

DOI: 10.16781/j.0258-879x.2019.05.0492

· 专题报道 ·

最小二乘支持向量机算法在中医临床脉图参数-血压预测模型的应用

杨晶东¹, 孙磊明^{1*}, 燕海霞²

1. 上海理工大学光电信息与计算机工程学院自主机器人实验室, 上海 200093

2. 上海中医药大学基础医学院中医诊断教研室, 上海 201203

[摘要] **目的** 提出基于最小二乘支持向量机(LSSVM)算法的学习模型, 以提高中医临床血压数据预测的准确度和效率。**方法** 将LSSVM学习模型应用于中医临床血压数据预测。用LSSVM等式约束代替支持向量机不等式约束, 将二次规划问题转化为线性方程求解问题, 降低计算复杂性, 加快算法收敛速度。收集320例患者的临床脉图参数及血压数据, 以其中300例样本作为训练样本, 训练得到LSSVM学习模型, 以其余20例样本作为测试数据, 用得到的LSSVM学习模型根据患者的脉图参数预测血压数据。**结果** 实验证明, LSSVM学习模型对血压数据有较好的预测准确度。其中基于多项式核函数的LSSVM学习模型较基于径向基核函数LSSVM学习模型表现出更好的学习和预测能力, 基于多项式核函数的LSSVM学习模型中收缩压、舒张压、平均动脉压预测结果的平均预测误差分别为7.88%、8.40%、6.67%, 低于基于径向基核函数的LSSVM学习模型的预测误差(分别为7.95%、9.70%、7.48%)。**结论** 本实验提出的基于LSSVM的学习模型仅通过患者的临床脉图参数就可预测患者血压数据, 对中医临床诊断有一定的参考价值。

[关键词] 中国传统医学; 脉图参数; 血压; 支持向量机; 核函数; 最小二乘法分析**[中图分类号]** R 241.1 **[文献标志码]** A **[文章编号]** 0258-879X(2019)05-0492-05

Application of least square support vector machine algorithm in clinical pulse diagram parameter-blood pressure prediction model of traditional Chinese medicine

YANG Jing-dong¹, SUN Lei-ming^{1*}, YAN Hai-xia²

1. Autonomous Robot Lab, School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China

2. Department of Traditional Chinese Medicine Diagnosis, Basic Medical College, Shanghai University of Traditional Chinese Medicine, Shanghai 201203, China

[Abstract] **Objective** To propose a learning model based on least square support vector machine (LSSVM) algorithm to improve the accuracy and efficiency for predicting clinical blood pressure data of traditional Chinese medicine (TCM). **Methods** The LSSVM learning model was used to predict the clinical blood pressure of TCM. By replacing the inequality constraints of support vector machine with LSSVM equality constraints, the quadratic programming problem was transformed into a linear equation solution problem to reduce computational complexity and speed up algorithm convergence. The clinical pulse diagram parameters and blood pressure data of 320 patients were collected. Three hundred of them were used as training samples, the remaining 20 samples were used as test data. The LSSVM learning model was used to predict blood pressure data according to the pulse diagram parameters of the patients. **Results** Experimental results showed that the LSSVM learning model had high prediction accuracy for blood pressure data. The LSSVM learning model based on polynomial kernel function had better learning and prediction abilities than the LSSVM learning model based on radial basis kernel function. The mean prediction errors of systolic blood pressure, diastolic blood pressure and mean arterial pressure obtained by the LSSVM learning model based on polynomial kernel function were 7.88%, 8.40% and 6.67%, respectively, which were lower than those obtained by the LSSVM learning model based on radial basis kernel function (7.95%, 9.70% and 7.48%, respectively). **Conclusion** The LSSVM learning model proposed in this experiment can be used to predict the blood pressure data of patients only by the clinical pulse diagram parameters, and is a good reference for clinical diagnosis of TCM.

[Key words] traditional Chinese medicine; pulse diagram parameters; blood pressure; support vector machine; kernel function; least-squares analysis

[Acad J Sec Mil Med Univ, 2019, 40(5): 492-496]

[收稿日期] 2018-10-25 **[接受日期]** 2018-11-16**[基金项目]** 国家自然科学基金(61374039), 上海市自然科学基金(15ZR1429100), 沪江基金(C14002). Supported by National Natural Science Foundation of China (61374039), Natural Science Foundation of Shanghai (15ZR1429100), and Hujiang Foundation (C14002).**[作者简介]** 杨晶东, 博士, 副教授, 硕士生导师. E-mail: eerfriend@yeah.net

*通信作者(Corresponding author). Tel: 021-55271064, E-mail: s842732406@gmail.com

将人工智能技术应用于中医临床诊断领域, 将可能有效提高中医临床脉图参数变化及血压的预测精度。目前常用的预测方法有过程神经网络^[1]、支持向量机 (support vector machine, SVM)^[2]等。朱群雄和甄玉山^[3]采用过程神经网络方法, 利用主成分分析表征参数间相关性, 将拟合过程作为神经网络输入, 建立软测量模型实现主变量的连续预测, 但该方法易陷入局部极小点并出现过拟合现象。SVM 方法基于结构风险最小化准则, 其拓扑结构由支持向量决定, 克服了人工神经网络 (artificial neural network, ANN) 过度依赖设计者经验的缺点, 兼顾了神经网络和灰度模型的优点^[4-5]。因此, 国内有些学者提出了基于粒子群优化 (particle swarm optimization, PSO) 算法的 SVM 预测模型。王保义等^[6]通过引入 Spark on YARN 内存计算平台并采用基于改进并行粒子群优化 (improved parallel particle swarm optimization, IPPSO) 算法, 对不确定参数进行优化。陈进东和潘丰^[7]通过改进粒子群中的学习因子对数据样本进行短期预测, 发现较标准 SVM 算法有较好的预测效果, 但该方法精度不足, 实时性较差, 且仅限于小样本数据分类测试。

最小二乘支持向量机 (least square support vector machine, LSSVM) 是标准 SVM 的一种扩展, 用等式约束代替 SVM 的不等式约束, 即将二次规划问题转化为线性方程求解问题, 降低计算的复杂性, 加快求解速度和提高抗干扰能力^[8-11]。本研究将 SVM 理论引入中医临床脉图参数及血压预测过程, 提出一种基于 LSSVM 学习模型, 分别采用多项式核函数和径向基核函数对患者临床脉图参数进行学习和训练, 并对患者血压进行分析与预测。

1 LSSVM 算法分析

命题: 如果用 SVM 的不等式约束代替等式约束, 将误差平方和损失函数作为训练集的经验损失, 那么可得到 LSSVM 函数估计为 $f(x) = \sum_{i=1}^m \alpha_i k(x, x^{(i)}) + b$ 。

证明: SVM 是通过一个非线性映射 $\varphi(x^{(i)})$ 将样本输入数据映射到一个高维特征空间, 从而将实际问题转化为一个带不等式约束的二次规划问题^[12-13]。假设一组训练样本为 $(x^{(i)}, y^{(i)})$, $i = 1, 2, \dots, m$, 其中 $x \in R^n$ 为系统样本输入, $y \in R$ 为系统输出, i 为样本序号。则 SVM 优化问题可形式化为

$$\begin{aligned} & \max_{\gamma, w, b} \\ & s.t. y^{(i)}(w^T x^{(i)} + b) \geq \gamma, i = 1, 2, \dots, m \\ & \|w\| = 1 \end{aligned} \quad (1)$$

式中 w 为超平面法向量, γ 为几何间隔, b 为截距, i 为样本序号。将目标函数优化修改为二次规划问题, 即目标函数改写成 $\min \frac{1}{2} \|w\|^2$ 。同时, 考虑到实际算法中可出现线性不可分情况而引入松弛变量 ζ_i (其中 i 为样本序号), 允许样本存在逼近误差^[14]。采用非线性映射后特征 $\varphi(x^{(i)})$ 代替原输入特征 $x^{(i)}$, 实现将输入空间中的非线性回归转化为高维特征空间中的线性回归, 因此得到 SVM 新的软间隔数学模型为

$$\begin{aligned} & \min_{\gamma, w, b} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^m \zeta_i \\ & s.t. y^{(i)}(w^T \varphi(x^{(i)}) + b) \geq 1 - \zeta_i, i = 1, 2, \dots, m \\ & \zeta_i \geq 0, i = 1, 2, \dots, m \end{aligned} \quad (2)$$

式中 w 为超平面法向量, γ 为几何间隔, b 为截距, ζ_i 为松弛变量, i 为样本序号。用 SVM 不等式约束代替等式约束, LSSVM 优化目标可表示为

$$\begin{aligned} & \min \frac{1}{2} \|w\|^2 + \frac{1}{2} C \sum_{i=1}^m \zeta_i^2 \\ & s.t. w^T \varphi(x^{(i)}) + b + \zeta_i = y^{(i)}, i = 1, 2, \dots, m \end{aligned} \quad (3)$$

式中 w 为超平面法向量, C 为误差惩罚系数, ζ_i 为松弛变量, i 为样本序号。构造拉格朗日函数为

$$\begin{aligned} L(w, b, \zeta, \alpha) = & \frac{1}{2} \|w\|^2 + \frac{1}{2} C \sum_{i=1}^m \zeta_i^2 \\ & - \sum_{i=1}^m \alpha_i (w^T \varphi(x^{(i)}) + b + \zeta_i - y^{(i)}) \end{aligned} \quad (4)$$

式中 α_i 为拉格朗日算子, w 为超平面法向量, γ 为几何间隔, b 为截距, ζ_i 为松弛变量, i 为样本序号。根据 KKT 条件得到下式

$$\begin{aligned} \frac{\partial}{\partial w} L(w, b, \zeta, \alpha) = & w - \sum_{i=1}^m \alpha_i \varphi(x^{(i)}) = 0 \\ \frac{\partial}{\partial b} L(w, b, \zeta, \alpha) = & - \sum_{i=1}^m \alpha_i = 0 \\ \frac{\partial}{\partial \zeta_i} L(w, b, \zeta, \alpha) = & C \zeta_i - \alpha_i = 0 \\ \frac{\partial}{\partial \alpha_i} L(w, b, \zeta, \alpha) = & -(w^T \varphi(x^{(i)}) + b + \zeta_i - y^{(i)}) = 0 \end{aligned}$$

消去 w 和 ζ_i 得到:

$$\begin{bmatrix} 0 & P^T \\ P & K + C^{-1}I \end{bmatrix} \begin{bmatrix} b \\ A \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix} \quad (5)$$

其中 $P = [1, 1, \dots, 1]^T$, $A = [\alpha_1, \alpha_2, \dots, \alpha_m]^T$, $y = [y_1, y_2, \dots, y_m]^T$, K 是一个核函数矩阵, $K_{ij} = \varphi(x^{(i)})^T \varphi(x^{(j)}) = k(x^{(i)}, x^{(j)})$, 得到 LSSVM 函数估计为

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

核函数 k 可选用径向基核函数

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

径向基核函数中的核参数 g 、惩罚参数 C 可选用梯度下降算法进行寻优, 核参数 g 和惩罚参数 C 更新公式为

$$g'_k = g_k - \alpha \frac{\partial f}{\partial g_k} \quad (8)$$

$$C'_k = C_k - \alpha \frac{\partial f}{\partial C_k} \quad (9)$$

式中 α 为学习率。核函数 k 也可选用多项式核函数

$$k(x, x^{(i)}) = ((x \cdot x^{(i)}) + 1)^d \quad (10)$$

LSSVM 模型采用交叉验证 Hold-Out 算法分割测试样本和训练样本, 该方法可应用于中医脉诊中的血压预测。

2 实验步骤

通过 LSSVM 学习模型对患者进行血压预测。脉图参数 (输入) 主要包括主波幅度 (h_1)、重波前波幅度 (h_3)、反映血管壁顺应性和动脉血管外周阻力状态参数 h_3/h_1 、反映动脉血管外周阻力状态参数 h_4/h_1 (h_4 为降中峡幅度)、主波上 1/3 处宽度 (w)、反映动脉管壁弹性和外周阻力大小参数 w/t (t 表示脉图起点到终止点的时值)、反应心脏射血功能参数 t_1/t (t_1 为脉图起点到主波峰点的时值) 等^[15]。血压参数 (输出) 包括舒张压 (diastolic blood pressure, DBP)、收缩压 (systolic blood pressure, SBP)、平均动脉压 (mean arterial pressure, MAP)。

定义 1 例患者的输入参数 (脉图) 和输出参数 (血压) 为 1 组样本。本研究收集 320 组患者的临床脉图参数和血压数据, 其中 300 组样本为训练样本, 20 组为测试样本。通过多项式核函数 (公式 10) 与径向基核函数 (公式 7) 分别训练得到 LSSVM 学习模型, 将样本的脉图参数作为模型输入的特征得到预测血压输出, 并将 20 组测试样本的 LSSVM 学习模型预测结果与临床测得患者实际血压进行比较。

LSSVM 预测过程包括: (1) 数据预处理。根据需要预测的样本数据特征选用所有样本中 1~300 组样本作为 LSSVM 训练样本, 选取 301~320 组样本作为测试样本。并将训练样本进行归一化预处理, 归一化方法选用 mapminmax 函数, 样本数据归一化到区间 [0,1]。(2) LSSVM

参数初始化。核函数选用多项式核函数 (公式 10) 与径向基核函数 (公式 7) 进行比较, 其中核参数 g 和惩罚系数 C 均初始化为 0.1。(3) 建立学习模型。根据 LSSVM 的算法编写程序输入训练样本 ($x^{(i)}, y^{(i)}$), 通过批量梯度下降迭代得到最优核参数 g 和惩罚参数 C , 根据所得参数建立 LSSVM 学习模型。随后输入预测样本的输入向量 $x^{(i)}$, 通过 LSSVM 学习模型预测出输出向量 $y^{(i)}$, 并与测试样本实际输出向量进行对比得到误差。

3 实验结果和分析

收集 320 组患者的临床脉图参数及血压数据, 对 300 组训练样本进行训练得到 LSSVM 学习模型, 对其余 20 组患者临床脉图参数及血压数据进行测试。对 LSSVM 学习模型使用批量梯度下降算法 (batch gradient descent) 训练最优参数, 每迭代一步, 均要用到训练集所有的数据, 如果训练样本的损失函数为凸函数就能保证收敛到最优值。由于本实验中训练样本规模小, 因此批量梯度算法使训练过程既能保证足够的精确度, 又有较好的实时性。图 1 所示离散点分别为 DBP (蓝色)、SBP (红色)、MAP (黑色) 实际数据及其拟合曲线, 图中可见 LSSVM 学习模型拟合曲线能够反映数据样本分布和特点。

针对训练所得的 LSSVM 学习模型, 对剩余 20 组样本进行测试。图 2A、2B 分别给出基于径向基核函数的 LSSVM 学习模型与基于多项式核函数的 LSSVM 学习模型输出对比曲线。图 2C、2D 分别给出基于径向基核函数的 LSSVM 学习模型与基于多项式核函数的 LSSVM 学习模型误差曲线。

从图 2A、2B 可见, 多数样本点预测结果与实际数据差距较小, 其中 1~10 号测试样本预测结果与实际结果更为接近。图 2C、2D 结果显示, 第 3、4、5、6 号样本的预测准确度较高。在本研究中采用基于多项式核函数的 LSSVM 学习模型中 SBP、DBP、MAP 预测结果的平均预测误差分别为 7.88%、8.40%、6.67%, 基于径向基核函数的 LSSVM 学习模型的预测误差分别为 7.95%、9.70%、7.48%, 与基于多项式核函数 LSSVM 学习模型相比, 平均预测误差分别增加了 0.89%、15.48%、12.14%。从训练时间分析, 2 种核函数的 LSSVM 学习模型训练时间均约为 0.37 s, 具有较好的实时性。

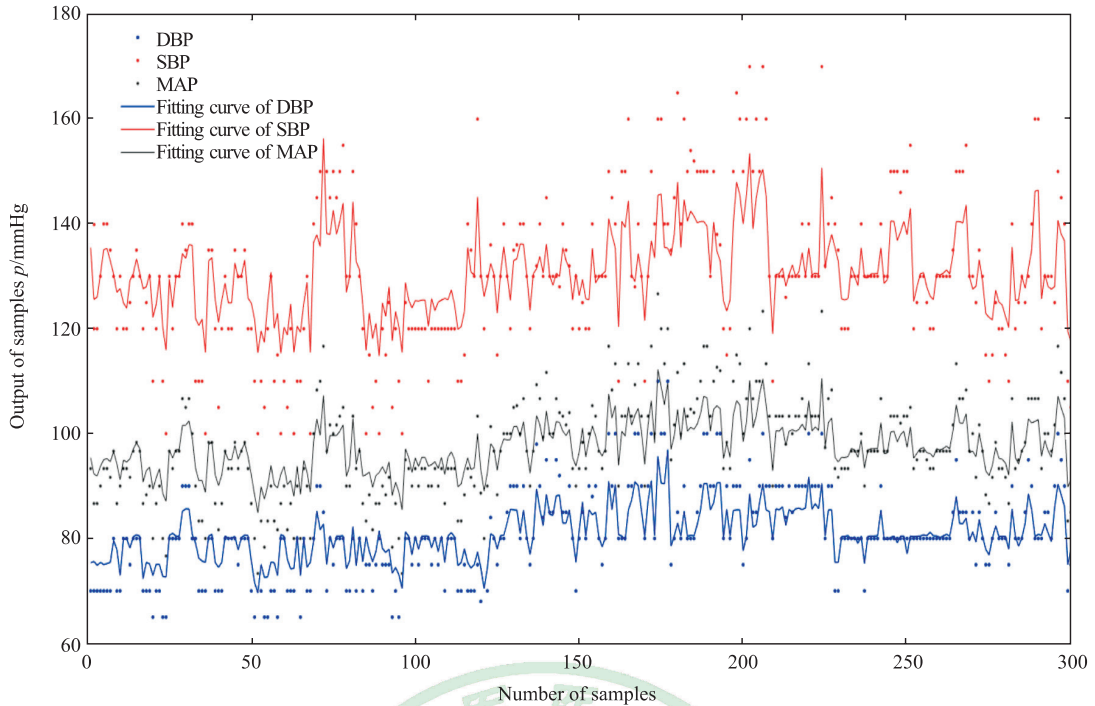


图 1 训练样本输出拟合曲线

Fig 1 Output fitting curves of training samples

1 mmHg=0.133 kPa. DBP: Diastolic blood pressure; SBP: Systolic blood pressure; MAP: Mean arterial pressure

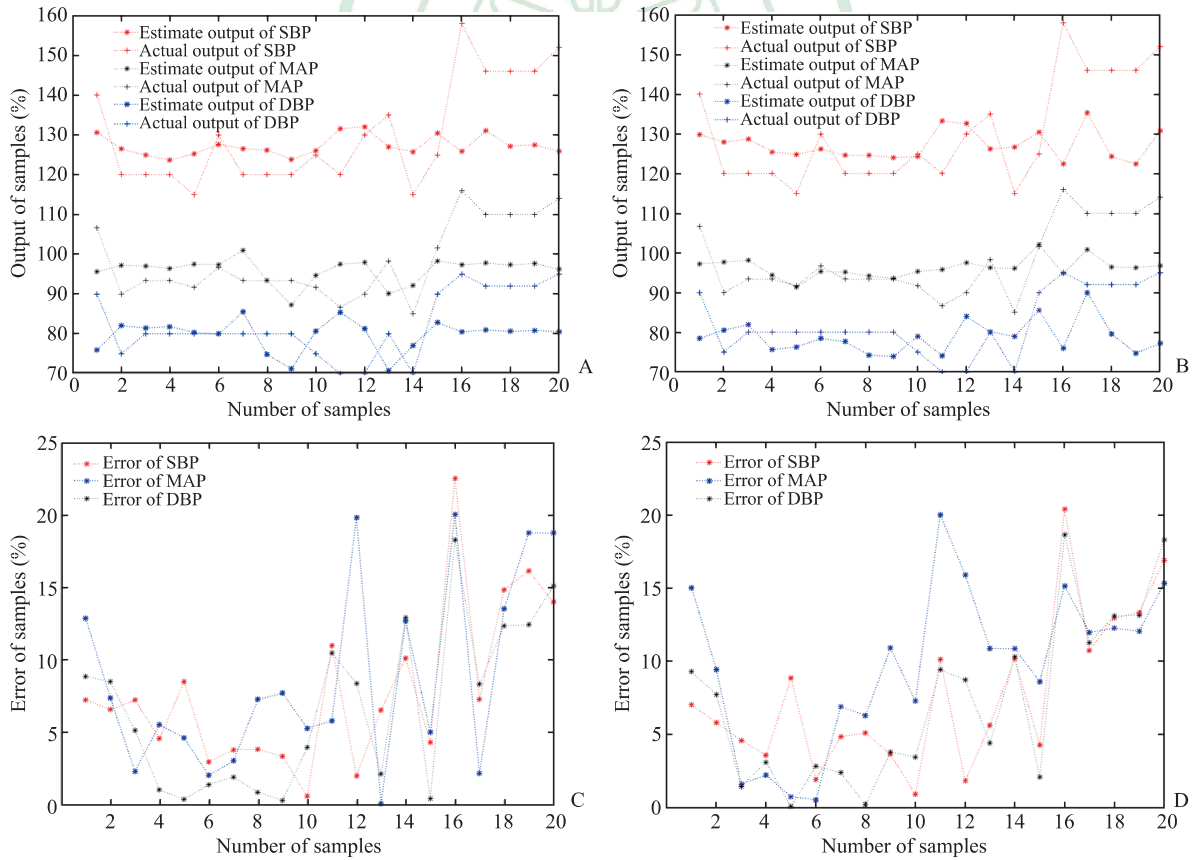


图 2 基于径向基核函数和多项式核函数 LSSVM 学习模型输出与误差曲线

Fig 2 Output and error curves of LSSVM learning model based on radial basis kernel function and polynomial kernel function

A: Output contrast of radial basis function kernel model; B: Output contrast of polynomial kernel model; C: Error curve of radial basis function kernel model; D: Error curve of polynomial kernel model. LSSVM: Least square support vector machine; DBP: Diastolic blood pressure; SBP: Systolic blood pressure; MAP: Mean arterial pressure

分析 LSSVM 学习模型预测的误差数据,基于多项式核函数的学习模型测试的样本具有较好的预测效果,这是由于梯度下降算法容易陷入局部最优,基于径向基核函数的 LSSVM 学习模型寻优核参数 g 和惩罚参数 C 时不能在有限迭代次数内获得全局最优参数,而多项式核函数属于全局核函数,对相距较远的样本点具有更好的预测效果,因此基于径向基核函数的 LSSVM 学习模型的预测误差较高。

综合对比,基于多项式核函数的 LSSVM 学习模型表现出对中医脉图参数-血压数据样本更好的学习和预测能力。多项式核函数适合于正交归一化的样本数据,属于全局核函数,相距较远的数据点对核函数值也会产生影响,本研究选用合适的指数参数 d 可有效减小高维度样本数据导致的较大计算量。对于本实验中的数据样本,基于多项式核函数 LSSVM 模型具有比径向基核函数模型更好的学习能力。由图 2C、2D 可见,基于多项式核函数 LSSVM 学习模型的预测误差均较基于径向基核函数学习模型低,但 2 种核函数 LSSVM 学习模型误差均整体呈升高趋势,这是因为本实验中样本处理选用标准交叉验证 Hold-Out 算法获得训练样本和测试样本,该算法易造成数据浪费,使测试阶段辨识度降低,部分数据出现较高误差。

4 结论

本研究提出一种基于 LSSVM 的学习模型,通过对临床脉图参数进行学习和训练,从而预测样本血压数据,并比较基于径向基和多项式核函数的 LSSVM 学习模型,发现基于多项式核函数的 LSSVM 学习模型较基于径向基核函数 LSSVM 学习模型表现出更好的学习和预测能力。今后我们将进一步改进并优化 LSSVM 学习模型及其参数。在临床应用中,我们提出的基于 LSSVM 学习模型仅通过患者的临床脉图参数就可预测患者血压数据,有助于减轻医师的工作量和患者负担,此外其也可

推广应用于其他临床数据的预测。

[参考文献]

- [1] 彭显刚,胡松峰,吕大勇. 基于 RBF 神经网络的短期负荷预测方法综述[J]. 电力系统保护与控制,2011,39:144-148.
- [2] 丁世飞,齐丙娟,谭红艳. 支持向量机理论与算法研究综述[J]. 电子科技大学学报,2011,40:2-10.
- [3] 朱群雄,甄玉山. 用于化工软测量的基于移动窗的过程神经网络[J]. 清华大学学报(自然科学版),2012,52:1165-1170.
- [4] SAIN S R, VAPNIK V N. Book review: the nature of statistical learning theory[J]. Technometrics, 1996, 38: 409.
- [5] VAPNIK V N. An overview of statistical learning theory[J]. IEEE Trans Neural Netw, 1999, 10: 988-999.
- [6] 王保义,王冬阳,张少敏. 基于 Spark 和 IPPSO_LSSVM 的短期分布式电力负荷预测算法[J]. 电力自动化设备,2016,36:117-122.
- [7] 陈进东,潘丰. 基于在线支持向量回归的非线性模型预测控制方法[J]. 控制与决策,2014,29:460-464.
- [8] SUYKENS J A K, VANDEWALLE J. Least squares support vector machine classifiers[J]. Neural Process Lett, 1999, 9: 293-300.
- [9] WANG H Q, SUN F C, CAI Y N, DING L G, CHEN N. An unbiased LSSVM model for classification and regression[J]. Soft Computing, 2010, 14: 171-180.
- [10] DE KRUIF B J, DE VRIES T A. Pruning error minimization in least squares support vector machines[J]. IEEE Trans Neural Netw, 2003, 14: 696-702.
- [11] 林升梁,刘志. 基于 RBF 核函数的支持向量机参数选择[J]. 浙江工业大学学报,2007,35:163-167.
- [12] 胡瑾秋,郭放,张来斌. 结合改进 PSO 算法和 LSSVM 的化工异常工况超早期监测预警研究[J]. 电子测量与仪器学报,2018,32:36-41.
- [13] LI L, SU H, CHU J. Modeling of isomerization of C8 aromatics by online least squares support vector machine[J]. Chin J Chem Eng, 2009: 17: 437-444.
- [14] 郑蓉建,潘丰. 基于 PLS-LSSVM 的谷氨酸发酵产物浓度预测建模[J]. 化工学报,2017,68:976-983.
- [15] 朱丽萍,吴宏进,张志枫,张瑜,许家佗,屠立平,等. 54 例不同性别健康大学生“寸口”脉象及脉图参数分析[J]. 上海中医药大学学报,2013,27:28-31.

[本文编辑] 杨亚红