连续值属性决策表中的知识获取方法

冯 林[1,2,3],王国胤[2],李天瑞[3]
(1. 四川师范大学计算机科学学院,四川成都 610066; 2. 重庆邮电大学计算机科学与技术研究所,重庆 400065;
3. 西南交通大学信息科学与技术学院,四川成都 610031)

摘 要: 提出了一种从连续值属性决策表中获取知识的方法 KACVA(Knowledge Acquisition from decision tables containing Continuous-Valued Attributes). 该方法将经典粗糙集理论对数据空间的等价划分转换为相似划分,把传统粗糙集理论中正域的表示方法扩充到连续值属性决策表中;通过计算连续值属性决策表中各条件聚类对决策类的分类能力,生成决策规则. 不同数据集的实验测试结果表明:对连续值属性决策表中的知识获取,KACVA 方法与传统的粗糙集相关知识获取方法及 C4.5 决策树分类方法相比,有更高的分类准确率.

关键词: 粗糙集; 属性约简; 知识获取; 离散化
中图分类号: TP18 **文献标识码**: A **文章编号**: 0372-2112 (2009) 11-2432-07

Knowledge Acquisition from Decision Tables Containing Continuous-Valued Attributes

FENG Lin[1,2,3], WANG Guo-yin[2], LI Tian-rui[3]
(1. *College of Computer Science, Sichuan Normal University, Chengdu, Sichuan* 610066, *China*;
2. *Institute of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing* 400065, *China*;
3. *School of Information Science and Technology, Southwest Jiaotong University, Chengdu, Sichuan* 610031, *China*)

Abstract: An approach for Knowledge Acquisition from decision tables containing Continuous-Valued Attributes (KACVA) is developed. The equivalence partition in the classical rough sets theory is converted into a similarity partition. A novel representation method of the positive region in decision tables with continuous-valued attributes is built. Through calculating classification abilities of each conditional cluster to decision classes, decision rules in decision tables containing continuous-valued attributes are generated. Experimental evaluation on different data sets shows that the KACVA algorithm has the better performance in the classification accuracy comparing with the knowledge acquisition approaches under classical rough sets theory and the decision tree approach, C4.5, in processing decision tables with continuous-valued attributes.

Key words: rough sets; attribute reduction; knowledge acquisition; discretization

1 引言

自波兰科学家 Pawlak 教授 1982 年提出粗糙集理论后,该理论作为一种处理不确定和含糊信息的重要数学工具,日益为人们接受并不断发展完善. 它已在属性约简[1,2]、决策规则生成[3]、视频检测[4]、数据融合[5]、故障诊断[6]等方面取得了较成功应用.

机器学习的目的是从数据中获取知识. 基于粗糙集理论的知识发现,主要借助于信息表(也称信息系统)这样一种有效的数据知识表达方式,它通过数据预处理、属性约简、规则生成等步骤建立分类知识系统. 属性约简是粗糙集理论的重要应用,也是其核心问题之一. 但是,经典粗糙集理论模型建立在不分明关系(等价关系)基础之上,它处理的属性值是清晰的离散值. 而对现实生活中广泛存在的连续属性取值却不能直接处理,如医疗信息系统中年龄、体温、血压等属性. 经典粗糙集理论处理这类性质的决策表中的属性约简,常采用将连续值数据进行离散化[7],再用传统的粗糙集方法求解属性约简,但这种先期离散化处理方法会导致一些信息的损失[8]. 为了解决这一问题,人们引入了模糊粗糙集模型[9~11]、邻域关系模型[6,12]来研究连续值决策表中的属性约简,并取得了很好的研究成果.

虽然人们对基于粗糙集理论的连续值决策表中的属性约简方法进行了大量研究,但对其知识获取方法却没有得到更广泛的关注. 本文将系统研究利用粗糙集理论获取连续值属性决策表中的分类规则问题. 首先,把

收稿日期:2008-07-04;修回日期:2009-01-04
基金项目:国家自然科学基金(No. 60773113,60873108);新世纪优秀人才支持计划(NCET);重庆市杰出青年科学基金(No. 2008BA2041);重庆市自然科学基金重点项目(No. 2008BA2041);四川师范大学科研基金(No. 06lk012)

Pawlak 粗糙集理论中的正域扩充到连续值属性决策表中;其次,针对连续值数据特点,利用统计学相关方法,提出了连续值属性重要性的度量准则. 由此提出了连续值属性决策表的知识获取方法. 这些方法能够直接处理决策表中的连续值数据,而不需要对其离散化处理. 对 UCI 数据库[13]中的多个连续值数据集,实验发现:适当设置算法中的某些参数,本文知识获取方法与传统的粗糙集相关知识获取方法[14~16]及 C4.5 决策树分类方法[17]相比,有更高的分类准确率.

2　基本概念

粗糙集理论对知识进行表达和处理的基本工具是决策表(也称决策信息系统). 它可以表示为 $S=(U,R,V,f)$,其中:U 是有限非空对象集合,也称论域,$R=C\cup D$ 是有限非空属性集合,$C\cap D=\varnothing$,$D\neq\varnothing$,子集 C 和 D 分别称为条件属性集和决策属性集,$V=\bigcup_{r\in R}V_r$ 是属性值的集合,V_r 表示属性 r 的值域,$f:U\times R\to V$ 是一个信息函数,它指定 U 中的每一个对象 x 的属性值.

对于任意的属性子集 $B\subseteq R$,不分明关系 $\mathrm{IND}(B)$ 定义为:

$$\mathrm{IND}(B)=\{(x,y)\mid (x,y)\in U\times U,\ \forall b\in B\ \forall x\in U\ \forall y\in U\ (f(x,b)=f(y,b))\} \tag{1}$$

对任意 $X\subseteq U$ 和不分明关系 IND(B),X 关于 B 的下近似集 $B_-(X)$ 和上近似集 $B^-(X)$ 分别定义如下:

$$\begin{cases} B_-(X)=\bigcup\{Y_i\mid Y_i\in U/\mathrm{IND}(B)\ \wedge Y_i\subseteq X\} \\ B^-(X)=\bigcup\{Y_i\mid Y_i\in U/\mathrm{IND}(B)\ \wedge Y_i\cap X\neq\varnothing\} \end{cases} \tag{2}$$

其中,$U/\mathrm{IND}(B)$ 是不分明关系 $\mathrm{IND}(B)$ 对论域 U 的划分.

X 的 B 正域 $\mathrm{POS}_B(X)$ 定义为:

$$\mathrm{POS}_B(X)=B_-(X) \tag{3}$$

本文中,$\forall c\in C$,如果 c 取连续实数值,则 S 称为连续值属性决策表 DTCVA(Decision Tables containing Continuous-Valued Attributes). 很显然,公式(1)、(2)和(3)不适合于 DTCVA. 为了扩展经典粗糙集理论相关概念在 DTCVA 中的表示,首先,我们定义一个相似关系描述 DTCVA 中各对象之间的相关程度.

定义 1　给定 $\mathrm{DTCVA}=(U,C\cup D,V,f)$,$B\subseteq C$ 和 $x,y\in U$,B 上的一个相似关系 $R(B)(x,y)$ 定义为:

$$R(B)(x,y)=\frac{1}{1+(\sum_{p=1}^{|B|}|f(x,a_p)-f(y,a_p)|)} \tag{4}$$

其中,$a_p\in B$. $R(B)(x,y)$ 的值域为 $(0,1]$,满足自反性、对称性.

本文实验部分,在使用公式(4)计算各对象的相似程度时,所有连续值属性用最大-最小值方法被标准化到 $[0,1]$ 区间,以减少因各属性量纲不一致对结果的影响.

为了叙述方便,文中引入如下符号:

$\theta_B(x)$ 表示对象 x 在属性子集 B 上满足公式(4)的相似类;

$\mu_{\theta_B(x)}(z)$ 表示对象 z 隶属于 $\theta_B(x)$ 的程度;

$\mu_X(z)$ 表示对象 z 隶属于集合 $X(X\subseteq U)$ 的程度,当 $z\in X$ 时,$\mu_X(z)=1$,否则 $\mu_X(z)=0$;

$I(\mu_{\theta_B(x)}(z),\mu_X(z))$ 表示 $\mu_{\theta_B(x)}(z)$ 蕴含于 $\mu_X(z)$ 的程度,它定义为 $\min(1,1-\mu_{\theta_B(x)}(z)+\mu_X(z))$;

$\psi_B(\alpha)=\{z\in U\mid I(\mu_{\theta_B(x)}(z),\mu_X(z))\geq\alpha\}$,$0\leq\alpha\leq 1$.

在公式(3)中,正域用下近似集来定义,即通过一个集合对另一个集合"精确"包含关系来确定. 在 DTCVA 中,对任意的 $X\subseteq U$,我们考虑计算论域上的对象 x 在特定的属性子集(知识空间) B 上对正域 $\mathrm{POS}_B(X)$ 的隶属程度. 首先,我们计算 x 在属性集 B 上的相似类 $\theta_B(x)$ 中每一个成员隶属度对 X 中相应成员隶属度的包含程度,只有那些包含程度大于或等于一定阈值 α 的成员,我们才予以考虑,由此我们给出如下定义.

定义 2　给定 DTCVA,$B\subseteq C$,$X\subseteq U$,$z\in U$,z 隶属于正域 $\mathrm{POS}_B(X)$ 的程度 $\mu_{\mathrm{POS}_B(X)}(z)$ 定义如下:

$$\mu_{\mathrm{POS}_B(X)}(z)=\frac{\mathrm{card}(\theta_B(z)\cap\psi_B(\alpha))}{\mathrm{card}(\theta_B(z))} \tag{5}$$

其中,$\mathrm{card}(\theta_B(z))=\sum_{x\in U_{\theta_B(x)}}$.

文献[9]中,Jensen 等人提出了基于模糊粗糙集的近似分类质量的概念,据此表示方法,我们给出 DTCVA 中近似分类质量的定义.

定义 3　给定 DTCVA,$B\subseteq C$,B 对决策 D 的近似分类质量 $\gamma_B(D)$ 定义为:

$$\gamma_B(D)=\frac{\sum_{X\in U/D}\sup_{X\in U/D}\mu_{\mathrm{POS}_B(X)}(x)}{|U|} \tag{6}$$

定理 1　在 DTCVA 中,$\forall B\subseteq C$,B 对决策 D 的近似分类质量 $\gamma_B(D)\in[0,1]$.

证明:由公式(5)知:$\forall x\in U$ 和 $X\in U/D$,有 $0\leq\mu_{\mathrm{POS}_B(X)}(x)\leq 1$ 成立.

于是,$0\leq\sup_{X\in U/D}\mu_{\mathrm{POS}_B(X)}(x)\leq 1$. 因此,在论域 U 上,不等式 $0\leq\frac{1}{|U|}\sum_{x\in U}\sup_{X\in U/D}\mu_{\mathrm{POS}_B(X)}(x)\leq 1$ 成立,即 $\gamma_B(D)\in[0,1]$.

3　DTCVA 中属性重要性的度量方法

针对连续值属性的特点,我们基于马氏距离计算 DTCVA 中属性重要性.

给定 DTCVA,设 $V_D=\{d_1,d_2,\cdots,d_l\}$,即 DTCVA 有 l 个决策类,进一步,设 D 把 DTCVA 分成 l 个决策表 $\mathrm{DTCVA}_i=(U_i,C\cup\{d_i\},V,f_i)$,其中,$U=\bigcup_{i=1}^{l}U_i$ 且 $\forall i\neq j, U_i\cap U_j=\varnothing$,设 $\{x_{i1},x_{i2},\cdots,x_{i|U_i|}\}$ 是 U_i 的对象集,在属性子集 $B=\{a_1,a_2,\cdots,a_m\}$ 下,定义 DTCVA 的协方差矩阵 W 如下:

$$W=\frac{1}{l}\sum_{i=1}^{l}\frac{1}{|U_i|}\sum_{q=1}^{|U_i|}[f_i(x_{iq},B)-\chi_i][f_i(x_{iq},B)-\chi_i]^T \tag{7}$$

式中 $f_i(x_{iq},B)$ 是一个信息函数值向量,它指出了 DTCVA_i 中的对象 x_{iq} 在属性 B 上的取值,即

$$f_i(x_{iq},B)=[f_i(x_{iq},a_1),f_i(x_{iq},a_2),\cdots,f_i(x_{iq},a_m)]^T \tag{8}$$

其中,$q=1,2,\cdots,|U_i|$.

χ_i 是 DTCVA_i 上各对象分别在条件属性 $a_1,a_2,\cdots,a_m$ 上的均值向量,即

$$\chi_i=[\frac{1}{|U_i|}\sum_{q=1}^{|U_i|}f_i(x_{iq},a_1),\frac{1}{|U_i|}\sum_{q=1}^{|U_i|}f_i(x_{iq},a_2),\cdots,\frac{1}{|U_i|}\sum_{q=1}^{|U_i|}f_i(x_{iq},a_m)]^T \tag{9}$$

T 表示矩阵的转置.

根据协方差矩阵 W,可定义条件属性集 B 上,决策类 d_i 与 d_j 之间的马氏距离 Φ_{ij}^B 为:

$$(\Phi_{ij}^B)^2=[\chi_i-\chi_j]^T W^{-1}[\chi_i-\chi_j] \tag{10}$$

其中,W^{-1} 是 W 的逆矩阵. 设在决策属性 D 上,Φ_{ij}^B 的均值为 m_B,在决策表 DTCVA 上,当 $l=2$ 时,属性集 B 对决策 D 的"可分离度"为:

$$\delta_B(D)=\Phi_{ij}^B \tag{11}$$

当 $l>2$ 时,属性集 B 对决策 D 的"可分离度"为:

$$\delta_B(D)=\sqrt{\frac{1}{C_l^2}\sum_{i=1}^{l}\sum_{j=1}^{l}(\Phi_{ij}^B-m_B)^2} \tag{12}$$

"可分离度"指出了在属性子集 B 上各决策类之间的分散程度,其中,C_l^2 表示在类别数 l 上的组合数.

定义 4 给定 DTCVA,$B\subseteq C$,属性子集 B 在属性集 C 中相对于决策 D 的重要性 $\mathrm{SIG}(B,C,D)$ 定义为:

$$\mathrm{SIG}(B,C,D)=\delta_C(D)-\delta_{C\setminus B}(D) \tag{13}$$

公式(13)指出了在属性全集 C 中,去掉属性子集 B 后,对各决策类之间"可分离性"的影响.

4 DTCVA 中的属性约简

首先,我们给出 DTCVA 中的属性约简定义.

定义 5 给定 $\mathrm{DTCVA}=(U,C\cup D,V,f)$,条件属性 C 关于决策 D 的约简 REDU 定义为 C 的一个属性子集,满足:

(1) $\gamma_C(D)=\gamma_{\mathrm{REDU}}(D)$;

(2) REDU 中去掉任何一个属性,(1)式不成立.

定义 5 理论上保证了 C 的一个最小子集 REDU 与 C 的近似分类质量相等. 但在实际应用中,使用公式(6)计算近似分类质量时,考虑到噪声和干扰的影响,要做到 $\gamma_C(D)$ 与 $\gamma_{\mathrm{REDU}}(D)$ 严格相等是困难的,我们可以设定一个很小的数 β,如果 $|\gamma_C(D)-\gamma_{\mathrm{REDU}}(D)|\leq\beta$ 成立,则认为 $\gamma_C(D)$ 与 $\gamma_{\mathrm{REDU}}(D)$ 近似相等.

下面给出启发式属性约简算法 HARCVA(Heuristic Attribute Reduction in Decision Tables Containing Continuous-Valued Attributes). 算法从条件属性集 C 开始,以第 3 节定义的属性重要性为启发式信息,逐次去掉重要性小且分类质量较差的属性,从而得到属性约简结果.

算法 1 HARCVA

输入:$\mathrm{DTCVA}=(U,C\cup D,V,f)$ 及阈值 β;

输出:DTCVA 中的属性约简 REDU.

Step1 按公式(6)计算 DTCVA 中决策 D 相对条件属性集 C 的近似分类质量 $\gamma_C(D)$;

Step2 按公式(13)计算每个条件属性的重要程度;

Step3 升序排列所有的条件属性,设排序后的结果为 $C=\{a_1,a_2,\cdots,a_n\}$;

Step4 令 REDU = C;

Step5 依顺序对 C 中的每个属性 a_i 重复以下操作:

5.1. 计算决策 D 相对于约简 REDU 在删除 a_i 后的近似分类质量 $\gamma_{\mathrm{REDU}-\{a_i\}}(D)$;

5.2. 如果 $|\gamma_C(D)-\gamma_{\mathrm{REDU}}(D)|\leq\beta$ 成立,则属性 a_i 可约,REDU = REDU - $\{a_i\}$,否则,属性 a_i 不能被约简,REDU 保持不变.

Step6 输出约简 REDU,算法结束.

算法时间复杂度分析:Step1 计算近似分类质量 $\gamma_C(\{d\})$ 的时间复杂度为 $O(|C||U|^2)$,Step2 中公式(7)的时间复杂度为 $O(|C|^2|U|)$,计算逆矩阵用 Strassen 算法的时间复杂度为 $O(|U|^{\log_2 7})$,公式(10)的时间复杂度为 $O(|C|^3)$,公式(12)的时间复杂度为 $O(l^2)$,因此,Step2 的时间复杂度为 $O(|U|^{\log_2 7}+|C|^2|U|+|C|^3)$,Step3 的时间复杂度为 $O(|C|\log|C|)$,Step5 时间复杂度为 $O(|C|^2|U|^2)$,因此,算法 1 的时间复杂度为:$O(|U|^{\log_2 7}+|C|^2|U|^2+|C|^3)$.

5 DTCVA 中的知识获取方法

经典粗糙集理论的知识获取方法以等价关系对论域的等价划分为基础. 黄金杰等人研究了模糊聚类对论域的划分问题[18],但数据集的属性个数并未约简. 针对约简后的连续值属性数据集,我们用 k-均值聚类方

法对论域聚类以替代经典粗糙集理论的等价类.

5.1　k-均值聚类

算法基本思想是:给定类的个数 k,将 n 个模式分到 k 个类 $G_1, G_2, \cdots, G_k$ 中去,使得类内对象之间的相似性最大,而类之间的相似性最小. 设 Class = $\{G_1, G_2, \cdots, G_k\}$,下面描述其算法.

算法 2　k-均值聚类算法

输入:约简后的连续值数据集,t 个样本,要聚类的类别数量 k;

输出:k 个类的聚类中心 $\mu_1, \mu_2, \cdots, \mu_k$.

Step1　初始化 $t, k, \mu_1, \mu_2, \cdots \mu_k$;

Step2　Do 按照最近邻 μ_i 分类 t 个样本

重新计算 μ_i;

Until μ_i 不再改变.

Step3　返回 $\mu_1, \mu_2, \cdots \mu_k$,算法结束.

算法 2 的时间复杂度为 $O(tk\vartheta|\mathrm{REDU}|)$,其中,$\vartheta$ 是迭代次数.

5.2　DTCVA 中的决策规则

算法 3　规则生成算法 KACVA

输入:DTCVA 及属性约简 REDU,规则可信度阈值 λ;

输出:　决策规则集 RuleSet.

Step1　RuleSet = $\{\}$, $U/D = \{Y_1, Y_2, \cdots Y_l\}$;

Step2　调用算法 2;

Step3　For $i = 1$ To k

3.1. For $j = 1$ To l

3.1.1. 计算 $c_{ij} = |G_i \cap Y_j| / |G_i|$;

3.1.2. 如果 $c_{ij} \geq \lambda$,则生成规则:RuleSet = RuleSet $\cup\{\mathrm{Des}(G_i) \rightarrow \mathrm{Des}(Y_j)(c_{ij})\}$,其中,$\forall G_i \in$Class 和 $x \in Y_j$, $\mathrm{Des}(G_i) = \bigwedge_{a \in \mathrm{REDU}}(a = \mu_i)$, $\mathrm{Des}(Y_j) = (D = f(x, D))$, c_{ij} 为规则的可信度因子;

Step4　输出决策规则 RuleSet,算法结束.

算法时间复杂度分析:该算法的时间复杂度主要由 Step2 与 Step3 决定. Step2 的时间复杂度为 $O(k\vartheta|U||\mathrm{REDU}|)$, Step3 的时间复杂度为 $O(kl)$,因此,算法 3 的时间复杂度为 $O(k\vartheta|U||\mathrm{REDU}|)$.

5.3　规则推理策略

算法 3 生成的每一条规则都是归纳问题空间中的一定对象集而得到,每条规则都有一定的代表性,它代表着一定范围的对象. 为了测试算法 3 知识获取方法的效果,定义合适的推理策略(即未知类对象的匹配策略)是必要的. 下面,我们给出对象对决策规则的匹配度的概念.

定义 6　对象 x 在属性 a_i 上与决策规则 R 中属性 a_i 的偏离系数定义为:

$$\kappa_i(f(x, a_i)) = |f(x, a_i) - \mu_i| \qquad (14)$$

此定义指出了对象 x 在属性 a_i 上的值远离决策规则 R 中属性 a_i 的中心点 μ_i 的大小.

定义 7　设决策规则 R 的长度为 LoR,则对象 x 对 R 的匹配度 MoD(x, R) 定义如下:

$$\mathrm{MoD}(x, R) = \frac{1}{\mathrm{LoR}}\sum_{i=1}^{\mathrm{LoR}}(1 - \kappa_i(f(x, a_i))) \qquad (15)$$

根据对象 x 对各决策规则的匹配度,可以综合确定对象 x 的类别,以此计算决策规则集对测试数据集的分类正确率. 算法描述如下:

算法 4　规则匹配算法

输入:测试数据集 TE 及规则集 RuleSet,并设 TE 的对象数为 $|T|$,RuleSet 的规则数为 $|R|$;

输出:测试正确率 Accuracy.

Step1　$\varepsilon = 0$; ε 存放正确分类的对象数目

Step2　For $i = 1$ To $|T|$

2.1. For $j = 1$ To $|R|$

2.1.1. 计算对象 x_i 与决策规则 R_j 的 MoD(x_i, R_j);

2.2. 选取使 MoD(x_i, R_j) 最大的 R_w,即 $R_w = \{R_j | \max(\mathrm{MoD}(x_i, R_j))\}, j = 1, 2, \cdots, |R|$ 作为对象 x_i 的分类结果 Result. 如果 R_w 有多条,则选取可信度因子最大的 R_w 的决策作为对象 x_i 的分类结果 Result;

2.3. 如 $f(x_i, D) =$ Result,则 $\varepsilon = \varepsilon + 1$;

Step3　输出 TestDataSet 的分类正确率 Accuracy = $\varepsilon/|T|$,算法结束.

算法 4 的时间复杂度为 $O(|T||R|)$.

6　实验及结果分析

为了验证前述 HARCVA 方法及 KACVA 方法的效果,我们在 UCI 机器学习数据库[13]上取 4 个连续值属性数据集进行实验. 数据集的基本情况见表 1. 实验分成 2 个部分.

表 1　实验数据集的基本特性

序号	数据集	样本数	条件属性数	决策类别数
1	Wine	178	13	3
2	Ecoli	336	7	8
3	Iris	150	4	3
4	New-Thyroid	215	5	3

(1) 实验 1　实验 1 的主要目的是验证 HARCVA 方法的效果. 因此,本文给出了利用 SVM 分类器的识别准确率来验证属性约简效果的方法,实验中采用 10 折交叉验证的方法分别在表 1 的 4 个数据集上分别实验.

实验步骤如下:

Step1　(*HARCVA* 方法). 使用本文中 *HARCVA* 方

法分别对表 1 中的数据集约简,并把约简后的数据集作为 *SVM* 分类器的输入,并输出识别结果;

Step2 (CEBARKNC 方法). 使用属性重要性方法[14]分别对表 1 中的数据集作离散化处理,用 CEBARKNC 方法[15]进行属性约简,并把约简后的数据集作为 SVM 分类器的输入,并输出识别结果;

Step3 (NoReduct 方法). 把表 1 中未约简的各数据集(原始数据集)作为 SVM 分类器的输入,并输出识别结果.

实验中 HARCVA 方法具体的参数设置为: Wine, Ecoli, Iris 和 New-Thyroid 数据集 α 的值分别为 0.7, 0.6, 0.7, 0.7, β 的值分别为 0.1, 0.05, 0.02, 0.01; 支持向量机的参数设置为 SVM Type: C_SVC, Kernel Function: RBF, Multiclass Method: one-against-one.

实验结果见表 2.

表 2 支持向量机 SVM 识别结果

数据集	HARCVA 方法		CEBARKNC 方法		NoReduct 方法	
	约简属性个数	分类准确率(%)	约简属性个数	分类准确率(%)	原始属性个数	分类准确率(%)
Wine	10	95.51	3	67.98	13	96.07
Ecoli	5	86.01	4	78.87	7	86.01
Iris	3	96.0	3	95.33	4	97.33
New-Thyroid	4	96.74	3	92.09	5	96.74

从表 2 中实验数据我们看出: 总体来讲, HARCVA 方法与 CEBARKNC 方法都能降低原始数据集中条件属性的个数. 一方面, 在 HARCVA 方法中, Ecoli 与 New-Thyroid 数据集用较少的属性获得了与 NoReduct 方法相同的分类准确率, 虽然 Wine 及 Iris 数据集在 HARCVA 方法中获得的分率准确率稍低于 NoReduct 方法, 但此时属性个数已减少; 另一方面, 对比 HARCVA 方法与 CEBARKNC 方法, 我们也看出: 离散化方法得到的属性个数最少, 但分类准确率低. 实验结果说明: 相对于经典粗糙集理论处理连续值属性决策表的属性约简, HARCVA 方法是有效的.

(2) 实验 2 实验 2 的主要目的是验证 KACVA 算法的有效性. 实验中采用了完全不相交的测试集和训练集. 即设总样本集为 DataSet, 训练样本集为 TR, 测试样本集为 TE, 其中, DataSet = TR ∪ TE, TR ∩ TE = ∅. 实验中设定 TR 与 TE 的比例分别为 2∶1, 3∶2, 4∶1, 目的是考察在不同样本数目的条件下, KACVA 方法的实验结果与启发式值约简[14]、文献[16]及 C4.5 决策树分类方法[17]进行对比. 每组实验都进行 10 次, 实验结果取 10 次实验的平均值.

实验步骤如下:

Step1 (方法 1). 使用属性重要性方法[14]对 TR 进行离散化处理, 用 CEBARKNC 方法[15]进行属性约简, 用启发式值约简方法[14]生成决策规则, 用 TE 进行样本测试, 输出分类准确率;

Step2 (方法 2). 用文献[16]中的方法生成决策规则, 用 *TE* 进行样本测试, 输出分类准确率;

Step3 (方法 3). 使用 C4.5 决策树分类方法[17]生成决策规则, 用 TE 进行样本测试, 输出分类准确率;

Step4 (本文方法). 使用 HARCVA 方法属性约简, KACVA 方法生成决策规则, 用 TE 进行样本测试, 输出分类准确率.

实验中具体的参数设置为: KACVA 方法中, Wine, Ecoli, Iris 和 New-Thyroid 数据集的聚类数分别为 8, 20, 8, 8, 各数据集的规则生成阈值 λ 均为 0.6, HARCVA 方法的参数设置同实验 1.

方法 1 与方法 2 都采用了重庆邮电大学计算机科学与技术研究所开发的 RIDAS 测试平台[19]来实现, 且在样本测试的时候都采用了少数优先的匹配策略[14].

实验结果见表 3 及表 4.

表 3 Ecoli 与 Iris 数据集的实验结果

训练方法	Ecoli 数据集的分类准确率(%)			Iris 数据集的分类准确率(%)		
	|*TR*| / |*TE*| 2∶1	|*TR*| / |*TE*| 3∶2	|*TR*| / |*TE*| 4∶1	|*TR*| / |*TE*| 2∶1	|*TR*| / |*TE*| 3∶2	|*TR*| / |*TE*| 4∶1
方法 1	54.91	51.05	57.58	92.92	88.83	92.33
方法 2	55.8	51.43	60.3	93.33	91.67	93.0
方法 3	79.66	80.74	80.31	92.45	94.06	95.42
本文方法	87.50	85.04	84.70	94.79	96.33	96.67

表 3 与表 4 的实验数据表明: KACVA 方法在表 1 中 4 个数据集上, 在多种不同训练集大小的条件下, 获得了最高的分类准确率; 另一方面, 我们也看到, 方法 1 与方法 2 相对于方法 3 与本文方法在不同训练集大小的

前提下,分类准确率较低.这说明本文提出的 KACVA 方法对于连续值属性决策表具有较强的数据概括能力;同时,经典粗糙集方法处理连续值属性决策表知识获取方法时,事先要对连续值属性作离散化处理,离散化过程造成了信息损失,导致了分类准确率的下降.实验结果说明 KACVA 方法是有效的.

表 4　New-Thyroid 与 Wine 数据集的实验结果

训练方法	New-Thyroid 数据集的分类准确率(%)			Wine 数据集的分类准确率(%)		
	|*TR*|/|*TE*| 2:1	|*TR*|/|*TE*| 3:2	|*TR*|/|*TE*| 4:1	*TR*|/|*TE*| 2:1	|*TR*|/|*TE*| 3:2	|*TR*|/|*TE*| 4:1
方法 1	77.78	80.58	81.86	41.83	48.59	41.11
方法 2	80.28	82.09	83.02	43.67	50.14	43.83
方法 3	91.54	91.28	92.77	93.2	91.1	92.21
本文方法	95.97	94.07	94.65	95.17	94.65	94.72

7　结论与下一步工作

粗糙集理论中对连续值属性的决策表中的知识获取方法是一个重要的课题.本文针对这一问题进行了研究,提出了连续值属性决策表中的知识获取方法.该方法不需要事先对连续值属性作离散化处理,从而有效减少了连续值属性离散化的信息损失.UCI 数据集上的实验结果表明,与经典粗糙集相关方法及 C4.5 决策树方法相比,本文方法能够较好地适应于连续值属性的决策表中知识获取,并具有较好的性能.基于粗糙集理论的连续值属性决策表中的高效知识获取方法是我们下一步研究方向.

参考文献:

[1] 邓大勇,黄厚宽,李向军.不一致决策系统中约简之间的比较[J].电子学报,2007,35(2):252 - 255.
Deng Da-yong, Huang Hou-kuan, LI Xiang-jun. Comparison of various types of reductions in inconsistent systems [J]. Acta Electronica Sinica, 2007, 35(2): 252 - 255. (in Chinese)

[2] 杨明.决策表中基于条件信息熵的近似约简[J].电子学报,2007,35(11):2156 - 2160.
Yang Ming. Approximate reduction based on conditional information entropy in decision tables [J]. Acta Electronica Sinica, 2007, 35(11): 2156 - 2160. (in Chinese)

[3] 常梨云,王国胤,吴渝.一种基于理论的属性约简及规则提取方法[J].软件学报,1999,10(11):1207 - 1211.
Chang Li-yun, Wang Guo-yin, Wu Yu. An approach for attribute reduction and rule generation based on rough set theory [J]. Journal of Software, 1999, 10(11): 1207 - 1211. (in Chinese)

[4] 韩冰,高新波,姬红兵.基于模糊粗糙集的新闻视频镜头边界检测方法[J].电子学报,2006,34(6):1085 - 1089.
Han Bing, Gao Xin-bo, Ji Hong-bing. A shot boundary detection method for news video based on rough-fuzzy sets [J]. Acta Electronica Sinica, 2006, 34(6): 1085 - 1089. (in Chinese)

[5] 徐捷,徐从富,耿卫东,潘云鹤等.基于粗糙集理论的动态目标识别及跟踪[J].电子学报,2002,30(4):605 - 607.
Xu Jie, Xu Cong-fu, Geng Wei-dong, Pan Yun-he. Dynamic objects identifying and tracing based on rough set theory [J]. Acta Electronica Sinica, 2002, 30(4): 605 - 607. (in Chinese)

[6] 肖迪,胡寿松.实域粗糙集理论及属性约简[J].自动化学报,2007,33(3):253 - 258.
Xiao Di, Hu Shou-song. Real rough set theory and attribute reduction [J]. Acta Automatica Sinica, 2007, 33(3): 253 - 258. (in Chinese)

[7] 李兴生,李德毅.一种基于密度分布函数聚类德属性离散化方法[J].系统仿真学报,2003,15(6):804 - 806.
Li Xin-sheng, Li De-yi. A new method based on clustering for discretization of continuous attributes [J]. Journal of System Simulation, 2003, 15(6): 804 - 806. (in Chinese)

[8] 王熙照,赵素云,王静红.基于 Rough 集理论的模糊值属性信息表简化方法[J].计算机研究与发展,2004,41(11):1974 - 1981.
Wang xi-zhao, Zhao su-yun, Wang Jing-hong. Simplification of information table with fuzzy-valued attributes based on rough sets [J]. Journal of Computer Research and Development, 2004, 41(11): 1974 - 1981. (in Chinese)

[9] Jensen R, Shen Q. Semantics-preserving dimensionality reduction: rough and fuzzy-rough-based approaches [J]. IEEE Transactions on Knowledge and Data Engineering, 2004, 16(12): 1457 - 1471.

[10] Alicja M R, Leszek R. Remarks on approximation quality in variable precision fuzzy rough sets model [A]. RSCTC2004 [C]. LNAI3066, Berlin, Heidelberg: Springer Verlag, 2004. 402 - 411.

[11] 冯林,王国胤.用于数据分析的变精度模糊粗糙模型[J].西南交通大学学报,2008,43(5):582 - 587.
Feng Lin, Wang Guo-yin. Variable precision fuzzy rough model for data analysis [J]. Journal of Southwest Jiaotong University, 2008, 43(5): 582 - 587. (in Chinese)

[12] 胡清华,于达仁,谢宗霞.基于邻域粒化和粗糙逼近的数值属性约简[J].软件学报,2008,19(3):640 - 649.
Hu Qing-hua, Yu Da-ren, Xie Zong-xia. Numerical attribute reduction based on neighborhood granulation and rough approximation [J]. Journal of Software, 2008, 19(3): 640 - 649.

(in Chinese)

[13] Blake C, Keogh E, Merz C J, et al. UCI repository of machine learning databases [DB/OL]. [2006-12-20]. http://www.ics.uci.edu/~mlearn/MLRepository.html.

[14] 王国胤. Rough 集理论与知识获取[M]. 西安:西安交通大学出版社, 2003. 150 - 152.
Wang Guo-yin. Rough set theory and acknowledge acquisition [M]. Xi'an: Xi'an Jiaotong University, 2003. 150 - 152. (in Chinese)

[15] 王国胤, 于洪, 杨大春. 基于条件信息熵的决策表约简[J]. 计算机学报, 2002, 25(7): 759 - 766.
Wang Guo-yin, Yu Hong, Yang Da-chun. Decision table reduction based on conditional information entropy [J]. Chinese Journal of Computers, 2002, 25(7): 759 - 766. (in Chinese)

[16] 王国胤, 何晓. 一种不确定条件下的自主式知识学习模型[J]. 软件学报, 2003, 14(6): 1096 - 1102.
Wang Guo-yin, He Xiao. A self-learning model under uncertain condition [J]. Journal of Software, 2003, 14(6): 1096 - 1102. (in Chinese)

[17] Quinlan J R. C4.5: programs for machine learning [M]. San Francisco: Morgan Kaufmann Publishers, 1993.

[18] 黄金杰, 武俊峰, 蔡云泽. 模糊粗糙集数据模型: 一种数据分析的新方法[J]. 计算机学报, 2005, 28(11): 1866 - 1874.
Huang Jin-jie, Wu Jun-feng, Cai Yun-ze. Fuzzy rough data model: a new technique for analyzing data [J]. Chinese Journal of Computer, 2005, 28(11): 1866 - 1874. (in Chinese)

[19] Wang G Y, Zheng Z, Zhang Y. RIDAS-a rough set based intelligent data analysis system [A]. Proceedings of ICML C2002 [C], Beijing: IEEE Press, 2002. 646 - 649.

作者简介:

冯 林 男, 1972 年 10 月出生于四川巴中. 博士研究生、副教授, 主要研究方向为: 粗糙集、粒计算、数据挖掘.
E-mail: scfengyc@126.com

王国胤 男, 1970 年生于重庆. 博士、教授、博士生导师、IEEE 高级会员. 主要研究方向为: 粗糙集、粒计算、智能信息系统、神经网络等. (通信作者) E-mail: wanggy@ieee.org

李天瑞 男, 1969 年生于福建莆田. 博士后、教授、博士生导师. 主要研究方向为: 粗糙集、粒计算、智能信息处理、数学模型等.