文章编号: 1001-0920 (2012) 04-0562-05

一种实域粗糙集模型及属性约简方法

冯 林, 原永乐, 苟仕蓉, 杨 军, 沈 莉

(四川师范大学 a. 计算机科学学院, b. 可视化计算与虚拟现实四川省重点实验室, 成都 610101)

摘 要: 通过对实域区间和决策值的重新划分, 对已经存在的属性广义重要度度量准则进行了扩展, 构建了对象空间上的广义邻域关系及广义邻域关系下的实域粗糙集模型, 并在此基础上提出了实域决策系统中属性约简方法(ARRDDS). 对不同数据集的实验测试结果表明, 与其他相关方法相比, ARRDDS 方法能够较好地处理决策表中实数域属性约简问题.

关键词: 广义重要度; 属性约简; 邻域关系; 实域粗糙集; 决策理论粗糙集模型

中图分类号: TP18 **文献标识码:** A

An approach for real domain rough sets model and attribute reduction

FENG Lin, YUAN Yong-le, GOU Shi-rong, YANG Jun, SHEN Li

(a. College of Computer Science, b. Sichuan Key Laboratory of Visualization Computing and Virtual Reality, Sichuan Normal University, Chengdu 610101, China. Correspondent: FENG Lin, E-mail: scfengyc@126.com)

Abstract: Through repartition for the real domain interval values and the decision classifications, an extended method of calculating the general important degree of an attribute and attribute subsets is proposed. A general neighborhood relation is established on objects space. Then a rough set model is constructed based on the general neighborhood relation. Furthermore, an approach for attribute reduction in a real domain decision system(ARRDDS) is developed. Results of experimental evaluation on different data sets show that ARRDDS has the better performance comparing with the related rough sets approaches of attribute reduction in processing decision tables with real-valued attributes.

Key words: general important degree; attribute reduction; neighborhood relation; real domain rough sets; decision-theoretic rough set

1 引 言

自波兰科学家 Pawlak 教授于 1982 年提出粗糙集理论以来, 该理论作为一种处理不确定和含糊信息的重要数学工具, 日益为人们所接受并得到不断发展和完善, 已在属性约简、决策规则生成、故障诊断等方面取得了较为成功的应用[1-3]. 属性约简是粗糙集理论的重要应用, 也是其核心问题之一[4]. 但是, 经典粗糙集理论模型建立在不分明关系(等价关系)的基础上, 所处理的属性值是清晰的离散值. 然而, 对于现实生活中广泛存在的连续的实数属性取值, 如医疗信息系统中的年龄、体温和血压等属性则不能直接处理. 经典粗糙集理论在处理具有这类性质的决策表中的属性约简时, 常采用将实数值数据进行离散化[5], 再用传统的粗糙集方法求解属性约简, 但这种先期离散化的处理方法会导致一些信息损失[6]. 为了解决这一问题, 人们引入了实域粗糙集理论[7]以及模糊粗糙集理论模型[8-9]. 其中文献 [7] 定义了属性的广义重要度和属性子集广义重要度计算方法, 以此度量样本空间中样本之间的相似性距离, 然后以广义欧式距离为基础建立样本空间中的广义近邻关系, 并给出了实域粗糙集理论的属性约简定义及其属性约简方法. 该文中的方法不需要对实数属性进行离散化处理, 从而减少了离散化过程对信息的损失.

虽然文献 [7] 对实数域决策表中的属性约简方法进行了深入探讨, 但该方法对于某些特定的实数域属性仍然不能较好地处理. 为此, 本文首先通过一个实例, 指出文献 [7] 提出的属性广义重要度度量方法需进一步扩展; 然后通过对实域区间及决策值的重新划分, 提出一种扩展的属性广义重要度度量准则, 并以此为基础构建了广义邻域关系下的实域粗糙集模型;

收稿日期: 2010-09-28; **修回日期:** 2011-10-04.

基金项目: 四川省教育厅科研基金项目(09ZC079).

作者简介: 冯林(1972−), 男, 副教授, 从事智能信息处理、粗糙集理论与方法等研究; 原永乐(1986−), 男, 硕士生, 从事粗糙集理论与应用的研究.

同时对该模型的性质进行了分析, 并基于此模型及性质提出了实域决策系统中属性约简定义及其属性约简方法(ARRDDS); 最后通过实验结果与分析表明了所提出的方法能够较好地处理决策表中实数域属性约简问题.

2　属性广义重要度的扩展方法

定理 1 [7]　给定决策系统

$$S=(U,C\bigcup\{d\},V),$$

$$U/\{d\}=\{d_1,d_2,\cdots,d_{r(d)}\},\ \exists a\in C,$$

设 $a(d_i)\subset V$ 表示属性 a 对应决策 d_i 的属性值子集区间, 如果有

$$a(d_i)\bigcap a(d_j)=\varnothing,\ i\neq j,\ i,j=1,2,\cdots,r(d),$$

则 $U/a=U/d$ 成立.

定义 1 [7]　给定决策系统

$$S=(U,C\bigcup\{d\},V),$$

$$U/\{d\}=\{d_1,d_2,\cdots,d_{r(d)}\},\ \forall a\in C,$$

定义属性 a 的广义重要度为

$$\sigma_g(a)=\begin{cases}1-\dfrac{1}{C^2_{r(d)}}\displaystyle\sum_{i\neq j,\ i,j=1}^{r(d)}\frac{a(d_i)\bigcap a(d_j)}{\max\ \mathrm{cross}(a(d))}, & \text{其他;}\\ 1,\ \forall i,j,a(d_i)\bigcap a(d_j)=\varnothing.\end{cases}\tag{1}$$

其中: $C^2_{r(d)}$ 表示 $r(d)$ 中取 2 的组合, $a(d_i)\bigcap a(d_j)\subset V$ 表示属性 a 对应决策值 d_i 的属性子集与对应决策值 d_j 的属性值子集的交集部分, $\max\ \mathrm{cross}(a(d))\subset V$ 表示属性 a 对应全部两两决策值的属性值子集的所有交集所包围的最大区间.

下面采用文献[7]中的例子分析上述定义的扩展方法.

例 1　实数决策表如表 1 所示, 文献[7]得到的广义属性重要度的计算结果如下:

$$\sigma_g(a_1)=0.533,\ \sigma_g(a_2)=0.333,\ \sigma_g(a_3)=1.$$

于是, 文献[7]得出如下结论: 属性 a_2 对应各个决策的值集重合很大, 因此它的广义重要度很小, 即对决策分类的帮助很小; 而属性 a_3 对应各个决策的值集均不相交, 因此它的广义重要度为 1, 即可凭借这一属性进行分类.

表 1　实数值决策表[7]

U	a_1	a_2	a_3	d
x_1	1.6	0.4	5.7	1
x_2	2.2	0.6	6.8	1
x_3	1.9	0.4	5.3	2
x_4	3.4	0.5	4.8	2
x_5	2.0	0.4	6.9	3
x_6	2.9	0.6	7.5	3

下面对表 1 增加 2 个对象 x_7 和 x_8, 如表 2 所示.

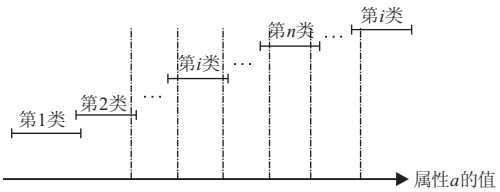
表 2　1 个实数值决策表

U	a_1	a_2	a_3	d
x_1	1.6	0.4	5.7	1
x_2	2.2	0.6	6.8	1
x_3	1.9	0.4	5.3	2
x_4	3.4	0.5	4.8	2
x_5	2.0	0.4	6.9	3
x_6	2.9	0.6	7.5	3
x_7	1.0	0.5	8.5	1
x_8	1.3	0.4	9.5	1

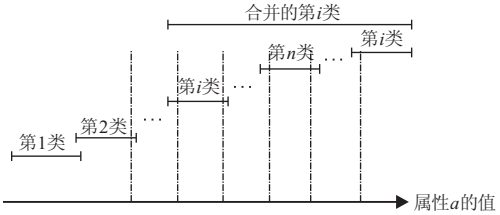
对于表 2, $U/\{d\}=\{\{x_1,x_2,x_7,x_8\},\{x_3,x_4\},\{x_5,x_6\}\}=\{d_1,d_2,d_3\}$. 属性 a_1,a_2 和 a_3 的广义属性重要度按定义 1 计算如下:

$$\sigma_g(a_1)=0.24,\ \sigma_g(a_2)=0.333,\ \sigma_g(a_3)=0.667.$$

对比表 1 和表 2 中的属性 a_3 可以发现, a_3 的广义属性重要度也应为 1, 这是因为增加的 2 个对象 x_7 和 x_8 区间为 $[8.5,9.5]$, 位于例 1 中 a_3 的区间之外, 它的决策值为 1. 这类似于在经典粗糙集模型中, 一个决策分类可以由多个条件属性等价类的并集组成, 而条件属性对决策分类并没有影响. 这种情况在现实生活中也有诸多实例, 比如在统计某一地区人的收入水平时, 决策分为高、中、低收入 3 类, 对其中某一类而言, 若某人的月收入为 [5 000, 6 000] 元, 则可以将它列入高收入人群. 然而, 一个人的月收入分类情况与多种因素有关, 如学历、工作年限、年龄、所从事的具体工作等. 因此, 对于某一特定人群, 可能他 (她) 的月收入为 [2 000, 3 000] 元便为高收入人群. 如果按定义 1, 则会将 [2 000, 6 000] 这个区间列入高收入人群. 显然, 这种简单合并区间并不恰当, 造成这种情况的原因是用相同的决策值对应的实数区间来衡量属性的重要度. 换言之, 定义 1 将图 1(a) 所示各类间的区间情况处理为图 1(b) 所示各类间的区间情况不完全合适. 下面对定义 1 进行扩展, 使之能够更合理地描述实值属性区间对决策分类的影响.



(a) 属性 a 的值与各决策分类之间的分布示意



(b) 合并的属性 a 的值与各决策分类之间的分布示意

图 1　实域属性 a 的取值与决策分类之间的关系

给定实域决策系统 $S=(U,C\bigcup\{d\},V,f)$ (下文中

将采用文献[10]中决策系统的定义, 如果$\forall a\in C, V_a$的取值为连续实数值, 则称S为一个实域决策系统). 不失一般性, 设$a(d_\hbar)$包含的区间为$a(d_i), a(d_{i+1}), \cdots, a(d_\lambda)$, 并记$(c_\hbar, r_\hbar), (c_i, r_i), (c_{i+1}, r_{i+1}), \cdots, (c_\lambda, r_\lambda)$分别为$a(d_\hbar), a(d_i), a(d_{i+1}), \cdots, a(d_\lambda)$的中心点和半径, 这样便将$a(d_\hbar)$分割成区间$[c_\hbar - r_\hbar, c_i - r_i)$和$(c_\lambda + r_\lambda, c_\hbar + r_\hbar]$. 于是, 可给出如下扩展的属性广义重要度度量方法:

定义 2 给定实域决策系统$S=(U, C\bigcup\{d\}, V, f)$, 对任意的$a\in C, U/\{d\}=\{d_1, d_2, \cdots, d_r\}, \forall a\in C$, 属性$a$的广义重要度定义如下:

$$\sigma(a)=\begin{cases}1-\dfrac{1}{C_r^2}\displaystyle\sum_{i\neq j,\, i,j=1}^{r}\dfrac{a(d_i)\bigcap a(d_j)}{\max\,\mathrm{cross}\,(a(d))}, \text{ 其他;}\\ 1-\dfrac{1}{C_k^2}\displaystyle\sum_{i\neq j,\, i,j=1}^{k}\dfrac{a(d_i')\bigcap a(d_j')}{\max\,\mathrm{cross}\,(a(d'))},\\ \quad \exists i,j, a(d_i)\bigcap a(d_j)=a(d_i)\text{ or }a(d_j).\end{cases}\tag{2}$$

如果$a(d_i)\bigcap a(d_j)=\varnothing$或$a(d_i')\bigcap a(d_j')=\varnothing$, 则定义它们的区间长度为0. 式(2)中: C_k^2表示从k中取2的组合, $a(d_i')\bigcap a(d_j')\subset V$表示属性$a$对应决策值$d_i'$的属性子集与对应决策值$d_j'$的属性值子集的交集部分, $\max\,\mathrm{cross}\,(a(d'))\subset V$表示属性$a$对应全部两两决策值的属性值子集的所有交集所包围的最大区间; k为属性a重新按$a(d_\hbar)$分割后对决策空间划分的区间个数; $d_1', d_2', \cdots, d_k'$为相应决策空间的决策值.

由表2及定义3, 可得到a_1的广义属性重要度为

$$\sigma(a_1)=$$
$$1-\frac{1}{C_4^2}\sum_{i\neq j,\, i,j=1}^{4}\frac{a(d_i')\bigcap a(d_j')}{[1.9, 2.2]}=$$
$$1-\frac{1}{6}\Big(\frac{[1.9, 2.2]+\varnothing+[2.0, 2.2]+\varnothing+\varnothing+\varnothing}{[1.9, 2.2]}\Big)=$$
$$0.72.$$

同理, $\sigma(a_2)=0.9, \sigma(a_3)=1$.

定理 2 给定实域决策系统$S=(U, C\bigcup\{d\}, V)$, 对任意的$a\in C$, 有以下结论成立:

1) $0<\sigma(a)\leqslant 1$;

2) $\sigma(a)$是$\sigma_g(a)$的扩展.

证明 1) 由定义2, 有

$$0\leqslant\frac{1}{C_r^2}\sum_{i\neq j,\, i,j=1}^{r}\frac{a(d_i)\bigcap a(d_j)}{\max\,\mathrm{cross}\,(a(d))}<1,$$

$$0\leqslant\frac{1}{C_k^2}\sum_{i\neq j,\, i,j=1}^{k}\frac{a(d_i')\bigcap a(d_j')}{\max\,\mathrm{cross}\,(a(d'))}<1,$$

于是, $0<\sigma(a)\leqslant 1$成立.

2) 在定义2中, 如果$a(d_i)\bigcap a(d_j)=\varnothing$或$a(d_i')\bigcap a(d_j')=\varnothing$, 则$\sigma(a)=1$. 另外, 若$k$是对属性$a$重新分割后对决策空间划分的区间个数, $d_1', d_2', \cdots, d_k'$为相应决策空间的决策值, 则当$d_1', d_2', \cdots, d_k'$退化为$d_1, d_2, \cdots, d_r$时, 有

$$\sigma(a)=1-\frac{1}{C_r^2}\sum_{i\neq j,\, i,j=1}^{r}\frac{a(d_i)\bigcap a(d_j)}{\max\,\mathrm{cross}\,(a(d))}.$$

于是, 定义2退化为定义1, 即$\sigma(a)$是$\sigma_g(a)$的扩展. □

3 实域粗糙集理论模型及分析

对于实数域决策系统, 可用广义欧式距离来度量论域中两个对象的相似程度.

定义 3 设实域决策系统$S=(U, C\bigcup\{d\}, V, f)$, $B\subseteq C$, 对任意的$x, y\in U$和$a_i\in B$, x和y在实域知识空间B上的广义重要度欧式距离$d_B(x, y)$定义如下:

$$d_B(x,y)=\sqrt{\sum_{i=1}^{|B|}\sigma(a_i)(f(x,a_i)-f(y,a_i))^2}\,.\tag{3}$$

显然, 如果$\sigma(a_i)=1$, 则定义3退化为普通的欧式距离公式.

定义 4 设实域决策系统$S=(U, C\bigcup\{d\}, V, f)$, $B\subseteq C$, 阈值$\delta\geqslant 0$, 对任意的$x\in U$, x在实域知识空间C上的邻域$N_B^\delta(x)$定义如下:

$$N_B^\delta(x)=\{y\in U|d_B(x,y)\leqslant\delta\}.\tag{4}$$

基于上述讨论, 下面给出实域粗糙集理论模型.

定义 5 设实域决策系统$S=(U, C\bigcup\{d\}, V, f)$, $B\subseteq C$, 对任意的$X\subseteq U$, X在实域知识空间B上的δ下近似、δ上近似定义如下:

$$\underline{B}^{\,\delta}(X)=\{x\in U|N_B^\delta(x)\subseteq X\},$$
$$\overline{B}^{\,\delta}(X)=\{x\in U|N_B^\delta(x)\bigcap X\neq\varnothing\}.\tag{5}$$

定义 6 设实域决策系统$S=(U, C\bigcup\{d\}, V, f)$, $B\subseteq C$, B对d的δ近似分类质量$\gamma_B^\delta(\{d\})$定义如下:

$$\gamma_B^\delta(\{d\})=\frac{\Big|\displaystyle\bigcup_{X_i\in U/\{d\}}\underline{B}^\delta(X_i)\Big|}{|U|}.\tag{6}$$

定理 3 设实域决策系统$S=(U, C\bigcup\{d\}, V, f)$, $\delta\geqslant 0$, 有以下性质成立:

1) 对于任意的$x\in U$和$A_1\subseteq A_2\subseteq\cdots\subseteq C$, 有

$$N_{A_1}^\delta(x)\supseteq N_{A_2}^\delta(x)\supseteq\cdots\supseteq N_C^\delta(x);$$

2) 对于任意的$x\subseteq U$和$A_1\subseteq A_2\subseteq\cdots\subseteq C$, 有

$$\underline{A}_1^\delta(X)\subseteq\underline{A}_2^\delta(X)\subseteq\cdots\subseteq\underline{C}^\delta(X);$$

3) 对于任意的$A_1\subseteq A_2\subseteq\cdots\subseteq C$, 有

$$\gamma_{A_1}^\delta(\{d\})\leqslant\gamma_{A_2}^\delta(\{d\})\leqslant\cdots\leqslant\gamma_C^\delta(\{d\});$$

4) 对于任意的$X_1\subseteq X_2\subseteq\cdots\subseteq C$, 有

$$\underline{C}^\delta(X_1)\subseteq\underline{C}^\delta(X_2)\subseteq\cdots\subseteq\underline{C}^\delta(U),$$
$$\bar{C}^\delta(X_1)\subseteq\bar{C}^\delta(X_2)\subseteq\cdots\subseteq\bar{C}^\delta(U).$$

证明　对于性质 1), 不失一般性, 先证 $N_{A_1}^{\delta}(x)\supseteq N_{A_2}^{\delta}(x)$. 设对于任意的 $y\in N_{A_2}^{\delta}(x)$, 由定义 5 有 $d_{A_1}(x,y)\leqslant d_{A_2}(x,y)$ 成立. 如果 $d_{A_2}(x,y)\leqslant\delta$, 则必有 $d_{A_1}(x,y)\leqslant\delta$, 即 $y\in N_{A_1}^{\delta}(x)$. 另外, 对于任意的 $y\in N_{A_1}^{\delta}(x)$, 根据定义 5, $y\in N_{A_2}^{\delta}(x)$ 不一定成立.

根据以上分析, 有 $N_{A_1}^{\delta}(x)\supseteq N_{A_2}^{\delta}(x)$ 成立. 同理, 可以得到结论 $N_{A_2}^{\delta}(x)\supseteq N_{A_3}^{\delta}(x)$. 以此类推, 有

$$N_{A_1}^{\delta}(x)\supseteq N_{A_2}^{\delta}(x)\supseteq\cdots\supseteq N_{C}^{\delta}(x)$$

成立.

性质 2)～性质 4) 可根据以上相关定义证明, 在此不再赘述. □

定理 3 中的性质 1)～性质 3) 分别指出了具有包含关系的属性子集之间的邻域、下近似集、近似分类质量的单调关系, 它们在属性约简算法中具有非常重要的作用, 而性质 4) 则指出了具有包含关系的对象子集相对于特定的条件属性空间上的上、下近似集之间的单调关系.

4　属性约简方法 ARRDDS

定义 7　设实域决策系统 $S=(U,C\bigcup\{d\},V,f)$, $B\subseteq C,\delta\geqslant 0$, B 称为 S 中一个属性约简, 当且仅当满足以下 2 个条件:

$$\gamma_B^{\delta}(\{d\})=\gamma_C^{\delta}(\{d\}),\tag{7a}$$

$$\forall b\in B(\gamma_{B\setminus\{b\}}^{\delta}(\{d\})\neq\gamma_B^{\delta}(\{d\})).\tag{7b}$$

定义 7 的条件 (7a) 保证一个属性约简与实域决策系统中所有条件属性具有相同的 δ 近似分类精度; 条件 (7b) 保证在一个属性约简中, 不存在多余的属性. 即定义 7 保证了 C 的一个最小属性子集 B 与 C 相对于决策 d 的 δ 近似分类能力相等.

从上述属性约简的定义可以看出, 它与文献 [7] 给出的属性约简定义的含义是完全不同的.

定义 8　设实域决策系统 $S=(U,C\bigcup\{d\},V,f)$, $B\subseteq C$, $\delta\geqslant 0, a\in C\setminus B$, $\text{ability}\,(\{a\},B,\{d\})$ 称为 a 相对于 B 对 d 的分类能力大小, 定义如下:

$$\text{ability}\,(\{a\},B,\{d\})=1-\frac{\left|\bigcup\limits_{X_i\in U/\{d\}}\underline{B}^{\delta}(X_i)\right|}{\left|\bigcup\limits_{X_i\in U/\{d\}}\underline{B\bigcup\{a\}}^{\delta}(X_i)\right|}.\tag{8}$$

显然, $0\leqslant\text{ability}\,(\{a\},B,\{d\})\leqslant 1$.

下面给出属性约简算法 ARRDDS.

算法 1　属性约简算法 ARRDDS.

输入: 实域决策系统 $S=(U,C\bigcup\{d\},V)$ 及阈值 δ;

输出: S 中的一个属性约简 R.

Step 1: $R\leftarrow\varnothing$, $1\leftarrow\left|\bigcup\limits_{X_i\in U/\{d\}}\underline{R}^{\delta}(X_i)\right|$, $T\leftarrow C$, $\gamma_R^{\delta}(\{d\})\leftarrow 0$.

Step 2: 对任意的 $x\in U$, 计算 C 上的邻域 $N_C^{\delta}(x)$.

Step 3: 计算 S 中决策 d 相对条件属性集 C 的 δ 近似分类质量 $\gamma_C^{\delta}(\{d\})$.

Step 4: While $\gamma_R^{\delta}(\{d\})<\gamma_C^{\delta}(\{d\})$ Do

Step 4.1: 计算 T 中每一个 a 的 $\text{ability}\,(\{a\},R,\{d\})$;

Step 4.2: 选择一个使 $\text{ability}\,(\{a\},R,\{d\})$ 取值最大的 a, $R\leftarrow R\bigcup\{a\}$;

Step 4.3: $T\leftarrow T\setminus\{a\}$;

End Do

Step 5: 输出 R, 结束.

定理 3 中性质 4) 保证如果 R 不是一个约简, 则必有 $\gamma_R^{\delta}(\{d\})<\gamma_C^{\delta}(\{d\})$ 成立. 因此, 该算法以空集为起点, 每次计算全部剩余的 $\text{ability}\,(\{a\},R,\{d\})$, 从中选择 $\text{ability}\,(\{a\},R,\{d\})$ 值最大的属性加入约简集合中, 直至循环条件结束.

下面分析算法 1 的时间复杂度: 算法 1 的时间复杂度主要由 Step 2, Step 3 和 Step 4 决定. Step 2 的时间复杂度为 $O(|C||U|^2)$, Step 3 的时间复杂度为 $O(|U|^2)$, Step 4 最坏情况下的时间复杂度为 $O(|C|^2|U|^2)$. 因此, 算法 1 的时间复杂度为 $O(|C|^2|U|^2)$.

5　实验结果与分析

为测试本文算法的效果, 选择 3 个数据集 (Wine 数据集[11]、Ecoli 数据集[11] 和一个 RANDDATA 数据集) 进行实验. Wine 数据集包括 3 个模式类, 每个样本包括 13 个实域条件属性, 每个条件属性被归一化为 $[0,1]$, 样本数为 178; Ecoli 数据集包括 8 个模式类, 每个样本包括 7 个实域条件属性, 每个条件属性被归一化为 [0,1], 样本数为 336; RANDDATA 是一个随机数据集, 使用随机数据集的目的是为了使它明显符合图 1(a) 中的各实数属性的区间特征, 它包括 3 个模式类, 每个样本包括 9 个实域条件属性, 随机产生的样本数为 1 000.

实验步骤如下:

Step 1: 使用本文 ARRDDS 方法、文献 [7] 中的方法和 HARCVA 方法[10]分别对 Wine 数据集、Ecoli 数据集和 RANDDATA 数据集进行属性约简, 分别记录其属性约简结果和约简运行时间, 并将约简后的数据集作为 SVM 分类器的输入, 采用 10 折交叉验证的方法, 并输出识别结果;

Step 2: 使用属性重要性方法[12]分别对 Wine 数据集、Ecoli 数据集和 RANDDATA 数据集作离散化处

理, 用CEBARKNC方法[13]进行属性约简, 记录其属性约简结果和约简运行时间, 并将约简后的数据集作为SVM分类器的输入, 采用10折交叉验证的方法, 并输出识别结果.

实验中, HARCVA方法中Wine和Ecoli数据集α的值分别为0.7和0.6, β的值分别为0.1和0.05; 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3, 表3中方法1, 方法2和方法3分别为文献[7], [10]和[13]中的方法.

表 3 多种粗糙集方法比较

数据集	实验内容	属性约简方法			
		本文方法	方法1	方法2	方法3
Wine	属性约简个数	5	11	10	3
	属性约简时间/s	3.1	2.5	4.3	3.6
	SVM分类准确率/%	98.31	96.63	95.51	67.98
Ecoli	属性约简个数	5	7	5	4
	属性约简时间/s	4.4	3.9	5.2	4.8
	SVM分类准确率/%	86.01	86.01	86.01	78.87
Randdata	属性约简个数	6	8	7	4
	属性约简时间/s	5.7	5.4	6.9	6.2
	SVM分类准确率/%	83.02	79.04	79.98	80.42

从表3可以看出: 本文ARRDDS方法与文献[7]约简方法相比, 虽然文献[7]约简方法需要的时间较少, 但保留的属性的个数多, 且分类准确率较低; 同时, 与经典HARCVA方法[10]及CEBARKNC方法[13]相比, 本文方法需要的时间最少, 且分类准确率最高. 另一方面, 从属性约简个数看, CEBARKNC方法虽然得到的属性个数少, 但分类准确率低. 综合以上分析, 实验结果表明本文方法能有效降低实域数据集中条件属性的维数, 取得较好的结果.

6 结 论

属性约简(也称特征选择)是近年来模式识别、数据挖掘、机器学习等领域研究的一个热点, 为此本文提出了一种基于实域粗糙集理论的属性约简新方法. 文中对现有文献属性及属性子集广义重要度的计算方法进行了扩展, 构建了实域信息系统上的邻域粗糙集模型, 并给出了实域信息系统中基于邻域粗糙集模型的属性约简方法ARRDDS. 与相关主要属性约简方法的对比实验表明, 本文给出的属性约简方法是有效的.

参考文献(References)

[1] Pawlak Z. Rough sets: Theoretical aspects of reasoning about data[M]. Berlin: Springer, 1991.

[2] 冯少荣, 张东站. 一种高效的增量式属性约简算法[J]. 控制与决策, 2011, 26(4): 495-500.
(Feng S R, Zhang D Z. Effective increment algorithm for attribute reduction[J]. Control and Decision, 2011, 26(4): 495-500.)

[3] Yao Y Y, Zhao Y. Attribute reduction in decision-theoretic rough set models[J]. Information Sciences, 2008, 178(17): 3356-3373.

[4] Qian Y H, Liang J Y, Pedrycz W, et al. Positive approximation: An accelerator for attribute reduction in rough set theory[J]. Artificial Intelligence, 2010, 174(9/10): 597-618.

[5] Jensen R, Shen Q. Fuzzy-rough sets assisted attribute selection[J]. IEEE Trans on Fuzzy Systems, 2007, 15(1): 73-89.

[6] Hu Q H, Yu D R, Liu J F, et al. Neighborhood rough set based heterogeneous feature subset selection[J]. Information Sciences, 2008, 178(18): 3577-3594.

[7] 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 D, Hu S S. Real rough set theory and attribute reduction[J]. Acta Automatica Sinica, 2007, 33(3): 253-258.)

[8] Jensen R, Shen Q. Fuzzy-rough sets assisted attribute selection[J]. IEEE Trans on Fuzzy Systems, 2007, 15(1): 73-89.

[9] Jensen R, Shen Q. Are more features better? A response to attributes reduction using fuzzy rough sets[J]. IEEE Trans on Fuzzy Systems, 2009, 17(6): 1456-1458.

[10] 冯林, 王国胤, 李天瑞. 连续值属性决策表中的知识获取方法[J]. 电子学报, 2009, 37(1): 2432-2438.
(Feng L, Wang G Y, Li T R. Knowledge acquisition from decision tables containing continuous-valued attributes[J]. Acta Electronica Sinic, 2009, 37(11): 2432-2438.)

[11] 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.

[12] Wang G Y, Zheng Z, Zhang Y. RIDAS — A rough set based intelligent data analysis system[C]. Proc of the 1st Int Conf on Machine Learning and Cybernetics. Beijing: IEEE Computer Society, 2002: 646-649.

[13] 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 G Y, Yu H, Yang D C. Decision table reduction based on conditional information entropy[J]. Chinese J of Computers, 2002, 25(7): 759-766.)