

文章编号:1673-5005(2009)05-0159-05

基于粗糙集的支持向量回归机混合算法

邓九英^{1,2}, 王钦若¹, 毛宗源³, 杜启亮³

(1. 广东工业大学 自动化科学与工程学院, 广东 广州 510090; 2. 广东教育学院, 广东 广州 510303;
3. 华南理工大学 自动化科学与工程学院, 广东 广州 510640)

摘要:利用粗糙集(RS)对不精确数据的处理能力,生成分类数据的边界集,替代原始样本作为训练集,减少训练集与获取的支持向量的数量,然后使用支持向量机的最小序列优化(SMO)算法改进回归学习机的性能。将粗糙集与SMO回归算法结合提出一种混合函数回归算法RS-SMO-RA。在常用SMO回归算法SMO-RA基础上,扩增一段简短的生成边界样本的算法程序。仿真结果表明,算法RS-SMO-RA的效率更高,且能够改进学习结果的性能。

关键词:支持向量回归机; SMO回归算法; 边界样本集; 粗糙集

中图分类号:TP 18

文献标识码:A

Support vector regression hybrid algorithm based on rough set

DENG Jiu-ying^{1,2}, WANG Qin-ruo¹, MAO Zong-yuan³, DU Qi-liang³

(1. School of Automatic Science & Engineering, Guangdong University of Technology, Guangdong 510090, China;
2. Guangdong Institute of Education, Guangzhou 510303, China;
3. College of Automatic Science & Engineering, South China University of Technology, Guangdong 510640, China)

Abstract: Rough set (RS) was utilized to analyze imprecise data and get the boundary set of the classified data. The boundary set can substitute the original inputs as a training subset, and the size of the training set and the gained support vectors are shorten. Then, the learning machine has solutions with high quality by sequential minimal optimization (SMO) algorithm of regression. Based on rough set and SMO algorithm of regression, a hybrid algorithm of RS-SMO-RA was presented for the enhanced capability of machine learning. For differentiating boundary samples, a simple and short module was added to the common algorithm of SMO regression, SMO-RA. The presented RS-SMO-RA algorithm is verified with high efficiency and performance.

Key words: support vector regression; algorithm of SMO regression; boundary set; rough set

支持向量机(SVM)的研究与应用在近几年有了较大的发展^[1-3],主要工作大多是关于SVM分类机,较为突出的是最小序列优化(SMO)算法^[4],尤其对大数据输入对象的学习,使用一次优化两个变量的二次规划的解析方法,使得算法运行时对存储空间与时间的需求减少。对SMO进行扩展,人们已经研究出使用SMO算法的回归机^[5-6],以及相关的改进措施与算法,促进了SVM回归机的研究与应用^[7-8]。原始SVM问题具有时间 $O(l^3)$ 与空间 $O((l+1)^2)$ 复杂度,在大样本条件下SVM学习机的性能

下降,存在求解比较困难的问题。把训练样本缩小至与SV分布的凸壳(边界)区域,可通过减少训练样本数来提高SVM回归机的性能^[9]。用粗糙集(RS)能快速准确区分出类边界区域中的样本,代替原始输入作为SVM学习机的训练子集。笔者把RS与使用SMO的回归机结合提出一种混合回归算法(RS-SMO-RA),仿真试验使用经典函数加上随机噪声作为输入模式,比较算法RS-SMO-RA与不用RS的SMO-RA算法的识别结果。

收稿日期:2008-12-20

基金项目:广东省科技厅科技攻关项目(2005B10201006);广州市科技攻关引导项目(2003Z3-D0091)

作者简介:邓九英(1962-),女(汉族),江西临川人,副教授,博士研究生,研究方向为智能计算、数据挖掘技术及仿真技术。

1 支持向量回归机

支持向量机的回归问题与分类问题相似,也是已知一个训练集^[10],如

$$T = \{(x_1, y_1), \dots, (x_l, y_l)\} \in (X, Y)^l. \quad (1)$$

其中 $x_i \in X = \mathbf{R}^n$ 是输入(或模式), $y_i \in Y = \mathbf{R}$ 是输出, $i = 1, \dots, l$ 。它与分类问题不同, y_i 并不限定取 -1 或 1 , 而可取任意实数。问题是给定一个新的模式 x , 根据训练集 T 推断它所对应的输出 y 。

1.1 支持向量分类机与回归机

对于线性可分成两类的问题, 根据给定的训练集 T , 其中 $x_i \in X$ 是输入指标, 或称模式, $y_i \in Y = \{1, -1\}$ 是输出指标, 问题就是对任意给定的一个新的模式 x , 根据训练集, 推断它所对应的输出 y 是 1 还是 -1 , 也就是寻找实值函数 $g(x)$, 以使用决策函数 $f(x) = \text{sgn}(g(x))$ 推断任一模式 x 对应的 y 数值。解决这种分类问题的方法就称为分类学习机。当 $g(x)$ 为线性函数时, 称为线性分类学习机; 当 $g(x)$ 为非线性函数时, 称为非线性分类学习机。

考察线性可分问题的特征, 则训练集 T 线性可分的充分必要条件是, T 的正类点集 $\{x_i | y_i = 1\}$ 的凸壳和负类点集 $\{x_i | y_i = -1\}$ 的凸壳不相交。假设已知图1中给出2维空间上的两类点, 分别做它们的凸壳, 找出这两个凸壳的最近点 c 和 d , 作线段 cd 的垂直平分线为平面分划直线, 或者在给出法方向 w 的情况下构造分划直线。对于适当给定的法方向, 有过 c 和 d 两点的两条极端直线, 应该选取使两条极端直线之间的“间隔”达到最大的法方向。

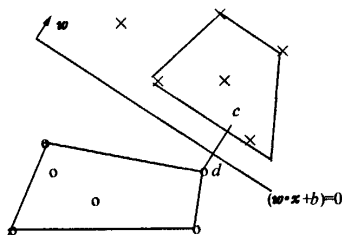


图1 最近点与分划平面直线

Fig.1 The most adjacent points and a bisector

考虑线性回归问题, 即所寻找的函数 $f(x)$ 为线性函数

$$y = f(x) = (w \cdot x) + b. \quad (2)$$

这个线性函数对应于空间 $\mathbf{R}^n \times \mathbf{R}$ 上的一个超平面。最简单的情况为 $n = 1$ (如图1), 线性回归问题是找一条与已知训练点偏离较小的直线。

定义1 设给定训练集 T , 并给定 $\varepsilon > 0$, 称一个超平面 $y = (w \cdot x) + b$ 为对于训练集 T 的硬 ε -带超平面, 如果该超平面的 ε -带包含了训练集 T 中所有的训练点, 即超平面满足

$$-\varepsilon \leq y_i - ((w \cdot x) + b) \leq \varepsilon, \quad i = 1, 2, \dots, l. \quad (3)$$

图2中的实线表示一个超平面, 两虚线之间的区域是该超平面的 ε -带。显然, 对由有限个训练点组成的训练集来说, 当 ε 充分大时, 硬 ε -带超平面总是存在的。寻找最小使硬 ε -带超平面存在的 ε 值为 $\varepsilon_{\min}$, 则是下列最优化问题:

$$\begin{cases} \min_{w, b} \varepsilon, \\ \text{s. t.} \quad \varepsilon \leq y_i - ((w \cdot x) + b) \leq \varepsilon, \quad i = 1, 2, \dots, l. \end{cases} \quad (4)$$

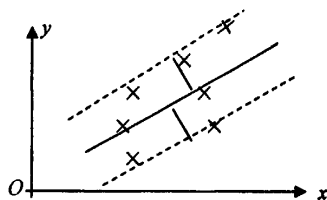


图2 硬 ε -带超平面

Fig.2 Hyperplane with hard ε -band

从给定的训练集 T 出发构造出两类点, 分别记它们为 D^+ 和 D^- , 即

$$D^+ = \{(x_i^T, y_i + \varepsilon)^T, i = 1, \dots, l\}, \quad (5)$$

$$D^- = \{(x_i^T, y_i - \varepsilon)^T, i = 1, \dots, l\}. \quad (6)$$

这样就形成了一个对正负两类点进行线性分划的问题。

定理1 设对给定的训练集 T 和 $\varepsilon > 0$, 考虑由式(5)和(6)定义的集合 D^+ 和集合 D^- , 则式(2)超平面是硬 ε -带超平面的充要条件是 D^+ 的凸壳和 D^- 的凸壳分别位于该超平面的两侧。

把训练集 T 变形, 令 $(x_{l+i}^T, y_{l+i} - \varepsilon)^T = (x_i^T, y_i - \varepsilon)^T, i = 1, \dots, l$, 则训练集式(1)可表为 $\{((x_1^T, y_1 + \varepsilon)^T; 1), \dots, ((x_l^T, y_l + \varepsilon)^T; 1), ((x_{l+1}^T, y_{l+1} - \varepsilon)^T; -1), \dots, ((x_{2l}^T, y_{2l} - \varepsilon)^T; -1)\}$. (7) 应用线性可分问题的最大间隔法, 超平面的权向量形式为 $(w^T, \eta)^T$, 则最优化问题为

$$\begin{cases} \min_{w, \eta, b} \frac{1}{2} \|w\|^2 + \frac{1}{2} \eta^2, \\ \text{s. t.} \quad z_i ((w \cdot x_i) + \eta (y_i + z_i \varepsilon) + b) \geq 1, \\ \quad i = 1, \dots, 2l. \end{cases} \quad (8)$$

其中

$$z_i = \begin{cases} 1, & i = 1, \dots, l; \\ -1, & i = l+1, \dots, 2l. \end{cases} \quad (9)$$

求解后构造出分划超平面 $(\mathbf{w}^* \cdot \mathbf{x}) + \eta^* y + b^* = 0$, 整理后得到线性回归函数

$$y = (\mathbf{w} \cdot \mathbf{x}) + b,$$

其中

$$\mathbf{w} = -\mathbf{w}^*/\eta^*, b = -\mathbf{w}^*/b^*.$$

最优化问题(8)的对偶问题为

$$\begin{cases} \min_{\alpha} \frac{1}{2} \sum_{i=1}^{2l} \sum_{j=1}^{2l} z_i z_j \alpha_i \alpha_j ((\mathbf{x}_i^T, \mathbf{y}_i + z_i \varepsilon)^T \cdot (\mathbf{x}_j^T, \mathbf{y}_j + z_j \varepsilon)^T) - \sum_{i=1}^{2l} \alpha_i, \\ \text{s. t. } \sum_{i=1}^{2l} z_i \alpha_i = 0, \alpha_i \geq 0, i = 1, \dots, 2l. \end{cases} \quad (10)$$

推广硬 ε -带超平面的最大间隔回归法,并引进核函数 $K(\mathbf{x}_i, \mathbf{x}_j)$ 用曲线代替直线,把问题转化为高维空间(Hilbert 空间)中的线性回归问题,构造出最优优化问题 $(\alpha^{(*)} \in \mathbf{R}^{2l})$ 为

$$\begin{cases} \min \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)k(\mathbf{x}_i, \mathbf{x}_j) + \varepsilon \sum_{i=1}^l (\alpha_i^* + \alpha_i) - \sum_{i=1}^l y_i(\alpha_i^* - \alpha_i), \\ \text{s. t. } \sum_{i=1}^l (\alpha_i - \alpha_i^*) \geq 0, \alpha_i^* \leq C. \end{cases} \quad (11)$$

式(11)中的 b 估算可任选位于开区间 $(0, C)$ 中的 α_j 或 α_k , 按下列方式计算:

若选 $\alpha_j \in (0, C)$,

$$b = y_j - \sum_{i=1}^l (\alpha_i^* - \alpha_i)k(\mathbf{x}_i, \mathbf{x}_j) + \varepsilon;$$

若选 $\alpha_k \in (0, C)$,

$$b = y_k - \sum_{i=1}^l (\alpha_i^* - \alpha_i)k(\mathbf{x}_i, \mathbf{x}_k) - \varepsilon.$$

1.2 用于函数回归的 SMO 算法

原始 SMO 算法是为了解决分类问题设计的,经过探讨与修改后得出了较简单有效的用于回归问题的 SMO 算法^[5]。

为了转换式(10), (11) 的形式,令 $\lambda_i = \alpha_i^* - \alpha_i$, $|\lambda_i| = \alpha_i^* + \alpha_i$, 其中 $-C \leq \lambda_i \leq C$, 使用简化符号 $k_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$, 并取 $k_{ij} = k_{ji}$, 改写决策函数和优化问题如下:

$$f(\mathbf{x}, \lambda, b) = \sum_{i=1}^l \lambda_i K(\mathbf{x}_i, \mathbf{x}) + b, \quad (12)$$

$$W(\lambda) = \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l \lambda_i \lambda_j k_{ij} + \varepsilon \sum_{i=1}^l |\lambda_i| - \sum_{i=1}^l y_i \lambda_i.$$

其约束条件为 $\sum \lambda_i = 0$ 。假设两个优化的参数为 λ_u 与 λ_v , 则 $s^* = \lambda_u + \lambda_v = \lambda_u^* + \lambda_v^*$, 其中上标“*”表示优化前的值, 需要循环优化不满足 Karush-Kuhn-Tucker(KKT) 条件的一对参量, 直至程序终止。用于回归的 KKT 条件为

$$|y_i - f_i| \begin{cases} < \varepsilon, \lambda_i = 0, \\ = \varepsilon, -C < \lambda_i \neq 0 < C, \\ > \varepsilon, |\lambda_i| = C. \end{cases}$$

推广应用到回归的 SMO 算法描述如下:

$$\textcircled{1} s^* = \lambda_u^* + \lambda_v^*;$$

$$\textcircled{2} \eta = k_{uu} + k_{vv} - 2k_{uv}, \Delta = 2\varepsilon/\eta;$$

$$\textcircled{3} \lambda_v = \lambda_v^* + 1/\eta (y_v - y_u + f_u^* - f_v^*);$$

$$\textcircled{4} \lambda_u = s^* - \lambda_v^*;$$

$$\textcircled{5} \text{if}(\lambda_u \cdot \lambda_v < 0)$$

$$\{ \text{if}(|\lambda_v| \geq \Delta \wedge |\lambda_u| \geq \Delta)$$

$$\lambda_v = \lambda_v - \text{sgn}(\lambda_v) \cdot \Delta;$$

else

$$\lambda_v = \text{step}(|\lambda_v| - |\lambda_u|)s^*;$$

$$\textcircled{6} L = \max(s^* - C, -C), H = \min(C, s^* + C);$$

$$\textcircled{7} \lambda_v = \min(\max(L, \lambda_v), H);$$

$$\textcircled{8} \lambda_u = s^* - \lambda_v.$$

选择任一符合条件的优化参数对 b 进行更新, 否则取两者的平均值(或不更新)。

$$b_u = y_u - f_u^* + (\lambda_u^* - \lambda_u)k_{uu} + (\lambda_v^* - \lambda_v)k_{uv} + b^*,$$

$$b_v = y_v - f_v^* + (\lambda_u^* - \lambda_u)k_{uv} + ((\lambda_v^* - \lambda_v)k_{vv} + b^*.$$

2 粗糙集与支持向量回归机的算法设计

SMO 回归算法的优势是能处理高维大数据量的对象,在具有稀疏解的情况下比较有效,否则效果变差或有较多支持向量时运行速度变慢。通常支持向量分布在数据集的边界域中,可采用粗糙集的方法寻找原始输入的边界集,该边界集作为回归学习机的训练子集,可提高回归学习机的性能。

2.1 粗糙集

给定信息系统 (U, A) , $A = C \cup D, C \cap D = \emptyset$, 其中 C 是条件属性集, D 是决策属性集^[11]。对于子集 $W \subseteq U$ 和属性 $a \in A$ 的等价关系 U/a , 定义 W 的

下近似为 $W^{(U/a)-} = \bigcup_{V \in U/a, V \subseteq W} V$, 也表示为 $S_a(W)$ (支持子集); 定义 W 的上近似为 $W^{(U/a)+} = \bigcup_{V \in U/a, V \cap W \neq \emptyset} V$, 或者称为 W 的正域 $\text{pos}(W) = W^{(U/a)-}$, W 的负域 $\text{neg}(W) = U - W^{(U/a)+}$, W 的边界域 $\text{bn}(W) = W^{(U/a)+} - W^{(U/a)-}$. 把 U 对所有属性进行分类为 $U/a_1, U/a_2, \dots, U/a_{|A|}$; 用 $\text{pos}(W)$ 表示等价类肯定属于 W 的 U 中元素组成的集合, 用 $\text{neg}(W)$ 表示等价类肯定不属于 W 的 U 中元素组成的集合, 用 $\text{bn}(W)$ 表示介于前两者之间的 U 中元素组成的集合, 则属性的边界集为各个属性边界集的交集, 即 $\text{bn}_A(W) = \text{bn}_{a_1}(W) \cap \dots \cap \text{bn}_{a_n}(W)$, $n = |A|$.

(13)

从式(9)可知决策属性 z 只与单个属性 y 有关, 其余输入 x 的分量通过式(12)对 y 产生影响, 没有对 z 产生直接的影响, 式(13)只须保留单一条件属性的边界集. 如果 y 是连续量时应做离散化处理, 考虑 ε 的作用尽量取离散间隔 $\delta \in [\varepsilon, 3\varepsilon]$ 较合理, 因为离散间隔的大小会影响边界集的规模与精确度.

2.2 使用粗糙集与 SMO 的混合回归算法设计

粗糙集的作用是生成输入模式的边界集, 并替代原始输入数据作为回归学习机的训练子集, 同样使用 SMO 回归算法进行变量优化, 构造出基于 RS 的 SMO 回归算法(RS-SMO-RA), 算法流程如图 3 所示.

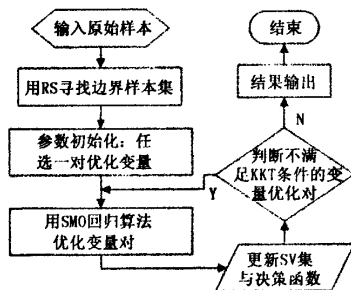


图 3 混合回归算法流程

Fig.3 Flowchart of RS-SMO-RA algorithm

程序运行的终止条件之一是设置足够大的总迭代次数, 另一个终止条件是支持向量 SV 的总数连续两次不变, 两个条件中任一个满足都表示退出循环.

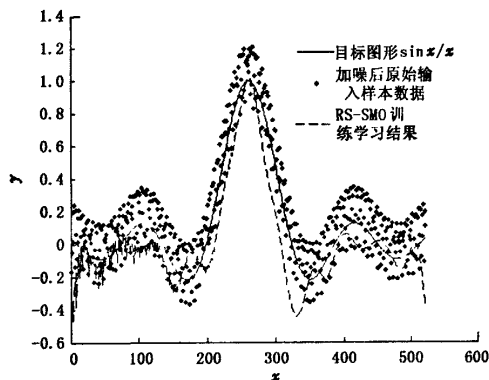
3 仿真结果

用 RS-SMO-RA 算法做试验, 并把结果与不用 RS 的 SMO-RA 算法(去掉图 2 中的“寻找边界集”模块)的试验结果比较, 试验对象的原形是函数 $\sin x/x$, $x \in [-13, 13]$, 均匀采样($\delta = 0.05$)获取 $l_1 = 521$ 个样本, 加上由

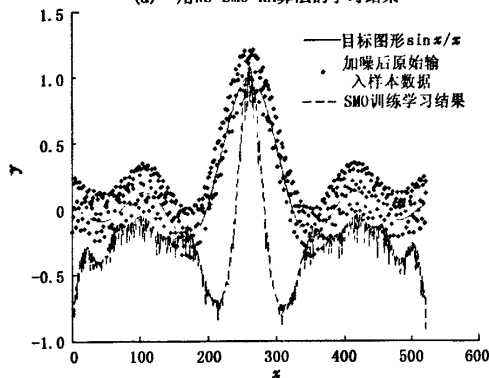
$$x_t^{(k)} = 3.78x_t^{(k-1)}(1 - x_t^{(k-1)})$$

产生的随机噪声 $(x_t - 0.5)/2$, 作为原始输入数据. 其中 $x_t^{(0)} = 0.1$, 用 RS 方法生成边界集的训练样本数 $l_2 = 469$ 个.

核函数的形式使用径向基 RBF 函数 $K(x_i, x_j) = \exp(-\|x_i - x_j\|^2)$; 试验时应根据对象选择参数 ε 和 C , ε 过大会无解, ε 过小时支持向量较多, 会降低回归学习机的性能, C 的取值影响支持向量构造出决策函数的平滑性, 图 4 和图 5 分别是两组不同 ε 和 C 值对应两种算法的回归结果, 它们的运行结果性能测试值如表 1 所示.



(a) 用RS-SMO-RA算法的学习结果



(b) 用SMO-RA算法的学习结果

图 4 参数 $\varepsilon = 0.015$ 和 $C = 0.28$ 时两种算法的试验结果

Fig.4 Experimental results of two methods with the parameters $\varepsilon = 0.015$ and $C = 0.28$

试验中 RBF 的平滑因子 $\sigma = 1$, 如果适当调整取不同的 σ 值, 则 C 和 ε 参数值应作相应的改变. 从试验结果可以看出, 算法 SMO-RA 运行得出较多的支持向量, 使得函数估计图的主峰过于陡峭, 图形的毛刺较多, 各项性能指标均不如算法 RS-SMO-RA 的运行结果.

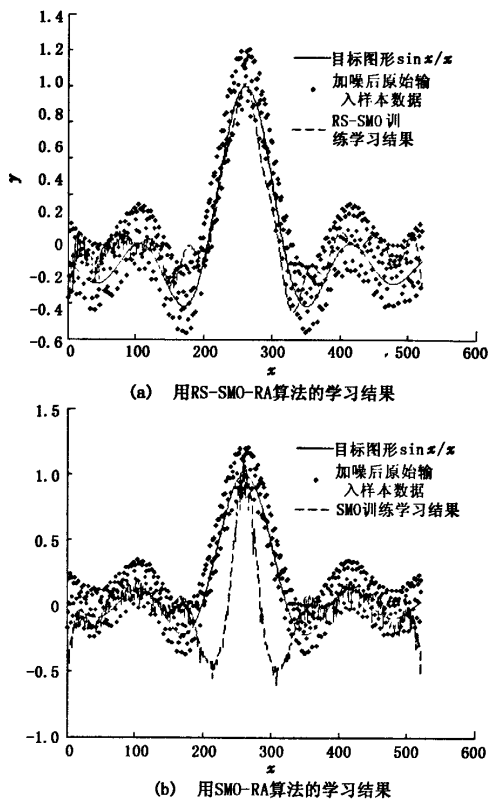


图5 参数 $\varepsilon = 0.03$ 和 $C = 0.25$ 时两种算法的试验结果

Fig.5 Experimental results of two methods with the parameters $\varepsilon = 0.03$ and $C = 0.25$

表1 两种算法的性能测试结果对比

Table 1 Performance results comparison of two methods

参数	RS-SMO-RA		SMO-RA	
$\varepsilon : C$	0.015 : 0.28; 0.03 : 0.25		0.015 : 0.28; 0.03 : 0.25	
SV 个数	210	187	233	201
循环 / 代	237	237	263	263
时间 / s	14	11	20	16

4 结束语

在大数据的系统识别中,算法 RS-SMO-RA 具有明显的优势,增加的程序段既简短又易于实现,算法实现与输入数据的维数高低没有直接的联系.构成决策函数的支持向量数越少,SVM 函数回归的结构风险越小,系统识别的结果性能比用 SMO-RA 算法有了较大的提高,而且采样点越密集(δ 越小),系统的识别结果会越好,RS-SMO-RA 算法的优势会加大.程序运行结果对参数 ε 和 C 的值变化比较敏感,应视应用对象选择合理的数值,结合核函数平滑因

子的设置则是参数组合优化问题.另外,如何选择优化变量 λ_u 与 λ_v 的样本和更新 b 的方法,对系统识别结果也有一定的影响,这需要进一步探讨.

参考文献:

[1] VAPNIK V N. Support vector method[J]. Lecture Notes in Computer Science, 1997(1):1327:263.

[2] VAPNIK Vladimir N. 统计学习理论[M]. 许建华, 张学工,译.北京:电子工业出版社,2004:6.

[3] 徐长航,陈国明,谢静.基于支持向量机和蒙特卡洛的结构可靠性分析方法及应用[J]. 中国石油大学学报:自然科学版,2008,32(4):103-108.

XU Chang-hang, CHEN Guo-ming, XIE Jing. Structural reliability analysis method based on support vector machines and Monte Carlo and its application[J]. Journal of China University of Petroleum (Edition of Natural Science), 2008,32(4):103-108.

[4] PLATT J. Sequential minimal optimization: a fast algorithm for training support vector machines[R/OL]. Tech Rep MSR-TR 98-14, Microsoft Res., Redmond, WA, Apr. 1998 [2007-09-30]. <http://research.microsoft.com/en-us/um/people/jplatt/smotr.pdf>.

[5] FLAKE Gary William. Steve Lawrence efficient SVM regression training with SMO [J]. Machine Learning, 2002,46:271-290.

[6] 杨杰,叶晨洲,全勇,等.支持向量机回归的简化 SMO 算法[J]. 红外与激光工程, 2004,10(33):371-375.

YANG Jie, YE Chen-zhou, QUAN Yong, et al. Simplified SMO algorithm for support vector regression[J]. Infrared and Laser Engineering, 2004,10(33):371-375.

[7] SHEVADE S K, KEERTHI S S, BHATTACHARYYA C, et al. Improvements to the SMO algorithm for SVM regression[J]. IEEE Transactions on Neural Networks, 2000,9(11):1188-1193.

[8] MATTERA D, PALMIERI F, HAYKIN S. An explicit algorithm for training support vector machines[J]. Signal Processing Letters, IEEE, 1999,9(6):243-245.

[9] 曾绍华,魏延,段廷才,等.一种支持向量逐步回归机算法研究[J]. 计算机工程与应用, 2007,43(8):78-81.

ZENG Shao-hua, WEI Yan, DUAN Ting-cai, et al. Research on algorithm of support vector stepwise regression [J]. Computer Engineering and Applications, 2007,43(8):78-81.

[10] 邓乃扬,田英杰.数据挖掘中的新方法——支持向量机[M]. 北京:科学出版社,2006:77-95.

[11] 张文修,吴伟志,梁吉业,等.粗糙集理论与方法[M].北京:科学出版社,2001:24-34.

(编辑 修荣荣)