

文章编号:1001-9081(2010)01-0156-03

基于相似关系粗糙集模型的数值属性约简算法

吴 敏

(合肥工业大学 电气与自动化工程学院, 合肥 230009)

(min.wu.hfut@gmail.com)

摘要:针对数值属性数据包含大量噪声而经典粗糙集方法易受噪声干扰的问题,提出一种属性度量指标综合衡量属性在样本上的差异性和相似性。以这种属性度量指标为启发式设计了相似关系粗糙集框架下的数值属性约简算法,并推广到经典粗糙集。在车牌字符集和UCI手写体数字字符集中和常用约简算法进行了比较,实验结果显示这种方法产生的约简属性可以导出规则数少并且具有较好分类能力的规则集。

关键词:字符识别;粗糙集;属性约简;特征选择;相似关系;数值属性

中图分类号: TP391 **文献标志码:**A

Algorithm of numerical attributes reduction based on similarity rough set

WU Min

(School of Electrical Engineering and Automation, Hefei University of Technology, Hefei Anhui 230009, China)

Abstract: As to the problem of interferential or noisy data reduction, an attribute significance evaluation principle was proposed based on the difference and similarity of attributes within objects. A numerical attributes reduction algorithm was constructed based on similarity rough set model, and it was extended to canonical rough set too. Experiments were carried out on two data sets, one is of license plate characters and the other is of UCI handwritten number. The experimental results show that the proposed algorithm can generate simpler but more powerful rule set than other reduction algorithms.

Key words: characters recognition; rough set; attributes reduction; feature selection; similarity relation; numerical attribute

0 引言

粗糙集理论是一种处理模糊和不确定性知识的数学工具^[1]。粗糙集属性约简是一种有效的特征选择方法,被广泛应用于知识库约简^[2]和特征选择^[3]。近年来,一些关于字符识别的文献使用粗糙集属性约简方法,删除冗余属性,搜索分辨力高的属性集合,然后利用约简的属性集构成分类器^[4]。这样可以降低分类器维度,并且保持分类能力。字符样本的属性一般是连续的数值型属性,受噪声影响大。基于等价关系的粗糙集处理数值属性数据时一般要经过离散化过程。根据信息论,离散化过程必然引入信息损失。另外,等价关系粗糙集属性约简仅考虑属性集合在不同类样本上的差异性,一些在同类异类对象之间变化频繁、易受干扰的特征也会被选择,这些特征会降低分类器的性能。为了克服经典粗糙集理论处理数值属性对象的不合理性,一些文献将经典等价关系粗糙集理论扩展到模糊粗糙集^[5-6]、相似关系^[7-8]和邻域关系^[9]粗糙集。胡清华等提出了基于邻域粗糙集模型的分类器^[10]和数值约简算法^[11]。本文在相似关系粗糙集框架中,提出一种结合属性综合分辨能力评价指标的数值属性约简算法。这种约简算法选择在同类样本间差异小的属性,可以克服粗糙集理论对噪声、干扰敏感的缺点。

1 相似关系粗糙集及属性约简

1.1 相似关系粗糙集模型

粗糙集理论将研究对象的集合称为论域 U , 对象的属性

值域为实数时论域空间为实数空间。相似粗糙集和邻域粗糙集通过在实数空间定义差异性度量,以此为基础建立相似关系和邻域关系,构成相似类和邻域粒子,从而实现实数空间概念的粒化。下面定义实数空间中的差异性度量。

定义1 论域中任意两个样本 $t_i, t_j (i \neq j)$ 在数值属性 a 上的相对差异性为:

$$diff(a, t_i, t_j) = \frac{|a(t_i) - a(t_j)|}{\max(V_a) - \min(V_a)} \quad (1)$$

其中 V_a 为属性 a 的值集合,类似地定义样本 t_i, t_j 在符号属性或离散化属性 a 上的差异性为:

$$diff(a, t_i, t_j) = \begin{cases} 1, & a(t_i) \neq a(t_j) \\ 0, & a(t_i) = a(t_j) \end{cases} \quad (2)$$

定义2 定义属性 a 上的相似关系如下:

$$SIM_a = \{(x, y) \in U \times U \mid diff(a, x, y) \leq t_a\} \quad (3)$$

属性集合 A 上的相似关系为:

$$SIM_A = \{(x, y) \in U \times U \mid \forall a \in A, diff(a, x, y) \leq t_a\} \quad (4)$$

其中 t_a 为相似性阈值, $t_a \in [0, 1]$ 。 $x, y \in U$ 在属性 a (属性集合 A) 上相似记为 $xSIM_a y$ ($xSIM_A y$)。

相似关系允许相似个体存在差异,是比等价关系宽松的二元关系,当相似性阈值 $t_a = 0$,相似关系退化为等价关系。

定义3 在决策表上定义个体 $x \in U$ 的相似类为:
 $SIM_A(x) = \{y \in U \mid ySIM_A x\}$ 。根据相似关系定义集合 $X \in U$ 的下近似集为:

$$SIM_A(X) = \{x \in U \mid SIM_A(x) \subseteq X\}.$$

上近似集为:

$$\overline{SIM_A}(X) = \bigcup_{x \in X} SIM_A(x)$$

显然 $\overline{SIM_A}(X) \subseteq X \subseteq \overline{\overline{SIM_A}}(X)$, X 的近似边界为:

$$BN(X) = \overline{SIM_A}(X) - \overline{\overline{SIM_A}}(X)$$

设决策属性 $\{d\}$ 将论域 U 划分为 $X = \{X_1, X_2, \dots, X_{r(d)}\}$, 集合

$$POS(SIM_A, \{d\}) = \bigcup_{i=1}^{r(d)} SIM_A(X_i)$$

称为相似关系下划分 X 的正区域。

相似关系粗集模型下定义相对约简如下。

定义4 属性集合 $R \in C$ 是相似关系意义下 C 的相对约简当且仅当:

- 1) $POS(SIM_R, \{d\}) = POS(SIM_C, \{d\})$;
- 2) $\forall R' \subset R, POS(SIM_{R'}, \{d\}) < POS(SIM_C, \{d\})$.

1.2 相似性阈值的影响

图1所示为二维空间两类样本 X_1 和 X_2 , 则 t_1 属于 X_1 的下近似, t_3 属于 X_2 的下近似, t_1 和 t_3 是 $X = \{X_1, X_2\}$ 的正区域, t_2 为边界样本。从图1可以看出如果增大相似性阈值 t_a , 则 t_1 也会成为边界, 因此正区域与 t_a 有很大关系。

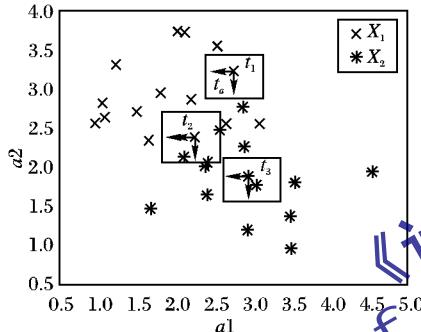


图1 相似关系粗糙集

正区域 $POS(SIM_A, \{d\})$ 关于相似性阈值 t_a 具有下列性质。

定理1 给定决策系统 $L = \langle U, C \cup \{d\} \rangle$, 属性子集 $A \in C$:

- (1) $\forall a \in A, 0 < t_a < \min_{\substack{t_i, t_j \in U \\ d(t_i) \neq d(t_j)}} diff(a, t_i, t_j) \Leftrightarrow POS(SIM_A, \{d\}) = \bigcup_i X_i$;
- (2) $\exists a \in A, t_a > \min_{\substack{t_i, t_j \in U \\ d(t_i) \neq d(t_j)}} diff(a, t_i, t_j) \Rightarrow POS(SIM_A, \{d\}) \subset \bigcup_i X_i$ 。

证明 (1) 成立 \Rightarrow (2) 成立, 因此仅证明(1)。

因为 $POS(SIM_A, \{d\}) \subseteq \bigcup_i X_i = U$, 只需证明 $\forall t_k \in U, \exists j, t_k \in \overline{SIM_A}(X_j)$ 。

设 $t_k \in X_j$, 对于 $\forall t_m \in SIM_A(t_k), \forall a \in A, diff(a, t_k, t_m) \leq t_a < \min_{\substack{t_i, t_j \in U \\ d(t_i) \neq d(t_j)}} diff(a, t_i, t_j)$ 。

所以 $d(t_m) = d(t_k) \Rightarrow t_m \in X_j, SIM_A(t_k) \subseteq X_j$, 即 $t_k \in SIM_A(X_j)$ 。证毕。

定理1说明哪些样本是属性子集 A 可清晰分辨的取决于人为设定的相似性阈值 $t_A = [t_{a1}, \dots, t_{a|A|}]$, t_A 是属性子集 A 可清晰分辨的不同类样本最小差异度, 随不同类样本允许差异度下限的增大, 属性子集 A 可清晰分辨的样本减少。另一方

面, 设定了相似性阈值 t_A , 在属性个数相同的条件下, $card(POS(SIM_A, \{d\}))$ 越大的属性集合 A 价值越高。

1.3 分辨信息表

为了衡量属性在同类、不同类对象上的相似、相异性, 这里引入分辨信息表 (Distinction Table, DT)。

定义5 设决策表 $T = \langle U, C \cup D, V, f \rangle$, 论域 $U = \{t_1, t_2, \dots, t_n\}$, 条件属性 $C = \{c_1, c_2, \dots, c_m\}$, 决策属性 $D = \{d\}$, 则分辨信息表 $DT = \langle U^*, C \cup D, V^*, f^* \rangle$, 其中 $U^* = \{(t_i, t_j)\}, i \neq j, i, j = 1, 2, \dots, n, (t_i, t_j)$ 简记为 t_{ij} , 属性值域 $V^* = [0, 1]$, 信息函数 $f^*: U^* \times R \rightarrow V^*, f^*$ 为差别函数 $diff$ 。

与二进制分辨矩阵^[12]相比, 分辨信息表增加了同类样本的属性差异信息, 这样, 可以从属性在同类样本上的相似性和不同类样本上的相异性两方面度量属性价值。

2 基于属性综合分辨能力的约简算法 RSDA

2.1 属性综合分辨力评价指标

根据好的属性集合在不同类样本上相异概率高而在同类样本上相似概率高的原则^[13], 利用定义1描述的差异性度量可以定义属性综合分辨力评价指标如下:

定义6 样本 $t_i, t_j \in U, i \neq j$, 设 a 为条件属性, a 的属性价值为:

$$\begin{aligned} Sig(a) = & \sum_{\substack{t_i, t_j \in U \\ t_i, t_j \text{ 不同类}}} diff(a, t_i, t_j) / \Delta(t_i, t_j) - \\ & \sum_{\substack{t_i, t_j \in U \\ t_i, t_j \text{ 同类}}} diff(a, t_i, t_j) / \Delta(t_i, t_j) \end{aligned}$$

其中 $\Delta(t_i, t_j)$ 为满足条件的样本对 (t_i, t_j) 的个数。

属性价值 Sig 的含义是: 若属性 a 区分开一对不同类样本, a 的贡献为差异度量 $diff(a, t_i, t_j)$, 差异越大贡献越大; 若 a 区分开一对同类样本, a 的惩罚为 $-diff(a, t_i, t_j)$, 差异越大惩罚越大。平均贡献与平均惩罚之和反映了属性区分样本的整体效应, 采用均值有利于避免个别包含噪声或错误的数据对属性评价的影响。由于属性不能清晰分辨相似样本, 计算属性价值时仅考虑不相似样本上的差异。

2.2 相似关系粗糙集的属性约简算法 S_RSDA

属性约简算法 S_RSDA 从空集出发, 选择剩余属性集合中 Sig 最大且至少能清晰分辨一对样本的属性加入约简集, 直到所有的样本对都至少能被约简集的一个属性清晰地分辨开。

S_RSDA 算法的步骤如下:

步骤1 根据原始数据集构造分辨信息表 DT , DT 中个体 t_{ij} 在属性 a 的值为:

$$a(t_{ij}) = \begin{cases} diff(a, t_i, t_j), & diff(a, t_i, t_j) \geq t_a \\ 0, & diff(a, t_i, t_j) < t_a \end{cases}$$

步骤2 $REDUCT = \emptyset$;

步骤3 对分辨信息表 DT 计算所有属性的综合价值 $Sig(a)$;

步骤4 选择属性价值最大的属性 $a^* = \operatorname{argmax}_k Sig(a_k)$;

If $\exists t_{ij} \in DT, d(t_i) \neq d(t_j) \& a^*(t_{ij}) \neq 0$,

将 a^* 加入 $REDUCT$,

去掉该属性所在的列和 $a^*(t_{ij}) \neq 0 \wedge d(t_i) \neq d(t_j)$ 的所有行;

Else

a^* 不能分辨余下的不同类样本对, 转步骤 6;

步骤 5 If 分辨信息表 DT 中没有不同类的样本对, 转步骤 6,

Else

转步骤 3;

步骤 6 Output REDUCT。

经典粗糙集 RSDA(C_RSDA) 算法步骤与 S_RSDA 相同, 不同之处在于分辨信息表 DT 的任一个体 t_i 在属性 a 上的值为:

$$a(t_{ij}) = \begin{cases} 1, & a(t_i) = a(t_j) \\ 0, & a(t_i) \neq a(t_j) \end{cases}$$

3 实验结果与分析

3.1 实验数据

3.1.1 车牌数字字符

实验样本是通过随机拍摄的车牌图像经字符自动分割算法获得, 样本字符均为经过位置和大小归一化处理的二值化图像。字符图像大小归一化为 64×64 像素, 然后分成 8×8 个网格, 统计每个网格中的黑像素个数作为该网格特征值。所有 64 个网格特征值组合起来构成粗网格特征, 作为字符的条件属性 $C = \{c_1, \dots, c_{64}\}$, 决策属性为字符类别 $d = \{0, 1, \dots, 9\}$ 。条件属性值离散化为 3 个区间。样本总数为 600, 分成 10 类, 每类 60 个样本, 随机分成训练样本和测试样本, 训练样本为 540 个, 测试样本为 60 个。



图 2 车牌数字字符样本

3.1.2 mfeat 手写体数字字符

UCI 数据库中的 mfeat 数据集提取自荷兰支票上的手写数字, 包括“0”至“9”10 个类别, 每类 200 个样本, 共 2000 个手写数字样本。样本字符是 15×16 像素二值图像, 见图 3。



图 3 根据 mfeat-pix 恢复的字符图像

本次实验选择 76 个 Fourier 系数和 64 个 Karhunen-Love 系数作为决策表条件属性构成数据集 mfeat_foukar, 数字类别作为决策属性。

3.2 C_RSDA 与其他等价关系粗糙集约简算法比较

实验先对数据集约简, 使用约简属性提取规则, 然后采用最近邻方法分类。由于等价关系粗糙集约简只能处理离散属性, 一些约简算法需计算正区域, 难以处理全部 mfeat 数据集, 本次实验从 mfeat 数据集随机选择了 900 个样本的训练集和 200 个样本的测试集, 将属性值等距离离散化为三个区间。实验在车牌字符集和 mfeat_foukar 字符集上测试了 C_RSDA 和常用等价关系约简算法的性能, 实验结果见表 1 和表 2。

C_RSDA 将属性同类相似性作为选择因素之一, 与仅以不同类对象分辨能力为标准选择属性的约简算法比较, 其约简集导出规则的覆盖率和泛化能力有显著的提高。在识别率方面, C_RSDA 搜索到的约简集识别率优于其他约简集, 说明同类相似性高的属性集合具有较高的分类能力。

表 1 C_RSDA 与其他约简算法性能比较: 约简率、识别率

约简算法	车牌字符		mfeat_foukar	
	属性个数	识别率	属性个数	识别率
Johnson*	9	0.94	* *	* *
Jelonek	10	0.88	16	0.570
MIBARK	9	0.96	16	0.620
J_C-filter ^[14]	10	0.96	38	0.570
C_RSDA	9	0.96	17	0.725
不约简	64	1.00	140	0.755

注: * 使用 ROSETTA V1.4.41 计算, ** 未能获得约简。

表 2 C_RSDA 与其他约简算法性能比较: 规则数

约简 算法	车牌字符	
	离散化区间为 2	离散化区间为 3
Johnson	82	219
Jelonek	96	235
MIBARK	81	214
J_C-filter	72	214
C_RSDA	82	154

3.3 C_RSDA 和 S_RSDA 约简性能比较

将数据集 mfeat_foukar 的属性值分别等距离离散化为 2~5 个区间, 然后应用 C_RSDA 和 S_RSDA 约简离散化属性, 并用 S_RSDA 约简未离散化属性。两种算法获得的约简属性对字符的识别效果记录于表 3。

表 3 C_RSDA 与 S_RSDA 约简集分类正确率及约简率

约简 算法	指标	离散化区间数				
		1	2	3	4	5
C_RSDA	识别率	—	0.825	0.725	0.805	0.83
C_RSDA	属性个数	—	16	17	12	9
S_RSDA	识别率	0.92*	0.825	0.735	0.785	0.785
S_RSDA	属性个数	50	16	18	12	10
不约简	识别率	0.93	0.875	0.755	0.895	0.895

注: * 阈值 $t_a = 1/3$; 离散区间数为 1 表示不离散化。

根据实验结果 C_RSDA 约简离散属性对象的性能优于 S_RSDA, 但 S_RSDA 不必对连续属性离散化, 是直接约简数值属性的方法, 不仅约简了大量冗余属性, 而且获得的约简具有接近全部属性的分类能力。

3.4 相似关系阈值对约简的影响

在 S_RSDA 约简算法中, 相似性阈值 t_a 是两个样本 t_i, t_j 属性值相对差异 $(|a(t_i) - a(t_j)| / V_a)$ 的控制参数。它控制了 S_RSDA 约简属性在不同类样本上的相对差异度下限。本次实验对每一个属性设定相同的阈值。图 4 显示在一定范围内增大 t_a , S_RSDA 约简属性个数增加, 近邻分类的正确率也随之提高。但是由于 RSDA 算法依赖不相似样本提供启发信息, 当 t_a 设置过大, 不相似样本会大大减少, 从而使算法选择的属性过少。

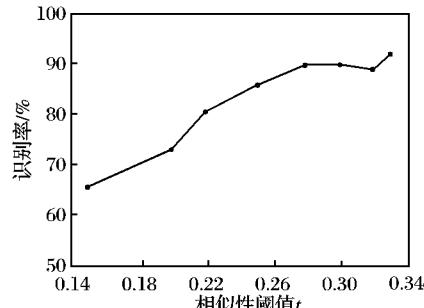


图 4 S_RSDA 约简集识别率与相似性阈值的关系曲线

(下转第 170 页)

0.025 和 0.46, 而 KF 的结果分别为 0.04 和 0.83, 非常明显地体现了 KPF 算法的跟踪精度要优于 PF 和 KF。对比跟踪结果的均方误差的方差, 可以得到同样的结果。

表 1 各算法均方根误差的均值及方差对比

算法	估计项	均方误差	
		均值	方差
KF	x 坐标	0.040 314	0.000 861 83
	y 坐标	0.828 100	0.224 710 00
PF	x 坐标	0.025 356	0.000 365 59
	y 坐标	0.460 670	0.046 141 00
KPF	x 坐标	0.022 465	0.000 242 75
	y 坐标	0.424 230	0.047 991 00

4 结语

本文提出一种新型的粒子滤波算法——KPF 算法。该算法弥补了基本粒子滤波算法没有融合观测值信息的缺点, 将每一个粒子在是用卡尔曼滤波进行更新, 融入新的观测值信息, 从而提高了粒子滤波算法的估计性能, 在 BOT 跟踪中表现出优于基本粒子滤波算法的性能。实验中针对 BOT 跟踪问题, 将 KPF 算法与 KF、PF 算法对比, 结果表明 KPF 算法性能明显优于其他两种算法。KPF 算法针对每一个粒子使用卡尔曼滤波器进行更新, 无疑增加了计算时间, 因此, 该算法的时间耗费比较重, 在今后的工作中将着重解决 KPF 算法的时间耗费问题。

参考文献:

- [1] BERGMAN N. Recursive Bayesian estimation: Navigation and tracking applications[D]. Sweden: Linkoping University, Department of Electrical Engineering, 1999.

(上接第 158 页)

4 结语

本文提出一种综合评价属性在样本上差异性和相似性的量化指标, 并应用于相似粗糙集属性约简, 可选择同类相似性较高的属性集合, 克服粗糙集理论对噪声、干扰敏感的缺点。RSDA 应用于字符数据集时, 在约简性能如规则数、识别率方面, 优于常用启发式约简算法 Johnson、Jelonek、MIBARK 以及 J_C _filter 算法。这说明粗糙集属性约简方法与特征选择的一些启发信息相结合, 不仅可以去除冗余而且能选择性能优良的特征集合, 从而降低了分类器的复杂度, 改善了分类性能。

参考文献:

- [1] PAWLAK Z. Rough sets[J]. Communications of the ACM, 1995, 38(11): 89–95.
[2] 苗夺谦, 胡桂荣. 知识约简的一种启发式算法[J]. 计算机研究与发展, 1999, 36(6): 681–684.
[3] GUYON I, ELISSEFF A. An introduction to variable and feature selection[EB/OL]. [2009-04-20]. <http://jmlr.csail.mit.edu/papers/volume3/guyon03a/guyon03a.pdf>.
[4] KIM D, BANG S Y. A handwritten numeral character classification using tolerant rough set[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(9): 923–937.
[5] ZADEH L A. Fuzzy logic = Computing with words [J]. IEEE Transactions on Fuzzy Systems, 1996, 4(2): 103–111.

- [2] 邓小龙, 谢剑英, 王林. 基于粒子滤波的仅有角测量的被动跟踪[J]. 上海交通大学学报, 2005, 39(5): 993–996.
[3] WELCH G, BISHOP G. An introduction to the Kalman filter[EB/OL]. [2009-04-25]. http://www.cs.unc.edu/~tracker/media/pdf/SIGGRAPH2001_CoursePack_08.pdf.
[4] ARULAMPALAM M S, MASKELL S, GORDON N, et al. A tutorial on particle filters for on-line nonlinear/non-Gaussian Bayesian tracking[J]. IEEE Transactions on Signal Processing, 2002, 50(2): 174–188.
[5] DOUCET A, FREITAS J F G, GORDON N J. Sequential Monte Carlo methods in practice[M]. New York: Springer-Verlag, 2001.
[6] GUSTAFSSON F, GUNNARSSON F, BERGMAN N, et al. Particle filters for positioning, navigation and tracking[J]. IEEE Transactions on Signal Processing, 2002, 50(2): 425–437.
[7] GORDON N J, SALMOND D J, SMITH A F M. Novel approach to nonlinear/non-Gaussian Bayesian state estimation[J]. IEE Proceedings on Radar and Signal Processing, 1993, 140(2): 107–113.
[8] CARPENTER J, CLIFFORD P, FEARNHEAD P. Improved particle filter for non-linear problems[J]. IEE Proceedings on Radar Sonar and Navigation, 1999, 146(1): 2–7.
[9] DOUCET A, GODSILL S, ANDRIEU C. On sequential Monte Carlo sampling methods for Bayesian filtering[J]. Statistics and Computing, 2000, 10(3): 197–208.
[10] CAPPE O, GODSILL S J, MOULINES E. An overview of existing methods and recent advances in sequential Monte Carlo[EB/OL]. [2009-04-25]. <http://perso.telecom-paristech.fr/~cappe/papers/06particle-cmg.pdf>.
[11] 王法胜, 赵清杰. 一种用于解决非线性滤波问题的新型粒子滤波算法[J]. 计算机学报, 2008, 31(2): 346–352.
[12] RADZIKOWSKA A M, KERRE E E. A comparative study of fuzzy rough sets[J]. Fuzzy Sets and Systems, 2002, 126(2): 137–155.
[13] STEPANIUK J. Similarity based rough sets and learning[C]// Proceedings of the Fourth International Workshop on Rough Sets, Fuzzy Sets, and Machine Discovery. Tokyo: [s. n.], 1996: 18–22.
[14] SLOWINSKI R, VANDERPOOTEN D. A generalized definition of rough approximations based on similarity[J]. IEEE Transactions on Knowledge and Data Engineering, 2000, 12(2): 331–336.
[15] LIN T, GRANULAR Y. Computing on binary relations I: Data mining and neighborhood systems[C]// Proceedings of the Rough Sets in Knowledge Discovery. Heidelberg: Physica-Verlag, 1998: 107–121.
[16] HU Q H, YU D R, XIE Z X. Neighborhood classifiers[J]. Expert Systems with Applications, 2008, 34(2): 866–876.
[17] 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.
[18] WANG G Y, WU Y, FISHER P S. Rule generation based on rough set theory[C]// Proceedings of SPIE — The International Society for Optical Engineering. New York: SPIE Press, 2000: 181–189.
[19] DASH M, LIU H. Feature selection for classification [EB/OL]. [2009-05-11]. <http://xin.cz3.nus.edu.sg/group/personal/ex/modules/DM/FeatureSelection.ppt>.
[20] 吴敏, 张崇巍. 结合一致性准则的粗糙集属性约简算法[J]. 合肥工业大学学报: 自然科学版, 2006, 29(7): 851–855.