

基于矢量基学习的最小二乘支持向量机建模

陈爱军^{1,2}, 宋执环¹, 李平¹

(1. 浙江大学 工业控制技术研究所, 浙江 杭州 310027; 2. 中国石油兰州石化公司 自动化研究院, 甘肃 兰州 730060)

摘要: 为使最小二乘支持向量机的解具有稀疏性, 本文提出了一种稀疏解算法—矢量基学习. 首先引入基矢量、基矢量集与矢量空间的概念, 并分析新样本矢量与矢量空间的夹角, 从而推导出该样本是否为基矢量的判断准则. 随着新样本的到来, 在线判别支持向量, 使LS-SVM的支持向量具有稀疏性. 提升LS-SVM动态建模的实时性, 本文进一步提出用于矢量基学习的增长记忆模式递推公式. 仿真分析及水处理厂的应用实例, 验证了该方法的可行性和有效性.

关键词: 最小二乘支持向量机; 矢量基; 稀疏性; 增长记忆模式; 支持向量

中图分类号: TP273.1

文献标识码: A

Modeling method of least squares support vector regression based on vector base learning

CHEN Ai-jun^{1,2}, SONG Zhi-huan¹, LI Ping¹

(1. Institute of Industry Control Technology, Zhejiang University, Hangzhou Zhejiang 310027, China;

2. Automation Institute of Lanzhou Petrochemical Corporation, Lanzhou Gansu 730060, China)

Abstract: To achieve a sparse solution for least squares support vector regression (LS-SVM), an algorithm called vector base learning (VBL) is proposed in this paper. Firstly, the concepts of base vector (BV), base vector set (BVS) and vector space are introduced. By calculating the angle between the new sample vector and the vector space, the criteria for determining whether the measurement vector is one of the BVS is then derived. This determination is carried out on-line for the coming new samples. This makes the solutions of LS-SVM having the feature of sparsity. To improve the modeling speed of LS-SVM, a recursive algorithm of increased memory mode for VBL algorithm is also proposed. Finally, simulation analysis and the modeling of a typical plant for water treatment clearly illustrated the validity and feasibility of the presented method.

Key words: least square support vector machine; vector base; sparsity; increased memory mode; support vector

1 引言(Introduction)

上世纪90年代, Vapnik V等人提出基于结构风险最小化的支持向量机(SVM)方法^[1], 与其他学习算法相比, 支持向量机具有非线性拟合、结构简单、全局最优和较好的泛化能力等优点, 在许多领域得到了应用. 然而对于支持向量机, 每增加一个新的样本, 需进行一次二次优化求解, 运算过于复杂则不具有实时性. 为提高求解速度, 研究者们提出了“分块法”、“分解法”以及最小二乘支持向量机(LS-SVM)^[2].

最小二乘支持向量回归估计将二次规划问题转变成线性方程组的求解, 简化计算的复杂性, 在函数估计、逼近和系统建模中得到了广泛应用. 由于每个样本对估计器都有贡献, LS-SVM失去了支持向量解的稀疏性优点, 样本量极大, 则模型的维

数也极大. 冗余信息的噪声被全部拟合到模型参数里, 削弱了模型的鲁棒性, 使辨识参数的推广能力下降. 因此, 从大量的训练样本中筛选出有意义的信息, 使LS-SVM的支持向量具有稀疏性, Suyken等人^[3~5]进行了算法研究. 文献[3]根据模型系数的最小贡献准则, 迭代缩减样本量, 获得稀疏性解; 文献[4]的方法类似于交叉检验; 文献[5]的评判准则为最小化模型的预报误差. 针对如何筛选支持向量, 上述文献提出了不同的方法, 但是运算量都很大, 不适用于实时性要求较高的应用场合.

受文献[7]启发, 本文在基矢量集(base vector sets, BVS)张成的矢量空间基础上, 通过判别新样本是否可由这个矢量空间近似表示, 提出一种新的最小二乘支持向量机稀疏解算法—矢量基学习算法(vector base learning, VBL). 该方法具有剔除样本冗余信

息、提升LS-SVM建模鲁棒性、运算量小的优点. 避免加入新样本需重新训练所有样本的问题, 本文进一步提出一种用于VBL动态建模的递推算法.

2 最小二乘支持向量回归估计(Least squares support vector regression)

对于一组输入样本 $\{x_i, y_i\}, i = 1, \dots, l, x_i \in \mathbb{R}^d$ 且 $y_i \in \mathbb{R}$, LS-SVM利用非线性映射, $\psi(\cdot) : X \rightarrow F$, 将训练数据集非线性地映射到一个高维特征空间, 从而使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, 设回归函数 $y = \langle w \cdot \psi(x) \rangle + b$.

在最小二乘支持向量回归估计中, 回归问题变为

$$\begin{aligned} \min \quad & \frac{1}{2} \|w\|^2 + \frac{1}{2} \gamma \sum_{i=1}^l e_i^2, \\ \text{s.t.} \quad & w^T \psi(x_i) + b + e_i = y_i, i = 1, \dots, l. \end{aligned} \quad (1)$$

其中: $e \in \mathbb{R}^{l \times 1}$ 为误差向量, γ 为正则化参数. 引入Lagrange乘子 $\lambda, \lambda \in \mathbb{R}^{l \times 1}$, 式(1)可化为

$$\begin{aligned} \min \quad J_{\text{LS-SVR}} = & \frac{1}{2} \|w\|^2 + \frac{1}{2} \gamma \sum_{i=1}^l e_i^2 - \\ & \sum_{i=1}^l \lambda_i (w^T \psi(x_i) + b + e_i - y_i). \end{aligned} \quad (2)$$

由KKT最优条件, 式(2)的解为

$$\begin{bmatrix} b \\ \vec{\lambda} \end{bmatrix} = \begin{bmatrix} 0 & \vec{1}^T \\ \vec{1} & Z Z^T + \frac{I}{\gamma} \end{bmatrix}^{-1} \times \begin{bmatrix} 0 \\ Y \end{bmatrix}. \quad (3)$$

其中: $Z = [\psi(x_1), \dots, \psi(x_l)]^T, \vec{\lambda} = [\lambda_1, \dots, \lambda_l]^T, Y = [y_1, \dots, y_l]^T, \vec{1}$ 为 $l \times 1$ 单位列向量. 非线性函数拟合的 $Z Z^T$ 内积运算, 可用满足Mercer条件的核函数 $K(x_i, y_j)$ 替代. 常见的核函数有线性函数、多项式函数、径向基函数、多层感知器函数.

由于 $w = \sum_{i=1}^l \lambda_i \psi(x_i)$, 因此非线性回归估计函数的表达式为

$$y = \sum_{i=1}^l \lambda_i k(x_i, x) + b. \quad (4)$$

3 LS-SVM 向量基学习算法(Vector-based learning of LS-SVM)

3.1 向量基的选择判据(Criteria for selecting base vector)

定义 1 设存在一组矢量 $v_i, i = 1, \dots, l, l \leq n$, 其高维空间的映射为 $\psi(v_i)$, 这些矢量的线性组合为

$$w = \delta(\psi^T(v_1), \dots, \psi^T(v_l)). \quad (5)$$

如果 w 线性无关, 则称矢量 v_i 为基矢量(base vector, BV).

定义 2 由基矢量 $v_i (i = 1, \dots, l)$ 组成的集合, 称为基矢量集(base vector set, BVS).

定义 3 在最小二乘支持向量建模中, 由支持向量基矢量集张成的子空间称为该模型的矢量空间.

现在考虑如何确定BVS的 l 个BV. 假设由支持向量构建的BVS已经有两个BV: $\psi(v_1), \psi(v_2)$, 分别对应图1的 a, b 两点. 这两个矢量张成一个矢量空间, 该空间可以表示为 $\Gamma = (\lambda_1, \lambda_2)(\psi(v_1), \psi(v_2))^T$, 假设另外两个新矢量(分别为图1的 c 和 d), 与 Γ 的夹角分别为 θ_1, θ_2 , 如果角度很小如 θ_1 , 则认为矢量 c 可由 a, b 近似线性表示, 而 d 的 θ_2 角太大, 表明用 a, b 近似线性表示 d 的误差太大, 因而将 d 引入到BVS中. 通过这种方式, 就确定了BVS以及矢量空间, 从而在这个矢量空间里, 最终得到期望的模型参数 λ . 因此, BVS的数量决定了LS-SVM模型系数 λ 的维数, 即模型的阶数. 通过新样本矢量与BVS张成的矢量空间的角度 θ 来确定其成为BV的条件, 为此需要引入下面的定理.

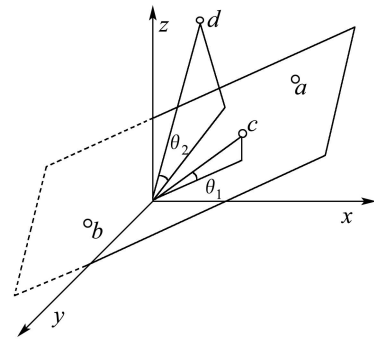


图1 平面矢量空间确定BV的示意图

Fig. 1 Illustration of determining BV by the plane vector space

定理 1 已知最小二乘支持向量回归模型由一组基矢量样本 $(v_1, y_1), \dots, (v_l, y_l)$ 构成, 如果这些基矢量集的映射 $\psi(v_i) (i = 1, \dots, l)$ 张成的矢量空间为 $\Gamma : \Gamma = (\lambda_1, \dots, \lambda_l)(\psi(v_1), \dots, \psi(v_l))^T$, 则对于一个新样本 $\psi(x_{n+1}) (l \leq n)$, 该样本与 Γ 的夹角 θ 可以表示为

$$\theta = \arcsin(\sqrt{\sigma_{\text{Lib}}/h_{n+1}}). \quad (6)$$

式中: $\sigma_{\text{Lib}} = h_{n+1} - H_{n+1}^T H_l^{-1} H_{n+1}$, $h_{n+1} = K(x_{n+1}, x_{n+1})$, 其中 $H_l = K(v_i, v_j)$, $H_{n+1} = K(x_{n+1}, v_i), i, j = 1, \dots, l, K(\cdot)$ 记为Kernel变换.

证 利用欧氏距离及矢量空间的概念, $\psi(x_{n+1})$ 在超平面的长度为 $\|\psi_{n+1}\|$, ψ_{n+1} 到各基矢量的距离为 $\|\psi_{n+1} - \sum_{i=1}^l \delta_i \psi(v_i)\|$, 则该矢量与空间 Γ 的夹角 θ 可表示为

$$\sin \theta = \min \frac{\|\psi(x_{n+1}) - \sum_{i=1}^l \delta_i \psi(v_i)\|}{\|\psi(x_{n+1})\|} =$$

$$\min \frac{\sqrt{(\psi(x_{n+1}) - \sum_{i=1}^l \delta_i \psi(v_i))^2}}{\sqrt{\psi^2(x_{l+1})}} =$$

$$\min \frac{\sqrt{\psi(x_{n+1}) \cdot \psi^T(x_{n+1}) - 2\psi(x_{n+1}) \sum_{i=1}^l \delta_i \psi^T(v_i)}}{\sqrt{\psi(x_{n+1}) \cdot \psi^T(x_{n+1})}} +$$

$$\frac{\sqrt{\sum_{i=1}^l \sum_{j=1}^l \delta_i \delta_j \psi(v_i) \psi^T(v_j)}}{\sqrt{\psi(x_{n+1}) \cdot \psi^T(x_{n+1})}},$$

采用Kernel变换, 则

$$\sin \theta = \min \sqrt{\frac{h_{n+1} - 2\delta^T H_{n+1} + \delta^T H_l \delta}{h_{n+1}}}.$$

式中: $h_{n+1} = K(x_{n+1}, x_{n+1})$, $H_{n+1} = K(x_{n+1}, v_i)$, $H_l = K(v_i, v_j)$, 其中 $i, j = 1, \dots, l$.

令 $\sigma_{\text{Lib}} = h_{n+1} - 2\delta^T H_{n+1} + \delta^T H_l \delta$, 对 σ_{Lib} 求偏导数, 则

$$\frac{\partial \sigma_{\text{Lib}}}{\partial \delta} = 2H_l \delta - 2H_{n+1} = 0, \text{ 从而 } \delta = H_l^{-1} H_{n+1}.$$

所以 $\theta = \arcsin(\sqrt{(h_{n+1} - H_{n+1}^T H_l^{-1} H_{n+1})/h_{n+1}})$. 证毕.

引入新的样本, 由式(6)计算 σ_{Lib} . 根据定理1证明知, 如果 $\sigma_{\text{Lib}} \geq h_{n+1} \sin^2 \theta$, 则说明这个新样本不可以由BV近似线性表示, 要将这个样本引入BVS, 反之则不引入. 因此, 运用 θ 角的大小可任意调整BVS的数量, 这样就可以减少 λ 的维数, 达到削减支持向量, 使最小二乘支持向量机的解具有稀疏性.

在基于VBL的LS-SVM动态建模中, 每增加一个新的样本需进行一次矩阵求逆, 模型的维数越高, 运算量也越大. 为此提出无需求逆的递推最小二乘回归估计(RLS-SVM), 该方法将提升LS-SVM动态建模的实时性.

3.2 递推的增长记忆模式算法(Recursive algorithm of increased memory mode)

引理 1 (Sherman-Woodbury)^[6] 设 X 一个 $(n+1) \times (n+1)$ 矩阵, 可分块表示为如下形式

$$X = \begin{bmatrix} A & u \\ u^T & \eta \end{bmatrix}. \quad (7)$$

其中: A 为 $n \times n$ 矩阵, u 为 $n \times 1$ 列向量, η 为标量. 则 X 的逆矩阵可表示为

$$X^{-1} = \begin{bmatrix} B & q \\ q^T & t \end{bmatrix}. \quad (8)$$

其中: $B = A^{-1} + tA^{-1}uu^TA^{-1}$, $q = -tA^{-1}u$, $t = 1/(\eta - u^TA^{-1}u)$, 公式的推导及证明见文[4].

假定第 $n+1$ 时刻有一个新的支持向量 (v_{n+1}, y_{n+1}) , 则式(3)变为

$$\begin{bmatrix} b \\ \tilde{\lambda} \\ \lambda_{n+1} \end{bmatrix} = \begin{bmatrix} 0 & \tilde{1}^T & 1 \\ \tilde{1} & \Omega + \frac{I}{\gamma} & p \\ 1 & p^T & \beta \end{bmatrix}^{-1} \times \begin{bmatrix} 0 \\ Y_{n \times 1} \\ y_{n+1} \end{bmatrix}. \quad (9)$$

其中: p 为 $n \times 1$ 列向量, $p_i = k(v_i, v_{n+1})$. $\beta = k(v_{n+1}, v_{n+1}) + 1/\gamma$, $i = 1, \dots, n$. 取

$$\begin{cases} A_{n+1} = \begin{bmatrix} 0 & \tilde{1}^T & 1 \\ \tilde{1} & \Omega + \frac{I}{\gamma} & p \\ 1 & p^T & \beta \end{bmatrix} = \begin{bmatrix} A_n & s \\ s^T & \beta \end{bmatrix}, \\ s^T = [1 \ p^T]. \end{cases} \quad (10)$$

则由引理1, A_{n+1} 的逆矩阵为

$$A_{n+1}^{-1} = \begin{bmatrix} B & q \\ q^T & t \end{bmatrix}. \quad (11)$$

其中: $B = A_n^{-1} + tA_n^{-1}ss^TA_n^{-1}$, $q = -tA_n^{-1}s$, $t = 1/(\beta - s^TA_n^{-1}s)$.

式(3)(9)(10)(11)即为递推最小二乘支持向量回归估计(RLS-SVM)的计算表达式. RLS-SVM方法有效的运用矩阵变换, 避免了加入新样本需重新求逆的问题, 从而提高了求解速度.

3.3 VBL算法的步骤(Steps of VBL algorithm)

步骤 1 确定支持向量机的Kernel函数类型及参数, 并由前两个样本建立矢量空间线的初始值, 取 $n = 2$.

步骤 2 如果有第 $n+1$ 个新样本 x_{n+1}, y_{n+1} 输入, 则根据定理1的结论式(6)计算 σ_{Lib} , h_{n+1} , H_{n+1} , 并判断 σ_{Lib} 是否满足 $\sigma_{\text{Lib}} \geq h_{n+1} \sin^2 \theta$, 是转步骤3, 否转步骤4.

步骤 3 将新样本加入到BVS中, 即第 $l+1$ 个基矢量的 v_{l+1} 等于 x_{n+1} , 利用增长模式算法(3) (9) (10)(11), 计算新的模型参数 $b, \tilde{\lambda}$, 且 $n = n+1$, $l = l+1$.

步骤 4 循环等待, 如有新样本转步骤2.

VBL算法从第2个样本开始, 随着样本的到来直接进行参数的辨识, 在线建立BVS, 并确定模型的阶次.

4 应用研究(Investigation of application)

4.1 仿真分析(Simulation analysis)

验证算法的辨识效果, 采用如下的仿真例子, 待辨识的对象具有如下形式:

$$y(k+1) = g[y(k), y(k-1)] + u(k) + e,$$

$$e \sim N(0, 0.5),$$

式中:

$$g[y(k), y(k-1)] = \frac{y(k)y(k-1)[y(k) + 2.5]}{1 + y(k)^2 + y(k-1)^2},$$

$$u(k) = \sin(2k\pi/25).$$

Kernel函数采用高斯函数: $k(x_i, x) = \exp(-\|x - x_i\|^2/\beta)$, $\beta = 0.001$, 正则化参数 $\gamma = 0.1$. 采用500个样本来训练最小二乘支持向量回归估计器, 图2和图3分别为LS-SVM和VBL的辨识结果, 为显示方便, 只画出前100个样本. 图中的实线为原始输出值, 点为受噪声污染的输出测量值, 点划线为辨识模型的输出, 可以看出两种方法都获得了较好的拟合效果. 其中VBL的辨识误差最小, 而且VBL算法更好的辨识了曲率拐点和曲率变化较大的点.

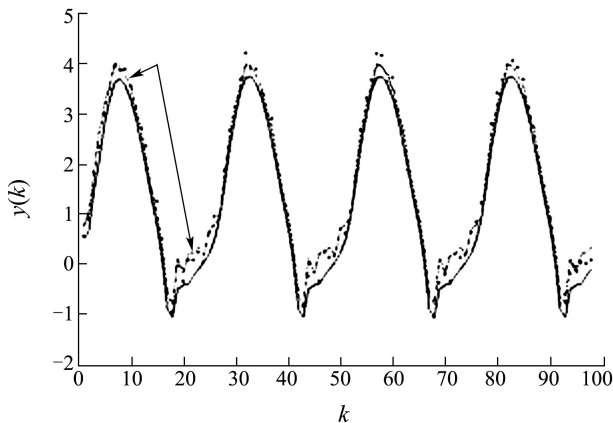


图2 LS-SVM辨识结果

Fig. 2 Results of LS-SVM identification

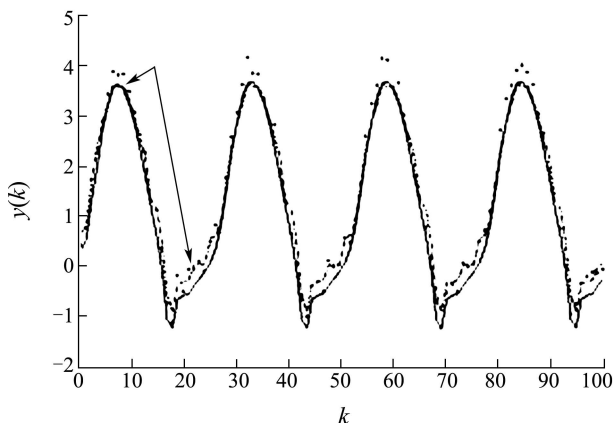


图3 VBL辨识结果

Fig. 3 Results of VBL identification

由于测量噪声的影响, 函数输出值不具有平滑性, 噪声对一些突变点影响的更明显(图中箭头所示). 样本全部用于建模, 则冗余信息中的噪声被全部拟合到模型参数, 降低了LS-SVM的鲁棒性. VBL算法能够较好的分离出冗余信息, 因此基

于VBL的LS-SVM辨识器, 具有更好的收敛性和辨识效果.

4.2 应用实例(Example of application)

验证VBL算法用于实际生产过程的可行性和有效性, 对一个自来水厂的生产过程进行了系统建模研究. 图4为一个自来水厂混凝加药沉淀系统的工艺流程略图. 自来水生产中的加矾混凝过程是水质净化的一个重要环节, 准确地投加所需混凝剂不仅是取得理想混凝效果的关键, 也是制水成本的主要组成部分.

混凝过程是一个复杂的物理、化学过程, 分析混凝过程可知, 影响混凝净化效果的几个主要因素有: pH值、浊度、TOC或COD、温度、流量和控制明矾投药的变频器频率. 混凝效果的指标为出水浊度, 通常忽略PH值、TOC指标以及温度等的影响, 仅考虑原水流量、原水浊度的影响. 因为混凝过程从混凝剂的投加, 到絮凝、沉淀、过滤, 大约需要40 min到120 min, 是一个非线性、大时滞时变系统, 所以难以通过其化学反应机理研究, 准确建立反应过程的数学模型.

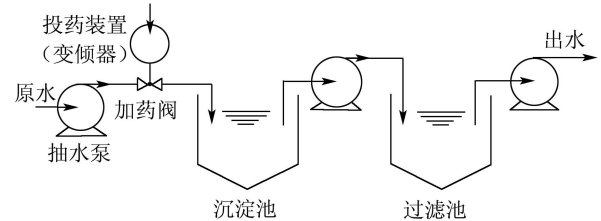


图4 加药混凝沉淀系统工艺流程

Fig. 4 Dosage coagulant sedimentation process

针对自来水加药系统的非线性和时变特性, 令 x_1 : 原水浊度; x_2 : 原水流量; u : 变频器频率; y : 出水浊度. t 时刻出水浊度的预测值 $y(t)$, 由表达式: $y(t) = f(x_1, x_2, u)$ 计算获得, f 为回归估计器的非线性函数.

实验数据取500组. 因为基于VBL算法的LS-SVM对建模样本有筛选功能, 所以500组样本都用于系统建模. 而广义LS-SVM建模, 则将样本分为训练样本和测试样本, 前300组数据用于非线性系统的初始化建模, 后200组数据用于测试.

实验结果如下:

1) 表1为基于VBL的LS-SVM与广义LS-SVM建模的预报误差统计分析结果, 图5, 6分别为两种建模方法的预报值及误差(为显示清晰只画出100个样本). 可见基于VBL的LS-SVM具有最小的预报误差和最好的辨识结果.

2) 引入递推算法后, VBL的建模和预报时间从9.1 s缩减至3.3 s, 递推算法大大提升了LS-SVM的动态建模实时性.

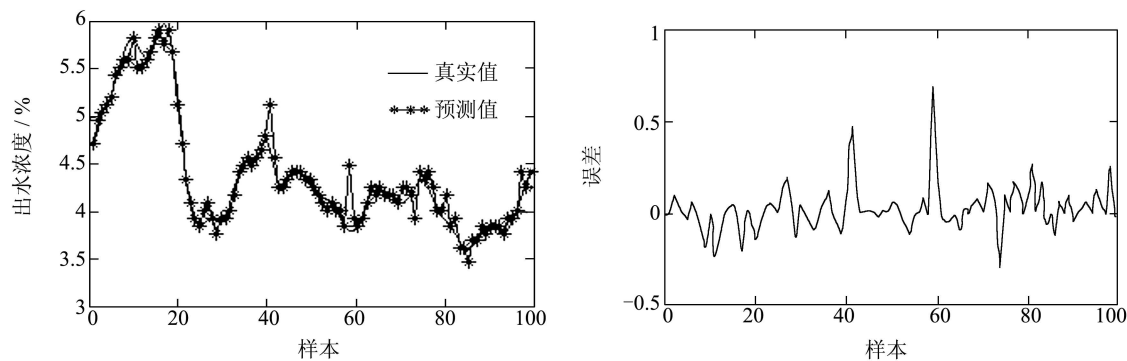


图5 基于VBL算法的RLS-SVM的仿真结果

Fig. 5 Simulation results based on the VBL algorithm of RLS-SVM

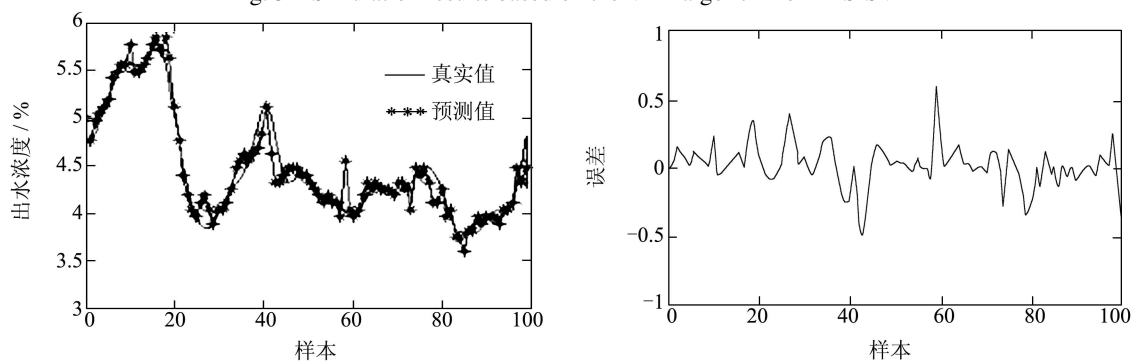


图6 传统LS-SVM的仿真结果

Fig. 6 Simulation results based on traditional LS-SVM

表1 预报误差的统计分析结果

Table 1 Statistical analysis results of prediction error

	最大绝对误差	平均相对误差	均方误差
基于VBL的LS-SVM	0.7123	0.0518	0.0963
传统LS-SVM	0.8033	0.0681	0.1048

实验结果表明:

1) LS-SVM用于非线性、时变特性系统的用时建模和预报, 有很好的应用效果;

2) 相对广义LS-SVM建模方法, 基于VBL算法的LS-SVM具有更佳的鲁棒性、收敛性和辨识效果;

3) 递推算法的引入, 可有效提升LS-SVM的动态建模速度.

5 结论(Conclusions)

矢量基学习算法的引入, 使LS-SVM的支持向量具有稀疏性, 提升了LS-SVM建模的鲁棒性、收敛性和辨识结果. 而且无需求逆的增长记忆模式, 避免LS-SVM加入新样本需重新训练所有样本的问题, 使LS-SVM更快的跟踪环境变化. 非线性函数的拟合和应用实例的仿真结果, 很好地验证了该方法的优点. 因此, 基于VBL算法的LS-SVM建模方法用于非线性系统的建模和辨识, 具有很大的应用潜力.

参考文献(References):

- [1] VAPNIK V N. *The Nature of Statistical Learning*[M]. Berlin: Springer, 1995.
- [2] SUYKENS J A K, VANDEWALE J. Least squares support vector machine classifiers[J]. *Neural Processing Letters*, 1999, 9(3): 293 – 300.
- [3] SUYKENS J A K, de BRABANTER J, LUKAS L, et al. Weighted least squares support vector machines: robustness and sparse approximation [J]. *Neurocomputing*, 2002, 48(1): 85 – 105.
- [4] JOZSEF V, GABOR H. A sparse least squares support vector machine classifier[C]//*Proc of 2004 IEEE Int Joint Conf on Neural Networks*. [S.l.]: [s.n.], 2004, 1: 25 – 29.
- [5] de KRUIF B J, DE VRIES T J A. Pruning error minimization in least squares support vector machines[J]. *Neural Networks*, 2003, 14(3): 696 – 702.
- [6] COWEN C C. *Linear Algebra for Engineering and Science*[M]. West Lafayette, Indiana: West Pickle Press, 1996.
- [7] 解应春. 基于Kernel学习机的建模与分类的应用算法研究[D]. 浙江 杭州: 浙江大学, 2003.
(XIE yingchun. *Application algorithm study on kernel machine for modeling and classifying*[D]. Hangzhou Zhejiang: Zhejiang University, 2003.)

作者简介:

陈爱军 (1974—), 男, 博士研究生, 研究方向为信号处理、过程建模、监测与控制, E-mail:chenaij@petrochina.com.cn;

宋执环 (1962—), 男, 教授, 博士生导师, 研究方向为工业过程建模与控制、信号处理;

李平 (1954—), 男, 教授, 博士生导师, 研究方向为智能控制、企业综合自动化、智能交通.