一种基于信息论的决策表连续属性离散化算法

岳海亮 闫德勤
(辽宁师范大学计算机与信息技术学院 大连 116081)

摘 要 连续属性离散化方法对后续阶段的机器学习和数据挖掘过程有着重要的意义。提出一种新的针对决策表的离散化算法,在该算法中,首先将信息熵用作判断标准,从候选断点集中选择合适的断点,然后删除一些冗余的断点来优化离散结果,在删除过程中为了尽可能保证决策表分类能力不变,使用不一致率对该过程进行控制。最后选取多组实验数据,使用当前流行的分类算法——支持向量机(SVM)对离散化后的数据进行分类预测,并与其它离散算法进行对比,结果表明本算法是有效的。

关键词 连续属性离散化,决策表,信息熵,不一致率
中图法分类号 TP18 **文献标识码** A

New Algorithm for Discretization Based on Information Entropy

YUE Hai-liang YAN De-Qin
(Department of Computer and Information Technology, Liaoning Normal University, Dalian 116081, China)

Abstract The discretization of continues attributes is always with great contribution to the followed process of machine learning or data mining. A new algorithm based on information entropy for discretization of decision table was proposed. Through inconsistency checking of decision table, we deleted some redundant cut points on the basis of preliminary discretization scheme. The experiments of classification of discreted data were performed by using SVM, and meanwhile compared with other algorithms, the presented algorithm is effective.

Keywords Discretization, Decision table, Information entropy, Inconsistency

1 引言

数据离散化技术在传统的机器学习中被当作边缘性课题而没有受到足够的重视,随着近年来数据挖掘的不断更新发展,数据离散化在数据挖掘技术中渐渐显现出其不可替代的重要性,在规则提取、特征分类等很多算法中,尤其是在应用粗集理论进行数据挖掘的研究中,连续属性数据必须进行离散化,因此越来越多的专家和学者开始关注数据离散化。

目前,对连续属性离散化的研究已经取得大量的成果,研究人员从不同领域提出了多种离散化算法,根据分类标准的不同,可以对离散化算法进行如下分类:有监督和无监督、静态和动态、全局和局部、自底向上和自顶向下,详细的分类介绍参看文献[1,2]。

本文结合信息论[3]中熵的概念,提出了一种新的针对决策表的有监督离散化算法,该方法对于解决大数据量问题有较高的计算效率,并且运用决策表的不一致率来控制离散化过程,保证了决策表的相容性不发生改变,有效地减少了离散化过程所造成的信息损失。

2 基本概念

2.1 决策表及相关定义

决策表[4]是数据挖掘中用于知识表达的一种有效工具,它表示当满足某些条件时,决策应当如何进行。它已经在诸如粗糙集等数据挖掘技术中被广泛地应用,本文的离散算法就是基于决策表提出的。下面给出决策表和决策表不一致率的定义。

定义1 一个决策表是由一四元组 $DT=(U,A,V,f)$ 构成的知识表达系统,其中 U 是论域,$A=C\cup D$ 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且 $C\cap D=\Phi$,V 是属性的取值范围构成的集合,f 是 $U\times A\rightarrow V$ 的映射,表示 U 中每个对象在各个属性上的取值。

为了简化算法的设计,在本文讨论中假设连续变量仅出现在条件属性中,决策属性值为离散的,该假设不失一般性。同时,对于多决策属性的决策表来说,通常的做法是把它转化为单决策属性进行处理[5],此时决策表的形式一般如表1所列。

表1 决策表的一般形式

对象	条件属性			决策属性
U	c_1	…	c_m	d
x_1	$u_{1,1}$	…	$u_{m,1}$	v_1
x_2	$u_{1,2}$	…	$u_{m,2}$	v_2
⋮	⋮	…	⋮	⋮
x_n	$u_{1,n}$	…	$u_{m,n}$	v_n

到稿日期:2009-05-14 返修日期:2009-08-19 本文受国家自然科学基金(60372071),中国科学院自动化研究所复杂系统与智能科学重点实验室开放课题基金(20070101)资助。

岳海亮(1984—),男,硕士生,主要研究方向为数据挖掘,E-mail:yhail26@163.com;闫德勤(1962—),男,博士,教授,主要研究方向为模式识别、数据挖掘和信息安全等。

其中，$U=\{x_1,x_2,\cdots x_n\}$，$C=\{c_1,c_2,\cdots,c_m\}$，$D=\{d\}$，$f(x_i,c_j)=u_{j,i}$，$f(x_i,d)=v_i$。

定义 2 设 R_i 是按照决策表 DT 中条件属性 c_i 的属性值相等确定的等价关系，在此等价关系下的等价类子集簇可表示为 $\{X_1,X_2,\cdots,X_s\}$，且 $X_1\cup X_2\cup\cdots\cup X_s=U$，对于某一个子集 X 而言，其实例个数用 $|X|$ 表示，其中决策属性为 j $(j=1,2,\cdots,r(d))$ 的实例个数为 k_j，则定义该集合的不一致率为：

$$\xi^X=\frac{|X|-|M_X|}{|X|} \tag{1}$$

其中，$|M_X|=\max\{k_j\}$，M_x 表示集合 X 中最大类的类标号。不一致率是数据预处理中一个很有效的准则，在文献[6]提出的数据预处理算法中，就使用不一致率来完成离散化和特征选择，获得了很好的效果。

在本文提出的算法中，使用不一致率来对初始选出的断点进行优化处理，通过定义一个不一致阈值来使离散化算法在保持高效的基础上对冗余断点进行删除处理。

2.2 离散化描述

设 DT 是如前所述的决策表，$U=\{x_1,x_2,\cdots,x_n\}$ 是论域，$A=C\cup\{d\}$，决策属性的值域为 $V_d=\{1,2,\cdots,r(d)\}$。对条件属性 $a\in C$，设其值域 $V_a=[l_a,r_a]$，其中有一组点 $l_a<c_1^a<c_2^a<\cdots<c_{m_a}^a<r_a$，则这一组点把条件属性值域划分为：

$$V_a=[l_a,c_1^a)\cup[c_1^a,c_2^a)\cup\cdots\cup[c_{m_a-1}^a,c_{m_a}^a)\cup[c_{m_a}^a,r_a]$$

即将属性 a 的取值分成了 m^a+1 个等价类，这里每个 $[c_i^a,c_{i+1}^a)$ 称为一个区间，每个 c_k^a 称为一个断点，属性 a 上全部断点的集合构成 a 的断点集 $P^a=\{c_k^a\}(1\leqslant k\leqslant m)$。离散化的目的就是对所有连续属性都找到适宜的断点集合 P，此时令：

$$f^P(x,a)=i\Leftrightarrow f(x,a)\in[c_k,c_{k+1})$$

则得到了一个新的决策表 (U,A,V^P,f^P)，即经过离散化后，原来的决策表被新的决策表所取代。

综上，可以把决策表连续属性离散化的本质归结为利用选取的断点来对条件属性构成的空间进行划分，即把连续的值空间划分成有限个区域。显然，划分的方法不同，离散化后得到的新决策表的粒度数是不同的，如果粒度数过高，则每个等价类中所含的个体数过少，在后续产生规则时的支持度降低，增加了决策系统的冗余度；如果粒度数过低，即造成了属性数据过离散化现象，使得正确分类的实例数所占整个实例空间的比例过低，导致不相容的信息增加[7]。可见，连续属性离散化存在一个度的问题，文献[8]指出，寻求最优离散化问题是 NP-hard 问题。

2.3 决策表的信息论描述及断点重要性定义

从上面的分析可知，决策表连续属性离散化就是要通过定义一组断点集，把连续值空间划分为有限空间，如何对断点进行选择是划分不同离散方法的重要标准之一，近期比较流行的 Chi2 算法就是把统计学理论引入离散化算法，并通过实验证明了这种方法的有效性[9]。在文献[10]中，作者把信息论观点引入决策表中以取代其代数观点，并证明了这两种观点在一定程度上的等价性，且在文献[11]中使用这一结论实现了决策表的属性约简并取得了很好的效果。本文基于决策表信息论观点提出一种新的断点重要性度量方式，以实现连续属性离散化。

定义 3 设决策表 DT 如前所述，$X\subseteq U$ 为子集，其实例个数为 $|X|$，其中决策属性为 $j(j=1,2,\cdots,r(d))$ 的实例个数为 k_j，则定义这个子集的信息熵[3]为：

$$E(X)=-\sum_{j=1}^{r(d)}p_j\log_2 p_j,\ p_j=\frac{k_j}{|X|} \tag{2}$$

其中，$0\leqslant E(X)\leqslant 1$。信息熵 $E(X)$ 表现了集合中信息的纯度，其值越小，说明集合 X 中个别决策属性值占主动地位，因此混乱程度越小。当集合中所有实例所属的类别个数都相等时，$E(X)=1$；当 X 中实例的决策属性值都相同时，$E(X)=0$。

在集合 $X\subseteq U$ 选取一断点为 c，决策属性值为 $j(j=1,2,\cdots,r(d))$ 的实例中，属于集合 X 且属性的值又小于断点值 c 的实例个数记为 $l_j^X(c)$，大于断点 c 的实例个数记为 $r_j^X(c)$。因此断点 c 可以将集合 X 分成两个子集 X_l 和 X_r，且由式(2)导出为：

$$E(X_l)=-\sum_{j=1}^{r(d)}p_j\log_2 p_j,\ p_j=\frac{l_j^X(c_i^a)}{|X_l|}$$

$$E(X_r)=-\sum_{j=1}^{r(d)}p_j\log_2 p_j,\ p_j=\frac{r_j^X(c_i^a)}{|X_r|}$$

此时由于有了断点 c_i^a，集合 X 的信息熵变为：

$$E^c(X)=\frac{|X_l|}{|X|}E(X_l)+\frac{|X_r|}{|X|}E(X_r) \tag{3}$$

则对于决策表 DT，假设 $L=\{Y_1,\cdots,Y_i,\cdots,Y_m\}$ 是决策表 (U,A,V,F) 已经被选取的断点的集合 P 依据属性 $a\in C$ 划分得到的等价类，则此时根据式(3)，整个决策表依据断点集合 P 的信息熵为：

$$E^U(P)=\frac{|Y_1|}{|U|}E(Y_1)+\cdots+\frac{|Y_i|}{|U|}E(Y_i)+\cdots+\frac{|Y_m|}{|U|}E(Y_m)$$

此时，从候选断点集合中选取一新的断点 c，假设该断点把 L 中的等价类 Y_i 分为两个部分 Y_i^l 和 Y_i^r，即断点集合 $P+\{c\}$ 将决策表划分为 $L^c=\{Y_1,\cdots,Y_i^l,Y_i^r,\cdots,Y_m\}$，同上可以计算出断点集合 $P+\{c\}$ 的信息熵 $E^U(P+\{c\})$。

定义 4 根据上述分析，定义从候选断点集中选出用于离散化的断点 c 的信息熵为：

$$E(c)=E^U(P)-E^U(P+\{c\}) \tag{4}$$

从定义可以看出，断点熵 $E(c)$ 是一差值，指加入一个断点后，依据新的断点集合形成的决策表信息熵是否有所减小，如果减小，即 $E(c)>0$，说明新断点 c 使等价类子集的决策属性值趋于单一，因此 $E(c)$ 体现了断点的重要性程度，$E(c)$ 越大表明断点 c 越重要。

3 算法描述

根据上面的理论分析，提出了本文算法 DIEIC(Discretization Based-on Information Entropy with Inconsistency Checking)，该算法属于局部离散算法，整个离散过程被分为以下 3 步。

3.1 计算候选断点集

在这一步中使用“邻值取均”算法[12]生成候选断点集，这一方法的好处在于其简单高效，且易于理解。设决策表 DT 如前所述，对每一个连续条件属性 $a\in C$，论域中其有限个不同的属性值经过排序后为：$l_a=v_0^a<v_1^a<\cdots<v_n^a=r_a$，则候选断点取为：

$$c_i^a=(v_{i-1}^a+v_i^a)/2\quad(i=1,2,\cdots,n)$$

所以连续条件属性 a 的候选断点集合为 $B^a=\{c_i^a\}$。

3.2 根据信息熵选取断点

为了从候选断点集 $B^a=\{c_i^a\}$ 中选择出合适的断点形成离散化断点集合 P^a（初始 P^a 为空），循环地使用式(4)，每次从 B^a 选出一个断点放入 P^a 中。在每次的选择过程中，要对 B^a 中所有的断点计算熵 $E(c)$，选择使 $E(c)$ 最大的那个断点放入 P^a，并从 B^a 中把该断点删除。

当候选断点集 B^a 中不再有使 $E(c)$ 为正的断点存在时，则终止，这样断点集 P^a 就是初始的离散断点集合，这就形成了一个初始的离散化决策表。

在处理时，每次选取一个连续的条件属性进行处理，属于局部寻优算法，虽然没有考虑属性间的相互关联性，但其效率要远远高于全局性离散化，同时用信息熵作为度量断点重要性的指标，可以不用预先设定区间个数，这体现了本文离散算法的自组织性。

3.3 基于不一致率合并区间

在上述步骤停止后，连续属性 $a\in C$ 的连续值空间被断点集合 P^a 划分成有限的离散值空间，但是这样的结果中存在两个不足：一是如果决策表中决策类别数较多，离散后的决策表粒度数会比较高，在后续的规则产生或分类处理时就体现不出离散数据的优越性；二是对于含有噪音数据的决策表，上述离散化往往出现一些孤立的区间，严重地影响了后续挖掘中形成的模型的质量。为了避免上述不足，提出了使用决策表不一致率来优化离散化过程。这里为属性 a 设定一不一致率阈值 $\xi^a_{\max}$。

设断点集合 P^a 把属性 $a\in C$ 划分为等价类簇 $L=\{Y_1,\cdots,Y_i,\cdots,Y_m\}$，根据式(1)可以计算出这些等价类的不一致率。设 ξ^{Y_i} 和 $\xi^{Y_{i+1}}$ 为相邻两个等价类 Y_i 和 Y_{i+1} 的不一致率，从而计算两个等价类合类的不一致率，即：

$$\xi^{Y_i\cup Y_{i+1}}=\frac{|Y_i\cup Y_{i+1}|-|M_{Y_i\cup Y_{i+1}}|}{|Y_i\cup Y_{i+1}|}\quad(1\leqslant i<m)$$

如果 $\xi^{Y_i}\leqslant\xi^a_{\max}$ 且 $\xi^{Y_{i+1}}\leqslant\xi^a_{\max}$，同时 $\xi^{Y_i\cup Y_{i+1}}\leqslant\xi^{Y_i}$ 且 $\xi^{Y_i\cup Y_{i+1}}\leqslant\xi^{Y_{i+1}}$，则合并 Y_i 和 Y_{i+1}，并从 P^a 中删除划分这两个等价类的断点。

通过以上 3 步处理，就可获得连续属性 $a\in C$ 的断点集 P^a，用同样的方法对决策表 DT 中的其他连续属性进行处理，最终可获得全部断点的集合 P。依据断点集合 P，带有连续属性的决策表 (U,A,V,f) 就被新的决策表 (U,A,V^P,f^P) 取代，离散化过程结束。

4 实验与分析

为验证上述算法的有效性并比较该算法与其他离散算法的性能，本文拟用 UCI 机器数据库[13]的数据进行实验，选用的数据基本信息如表 2 所列（更详细的数据介绍参看 UCI 数据库）。

表 2 数据信息表

数据集名	连续属性数	类别数	实例数
Iris	4	3	150
Wine	9	7	214
Pima	8	2	768
Sonar	60	2	208

作为对比，对上述 4 个数据集使用本文提出的算法（DIEIC）和其他算法（Equal_F[1]，Ext_Chi2[9]，CAIM[14]）进行离散化处理。

Equal_F 指等频算法，是典型的无监督数据离散化方法，以其高效性著称；Ext_Chi2 是对基于统计学理论提出的有监督的、自底向上的离散算法 Chi2 的改进；CAIM 是目前比较新的算法，该算法定义了属性和类之间的依赖度，并以此作为离散化过程的判断标准；本文提出的算法是基于信息论中信息熵的概念并与决策表本身的性质进行结合形成的，属于有监督的自顶向下的离散算法。为了完成对比，在 VC++6.0 开发环境中实现了这 4 种算法，并对上述 4 个数据集利用实现的算法进行离散化处理。

为了获得对比效果，本文使用当前流行的数据挖掘分类技术 SVM[15]（Support vector machine，支持向量机）对离散后的数据进行分类。DMBench 1.0 ahpha[16]是上海交通大学图像处理与模式识别研究所开发的用于数据挖掘的软件，软件中实现了 SVM 算法。

实验过程中对离散后的每个数据集随机选取 80%作为训练集，其余 20%作为测试集。选用 SVM 的分类方式为“一对多(1-V-r)”，模型类型选为 C-SVC，核函数类型选为 RBF 函数，惩罚因子 C 搜索范围为[1,100]，核函数参数 γ 搜索范围设为[0.05,0.5]。由于核函数依赖于输入样本向量的内积，因此大的属性值容易导致计算复杂，训练时间较长，为避免上述情况发生，将决策表中所有条件属性值进行如下归一化处理：

$$\bar{x}_i=2\frac{x_i-\min(x_i)}{\max(x_i)-\min(x_i)}-1$$

归一化后的属性值 $x_i\in[-1,+1]$。训练样本与测试样本采用相同的归一化方法。完成上述设定后进行分类处理，获得的预测精度如表 3 所列。

表 3 SVM(1-V-r)分类预测实验结果

比较指标	数据集名	算法			
		Equal_F	Ext_Chi2	CAIM	DIEIC
预测精度(%)	Iris	93.3	100.0	90.0	93.3
	Wine	94.4	95.6	94.4	96.8
	Pima	72.1	65.2	70.6	73.7
	Sonar	47.6	60.6	76.2	75.3

从实验结果可知，本文提出的算法（DIEIC）在处理类别数多且样本量大的数据时较其他算法有一定的优势。Ext_Chi2 在处理 Iris 可以达到 100%的预测精度，这主要是因为 Chi2 系列算法是自底向上的，对处理小样本数据有很好的效果，但对于大量实例的数据而言，Chi2 系列算法不但处理速度慢而且效果也比自顶向下的数据要差。从表中还可以看到，对于处理属性多的数据，本文的算法与 CAIM 算法性能相近，但本文算法在计算量上明显比 CAIM 小很多，从数据预处理的标准来看，本文算法比 CAIM 的实用价值大。

结束语 数据挖掘用于在海量数据中发现知识。数据离散化技术作为其数据预处理过程，随着数据挖掘的发展而成为该领域的重要研究课题之一。好的离散化方法要达到 3 个标准[17]：一是尽量避免或减少原数据集中的信息损失；二是离散区间数（或断点数）尽量少，以提高后续分类学习算法的效率；三是作为数据挖掘的预处理步骤，离散算法应简单而有效，运算量小，易于实现。由于决策表提出的离散算法步骤简

（下转第 237 页）

j 代($j<k$)后进行一次分配和收集操作。每次分配操作的通讯代价为 h_i,每次收集操作的通讯代价也为 h_i。当群体进化 k 代时的通讯总代价为 $2h_ik/j$。基于 Agent 的并行遗传算法中,每个子群体独立进化 j 代($j<k$)后进行一次通讯,每一代均产生了比黑板中 B-Agent 更优的 B-Agent。每个子群体的每一次通讯都接收一个来自黑板的当前全局 B-Agent,并向黑板发送一个自己的 B-Agent。若传送一个进化个体为 1 个单位的通讯代价,则每个子群体同时进化 k 代时通讯代价总和为 $2hk/j$。由此可见,同样情况下,基于 Agent 的并行遗传算法的通讯代价只有经典粗粒度孤岛并行遗传算法的 $1/i$。

另外,一个子群体从上一次访问公共存储区中的 B-Agent 到下一次访问公共存储区中的 B-Agent 之间,很可能公共存储区中的 B-Agent 被更新了多次,子群体实际只访问了上一次访问时公共存储区中的 B-Agent 和下一次访问时公共存储区中的 B-Agent,跳过了中间被更新了多次的公共存储区中的 B-Agent 的访问,因而避免了一些不必要的通讯代价。因此,基于 Agent 的并行遗传算法模型的通讯代价比其它粗粒度并行模型要小,比细粒度并行模型的通讯代价更小。

结束语 在遗传算法中引入 Agent 之后,可充分发挥遗传算法的群体搜索策略的优势。运用多 Agent 系统智能地实施遗传算法,能从随机的遗传过程中获取表征进化状态的信息,智能地调度遗传操作,较好地克服早熟问题,在个体空间中有效地实现全局最优搜索。基于多 Agent 的并行遗传算法的最大特点是增强了子群体的自主性,极大地减少了子群体之间的通讯开销。各子群体的 Agent 之间交换的信息始终是当前整个群体中的最优个体,一个子群体所引进的最优个体有利于改善其新一代个体的特质,有利于该子群体加快收敛速度。石油勘探属性优化的应用实例表明,它具有更好的性能,能高效地求解地震勘探数据,处理最优化问题。

参 考 文 献

[1] 杨光正,等. 一种薄层砂岩的分类方法[J]. 模式识别与人工智能,1994,7(4):312-316

[2] Srinivas M, Patnaik L M. Adaptive Probabilities of Crossover and Mutation in Genetic Algorithms[J]. IEEE Transaction on System, Man and Cybernetics, 1994, 24(4): 656-667

[3] Hartvigsen G, Johnsan D. Corperation in a distributed artificial intelligence environment the stormcast application[J]. Engineering Apply of Artificial Intelligence, 1990, 3(3): 229-237

[4] Onn S, Tennenholtz M. Determination of social laws for multi-Agent mobilization[J]. Artificial Intelligence, 1997, 95(1): 155-167

[5] Chaib-draa B, Millot P. A framework for cooperative work: An approach based on the intentionlity[J]. AI in Engineering, 1990 (4): 199-205

[6] Findler N V. Multiagent coordination and cooperation in a distributed dynamic environment with limited resoures[J]. Artificial Intelligence in Engineering, 1995, 9: 229-238

[7] Zhang C. Cooperation under uncertainty in distributed expert system[J]. AI, 1992, 56: 21-69

[8] 丁明勇,代春艳,杨永斌. 一种基于 Multi-Agent 的工程项目智能评标模型[J]. 计算机科学,2008,35(1):224-226

[9] 梁军,程显毅. 基于混合蚁群遗传算法的 Agent 联盟求解[J]. 计算机科学,2009,36(4):227-231

(上接第 233 页)

单,为避免信息量损失,把信息熵引入处理过程,满足离散化的标准,处理过程中每一步都使用最简单高效的处理算法来实现。通过实验也证明该算法对数据量与类别数多的数据有较好的性能及较高的实用价值。

参 考 文 献

[1] Hussain F, Liu H, Tan C L, et al. Discretization: an Enabling Technique[J]. Data Mining and Knowledge Discovery, 2002, 6 (4): 393-423

[2] Dougherty J, Kohavi R, Sahami M. Supervised and Unsupervised Discretization of continuous feature[C]// Machine learning: Proc. 12th Int'l Conf. 1995: 194-202

[3] 孟庆生. 信息论[M]. 西安:西安交通大学出版社,1986:18-30

[4] 王东锴. 决策表中的知识发现研究[D]. 合肥:中国科学技术大学,2002

[5] 王国胤. Rough 集理论和知识获取[M]. 西安:西安交通大学出版社,2001

[6] Ribeiro M X, Ferreira M R P, et al. Data pre-processing: a new algorithm for feature selection and data discretization[C]// Conference on Soft computing as Transdisciplinary Science and Technology: Proc. 5th (CSTST'2008). Cergy-Pontoise, France, 2008: 252-257

[7] 高原,耿国华,周明全,等. 一种改进的快速数据离散化算法[J]. 小型微型计算机系统,2009,30(2):279-282

[8] Nguyen S H. Discretiazaion of Real Value Attributes: A Boolean reasoning approaches[D]. Warsaw: Warsaw University, Poland, 1997

[9] Su Chao-Ton, Hsu Jyh-Hwa. An Extended Chi2 Algorithm for Discretization of Real Value Attributes[J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(3): 437-441

[10] 苗夺谦,王珏. 粗糙集理论中概念与运算的信息表示[J]. 软件学报,1999,10(2):113-116

[11] 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):1-8

[12] 谢宏,程浩忠,牛东晓. 基于信息熵的粗糙集连续属性离散化算法[J]. 计算机学报,2005,28(9):1570-1574

[13] Hettich S, Bay S D. The UCI KDD Archive [DB/OL]. http://kdd. ics. uci. edu/, 1999

[14] Kurgan L A, Cios K J. CAIM Discretization Algorithm[J]. IEEE Transactions on Knowledge and Data Engineering, 2004, 16(2): 145-153

[15] 李国正,王猛,等. 支持向量机导论[M]. 北京:电子工业出版社,2000

[16] 刘国平. 支持向量机若干问题以及数据挖掘平台的研究[D]. 上海:上海交通大学,2005

[17] 赵静娴,倪春鹏,等. 一种高效的连续属性离散化算法[J]. 系统工程与电子技术,2009,31(1):195-199