应用有效信息比率的离散化算法

诸文智，王靖程，张彦斌，贾立新
（西安交通大学电气工程学院，710049，西安）

摘要：针对目前离散化信息量度无法准确表征数据离散后有效分类信息量的问题，提出了一种基于有效信息比率的离散化算法.在构建离散化方案相依表的基础上，分析了离散区间内类属性分布与分类信息蕴含量间的关系，并根据类属性分布信息引入有效信息比率，用于表征各离散区间内有效分类信息量.然后，依据离散化方案的离散区间数及其有效信息比率，设计出表征离散化方案划分质量的离散化评价指标，从而提高了数据的离散化效果.仿真实验和实际应用的结果表明，该算法离散化后在有效分类信息量和分类预测精度上高于主流基于信息论的离散化算法.

关键词：离散化；信息；有效信息比率
中图分类号：TP274 **文献标志码**：A **文章编号**：0253-987X(2011)04-0012-06

Discretization Algorithm Based on Effective Information Ratio

ZHU Wenzhi, WANG Jingcheng, ZHANG Yanbin, JIA Lixin
(School of Electrical Engineering, Xi′an Jiaotong University, Xi′an 710049, China)

Abstract: Since the current information measures of discretization can not accurately reflect the degree of the effective class information in the discretized dataset, a discretization algorithm based on effective information ratio is presented, and a contingency table of corresponding discretization scheme is constructed. According to the analyses of the relationship between the class distribution and the remaining class information, the effective information ratio based on the class distribution is analyzed to indicate the degree of effective information in each discretized interval. An improved discretization criterion is generated to evaluate the quality of the discretization scheme following the number of discrete internal and the effective information ratio. The simulation and applications illustrate the more effective class information and the higher classification accuracy than the other information-based solutions.

Keywords: discretization; information; effective information ratio

离散化即将连续属性的值域划分为有限数量的区间，并赋予每个区间相应的离散数字标号.离散化方法作为知识发现技术的一个重要环节，通常可改善挖掘出的知识的可理解性[1]，离散化过程可能造成某些相关信息丢失，导致所发现的知识精确度降低.因此，离散化方法应在保持分类预测精度的基础上，兼顾离散化后信息完整性和数据简单性.离散化算法以是否考虑数据中类属性信息分为两类：无监督类型和有监督类型.早期简单的离散化方法，如等宽度法和等频率法[2]等属于无监督类型.此类算法忽略了类属性与连续属性之间的关系，以用户自定义的区间宽度或区间内实例数量，生硬地对连续属性进行划分，虽然算法执行速度较快，但离散化效果不佳，所以需要研究有监督离散化方法.

有监督离散化算法根据离散化评价指标的不同，又可分为基于统计的算法（如 Chi2[3] 和 Khi-

收稿日期：2010-09-11. 作者简介：诸文智(1982—)，男，博士生；张彦斌(联系人)，男，教授，博士生导师. 基金项目：国家自然科学基金资助项目(61005058)；教育部高等学校博士学科点专项科研基金资助项目(20090201110019，20070698059).
网络出版时间：2011-01-17 网络出版地址：http://www.cnki.net/kcms/detail/61.1069.T.20110117.1046.014.html

ops[4]等）和基于信息论的算法（如最大熵[5]和CADD[6]等）. 近年来，基于信息论的算法发展很快，主要有：基于最小描述长度算法（MDLP）[7]应用信息熵量度，并提出最小描述长度准则作为离散化评价指标；类-属性相依性最大算法（CAIM）[8]和类-属性相依系数算法（CACC）[9]在信息熵量度的基础上，提出了各自的类-属性相依性量度作为离散化评价指标. 其中，CAIM 算法偏重于获得最小划分区间数，而 CACC 算法偏重于获得最大类-属性相依性. 但是，上述算法中离散化评价指标均无法表征数据离散后的有效分类信息量，造成所获得的离散化方案均非最优.

本文提出一种基于有效信息比率的离散化算法，在构建连续属性对应离散化相依表的基础上，从信息论的角度综合分析了离散化区间中类属性的概率分布与信息蕴含量的关系，进而提出基于有效信息比率的离散化评价指标. 本文算法以保持数据分类能力为前提，得到了分类信息更加完整且离散划分更加合理的离散化方案.

1　离散化问题描述

知识表示是数据处理的首要问题. 为了便于数据的信息化处理，本文使用列联表对数据实例进行划分，其 4 元组形式化定义如下

$$T = (U, F \cup C, V, f)$$

式中：$U=\{x_1, x_2, \cdots, x_M\}$为数据对象的非空有限集合，称为论域；$F=\{a \mid a \in F\}$称为条件属性集，每个 $a_i \in F(1 \leqslant i \leqslant m)$称为 F 的一个条件属性；$C=\{d \mid d \in C\}$称为类属性，并有 $F \cap C=\varnothing$，$F \neq \varnothing$，$C \neq \varnothing$；$V=\bigcup V_a$（$\forall a \in F \cup C$）是信息函数 f 的值域；$f=\{f_a: U \to V_a, \forall a \in F \cup C\}$表示列联表的信息函数，$f_a$ 为属性 a 的信息函数.

根据上述定义，数据集可以表示为论域元素数量为 M 的列联表，其中任一对象实例只属于 s 种类别（类属性值）中的某一类. 设 $a \in F$ 为数据集中任一连续属性，其上存在一离散化方案 D，将连续属性的值域划分为 n 个互不相交的区间

$$D: \{[b_0, b_1], (b_1, b_2], \cdots, (b_{n-1}, b_n]\}$$

其中，属性 a 的值域 $V_a=[b_0, b_n]$. 方案 D 中的值以升序方式排列，并组成对应的断点集 $C_{P_D}=\{b_0, b_1, \cdots, b_n\}$. 由于断点集和离散化方案是一一对应的，可以使用两者中任意一种表示相应连续属性的离散化结果. 根据上述定义建立属性 a 的离散化方案 D 对应的相依表，如表 1 所示.

表 1 中元素 q_{fi} 表示在当前离散化方案 D 下，数据集中类属性值为 c_f，且为属性 a 的值落在离散化区间 i 内的实例数量. 表中第 f 行元素的总和为 q_{f+}，它表示数据集中类属性值为 c_f 的实例总数. 相应地，表中第 i 列元素的总和为 q_{+i}，它表示数据集中为属性 a 的值落在离散化区间 i 内的实例总数. 根据属性 a 的值域和离散化方案 D，将实例划分至 n 个离散化区间，并在每一区间内形成一定类属性分布. 因此，离散化方案 D 的变化将改变每一离散化区间内类属性的分布信息.

离散化过程的实质是通过选取断点形成离散化方案，完成条件属性集的空间划分. 其中如何设计好的离散化评价指标，用于指导断点选择，改善离散化效果，已成为离散化研究中的关键问题之一.

2　基于有效信息比率的离散化算法

传统基于信息论的离散化算法直接应用信息熵作为离散化的评价指标，这类算法有一定的不足：在已有分类信息的基础上，信息熵指标有偏向较细划分的倾向，为使离散化后每一区间上的分类信息尽

表 1　属性 a 的离散化方案 D 对应的相依表

类别	区间					类别实例总数
	$[b_0, b_1]$	…	$(b_{i-1}, b_i]$	…	$(b_{n-1}, b_n]$	
c_1	q_{11}	…	q_{1i}	…	q_{1n}	q_{1+}
⋮	⋮		⋮		⋮	⋮
c_f	q_{f1}	…	q_{fi}	…	q_{fn}	q_{f+}
⋮	⋮		⋮		⋮	⋮
c_s	q_{s1}	…	q_{si}	…	q_{sn}	q_{s+}
区间内实例总数	q_{+1}	…	q_{+i}	…	q_{+n}	M

可能确定,算法将对离散化区间不断细化,区间数目不断增多.虽然分类信息蕴含量越来越大,但离散化结果却更加复杂,不利于数据简化及后续的知识提取过程.经此种离散化处理后的数据所生成的分类模型,增强了训练数据的拟合能力,但对测试数据的预测能力减弱了.因此,本文在信息熵的基础上,定义了离散化区间上的有效信息比率,并基于有效信息比率,提出一种新的离散化评价指标,其离散化目标如下:

(1)保持并改善数据分类能力;

(2)合理化缩减离散化区间数量,平衡离散化后信息完整性和数据简单性;

(3)离散化评价函数运算量小,易于实现.

2.1 有效信息比率定义

对表1所示的离散化方案 D,定义离散化区间 i 内类属性分布的条件概率向量为 $\boldsymbol{P}^i=(P_1^i,\cdots,P_f^i,\cdots,P_s^i)$,其中 $P_f^i=\frac{q_{fi}}{q_{+i}}$,且 $\sum_{f=1}^{s}P_f^i=1$,则区间 i 上的条件熵表示为

$$H_{Di}(d\mid a)=-\sum_{f=1}^{s}P_f^i\,\mathrm{lb}P_f^i \tag{1}$$

$H_{Di}(d|a)$ 表征了区间 i 内类属性分布的有序度,值越小表示区间内类属性分布有序度越高.当类属性分布的条件概率相等时,即 $P_1^i=\cdots=P_f^i=\cdots=P_s^i=1/s$,其条件熵达到最大值 $\mathrm{lb}(s)$,区间 i 内类属性分布有序度降至最低,其分类信息蕴含量也最小.我们定义此时的类属性分布为信息量基准分布.相反,当区间 i 内实例都属于同一类时,即有 $\forall P_f^i=1$ 且 $\sum_{f=1}^{s}P_f^i=1$,则其条件熵达到最小值0,区间 i 内类属性分布有序度达到最高,其分类信息蕴含量也最大.

基于上述分析,对属性 a 上的离散化方案 D,本文定义区间 i 内类属性分布与信息量基准分布对应条件熵之间偏离的程度为区间 i 上的有效信息比率,并将其作为表征数据离散化后有效分类信息量的度量,如下所示

$$R_D(p^i)=1-H_{Di}(d\mid a)/\mathrm{lb}(s) \tag{2}$$

2.2 离散化评价指标

在已有分类信息的基础上,本文所提离散化评价指标度量了连续属性 a 上离散化方案 D 的划分质量,其定义如下

$$J_D(d,a)=\sum_{i=1}^{n}\frac{m_{\mathrm{ax},i}}{q_{+i}}R_D(p^i)/\mathrm{lb}(n) \tag{3}$$

式中:n 表示离散区间数;$m_{\mathrm{ax},i}$ 表示相依表中区间 i 对应列的最大值;$R_D(p^i)$ 表示区间 i 上的有效分类信息量.

对所提离散划评价指标分析如下.

(1)J_D 越大,连续属性离散化分后有效分类信息量越大,离散化方案 D 的划分质量越好.

(2)定义相依表第 i 列中最大值对应类别为区间 i 上的主导类,其条件概率 $m_{\mathrm{ax},i}/q_{+i}$ 使离散化算法倾向于最大化各离散区间中的主导类,而有效信息比率 $R_D(p^i)$ 则将各区间内类属性分布信息引入算法中.

(3)将 $m_{\mathrm{ax},i}/q_{+i}$ 作为有效信息比率 $R_D(p^i)$ 的加权值,即有权重地看待类属性分布与划分质量之间的关系,这样评价函数将更加全面地反映离散化方案的划分质量.

(4)评价指标必须将离散区间数 n 限定在一个合理的范围内,如划分过细,将导致过拟合问题,如划分过粗,将造成离散结果不合理,因此本文应用 $\mathrm{lb}(n)$ 代替 n 作为方程算子.

(5)J 的值域为 $[0,n/\mathrm{lb}(n)]$.对离散化方案 D 而言,当所有划分区间上的类属性分布为信息量基准分布时,J 取得最小值0;相反,当所有划分区间内实例都属于某一类时,J 取得最大值 $n/\mathrm{lb}(n)$.

2.3 算法描述

输入:训练数据集 $T_D=\{\bigcup_{i=1}^{m}A_i,d\}$,其中连续属性数量为 m,实例总数为 M,类属性值数量为 s.对每一连续属性 A_i,算法步骤如下:

(1)对样本实例在属性 A_i 上的取值 V_{A_i} 进行升序排列;

(2)获得样本实例在属性 A_i 上的最小值 d_0 和最大值 d_n;

(3)计算排序后 V_{A_i} 集中两两相邻元素的中点,组成断点候选集 C_{P_c};

(4)初始化断点集 C_P 为 $[d_0,d_n]$,全局最大值 J 为0;

(5)设划分区间数 $n=1$;

(6)将断点候选集 C_{P_c} 中不属于断点集 C_P 的断点值尝试性地添加到 C_P 中,并根据式(3)计算对应离散化方案的 J 值;

(7)选择产生最大 J 的断点,添加到 C_P 中;

(8)判断当前离散化方案是否满足 J 大于全局最大值或者 $n<s$,满足则更新离散化方案 D 与当前断点集 C_P 一致,并设 $n=n+1$,转步骤(6),不满足则算法结束.

输出：在连续属性 A_i 上离散区间数为 n 的离散化方案 D.

3　实验结果及分析

3.1　仿真实验

为了验证所提出的基于有效信息比率的离散化算法性能，选择分类决策问题中公认的性能评估方法 C4.5[10] 进行分类实验，比较了本文算法(EIRDC)、MDLP、CAIM 和 CACC 等 4 种方法在 8 个机器学习数据库[11]上的仿真实验结果. 4 种算法以各自离散化评价指标为准，无需另外进行参数设定. 对训练集应用 4 种离散化算法，获得各算法对应的离散化方案、区间数和训练集. 应用文献[8]中的相依性量度，对比不同算法获得的离散化方案表达式如下

$$c_{\text{air}}=\sum_{i=1}^{n}\sum_{f=1}^{s}p_{fi}\text{lb}\frac{p_{fi}}{p_{f+}p_{+i}}\Big/\sum_{i=1}^{n}\sum_{f=1}^{s}p_{fi}\text{lb}\frac{1}{p_{fi}}\quad(4)$$

利用所得离散化方案对验证集进行离散，获得各算法对应离散后的验证集. 最终，以离散化后的训练集进行决策树模型训练，并使用离散化后的验证集进行模型检验，得到各离散化方法的分类预测精度. 使用 C4.5 算法所得决策树分类模型的预测精度和规则数来评估算法分类效果，使用离散化方案对应的 c_{air} 值和离散区间数来评估离散化效果. 实验采用十折交叉检验法，以数据库 90%的数据作为训练集，其余 10%的数据作为验证集. 表 2 给出了各数据库的基本情况，对于 thy 数据库上的离散属性，本文不做处理直接应用. 表 3 给出了仿真实验结果的相关数据，所有实验结果均为 10 次实验的平均值.

表 2　数据库相关信息

编号	数据库	实例数量	属性数量	连续属性数量	类别数量
1	vehicle	846	18	18	4
2	iris	150	4	4	3
3	pima	768	8	8	2
4	ionos	351	33	32	2
5	thy	7 200	21	6	3
6	page-blocks	5 473	10	10	5
7	pendigit	10 992	16	16	10
8	wave	5 000	21	21	3

为了便于比较实验结果，本文将相关算法在同一评估标准上进行位置排序，并统计各算法在 8 个数据库上的平均排位，如表 3 中列平均排位所示. 算法在预测精度与 c_{air} 均值上对应的值越大，算法的排位值越大；相反，算法在 C4.5 规则数和离散区间均数上对应的值越大，算法的排位值越大.

表 3　仿真实验结果

数据库		vehicle	iris	pima	ionos	thy	page-blocks	pendigit	wave	平均排位
预测精度/%	EIRDC	70.86*	93.87*	77.34*	94.47*	99.42	97.17*	89.43*	77.60*	1.13*
	MDLP	70.68	93.87*	76.77	89.49	99.48*	96.73	88.73	75.31	2.63
	CAIM	70.81	93.87*	74.49	90.66	98.66	96.60	88.93	76.75	2.88
	CACC	69.65	93.87*	76.58	91.77	99.08	96.36	89.12	77.56	2.63
C4.5 规则数	EIRDC	130.9	3.5*	8.0*	9.8*	17.0	94.0	1 746.2*	347.2*	1.63*
	MDLP	143.6	3.5*	12.1	20.4	48.7	117.7	2 023.6	452.5	3.38
	CAIM	119.8*	3.5*	8.4	11.2	18.3	90.9	1 909.1	352.7	2.13
	CACC	126.9	3.5*	13.5	11.2	13.3*	73.2*	1 887.0	348.2	1.88
c_{air} 均值	EIRDC	0.095*	0.555*	0.029*	0.135*	0.038	0.148*	0.137*	0.074*	1.25*
	MDLP	0.078	0.480	0.027	0.131	0.033	0.087	0.128	0.063	3.25
	CAIM	0.088	0.542	0.025	0.087	0.041	0.140	0.137*	0.071	2.625
	CACC	0.094	0.550	0.028	0.131	0.048*	0.145	0.137*	0.072	1.625
离散区间均数	EIRDC	4.03	3.0*	2.0*	2.6	3.0*	5.0*	10.0*	4.2	1.38
	MDLP	4.1	3.0*	2.2	4.4	4.8	7.1	10.3	5.1	2.75
	CAIM	4.0*	3.0*	2.0*	2.0*	3.0*	5.0*	10.0*	3.0*	1*
	CACC	4.09	3.0*	2.3	3.4	3.2	5.0*	10.0*	4.3	2.13

注：上角加“*”的数据为各数据库中相对于实验评估标准的最佳值.

由表 3 可以看出，EIRDC 算法在 iris 和 thy 等 6 个数据库上获得了最高预测精度，在 iris 上获得与其他 3 种算法相同的预测精度，在 thy 上也获得了接近最高的预测精度. 从预测精度的平均排位来看，EIRDC 算法最佳，CACC 与 MDLP 算法稍次之，而 CAIM 算法排位靠后. 这说明 EIRDC 算法对离散区间的划分较合理，有利于保持并改善离散化后数据的分类能力. 就决策树分类模型的规则数而言，由于 4 种算法根据各自的评价指标对连续属性进行划分，离散区间数不定，则在多维数据库中容易造成分类决策树规则的骤增（如 vehicle 和 pendigit 等）. 即使如此，相比其他 3 种算法，EIRDC 算法在 C4.5 规则数上的平均排位仍然最佳. 总的来说，本文算法在提高分类精度的基础上，保持了相对较少的决策树规则数.

由于所比较的的方法均基于信息论，因此本文通过计算各数据库离散化后对应的 c_{air} 均值与离散区间均数来评价各算法生成的离散化方案. 其中 c_{air} 是离散后数据分类信息的度量值，其值越大，数据离散后分类信息损失量越小. 从实验结果上看：MDLP 算法在 c_{air} 和离散区间均数上平均排位均在最后. 相比 CAIM 与 CACC 算法，EIRDC 算法在 iris、page-blocks 和 pendigit 上的离散化方案以相同的离散区间数获得了最大的 c_{air}；EIRDC 算法在 vehicle、pima、ionos 和 wave 上的离散化方案，其离散区间数小于 CACC 算法，接近于 CAIM 算法，且获得了最大 c_{air}；由于 thy 的不平衡程度较重，EIRDC 算法的应用效果不佳，但其实验结果与其他数据库相比相差不大，仍可接受. 不难看出，实验结果符合上节分析，本文所构造的离散化评价方案较好地平衡了离散后分类信息量和离散区间数对离散结果的影响，在保证离散后数据分类能力的基础上，取得了较好的离散效果.

4 种算法在 8 个数据库上的平均离散时间如图 1 所示. 从图中可以看出：数据集维数较少时，4 种算法的平均离散时间差别不大；在多维数据库 vehicle、ionos、thy 和 wave 中，MDLP 算法平均离散时间略低于 CAIM、CACC 和 EIRDC 算法. 这说明本文算法在明显提高离散化效果的同时，其算法运行时间与其他算法相当.

3.2 磨矿浓度预测结果

磨矿浓度的准确检测是选矿厂球磨机自动控制得以有效实施的关键，在实际工业中，由于现场环境恶劣、影响因素众多等原因一直未能得以解决[12].

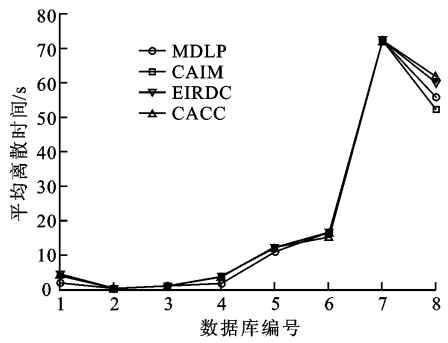

图 1 4 种算法在 8 个数据库上的平均离散时间

经长期研究，本文在选矿厂磨矿分级系统工业数据库中选择给矿量、返砂水量、排矿水量、电耳电流、磨机电流及分级机电流等作为条件属性，以磨矿浓度作为类属性，基于粗糙集方法和模糊系统建立基于推理规则的磨矿浓度预测模型，如下所示

$$y_{pd} = f(M, W_{rs}, W_{ca}, I_{ee}, I_m, I_c) \tag{5}$$

式中：y_{pd} 为磨矿浓度；M、W_{rs}、W_{ca}、I_{ee}、I_m 和 I_c 分别表示给矿量、返砂水量、排矿水量、电耳电流、磨机电流和分级机电流.

选择某选矿厂磨矿分级系统一段磨机进行实验，以磨矿浓度处于高浓度（82±0.65）%、正常浓度（80±0.60）%及低浓度（79±0.47）%等稳定工况下的 2 100 组数据构成现场实验数据库. 首先应用上述 3 种性能相近的 CAIM、CACC、EIRDC 算法对训练集进行离散处理，然后采用贪心策略的决策表简化过程[13]，建立各算法对应的磨矿浓度预测推理规则库，并运用三角模糊器、乘积推理机和中心平均解模糊器构建模糊系统，对验证集进行预测. 实验同样采用十折交叉检验法，结果如表 4 所示，所有结果均为 10 次实验的平均值.

表 4 磨矿浓度预测实验结果

算法	分类正确率/%	磨矿浓度预测值/%		
		高浓度工况	正常浓度工况	低浓度工况
CAIM	92.85	81.77±0.76	80.40±0.74	78.80±0.77
CACC	93.95	81.84±0.68	80.14±0.66	79.21±0.70
EIRDC	94.16	81.91±0.53	79.92±0.41	79.15±0.52

由表 4 易知，使用 EIRDC 算法获得了最高的分类正确率，并使模糊系统在 3 种不同工况下，获得了较另两种算法精度更高的磨矿浓度预测值.

综合上述实验结果，本文所提离散化算法的平均算法性能优于 CAIM 算法、CACC 算法和 MDLP 算法.

4 结 论

本文提出了一种基于有效信息比率的离散化算法. 该方法充分考虑离散化区间内类属性的分布信息以及离散区间数对离散化结果的影响，在合理化缩减离散化区间的基础上，选择具有最大有效分类信息比率的离散化方案，提高了离散化后数据库的分类预测精度. 仿真实验和实际应用的结果表明，EIRDC 平均算法性能优于目前主流基于信息论的离散化方法. 但是，在高维数据库中的离散化时间有所增加，后续工作可以介入粗糙集属性约简来改进 EIRDC 算法.

参考文献：

[1] 杨炳儒. 基于内在认知机理的知识发现理论[M]. 北京：国防工业出版社，2009：42-43.

[2] CHIU D K Y, WONG A K C, CHEUNG B. Information discovery through hierarchical maximum entropy discretization and synthesis[C]//Knowledge Discovery in Databases. Cambridge, Massachusetts, USA: MIT Press, 1991:125-140.

[3] QU Wenyu, YAN Deqian, SANG Yu. A novel chi2 algorithm for discretization of continuous attributes[C]//Progress in WWW Research and Development 10th Asia-Pacific Web Conference. Heidelberg, Germany: Springer-Verlag, 2008:560-571.

[4] BOULLE M. Khiops: a statistical discretization method of continuous attributes[J]. Machine Learning, 2004, 55(1): 53-59.

[5] KUMAR A, ZHANG D. Hand-geometry recognition using entropy-based discretization[J]. IEEE Transactions on Information Forensics and Security, 2007, 2(2): 181-187.

[6] SIMOVICI D A, JAROSZEWICZ S. An axiomatization of partition entropy[J]. IEEE Transactions on Information Theory, 2002, 48(7): 2138-2142.

[7] LE D D, SATOH S. Ent-boost: boosting using entropy measures for robust object detection[J]. Pattern Recognition Letters, 2007, 28(9): 1083-1090.

[8] KURGAN L A, CIOS K J. CAIM discretization algorithm[J]. IEEE Transactions on Knowledge and Data Engineering, 2004, 16(2): 145-153.

[9] TSAI C J, LEE C I, YANG W P. A discretization algorithm based on class-attribute contingency coefficient [J]. Information Sciences, 2008, 178(3): 714-731.

[10] QUINLAN J R. C4.5: programs for machine learning [M]. San Mateo, CA, USA: Morgan Kaufmann Publishers, 1993:17-25.

[11] FRANK A, ASUNCION A. UCI machine learning repository [EB/OL]. [2010-08-22]. http://archive.ics.uci.edu/ml.

[12] TIE Ming, YUE Heng, CHAI Tianyou. A hybrid intelligent soft-sensor model for dynamic particle size estimation in grinding circuits [J]. Lecture Notes in Computer Science, 2005, 3498: 871-876.

[13] POLKOWSKI L, TSUMOTO S, LIN T Y. Rough set methods and applications: new developments in knowledge discovery in information systems [M]. Heidelberg. Germany: Physica-Verlag GmbH, 2000: 49-88.

[本刊相关文献链接]

人眼视觉特性与粗糙集结合的 X 射线图像增强算法. 西安交通大学学报，2009，43(06)：48-51.

粗糙元特征对风洞中大气表面层模拟的影响. 西安交通大学学报，2009，43(03)：87-91.

一种基于分辨函数的属性约简算法及其应用. 西安交通大学学报，2008，42(12)：1455-1458.

一种 Roberts 自适应边缘检测方法. 西安交通大学学报，2008，42(10)：1240-1244.

微通道内流的微尺度粒子图像测速技术实验研究. 西安交通大学学报，2007，41(11)：1355-1359.

粗糙集属性约简判别分析方法及其应用. 西安交通大学学报，2007，41(08)：939-943.

粗糙集高效遗传约简算法. 西安交通大学学报，2007，41(04)：444-447.

采用 Galerkin 离散方法的 T-小波边界元法. 西安交通大学学报，2010，44(12)：99-104.

采用离散粒子群算法的网格任务安全级调度. 西安交通大学学报，2010，44(06)：21-26.

DT-PTV 离散相伪矢量的概率择优式剔除算法. 西安交通大学学报，2009，43(05)：119-123.

（编辑　杜秀杰）