策表局部最大不确定性来度量决策表的不确定性.

对于决策表 $S=\langle U,R,V,f\rangle,R=C\cup D,C$ 为条件属性集,D 为决策属性集,分类 $E_i\in U|IND(C)(i=1,\ldots,m)$为条件分类,$X_j\in U|IND(D)(j=1,\ldots,n)$为决策分类,则对于任意条件分类 E_i,对应有集合 T_i,满足:

$$T_i=\max\{E_i\cap X_j|X_j\in U|IND(D)\}.$$

因此,对于各条件分类集合 $E_1,\ldots,E_i,\ldots,E_m$ 都分别存在对应的 $T_1,\ldots,T_i,\ldots,T_m$.

定义 5(决策表整体确定性和不确定性). 给定决策表 $S=\langle U,R,V,f\rangle,E_1,\ldots,E_i,\ldots,E_m$ 是所有的条件分类,那么决策表整体确定性定义为 $\mu_c=\dfrac{\sum_{i=1}^{m}|T_i|}{|U|}$;决策表整体不确定性定义为 $\mu_{uc}=1-\mu_c=1-\dfrac{\sum_{i=1}^{m}|T_i|}{|U|}$.

定义 6(条件分类对决策分类的确定性程度[6]**)**. 给定决策表 $S=\langle U,R,V,f\rangle,R=C\cup D,C$ 为条件属性集,D 为决策属性集,分类 $E_i\in U|IND(C)(i=1,\ldots,m)$ 为条件分类,$X_j\in U|IND(D)(j=1,\ldots,n)$ 为决策分类,则任意条件分类 $E_i\in U|IND(C)$对于决策属性分类的确定性程度定义为$\kappa(E_i)=\max\{|E_i\cap X_j|/|E_i|\ |X_j\in U|IND(D)\}=|T_i|/|E_i|$.

定义 7(决策表局部最小确定性和局部最大不确定性). 给定决策表 $S=\langle U,R,V,f\rangle,\kappa(E_1),\ldots,\kappa(E_i),\ldots,\kappa(E_m)$是条件分类对决策分类的确定性程度,则决策表局部最小确定性定义为$\alpha_c=\min\{\kappa(E_1),\ldots,\kappa(E_i),\ldots,\kappa(E_m)\}$;决策表局部最大不确定性定义为$\alpha_{uc}=1-\alpha_c$.

决策表整体不确定性反映了决策表的整体冲突情况,决策表局部最大不确定性反映了决策表各条件分类中的最大冲突情况.

定理 1. 由决策表 S 得到的规则集 F,在决策表能够充分反映领域样本数据的情况下,对从决策表中获取的规则知识进行测试的最大可能正确率η等于决策表整体确定性μ_c,即有 $\eta=\mu_c=\dfrac{\sum_{i=1}^{m}|T_i|}{|U|}$,其中 m 为决策表条件分类数.

由前面的定义易证这一定理.

例 2:对如表 1 所示的决策表条件分类 E_1,E_2,E_3,E_4,E_5 对应有集合 T_1,T_2,T_3,T_4,T_5,且满足 $T_1=E_1,T_2=E_2,T_3=E_3,T_4=E_4,T_5=E_{5.1}$,则有

$$\kappa(E_1)=1,\kappa(E_2)=1,\kappa(E_3)=1,\kappa(E_4)=1,\kappa(E_5)=0.8,$$
$$\eta=\mu_c=(|T_1|+|T_2|+|T_3|+|T_4|+|T_5|)/|U|=0.99,$$
$$\alpha_c=\min\{\kappa(E_1),\kappa(E_2),\kappa(E_3),\kappa(E_4),\kappa(E_5)\}=\min\{1,1,1,1,0.8\}=0.8.$$

2.2 决策规则的不确定性度量

对于规则来说,我们可以用可信度来表示和度量其不确定性.下面我们给出可信度的定义:

定义 8(可信度). 对于决策表 $S=\langle U,R,V,f\rangle,R=C\cup D$ 是属性集合,子集 C 和 D 分别为条件属性集和决策属性集,决策规则 $A\rightarrow B$ 的可信度 $CF(A\rightarrow B)$定义为 $CF(A\rightarrow B)=|X\cap Y|/|X|$,其中,集合 X 为条件属性值满足公式 A 的样本的集合,集合 Y 为决策属性值满足公式 B 的样本的集合.

定义 9(最小可信度). 对于决策表 $S=\langle U,R,V,f\rangle$,生成规则集 F(设 F 有 n 条规则),属于 F 的规则为$f_1,\ldots,f_i,\ldots,f_n$,对应的可信度为 $cf_1,\ldots,cf_i,\ldots,cf_n$,则规则集 F 的最小可信度为$\beta=\min\{cf_1,\ldots,cf_i,\ldots,cf_n\}$.

在 Skowron 的缺省规则获取算法中,规则集的生成受阈值控制,对规则集中的任意一条规则,其可信度要大于等于阈值,即有规则集的最小可信度大于等于阈值[4].这个阈值的选取,是根据对领域问题的认识而人为地选取的,这也正是该算法在学习过程中对先验知识的依赖.

2.3 决策表与规则集不确定性关系

定理 2. 对决策表 $S=\langle U,R,V,f\rangle$,其中 $R=C\cup D$,我们取决策表局部最小确定性为阈值来控制规则集 F 的生成,

© 中国科学院软件研究所 http://www.jos.org.cn

用 F 对样本数据集进行测试,理论上可以得到对样本数据集测试的最大正确率.

证明:首先根据条件属性集 C 计算决策表 S 的条件分类:$E_{(k,C)}\in U|IND(C),k=1,\ldots,|U|IND(C)|$.

对于任意 $E_{(k,C)}$,对应地存在 $T_{(k,C)}=\max\{E_{(k,C)}\cap X_i|X_i\in U|IND(D)\}$,且对任意条件分类 $E_{(k,C)}$,存在 $X_j\in U|IND(D)$,并满足 $E_{(k,C)}\cap X_j=\max\{E_{(k,C)}\cap X_i|X_i\in U|IND(D)\}=T_{(k,C)}$.

因此$|E_{(k,C)}\cap X_j|/|E_{(k,C)}|=|T_{(k,C)}|/|E_{(k,C)}|$是规则 $\mathrm{Des}(E_{(k,C)},C)\rightarrow\mathrm{Des}(X_j,D)$的可信度因子,其中,$\mathrm{Des}(E_{(k,C)},C)$为用条件属性集 C 来表示条件分类 $E_{(k,C)}$的公式,$\mathrm{Des}(X_j,D)$为用决策属性集 D 来表示决策分类 X_j 的公式.由此可知,$|T_{(k,C)}|/|E_{(k,C)}|=\kappa(E_{(k,C)})$.

又因为$\alpha_c=\min\{\kappa(E_{(1,C)}),\ldots,\kappa(E_{(k,C)}),\ldots,\kappa(E_{(|U|IND(C)|,C)})\}$,所以$|T_{(k,C)}|/|E_{(k,C)}|\geq\alpha_c$.

因此,我们可以得到规则 $\mathrm{Des}(E_{(k,C)},C)\rightarrow\mathrm{Des}(X_j,D)|\ |T_{(k,C)}|/|E_{(k,C)}|$,其中$|T_{(k,C)}|/|E_{(k,C)}|$是该规则的可信度因子.

如果用这条规则测试 $E_{(k,C)}$中的数据,可以得到正确结果的样本集合为 $T_{(k,C)}=\max\{E_{(k,C)}\cap X_i|X_i\in U|IND(D)\}$.

因此,测试整个决策表得到正确结果的样本集合为 $\sum_{k=1}^{|U/IND(C)|}T_{(k,C)}$.

所以,样本数据集测试正确率为 $\eta=\dfrac{\sum_{k=1}^{|U/IND(C)|}|T_{(k,C)}|}{|U|}$.

根据定理 1,定理 2 得证. □

根据定理 2 得知,在决策表充分反映领域样本数据的情况下,取决策表局部最小确定性α_c作为阈值,理论上可以得到对样本数据集测试的最大正确率.这样,我们就可以实现由决策表的不确定性来自动控制规则集的生成过程.这里,所得规则集的实际测试效果如何,还依赖于测试过程中的推理策略(冲突消解策略).

3 数据自主式知识获取算法

阈值的选取就是 Skowron 的缺省规则获取算法中的先验知识.如果能够根据决策表中的数据自动分析得到该阈值,即可实现数据自主式的知识获取.根据第 2 节中的结论,我们可以通过计算决策表局部最小确定性α_c来获取该阈值.这样,通过改进 Skowron 的缺省规则获取算法,我们提出自主式知识获取算法.具体算法如下:

算法 1. 数据自主式知识获取算法.

输入:决策表 $S=\langle U,R,V,f\rangle$,其中 $R=C\cup D$,U 是决策表中个体(或称为元素,样本)的全集,R 是每个个体的属性集,包括条件属性集 C 和决策属性集 D.

输出:缺省规则集.

第 1 步.根据条件属性集 C 计算决策表 S 的不分明关系,即条件属性集 C 对决策表 S 的条件分类:$E_{(k,C)}\in U|IND(C),k=1,\ldots,|U|IND(C)|$;用定义 6 和定义 7 计算出决策表局部最小确定性α_c,以α_c为控制规则生成的阈值.

第 2 步.如果某个划分 $E_{(k,C)}$对特定决策(如 X_j)的成员度超过α_c,则根据决策表 S 的分明矩阵产生相应的缺省规则,即

如果$|E_{(k,C)}\cap X_j|/|E_{(k,C)}|\geq\alpha_c$,则得到规则

$$\text{Rule: } \mathrm{Des}(E_{(k,C)},C)\rightarrow\mathrm{Des}(X_j,D)|\ |E_{(k,C)}\cap X_j|/|E_{(k,C)}|,$$

其中,$|E_{(k,C)}\cap X_j|/|E_{(k,C)}|$是规则 $\mathrm{Des}(E_{(k,C)},C)\rightarrow\mathrm{Des}(X_j,D)$的可信度因子.

第 3 步.将决策表 S 加入决策表集合ψ,即$\psi=\{S\}$.

第 4 步.如果$\psi=\varnothing$,则结束;否则,从ψ中取出一个决策表 $S^*=\langle U,R^*,V^*,f^*\rangle$,计算其属性核 $Core(C^*)$.通过删除某一核属性(如 C_{Cut})可以得到条件属性上的投影 $C_{Pr}=C^*-C_{Cut}$,其中 $r=1,\ldots,|Core(C^*)|$,C^*为该决策表的条件属性集合,C_{Cut}是被删掉的核条件属性.对每个投影 C_{Pr}作如下处理:

① 如果 $C_{Pr}=\varnothing$,则不对该投影做任何操作;否则,做下面 4 步操作.

② 将投影得到的新决策表 $S'=\langle U,R',V',f'\rangle$加入$\psi$,$\psi=\psi\cup\{S'\}$,其中 $R'=C_{Pr}\cup D$;

③ 根据条件属性计算投影 C_{Pr}的不分明关系,即条件属性对该投影决策表 S'的划分

© 中国科学院软件研究所 http://www.jos.org.cn

$$E_{(k,C_{Pr})}\left(E_{(k,C_{Pr})}\in U\mid IND(C_{Pr}),k=1,...,\left|U\mid IND(C_{Pr})\right|\right).$$

④ 如果某个划分 $E_{(k,C_{Pr})}$ 对特定决策(如 X_j)的成员度超过α_c,则根据决策表 S'的分明矩阵产生相应的缺省规则,即如果 $\left|E_{(k,C_{Pr})}\cap X_j\right|\Big/\left|E_{(k,C_{Pr})}\right|\geq\alpha_c$,则得到规则

$$\text{Rule}':\text{Des}\left(E_{(k,C_{Pr})},C_{Pr}\right)\rightarrow\text{Des}\left(X_j,D\right)\left\|E_{(k,C_{Pr})}\cap X_j\right|\Big/\left|E_{(k,C_{Pr})}\right|.$$

⑤ 为每条缺省规则 Rule′构造封锁该规则的事实:

若存在 E_i,E_i 属于 $U|IND(C)$,并且 E_i 是 $E_{(k,C_{Pr})}$ 的子集,并且 $E_i\cap X_j=\varnothing$,则形成事实:

$$F'\text{:Des}(E_i,C_{Cut})\rightarrow\text{NOT(Rule}').$$

第 5 步.转第 4 步.

Skowron的缺省规则获取算法得到的是不确定性决策规则.它按照规则可信度因子从高到低的顺序生成规则.阈值越大,得到的规则越少;反之,阈值越小,得到的规则越多.上述数据自主式知识获取算法首先计算决策表的局部最小确定性α_c,并以α_c 为阈值来控制规则的生成过程,这样就可以避免产生不必要的冗余规则.下面通过一个例子加以说明.

例 3:将如表 1 所示的决策表分别通过 Skowron 算法(阈值选 0.55,即 Skowron 教授在文献[4]中为本例选的阈值)和本文的数据自主式知识获取算法得到规则集(在数据自主式知识获取算法中,阈值α_c 的计算结果为 0.8)rules_1 和 rules_2 如下:

rules_1: R_1: $a_1c_3\rightarrow d_1|1.0$ R_2: $a_1c_1\rightarrow d_2|1.0$ R_3: $b_2c_1\rightarrow d_2|1.0$

R_4: $a_2\rightarrow d_2|1.0$ R_5: $b_3\rightarrow d_2|1.0$ R_6: $a_3\rightarrow d_3|0.8$

R_7: $b_5\rightarrow d_3|0.8$ $R_8(C-\{a\})$: $b_2c_3\rightarrow d_1|0.62$ $R_9(C-\{c\})$: $a_1\rightarrow d_1|0.91$

$R_{10}(C-\{a,b\})$: $c_3\rightarrow d_1|0.56$ $R_{11}(C-\{a,c\})$: $b_2\rightarrow d_1|0.59$ $F_1(C-\{a\})$: $a_2\rightarrow\text{NOT}(R_8)$

$F_2(C-\{c\})$: $c_1\rightarrow\text{NOT}(R_9)$ $F_3(C-\{a,b\})$: $b_3\rightarrow\text{NOT}(R_{10})$ $F_4(C-\{a,c\})$: $a_2\rightarrow\text{NOT}(R_{11})$, $c_1\rightarrow\text{NOT}(R_{11})$

rules_2: R_1: $a_1c_3\rightarrow d_1|1.0$ R_2: $a_1c_1\rightarrow d_2|1.0$ R_3: $b_2c_1\rightarrow d_2|1.0$

R_4: $a_2\rightarrow d_2|1.0$ R_5: $b_3\rightarrow d_2|1.0$ R_6: $a_3\rightarrow d_3|0.8$

R_7: $b_5\rightarrow d_3|0.8$ R_8: $a_1\rightarrow d_1|0.91$ $F_1(C-\{c\})$: $c_1\rightarrow\text{NOT}(R_8)$

两个规则集通过少数优先策略[7]测试决策表中的数据,得到测试正确率均为 99%,但是 rules_1 的规则数目多,rules_2 的规则数目少,显然 rules_2 的效率高.

4 仿真实验

为了验证前述的数据自主式知识获取算法的效果,我们做了大量的数据测试工作.为了体现测试的客观性,测试工作一部分由我们完成,一部分请项目组外的李志君同学完成.

在测试实验中,我们采用少数优先[7]的规则选取冲突消解策略.我们从整个数据集中随机地抽取其中 50%的数据作为训练数据集,用本文的数据自主式知识获取算法学习得到决策规则,然后用整个数据集对所得到的规则集进行测试.实验步骤如下:

第 1 步.将训练数据集利用"基于属性重要性离散化算法[8]"进行离散化处理.

第 2 步.计算出决策表局部最小确定性α_c,以α_c 为阈值,通过自主式知识获取算法生成规则集,用该规则集对整个数据集进行测试.

第 3 步.从 1.0 到 0 取 10 个均分点,分别用这 11 个值作为阈值,用 Skowron 的缺省规则获取算法生成规则集,用这些规则集分别来测试整个数据集,得到测试正确率.

测试实验结果见表 2.

从表 2 我们可以清楚地看出,在阈值从 1 减小到α_c 的过程中,正确识别率逐步增大,且变化较快,而当阈值继续减小时,正确识别率变化不明显.我们取α_c 作为阈值控制生成的规则集对样本的测试正确率是很好的,即使没有得到最大正确率,其值也是在最大正确率取值的附近.而且α_c 取值比较接近整个测试正确率的转折区域,在保证有较高正确率的同时也不会产生过多的规则.此外,从表 2 也可以看出,α_c 的取值是变化的,不同的数据集可能有不同的值,这是由领域数据的不同特性决定的.实验数据表明,这种通过取决策表局部最小确定性值α_c 为阈值

© 中国科学院软件研究所 http://www.jos.org.cn

控制规则集生成的方法是有效的.

Table 2 Simulation results

表 2 数据测试结果

Data set	Number of training samples	Number of testing samples	α_c	Threshold / Correct recognition rate (%) 1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.0	α_c
Tic-tac-toe	479	958	0.5	15	15	20	57	69	71	69	67	67	67	67	71
Buf	172	345	0.5	5	5	17	21	45	59	59	57	54	54	54	59
Hayes-roth3	66	132	0.5	73	73	75	83	87	87	87	87	87	87	87	87
Iris-1	75	150	0.42	0	0	11	35	35	35	55	55	55	55	55	55
Iris	75	150	0.5	19	19	19	25	32	49	56	56	49	49	49	49
Car	487	974	0.5	60	60	61	79	79	80	80	78	78	78	78	80
Balance-scale	312	625	0.5	8	11	44	56	65	70	70	70	65	55	52	70
Krkopt	517	1035	0.5	42	57	80	83	87	86	86	86	86	81	81	86
House-votes	217	435	0.5	6	54	54	86	86	86	86	86	86	86	86	86
Yeast	185	371	0.25	10	10	10	18	18	35	41	44	40	39	39	40
Tic-tac-5000	1000	5000	0.5	51	60	71	80	82	82	62	81	80	78	78	82
Heart10000	1500	10000	0.5	14	47	73	83	83	84	84	82	80	78	78	84
Hayes-1	66	132	0.42	9	9	9	9	9	9	48	48	48	32	32	48
Glass	107	214	0.33	14	14	14	14	17	41	41	41	41	41	41	41
Ecolt	168	336	0.43	26	26	31	41	52	63	63	65	64	64	64	63
Postoperative	45	90	0.5	36	36	50	73	73	68	68	63	63	63	83	68
Heart	135	270	0.5	15	54	64	67	71	82	82	82	74	74	71	82
Lenses	12	24	0.38	50	50	50	50	50	50	50	67	67	67	67	67
Flare	161	323	0.54	28	32	43	75	74	74	74	74	74	65	65	74
Monk-3	216	432	0.5	34	34	34	36	48	50	50	50	50	50	50	50
Adult+Stretch	10	20	0.5	60	60	60	70	70	80	80	80	80	80	80	80
New-thyroid	107	215	0.5	51	51	64	72	79	81	81	81	81	81	81	81

注:在仿真实验中,为了说明本文的方法对不确定数据的处理能力,我们从实验数据集中随机删除一些属性,以增加数据的不确定性,因此,表 2 中一些数据测试正确率不是很高.表中的 22 个数据集前 12 个由我们完成,后 10 个由课题组外的李志君同学完成.

对于每个实验数据集,我们都可以画出正确识别率随阈值变化的曲线.图 1 就是对 Ecolt 数据集(实线)和 Lenses 数据集(虚线)进行测试得到的正确识别率随阈值变化的曲线.从图中可以看出,在阈值从 1 减小到α_c的过程中,正确识别率逐步增大,且变化较快,而当阈值继续减小时,正确识别率变化就不明显了,变化幅度很小.在规则生成的过程中,既希望提高正确率,又希望得到的规则数目适当.由此我们可以看出,对于这两个数据集,我们所确定的阈值α_c 是很好的,既保证了样本识别的正确率,又使得规则数不会太多.限于篇幅,我们不能在此画出所有测试数据集的正确识别率随阈值变化的曲线,有兴趣的读者可以根据表 2 所示的结果画出这些曲线,其结果是与图 1 类似的.这说明,本文的算法中所选取的阈值是合理的,出现在正确识别率随阈值变化的曲线的转折点附近.

© 中国科学院软件研究所 http://www.jos.org.cn

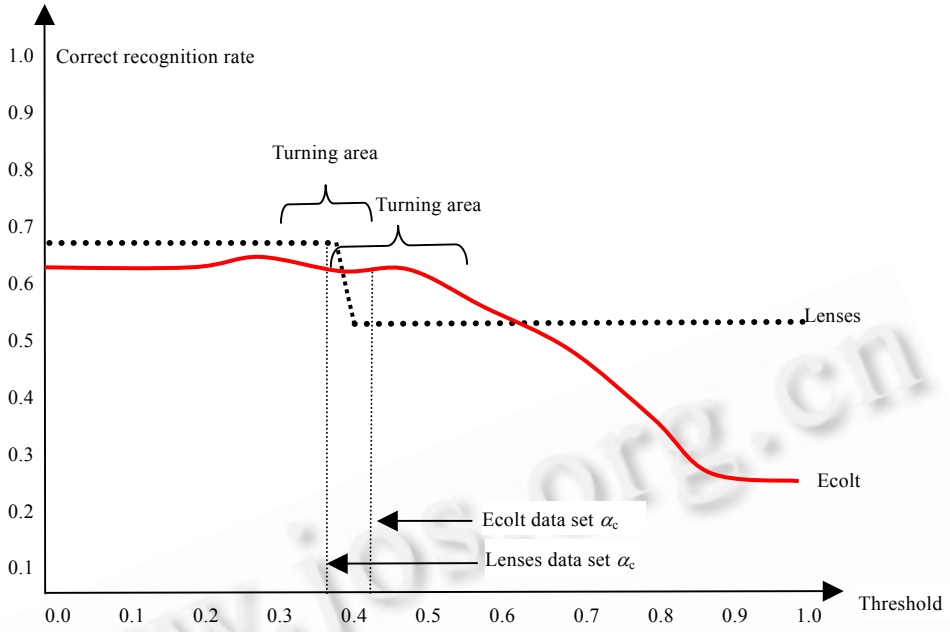

Fig.1 Simulation result of dataset Ecolt and Lenses
图 1 Ecolt 和 Lenses 数据集测试结果

5 结束语

Rough 集理论中对不确定性问题的研究是一个重要的课题,自主式机器学习也是当前机器学习理论研究中的一个热点问题.本文针对决策表和规则集的不确定性度量问题进行了研究,提出了一种度量决策表和规则集的不确定性的方法,通过采用决策表局部最小确定性值α_c 为阈值来控制规则集的生成过程,成功地实现了在不确定性的条件下完全由原始数据自主控制的机器学习方法,实现了不确定性条件下知识学习的自主性.仿真实验结果也表明了该方法是有效的,阈值α_c 的选取是合理的.

致谢 李志君同学做了大量的测试工作,特此表示感谢.

References:

[1] Pawlak Z. Rough set. International Journal of Computer and Information Sciences, 1982,11(5):341~356.
[2] Pawlak Z, Grzymala-Busse J, Slowinski R, Ziarko W. Rough sets. Communications of the ACM, 1995,38(11):89~95.
[3] Pawlak Z. Vagueness—A rough set view. In: Mycielski J, Rozenberg G, Salomaa A, eds. Structures in Logic and Computer Science: A Selection of Essays in Honor of A. Berlin: Springer-Verlag, 1997. 106~117.
[4] Mollestad T, Skowron A. A rough set framework for data mining of propositional default rules. In: Ras ZW, Michalewicz M, eds. Foundations of Intelligent Systems of the 9th International Symposium (ISMIS'96). Berlin: Springer-Verlag, 1996. 448~457.
[5] Wang GY. Uncertainty measurement of decision table information systems. Computer Science, 2001,28(5):23~26 (in Chinese with English abstract).
[6] Wang GY. Rough Set Theory and Knowledge Acquisition. Xi'an: Xi'an Jiaotong University Press, 2001 (in Chinese).
[7] Wang GY, Wu Y, Liu F. Generating rules and reasoning under inconsistencies. In: Proceedings of the IEEE International Conference on Industrial Electronics, Control and Instrumentation. Nagoya, 2000. 2536~2541. http://www.nuee.nagoya-u.ac.jp/institute/IECON2K/.
[8] Hou LJ, Wang GY, Nie N, Wu Y. Discretization in rough set theory. Computer Science, 2000,27(12):89~94 (in Chinese with English abstract).

附中文参考文献:

[5] 王国胤.决策信息系统中的不确定性度量.计算机科学,2001,28(5):23~26.
[6] 王国胤.Rough 集理论与知识获取.西安:西安交通大学出版社,2001.
[8] 侯利娟,王国胤,聂能,吴渝.粗糙集理论中的离散化问题.计算机科学,2000,27(12):89~94.

© 中国科学院软件研究所 http://www.jos.org.cn