

基于 TS-SVM 模型的水安全评价

汪嘉杨^{1,2}, 王文圣², 李祚泳^{1,2}, 张 碧¹

(1. 成都信息工程学院, 四川 成都 610041; 2. 四川大学水电学院, 四川 成都 610065)

摘要 核函数参数和误差惩罚因子的选择对支持向量机模型(SVM)的精度有较大影响。为充分发挥 SVM 的性能,提出基于禁忌搜索算法优化的支持向量机模型(TS-SVM)。将 TS-SVM 应用于山东省水安全指数计算,得到各个水安全利用方案的评价指数值和安全等级。对比分析表明,TS-SVM 具有较强的学习能力,能给出合理的评价结果,是一种有效的评价模型。

关键词 支持向量机;禁忌搜索算法;水安全评价;安全指数

中图分类号:TV213.4 文献标识码:A 文章编号:1004-693X(2010)02-0001-04

Water safety assessment based on TS-SVM model

WANG Jia-yang^{1,2}, WANG Wen-sheng², LI Zuo-yong^{1,2}, ZHANG Bi¹

(1. Chengdu University of Information Technology, Chengdu 610041, China; 2. School of Hydraulic Engineering, Sichuan University, Chengdu 610065, China)

Abstract: The selection of the kernel function parameter and error penalty factor affected the precision of the support vector machine (SVM) significantly. In order to give full play to the SVM model, a TS-SVM model in which parameters were optimized by a taboo search algorithm (TS) was proposed. By using this model to calculate the evaluation indexes of the water safety of Shandong Province, safety indexes and safety grades of every utilization project were obtained. The results show that the TS-SVM model, which has strong learning ability, can yield rational results and is a practicable model for water safety assessment.

Key words: support vector machine; taboo search algorithm; water safety assessment; safety index

随着社会经济的发展,水安全问题越来越成为全世界关注的焦点,水安全已不再单是水资源问题,还包括它引发的其他问题,如经济安全、社会安全问题,甚至上升到了国家安全的高度。在水安全研究中,水安全的范畴、水安全的度量、水安全评价和水安全保障体系的建设等都是其研究的主要内容。但国内外对水安全的研究才刚刚起步,定性研究居多,在系统性和量化方面显得不足^[1],因此需要尽快从水安全评价、水安全度量和水安全保障体系等方面入手,建立一套完整的水安全评价和管理的理论体系。

由于水资源和水环境系统存在大量的不确定因素,水安全评价方法仍处于探索、发展阶段。传统水

安全评价方法有层次分析法、集对分析法、属性识别法、投影寻踪方法等^[2-5]。支持向量机(support vector machine, SVM)能够根据有限样本信息在模型的复杂性和学习能力之间寻求最佳折中,获得因变量和自变量之间非常复杂的映射关系,与传统的基于经验风险最小化的神经网络方法相比,具有学习速度快、全局最优和推广能力强的优点,作出的评价和拟合结果大多明显好于其他的模型和方法,因此得到较为广泛的应用^[6-9]。然而,支持向量机作为统计学习理论的一项核心内容,其应用远未达到理论期望的效果,支持向量机参数的选取问题是影响其应用的一个重要原因,为了充分发挥 SVM 的性能,必须选择合适的参数。研究中发现,核函数参数 σ 和误差惩罚因子 C

对 SVM 的精度有较大影响,对机器学习性能至关重要^[8]。为此,笔者将禁忌搜索(taboo search, TS)算法用于 SVM 模型的参数优化,提出 TS-SVM 耦合模型,并将耦合模型应用于山东省水安全指数计算,相应得到区域各个水安全利用方案的安全评价等级。

1 支持向量机的基本原理

支持向量机常用于分类及回归分析,回归支持向量机是在分类方法的基础上,通过引进适当的损失函数推广而来的。给定 l 个样本数据 $\{x_k, y_k\} \quad k = 1, 2, \dots, n$, 其中 $x_k \in R^m$ 为 m 维输入因子(自变量), $y_k \in R$ 为输出因子(因变量),则函数逼近问题就是要寻找一个函数 f , 使之通过 x 能获得最优的 y 。函数 f 一般是一个非线性函数。利用非线性映射 $\varphi(x)$ 将训练数据集非线性地映射到一个高维特征空间(Hilbert 空间),使得在输入空间中的非线性函数估计问题转化为高维特征空间中的线性函数估计问题^[8]。要求拟合的函数形式为:

$$f(x) = w\varphi(x) + b \quad (1)$$

式中: w 为权向量; $\varphi(x)$ 为输入空间到高维特征空间的非线性映射; b 为偏置量。

根据结构风险化最小原则,即要寻求最优回归超平面,使风险函数 R_{reg} 最小。

$$R_{\text{reg}} = \frac{1}{2} \|w\|^2 + CR_{\text{emp}}^{\epsilon} \quad (2)$$

式中: $\|w\|^2$ 为描述函数 $f(x)$ 复杂度的项; C 为设定的惩罚因子(调整参数),用于控制错分样本惩罚的程度,实现错分样本数与模型复杂度之间的折中; $R_{\text{emp}}^{\epsilon}$ 为经验风险。再引进 ϵ 不灵敏损失函数:

$$|y - f(x)|_{\epsilon} = \max\{0, |y - f(x)| - \epsilon\} \quad (3)$$

它意味着不惩罚偏差小于 ϵ 的误差项,取经验风险为:

$$R_{\text{emp}}^{\epsilon} = \frac{1}{l} \sum_{i=1}^l |y - f(x)|_{\epsilon} \quad (4)$$

则支持向量机的回归问题就等价于解决一个二次规划问题。将其最优化问题利用对偶原理、拉格朗日乘子法和核技术,表示为:

$$\begin{aligned} \max_{\alpha, \alpha^*} J = & -\frac{1}{2} \sum_{i,j=1}^l (\alpha_i - \alpha_i^*) (\alpha_j - \alpha_j^*) k(x_i, x_j) - \\ & \epsilon \sum_{i=1}^l (\alpha_i + \alpha_i^*) + \sum_{i=1}^l y_i (\alpha_i - \alpha_i^*) \\ \text{s.t.} \quad & \begin{cases} \sum_{i=1}^l (\alpha_i - \alpha_i^*) = 0 \\ \alpha_i, \alpha_i^* \in [0, C] \end{cases} \end{aligned} \quad (5)$$

其中, $C > 0$ 。利用最优化理论中的二次型规划方法,通过求解式(5)可以求得参数 α_i, α_i^* , 利用泛函

分析的 KKT 条件可以求得参数 b , 便可以求出拟合样本集的估计函数 $f(x)$ 的解析表达式,最后回归支持向量机的输出为:

$$f(x) = \sum_{i=1}^l (\alpha_i - \alpha_i^*) k(x_i, x) + b \quad (6)$$

可以看出对应系数 $\alpha_i - \alpha_i^* \neq 0$ 的样本 (x_i, y_i) 都是支持向量。核函数 $k(x_i, x)$ 是满足 Mercer 条件的任意对称函数,常用的核函数是径向基函数:

$$k(x, x_i) = \exp\left(-\frac{\|x - x_i\|^2}{2\sigma^2}\right) \quad (7)$$

式中: σ 为径向基核函数的宽度参数。

2 基于禁忌搜索算法的支持向量机模型(TS-SVM)

式(7)中径向基函数的宽度参数 σ 精确定义了高维特征空间非线性映射函数的结构,因而控制了最终解的复杂性,从而影响高维特征空间上的线性回归。当 σ 减小时,回归函数的复杂度增加,容易造成过拟合, σ 值过大或过小都会使系统的泛化性能变差。因此适当选择 σ 对回归模型的泛化能力很关键。式(5)中惩罚因子 C 是调节训练误差和模型复杂度之间的折中,间接地影响回归模型的泛化能力, C 值过大或过小都会因过学习或欠学习使系统的泛化性能变差^[8]。

传统的参数选取多是在实验中采用反复试验的方法,该方法需要试验者的先验知识作为指导,并且需要较高的时间代价,因此传统的参数选取方法不太适应支持向量机理论的发展。针对参数选取问题,一些学者从不同的角度进行了研究和探讨^[10-11]。Wang 等^[12]根据 Fisher 判别函数得出了在分类问题中计算 σ 的经验公式,根据 Scale space 理论给出了在回归问题中计算 σ 取值的方法,这为 σ 的优化选择提供了一种可选的途径。Carl 利用实验设计方法获得合适的参数取值^[13],可以显著地减少搜索代价。

优化算法的引入克服了 SVM 参数选择的盲目性,提高支持向量机的推广预测能力。与目前普遍应用的通过交叉验证试算确定 SVM 参数的方法相比,优化算法优选参数具有更明确的理论指导意义。因此,一些学者尝试采用现代优化算法(比如粒子群算法、遗传算法、免疫进化算法等)对 SVM 模型的参数进行优化求解^[7-9]。笔者采用禁忌搜索算法来搜索最佳的数 σ 和 C ,建立了 TS-SVM 耦合模型。禁忌搜索算法是一种启发式全局逐步寻优搜索算法,具体内容可见文献^[14-15]。建立 TS-SVM 模型的步骤如下。

a. TS 初始化。设定禁忌列表长度、待优化参数取值范围、邻域生成方式、最大循环次数等。

b. 计算适应度大小。从用于训练的样本数据中划分出一小部分(一般取全部训练样本的 10%)作为验证样本集。用剩余的训练样本训练 SVM 模型,达到拟合要求后,用获得的 SVM 模型对划分出的验证样本进行预测,采用验证样本的预测误差平方和作为适应度值。

c. 更新 TS 算法的禁忌列表、最优值。

d. 检查结束条件,若满足,则结束寻优,否则 $t = t + 1$,转至步骤 b。结束条件为寻优达到最大进化代数或适应度值小于给定精度。

e. 将 TS 算法寻优得到的参数向量 (C, σ) 赋予 SVM,用 SVM 模型对全体训练样本进行学习,并进行测试集(检验集)样本数据的预测。

3 基于 TS-SVM 模型的函数拟合分析

为了对 TS-SVM 耦合模型的计算精度进行检验,以分析模型的学习能力和学习精度,将其应用到 3 个典型的复杂函数拟合中。多峰正函数、Schaffer 函数和 Camel 函数的表达式分别如式(8)~(10)所示。

$$y_1 = e^{-0.001x} \cos^2(0.8x) \quad x \in [0, 18] \quad (8)$$

$$y_2 = 0.5 + \frac{(\sin(\sqrt{x_1^2 + x_2^2}))^2 - 0.5}{1.0 + 0.001(x_1^2 + x_2^2)}$$

$$- 10 \leq x_1, x_2 \leq 10 \quad (9)$$

$$y_3 = \left(4 - 2.1x_1^2 + \frac{x_1^4}{3}\right)x_1^2 + x_1x_2 + (-4 + 4x_2^2)x_2^2 \quad -2 \leq x_1, x_2 \leq 2 \quad (10)$$

分别对各函数的自变量区间范围内生成 1 000 个自变量数值点,采用 TS 算法优化确定 SVM 模型中的惩罚因子 C 和核函数参数 σ ,得到优化后的用于函数拟合的 TS-SVM 耦合模型。优化得到的惩罚因子 C 和核函数参数 σ 分别为:多峰正函数: $C =$

196.3242, $\sigma = 0.2036$;Schaffer 函数: $C = 115.0670$, $\sigma = 0.1310$;Camel 函数: $C = 145.1777$, $\sigma = 0.0857$ 。再在函数区间范围内生成 100 个不同的点作为函数拟合的自变量数值,采用 TS-SVM 耦合模型计算得到函数拟合值,3 个典型函数的拟合结果如图 1 所示。从图 1 可以看出,TS-SVM 耦合模型计算得到的函数拟合值与函数的理论值非常接近,拟合精度高,表明 TS-SVM 耦合模型学习能力强、计算精度高。

4 基于 TS-SVM 的山东省水安全指数预测实例

以山东省水安全状况评价为例,通过 TS-SVM 计算区域水安全指数,得到各个水安全利用方案的安全等级。水安全评价指标体系的选择参考了文献[4],将水安全评价的指标体系划分为 5 个安全评价子系统,共 23 个评价指标,如表 1 所示。笔者建立 TS-SVM 耦合模型由 TS 模块和 SVM 模块两部分耦合而成,其中 SVM 模块采用的 LS-SVMlab^[16]模式识别与回归软件,在计算机上用 Matlab 编程实现。参数设置为:禁忌列表长度 100,最大迭代次数 1 000,惩罚因子 C 的搜索范围定为[0.001, 200],RBF 核函数参数 σ 的搜索范围定为[0.001, 100]。基于 TS-SVM 的山东省水安全评价步骤如下。

a. 分别对每个评价子系统,在各安全评价等级范围内随机生成 10 组数据,共 50 组评价数据,经过规范化处理后,作为 TS-SVM 的训练输入样本,并设定相应各等级的指数目标值 I_k (如表 1 所示)作为输出样本。用 TS-SVM 方法寻找各个子系统最优参数向量 (C, σ) 取值如表 2 所示。

b. 用训练好的 SVM 模型对方案 1、方案 2 和方案 3(各方案数据如表 1 所示)进行检验,得到各方案的安全指数值 I 。

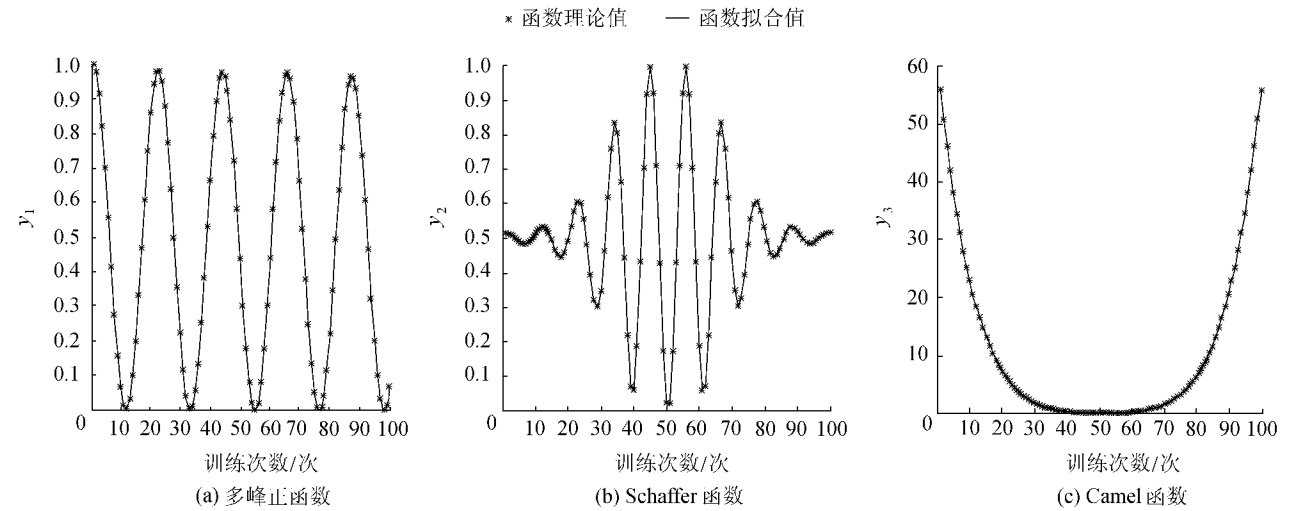


图 1 函数拟合结果

表 1 山东省水安全评价指标分级标准及可行方案

评价指标	生命安全			经济安全					生态安全				
	人均水资源量/m ³	人均生活用水量/(L·d ⁻¹)	饮用水水质达标率/%	人均GDP/元	工业万元产值取水量/m ³	工业用水重复利用率/%	人均粮食/(kg·a ⁻¹)	人均有效灌溉面积/hm ²	地表水水质达标率/%	地下水水质达标率/%	客水水质达标率/%	污径比/%	土壤盐渍化率/%
I 非常安全	≥1000	≥200	≥99	≥14852	<10	≥85	≥500	≥0.07	≥90	≥95	≥90	≤3.2	≤1
II 安全	1000~750	200~160	99~96	14852~12064	10~23	85~75	500~433	0.07~0.06	90~75	95~90	90~75	3.2~5.5	1~4
III 基本安全	750~500	160~120	96~93	12046~9276	23~37	75~65	433~366	0.66~0.04	75~60	90~85	75~60	5.5~7.7	4~7
IV 不安全	500~250	120~80	93~90	9276~6488	37~50	65~55	366~300	0.04~0.03	60~45	85~80	60~45	7.7~10	7~10
V 极不安全	<250	<80	<90	<6488	>50	<55	<300	<0.03	<45	<80	<45	>10	>10
方案1	350	80	95	11700	16	70	400	0.05	0.80	1	1	0.44	1
方案2	347	75	93	11660	15	68	390	0.05	0.75	0.9	0.96	0.3	0.6
方案3	344	71	90	11618	15	65	377	0.05	0.70	0.7	0.95	0.1	0.5

评价指标	社会安全					管理安全					目标值 <i>I</i> ₀
	干旱受灾面积率/%	洪涝受灾面积率/%	水灾害损失占GDP比例/%	河道防洪达标率/%	城市防洪达标率/%	水利投资占GDP比例/%	城市管网供水利用率/%	灌溉水利用系数	水利现代化投资比例/%	管理人员学历结构系数	
I 非常安全	≤5	≤2	≤1	≥95	≥95	≥1.0	≥95	≥0.9	≥5.5	≥0.9	0.1
II 安全	5~13.3	2~4.67	1~2.3	95~87	95~87	1.0~0.74	95~91.7	0.9~0.77	5.5~4	0.9~0.8	0.2
III 基本安全	13.3~21.6	4.67~6.34	2.3~3.6	87~79	87~79	0.74~0.47	91.7~88.4	0.77~0.64	4~2.5	0.8~0.7	0.3
IV 不安全	21.6~30	6.34~10	3.6~5	79~70	79~70	0.47~0.2	88.4~85	0.64~0.5	2.5~1	0.7~0.6	0.4
V 极不安全	>30	>10	>5	<70	<70	<0.2	<85	<0.5	<1	<0.6	0.5
方案1	49	0.053	2.48	40	70	0.25	85	0.6	2	0.8	
方案2	36	0.04	2.15	30	60	0.23	82	0.58	1.5	0.7	
方案3	20	0.02	1.8	20	50	0.22	80	0.55	1	0.6	

注（1）饮用水水质达标率表示饮用水中能达到国家制定的饮用水卫生标准的百分比（2）地表水水质达标率表示地表水中Ⅳ类水及优于Ⅳ类水水量占地表水总量的百分比（3）地下水水质达标率表示地下水中Ⅲ类水及优于Ⅲ类水水量占地下水总量的百分比（4）客水水质达标率表示客水中Ⅲ类水及优于Ⅲ类水水量占客水总量的百分比。

表 2 优化得到的支持向量机参数值及水安全评价结果

子系统	生命安全	经济安全	生态安全	社会安全	管理安全	综合水安全	本文等级	文献 4 等级	
评价结果	方案 1	0.3264	0.2798	0.2758	0.2738	0.3131	0.29378	Ⅲ(基本安全)	Ⅲ
	方案 2	0.3883	0.2834	0.2758	0.2738	0.3766	0.31958	Ⅳ(不安全)	Ⅳ
	方案 3	0.4784	0.2891	0.2758	0.2738	0.3892	0.34126	Ⅳ(不安全)	Ⅳ
惩罚因子 C	193.7017	108.1204	82.3047	116.4205	134.0174				
核函数参数 σ	1.5478	2.2883	1.8396	0.9716	2.0440				

c. 将各个子系统的安全指数值平均,得到综合水安全指数值和对应的评价等级,见表2。表2中还列出了属性识别法^[4]的评价结果,可见各方案的评价等级和文献[4]的评价结果是相同的。

5 结 论

a. 在对支持向量机参数性能进行分析的基础上,提出了TS-SVM耦合模型,用于水安全评价指标的计算。采用TS算法来优化SVM的参数,克服了模型参数选择的盲目性,与目前普遍应用的通过交叉验证试算确定参数的方法相比,用TS算法优选参

数具有更明确的理论指导。

b. 笔者仅采用径向基函数作为核函数进行研究,实际上核函数选择的不同对建立SVM模型有直接影响,因而选择其他类型核函数进行比较的研究还有待进一步深入。

c. 通过TS-SVM模型计算出的水安全指数,能衡量区域水安全水平,反映区域水安全客观状况。对山东省水安全指数计算的实例分析,表明了TS-SVM模型的实用性,为水安全评价和预测提供了一条有效的途径。

(下转第9页)

温 3~8 月降低、9 月至次年 1 月升高 ;丹江口水库建成后出库水质没有明显的变化 ,汉江中下游水质恶化和发生“ 水华 ”的主要原因是污染增加 ,次要原因是水量减少 ;丹江口水库建成后中下游泥沙量锐减导致鱼类资源减少 ,而流速降低、流量减少也使喜急流的鱼类栖息环境变化 ,导致数量减少。南水北调中线工程建成运行后必将对汉江中下游水文生态产生新的影响 ,汉江中下游江段流量将比现状减少 ,流速变缓 ,对沿岸点源、面源污染的稀释自净能力降低 ,生态环境供水量不足 ,从而导致水质恶化 ;“ 水华 ”现象有可能进一步加重 ,鱼类资源进一步减少。随着汉江流域经济的持续增长和人口的增加 ,从长远来看 ,只有引江济汉工程和汉江流域水污染控制工程的建成才能从根本上提高汉江的水环境容量 ,改善水质 ,减少“ 水华 ”的发生概率 ,满足工农业发展对水资源的需求。

参考文献：

[1] 谢平 ,夏军 ,窦明 ,等 .南水北调中线工程对汉江中下游水华的影响及对策研究[J].自然资源学报 ,2004 ,19(4):418-422.

(上接第 4 页)

参考文献：

[1] 阮本清 ,魏传江 .首都圈水资源安全保障体系建设[M].北京 :科学出版社 ,2004 :50-52.

[2] 樊彦芳 ,刘凌 ,陈星 ,等 .层次分析法在水环境安全综合评价中的应用[J].河海大学学报 :自然科学版 ,2004 ,32(5) :512-514.

[3] 何焰 ,由文辉 ,吴健 .上海市水环境生态安全评价[J].水资源保护 ,2006 ,22(6) :18-20 ,27.

[4] 张龙云 ,曹升乐 ,杨尚阳 .属性识别理论在水安全评价中的应用研究[J].山东大学学报 :工学版 ,2006 ,36(5) :70-72.

[5] WANG Wen-sheng ,JIN Ju-liang ,DING Jing ,et al . A new approach to water resources system assessment 'set pair analysis method'[J]. Science in China Series E :Technological Sciences ,2009 ,32(10) :3004-3016.

[6] MUKHERJEE S ,OSUMA E ,GIROSI F .Nonlinear prediction of chaotic time series using support vector machines[C]/Proc of the IEEE Workshop on Neural Networks for Signal Processing .Ameliz Island ,1997 :511-520.

[7] XU P ,CHAN A K .An efficient algorithm on multi-class support vector machine model selection[C]/Proceedings of the International Joint Conference on Neural Networks 2003 .Portland ,2003 :3229-3232.

[8] 熊建秋 ,邹长武 ,李祚泳 .基于免疫进化支持向量机的年

[2] 窦明 ,谢平 ,夏军 ,等 .汉江水华问题研究[J].水科学进展 ,2002 ,13(15) :557-561.

[3] 徐德龙 ,沙志贵 .南水北调对汉江中下游水文特性的影响[C]/湖北科技论坛 .武汉 :湖北省水利学会 ,2001 :37-39.

[4] 石自堂 ,冯红春 ,陈晓群 ,等 .南水北调中线工程对汉江中下游的影响及其对策[C]/湖北科技论坛 .武汉 :湖北省水利学会 ,2001 :5-7.

[5] 李晓路 ,胡振鹏 ,张文捷 .大坝下游河道水温变化规律及其影响[J].江西水利科技 ,1995 ,21(3) :167-173.

[6] 刘大银 ,周晓刚 ,汪翰 ,等 .汉江中游水污染规律与控制方案研究[J].长江流域资源与环境 ,2001 ,10(4) :365-371.

[7] 邬红娟 ,彭建华 ,韩德举 ,等 .丹江口水库浮游植物及其演变[J].湖泊科学 ,2005 ,17(4) :366-372.

[8] 徐新伟 ,吴中华 ,于丹 ,等 .汉江中下游水生植物多样性及南水北调工程及其影响[J].生态学报 ,2002 ,22(11) :1933-1938.

[9] 李修峰 ,黄道明 ,谢文星 ,等 .汉江中游鱼类资源现状[J].湖泊科学 ,2005 ,17(4) :366-372.

[10] 方芳 ,陈国湖 .调水对汉江中下游水质和水环境容量影响研究[J].环境科学与技术 ,2003 ,26(1) :14-15.

(收稿日期 :2008-11-26 编辑 :徐 娟)

用电量预测[J].四川大学学报 :工程科学版 ,2006 ,38(2) :6-10.

[9] 张红梅 ,卫志农 ,龚灯才 ,等 .基于粒子群支持向量机的短期电力负荷预测[J].继电器 ,2006 ,34(3) :28-31.

[10] CHERKASSKY V ,MA Yun-qian .Practical selection of SVM parameters and noise estimation for SVM regression[J]. Neural Networks ,2004 ,17(1) :113-126.

[11] 董春曦 ,饶鲜 ,杨绍全 ,等 .支持向量机参数选择方法研究[J].系统工程与电子技术 ,2004 ,28(8) :1117-1120.

[12] WANG Wen-jian ,XU Zong-ben ,LU Wei-zhen ,et al . Determination of the spread parameter in Gaussian kernel for classification and regression[J]. Neurocomputer ,2003 ,55 :643-663.

[13] STAELIN C .Parameter selection for support vector machines [DB/OL]. [2003 - 10 - 10] <http://www.hpl.hp.com/techreports/2002/HPL-2002-354R1.html>.

[14] ASIC M D , KOVACEVIE-VUJCIC V V . Taboo search methodology in global optimization[J]. Computers and Mathematics with Application ,1999 ,33(3) :125-133.

[15] 汪嘉杨 ,李祚泳 ,倪长健 ,等 .基于混合禁忌搜索算法的水位流量关系拟合[J].系统工程 ,2006 ,24(6) :107-110.

[16] PELCHMANS K ,SUYKENS J A .GESTEL T V ,et al .LS-SVM lab :a MATLAB/C toolbox for least squares support vector machines[EB/OL].[2002]. <http://www.esat.kuleuven.ac.be/sista/lssvmlab/>.

(收稿日期 :2008-11-04 编辑 :徐 娟)