

文章编号: 1001-0920(2012)11-1745-06

时序峰值预测的最小二乘支持向量回归模型

袁从贵^{1,2}, 张新政¹

(1. 广东工业大学 自动化学院, 广州 511442; 2. 东莞职业技术学院 电子工程系, 广东 东莞 523808)

摘要: 针对最小二乘支持向量回归模型中, 呈稀疏分布的时序峰值样本拟合预测误差偏大的问题, 基于加权最小二乘思想, 提出一种新的用于时序峰值预测的最小二乘支持向量回归模型. 根据样本分布密度和输出期望幅值, 优化了经验风险控制目标. 解得模型的拟合预测误差不受样本分布的影响, 而且在保持整体样本拟合预测精度的同时, 对峰值样本的拟合预测精度有了显著提高. Lorenz 时序预测和电力负荷预测的仿真结果表明了模型的有效性.

关键词: 峰值预测; 支持向量回归; 加权最小二乘; 核密度估计

中图分类号: TP181

文献标志码: A

Least squares support vector regression for prediction of peak samples in time series

YUAN Cong-gui^{1,2}, ZHANG Xin-zheng¹

(1. School of Automation, Guangdong University of Technology, Guangzhou 511442, China; 2. Department of Electronic Engineering, Dongguan Polytechnic College, Dongguan 523808, China. Correspondent: YUAN Cong-gui, E-mail: glolly@126.com)

Abstract: The sparse distributed peak samples in time series are poorly fitted in the least squares support vector regression model. Therefore, a new least squares support vector regression model is proposed based on weighted least squares method and used to predict peak samples in time series. In this model, the structural risk objective is optimized by the distribution density and the amplitude of expected output. The fitting errors of model are not influenced by the distribution of samples, and the fitting and prediction accuracy of the peak samples is improved significantly with the holistic accuracy maintained simultaneously. The simulation results on the Lorenz time series prediction and load prediction in power system show the effectiveness of the model.

Key words: prediction of peak samples; support vector regression; weighted least squares; kernel density estimator

1 引言

时间序列是对所研究系统历史行为的客观记录, 它包含系统结构特征及其运行规律. 实际问题中的复杂动力系统表现出多变量动态演化行为和多层次结构, 系统的结构参数和边界条件具有时变性、复杂性和不确定性, 同时由于所获取的系统信息往往不够完备, 系统难以用精确的解析数学模型来辨识, 一般通过观测或实验获取的时间序列进行分析.

传统的时间序列建模方法, 如回归分析、神经网络等注重于对整体样本的拟合逼近, 遵循的是经验风险最小化原则, 要求分析得到的模型对于经验样本的拟合误差最小, 而不能优化模型的泛化推广能力.

Suykens^[1]在支持向量机基础上提出的最小二乘支持向量回归模型, 按结构风险最小化目标优化模型参数, 同时考虑了模型对经验样本的逼近精度和模型的泛化能力, 已广泛地用于金融时序预测、电力负荷预测、非线性时序分析等^[2-4]领域, 获得了比传统时序建模方法更好的拟合预测性能.

与传统时序预测模型一样, 最小二乘支持向量回归模型参数贮存的信息更多的是反映样本分布密度较高的时序变化规律, 呈稀疏分布的峰值样本预测误差明显大于密集分布区域时序样本的预测误差. 实际上, 对于城市用水预测、环境污染预测和电力负荷预测等诸多问题^[5-7], 峰值样本的预测精度更为重要.

收稿日期: 2011-05-16; 修回日期: 2011-08-31.

基金项目: 国家自然科学基金项目(61074185); 广东省中国科学院全面战略合作项目(2010B090301042); 广州科技计划项目(2011J4300079).

作者简介: 袁从贵(1978-), 男, 讲师, 博士生, 从事数据驱动建模、非线性系统控制等研究; 张新政(1955-), 女, 教授, 博士生导师, 从事复杂系统建模与控制等研究.

Suykens^[8]、Cui^[9]、Quan^[10]等基于加权最小二乘的思想,提出了多种最小二乘支持向量机模型,用于降低模型对异常样本的敏感性.这些模型采用的加权方法是对拟合误差大的样本赋予一个小的权值(如Huber权重),甚至近似为零的权值(如Bisquare权重、Suykens权重).

为了有效拟合时序峰值过程,解决最小二乘支持向量回归模型中峰值样本的拟合预测误差偏大的问题,以适合时序峰值预测应用的需要,本文基于加权最小二乘思想,提出了一种新的利用样本分布密度和模型输出期望幅值加权的最小二乘支持向量回归模型.所给出的加权方法能够消除样本分布对模型性能的影响,显著提高时序峰值预测精度.通过Lorenz时序预测和纽约电力负荷预测仿真实例,将本文模型与标准最小二乘支持向量回归模型进行了比较,所得结果表明了本文模型的可行性和有效性.

2 最小二乘支持向量回归分析

2.1 最小二乘支持向量回归模型

给定模型回归拟合时间序列样本集 $\{X_t, y_t\}_{t=1}^N$,其中输入 $X_t \in R^n$,输出 $y_t \in R$.通过非线性映射 $\varphi(\cdot): R^n \rightarrow R^{n_h}$,将样本从输入空间映射到高维特征空间并进行线性回归,有

$$y(X) = \omega^T \varphi(X) + b. \quad (1)$$

其中: ω 和 b 为待辨识的回归参数, $\omega \in R^{n_h}$, $b \in R$.根据结构化风险最小化原则, ω, b 寻优的目标函数为

$$R = \frac{1}{2} \omega^T \omega + \frac{\gamma}{2} R_{\text{emp}}. \quad (2)$$

其中: $\omega^T \omega$ 项用于控制逼近函数的复杂度,优化模型的泛化能力; R_{emp} 为样本拟合经验风险; $\gamma > 0$ 为调整参数, γ 越大,对学习样本的拟合程度越高,经验风险越小.取拟合误差的二次项和作为 R_{emp} ,得到如下优化问题^[1]:

$$\begin{aligned} \min_{\omega, b, e} J(\omega, e) &= \frac{1}{2} \omega^T \omega + \frac{1}{2} \gamma \sum_{t=1}^N e_t^2; \\ \text{s.t. } y_t &= \omega^T \varphi(X_t) + b + e_t, \quad t = 1, 2, \dots, N. \end{aligned} \quad (3)$$

其中 $e_t \in R$ 为样本拟合误差变量.为求解优化问题(3),定义拉格朗日函数如下:

$$\begin{aligned} L(\omega, b, e; \alpha) &= \\ J(\omega, e) &- \sum_{t=1}^N \alpha_t \{\omega^T \varphi(x_t) + b + e_t - y_t\}, \end{aligned} \quad (4)$$

其中 α 为拉格朗日乘子.分别对 ω, b, e_t 和 α_t 求偏导数,令偏导数为零并消除 ω 和 e_t ,得到如下线性方程组:

$$\begin{bmatrix} 0 & L^T \\ L & \Omega + \gamma^{-1} I \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ Y \end{bmatrix}. \quad (5)$$

其中: $L = [1; \dots; 1]$; $\alpha = [\alpha_1; \dots; \alpha_N]$; $\Omega_{t,j} = \varphi(X_t)^T \varphi(X_j)$, $t, j = 1, \dots, N$.选择核函数 $k(X_t, X_j) = \varphi(X_t)^T \varphi(X_j)$,满足Mercer条件.最后解得最小二乘支持向量回归模型

$$y = \sum_{t=1}^N \alpha_t k(X_t, X) + b. \quad (6)$$

2.2 样本分布对模型的影响

不妨设时序数据集 $\{X_t, y_t\}_{t=1}^N$ 样本独立同分布,其联合分布为 $F(\cdot)$.样本 (X_t, y_t) 根据联合分布 $F(X, y) = F(X)F(y|X)$ 随机取样得到.给定回归拟合的备选模型集 $\{\psi(X, p) | p \in \Lambda\}$, p 为模型参数.在式(6)中, p 包含回归参数 α, b 和核参数.定义度量经验样本输出 y 与回归模型响应 $\psi(X, p)$ 之间差异的损失函数为 $L(y, \psi(X, p))$.经验样本的产生带有随机性,模型回归拟合误差必须考虑损失函数的数学期望值,即

$$R(p) = \int L(y, \psi(X, p)) dF(X, y). \quad (7)$$

$R(p)$ 称为风险泛函,它是体现回归模型对经验样本响应程度的指标.取拟合样本误差和来控制经验风险,即样本经验风险泛函为

$$R'_{\text{emp}}(p) = N \int L(y, \psi(X, p)) dF(X, y). \quad (8)$$

由于分布函数 $F(\cdot)$ 是未知的,式(8)的经验风险无法实际计算.根据已有的经验样本集,可以建立如下样本经验风险泛函期望:

$$\begin{aligned} R_{\text{emp}}(p) &= E(R'_{\text{emp}}) = \\ E\left(N \int L(y, \psi(X, p)) dF(X, y)\right) &= \\ N \sum_{t=1}^N L(y_t, \psi(X_t, p)) f(X_t, y_t) \delta_{X \times y}. \end{aligned} \quad (9)$$

其中: $\delta_{X \times y}$ 表示样本分布空间的单位超立方, $f(\cdot)$ 为样本分布密度.在标准最小二乘支持向量回归模型中,损失函数 $L(\cdot)$ 取整体样本拟合误差的二次项和,所以其样本经验风险泛函期望为

$$R_{\text{emp1}}(p) = N \sum_{t=1}^N e_t^2 f(X_t, y_t) \delta_{X \times y}. \quad (10)$$

从式(10)可以看出,最小二乘支持向量回归模型的经验风险控制项隐含有样本的分布密度加权 $f(\cdot)$.分布密度高的样本具有较大的拟合误差权重,因此,模型经验风险控制及参数优化倾向于减小密集分布样本的拟合预测误差,从而导致呈稀疏分布的峰值样本预测误差偏大.

3 时序峰值预测的最小二乘支持向量回归模型

建立适用于时序峰值预测的最小二乘支持向量回归模型的基本思想是,使模型参数学习倾向于减小

模型期望输出较大值的误差, 提高算法对时序峰值过程的拟合预测精度. 这里主要是基于加权最小二乘思想, 通过在经验风险损失函数中引入合理的拟合误差加权修正系数, 增大峰值样本拟合误差权重的方法来实现. 模型首先通过非参数核密度方法估计得到样本分布密度; 然后根据样本分布密度和模型期望输出幅值得到拟合误差加权修正系数, 优化样本, 拟合经验风险; 最后解得峰值时序预测的最小二乘支持向量回归模型.

3.1 样本分布密度估计

样本变量的联合分布密度是未知的, 只能根据经验样本估计得到. 核密度估计作为一种非参数分布密度估计方法, 它不依赖于样本分布的先验知识, 对样本分布不附加任何假定, 是一种从数据样本本身出发, 研究样本分布特征的方法.

设 $w(\cdot)$ 为 R^d 上一个已知的核函数, 满足

$$\sup_{-\infty < \xi < +\infty} |w(\xi)| < +\infty, \quad w(\xi) = w(-\xi), \quad (11)$$

$$w(\xi) \geq 0, \quad \int_{-\infty}^{+\infty} w(\xi) d\xi = 1, \quad (12)$$

$$\lim_{|\xi| \rightarrow \infty} |uw(\xi)| = 0, \quad (13)$$

则多变量样本分布密度的 Parzen 核密度估计^[11]为

$$\hat{f}(\xi_k) = \frac{1}{Nh^d} \sum_{i=1}^N w\left(\frac{\xi_k - \xi_i}{h}\right). \quad (14)$$

其中: $\xi_k \in R^d$; $h > 0$ 为核平滑参数. 设 $f(\cdot)$ 在 ξ_k 连续, 当 $N \rightarrow +\infty$, $h \rightarrow +0$ 及 $Nh \rightarrow +\infty$ 时, $\hat{f}(\xi_k)$ 依概率收敛到 $f(\xi_k)$.

拉普拉斯核密度估计能以一个合适的精度逼近样本的联合分布密度, 而且具有较小的计算复杂度^[12]. 采用多变量拉普拉斯函数核替代式 (14) 中的 $w(\cdot)$, 得到样本分布的拉普拉斯核密度估计为

$$\hat{f}(\xi_k) = \frac{1}{N(2\lambda)^d} \sum_{i=1}^N \prod_{j=1}^d e^{-\frac{1}{\lambda} |\xi_{k,j} - \xi_{i,j}|} = \frac{1}{N(2\lambda)^d} \sum_{i=1}^N \exp\left(-\frac{1}{\lambda} \sum_{j=1}^d |\xi_{k,j} - \xi_{i,j}|\right). \quad (15)$$

其中 λ 为拉普拉斯核平滑参数, 用于平滑不同距离样本对 ξ_k 处样本分布密度估计的影响, 其值可以参考 Silverman^[11]给出的优化选择方法获取.

当样本为重拖尾分布时, 特别是小样本情况下, 如果固定平滑参数, 则在核密度估计得到的分布密度的尾部会出现伪噪声分布现象. 增大核平滑参数的方法虽然可以平滑这一伪噪声分布, 但核平滑参数过大会掩盖高密度分布区域样本的局部分布特征. 此时, 采用变参数估计方法, 即在样本分布密集区域取较小

的核平滑参数, 而在样本低密度分布区域取较大的核平滑参数, 可以得到更好的样本分布密度估计结果.

3.2 拟合误差加权修正系数设计

根据式 (9) 和 (10), 要消除样本分布对标准最小二乘支持下向量回归模型的影响, 避免呈稀疏分布的峰值样本拟合预测误差偏大的问题, 可以通过采用样本分布密度的倒数加权模型损失函数 $L(\cdot)$ 的方法来实现. 同时, 为了进一步提高时序峰值预测精度, 在损失函数中引入模型期望输出幅值加权因子, 以增大峰值样本拟合误差权重. 据此, 包含样本分布密度和幅值加权的误差加权修正系数 v_t 定义如下:

$$v_t = f^{-1}(X_t, y_t) \times \mu_t, \quad (16)$$

其中 μ_t 为模型输出期望幅值加权因子, 定义为

$$\mu_t = \exp\left(C \times \frac{y_t - y_{\min}}{y_{\max} - y_{\min}}\right). \quad (17)$$

这里 $C \geq 0$, 为幅值加权调整系数. C 越大, 对峰值样本的加权权重越大. 当 $C = 0$ 时, 仅对样本拟合误差进行分布密度加权.

将 v_t 加权的样本拟合误差的二次项和替代式 (9) 中的损失函数 $L(\cdot)$, 得到时序峰值预测的最小二乘支持向量回归模型的样本经验风险泛函期望

$$R_{\text{emp}_2}(p) = N \sum_{t=1}^N \exp\left(C \times \frac{y_t - y_{\min}}{y_{\max} - y_{\min}}\right) e_t^2 \delta_{X \times y}. \quad (18)$$

由上式可知, 经过误差加权修正, 样本经验风险泛函期望不再包含样本分布因子, 模型风险优化目标不受样本分布的影响, 从而避免了峰值样本的稀疏性引起的模型拟合误差偏大的问题. 另外, 峰值样本具有相对较大的拟合误差权重, 模型风险优化的结果有利于减小峰值样本的拟合预测误差.

3.3 加权最小二乘支持向量回归模型

在回归拟合时序样本集中增加样本拟合误差加权修正系数 v_t , 使原样本转换成 $\{X_t, y_t, v_t\}_{t=1}^N$. 优化目标函数

$$\min_{\omega^*, b^*, e^*} J(\omega^*, e^*) = \frac{1}{2} \omega^{*T} \omega^* + \frac{1}{2} \gamma \sum_{t=1}^N v_t (e_t^*)^2; \quad (19)$$

s.t. $y_t = \omega^{*T} \varphi(X_t) + b^* + e_t^*, \quad t = 1, 2, \dots, N.$

利用拉格朗日乘子求解, 有

$$L(\omega^*, b^*, e^*; \alpha^*) = J(\omega^*, e^*) - \sum_{t=1}^N \alpha_t^* \{\omega^{*T} \varphi(X_t) + b^* + e_t^* - y_t\}, \quad (20)$$

其中 $*$ 表示未知变量. 分别对 $\omega^*, b^*, e_t^*, \alpha_t^*$ 求偏导数并令偏导数为零, 得

$$\begin{cases} \omega^* = \sum_{t=1}^N \alpha_t^* \varphi(X_t); \\ \sum_{t=1}^N \alpha_t^* = 0; \\ \alpha_t^* = \gamma v_t e_t^*, t = 1, \dots, N; \\ \omega^{*T} \varphi(X_t) + b^* + e_t^* - y_t = 0, t = 1, \dots, N. \end{cases} \quad (21)$$

设 $\Psi = \Omega + \frac{1}{\gamma} \text{diag}(\frac{1}{v_1}; \dots; \frac{1}{v_N})$, $\alpha^* = [\alpha_1^*; \dots; \alpha_N^*]$. 注意到 Ψ 为实对称正定矩阵, 解线性方程组 (21) 可得适用于时序峰值预测的最小二乘支持向量回归模型

$$y_i = \sum_{t=1}^N \alpha_t^* k(X_t, X_i) + b^*. \quad (22)$$

其中

$$\begin{cases} b^* = L^T \Psi^{-1} L^{-1} L^T \Psi^{-1} Y, \\ \alpha^* = \Psi^{-1} (Y - Lb^*). \end{cases} \quad (23)$$

模型拟合误差

$$e_t^* = \frac{\sum_{i=1}^N \Psi_{it} (y_i - b)}{\gamma v_t \sum_{i=1}^N \Psi_{it} k(X_i, X_t) + \Psi_{tt}}. \quad (24)$$

这里 Ψ_{it} 表示 Ψ 第 i 行、第 j 列元素的代数余子式. 因为 $\Psi_{it}, \gamma, v_t > 0$, 且 Ψ_{it} 与 v_t 无关, 所以样本拟合误差 e_t^* 随误差加权修正系数 v_t 的增大而减小.

4 应用仿真

为了验证本文提出模型的有效性, 下面采用 Lorenz 时序^[13]预测和电力负荷^[14]预测两个实例进行仿真验证.

4.1 Lorenz 时间序列预测

Lorenz 系统是美国气象学家洛伦兹简化大气环流模型得到的三变量一阶微分方程组, 是基于流体力学中的 Navier-Stokes 方程、热传导方程和连续性方程构建的三维动力耗散系统, 其含时间参数的形式为

$$\begin{cases} \frac{d(x(t))}{dt} = \sigma(y(t) - x(t)), \\ \frac{d(y(t))}{dt} = \rho x(t) - y(t) - x(t)z(t), \\ \frac{d(z(t))}{dt} = x(t)y(t) - \beta z(t). \end{cases} \quad (25)$$

其中: σ, ρ 和 β 为正数, 分别称为 Prandtl 数、Rayleigh 数和比例常数; x 为对流强度; y 正比于上升与下降流温差; z 为垂直方向的温度梯度. 作为一个典型的时序预测例子, 在很多研究中 σ 取 10, β 取 8/3, 而 ρ 则为可调参数. 系统在 $\rho = 28$ 时表现出混沌特性, ρ 为其他值时会显示出组结的周期轨道. 在下面的仿真中, $\rho = 28, (x_0, y_0, z_0) = (15.34, 13.68, 37.91)$.

同时采用最小二乘支持向量回归模型 (LS-SVR) 和时序峰值预测模型 (TSPLS-SVR) 对 z 进行超前 8 步单变量回归预测, 输入延迟分别为 8, 25, 68, 57, 9^[15]. 训练样本 800 组, 测试样本 120 组. 其中 TSPLS-SVR 模型的样本幅值加权调整系数 C 取零, 即误差修正权重 v_t 取 $f^{-1}(\cdot)$, 以检验样本分布密度加权的结果. 为了便于比较, 两个模型均使用 RBF 核函数, 调整参数 $\gamma = 1$, 核参数采用广义交叉验证寻优得到.

两种模型的测试样本预测结果如图 1 所示. 从图中可以看出, 在峰值附近, TSPLS-SVR 模型的预测结果明显优于 LS-SVR 模型. 这是因为样本分布密度加权加大了峰值样本拟合误差权重, 从而提高了峰值样本的拟合精度.

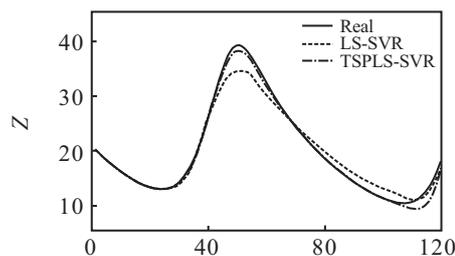


图 1 Lorenz 时序预测结果

样本分布密度与模型回归拟合预测误差绝对值的对比见图 2. 为了消除拟合预测误差的随机性影响, 对误差进行了平滑处理.

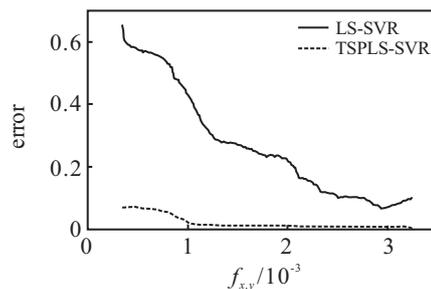


图 2 样本分布密度对模型拟合预测误差的影响

图 2 表明, 样本分布密度对标准 LS-SVR 模型回归拟合预测精度的影响明显. 分布密度小的样本拟合预测误差较大, 分布密度大的样本模型拟合预测误差较小. 经过分布密度加权优化后, 得到的 TSPLS-SVR 模型的拟合预测误差基本不受样本分布密度的影响.

4.2 电力负荷预测

电力负荷预测是配电网规划的基础, 也是电力系统运行、控制所必不可少的一部分. 在实践中, 无论是制定电力系统规划还是实现电力系统运行自动化, 进行相应的负荷预报都非常重要. 对于发电厂或电力建设公司而言, 负荷预测是其制定发电计划、检修计划和电价制定的基础. 对于供电公司或电力调度公司而言, 电力负荷预测是其制定购电计划, 对地区电网进

行整合规划或协调计划的依据, 在确保供电范围内的电网安全、可靠运行的同时, 可获得最大的经济效益。

NYISO (New York independent system operator) 是纽约电力系统的核心, 负责纽约近 11 000 英里高压传输线、超过 500 个发电机组的运行, 管理和监控整个纽约州电力市场。下面分别采用 LS-SVR 和 TSPLS-SVR 模型提前 24 h 预测纽约市电力负荷, 输入延迟分别为 168, 24, 144, 162, 172^[15]。仿真数据来自于 NYISO 公布的 2004 年 12 月监测值, 200 组训练样本, 75 组测试样本。两种模型均采用 RBF 核, 用广义交叉寻优算法对核参数进行优化选择, 并取多次仿真结果均值以消除参数优化对仿真测试的影响。加权调整系数 $C = 4$, 误差加权修正系数 $v_t = f^{-1}(\cdot) \times e^{\frac{4y_t - 16.208}{3.235}}$ 。峰值样本预测结果如图 3 所示。从图中可以看出, TSPLS-SVR 电力负荷峰值预测的结果明显优于 LS-SVR。

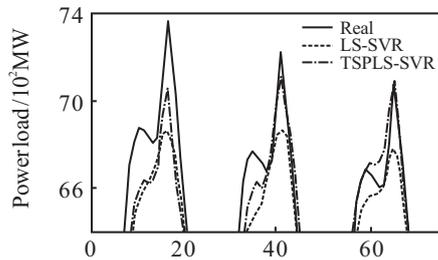


图 3 纽约电力负荷峰值预测结果

表 1 给出了两个模型预测的平均绝对误差 (MAE) 和平均绝对百分比误差 (MAPE)。其中峰值样本取自于整体测试样本数据集, 比例为 20%。

表 1 纽约电力负荷预测误差

预测模型	MAE/MW		MAPE/%	
	整体样本	峰值样本	整体样本	峰值样本
LS-SVR	174.55	270.71	2.99	3.47
TSPLS-SVR	170.43	164.46	2.99	2.35

分析表 1 可以得出如下结论: 1) LS-SVR 对峰值样本的平均预测误差明显大于整体样本; 2) 两种模型对全部测试样本的平均预测精度相近; 3) TSPLS-SVR 预测峰值样本的相对误差比 LS-SVR 降低了 30%。这是因为 LS-SVR 优化目标中经验风险隐含密度加权, 导致呈稀疏分布的峰值样本预测精度低于整体样本的预测精度。两种模型都是采用误差的二次项控制经验风险, 所有测试样本的平均预测误差比较接近。TSPLS-SVR 调整了不同分布样本的经验风险, 使得峰值样本具有更高的误差权重, 有效提高了这部分样本的预测精度。

为了考察加权调整系数 C 对模型预测精度的影响, 在 C 取不同值的情况下对模型进行仿真测试, 预测结果见表 2。

表 2 C 不同取值时对 TSPLS-SVR 模型预测误差的影响

C	MAE/MW		MAPE/%	
	整体样本	峰值样本	整体样本	峰值样本
0	169.97	217.43	2.94	3.10
1	171.29	193.97	2.97	2.77
2	173.04	179.00	3.01	2.56
3	172.27	168.25	3.01	2.41
4	170.43	164.46	2.99	2.35
5	169.26	167.70	2.98	2.40

从表 2 可以看出, 峰值样本的平均预测误差随着 C 的增大而明显降低。仿真时发现当 C 取值偏大时, 会出现过拟合现象。这是因为相对于其他样本, 随着 C 的增大, 峰值样本具有更大的拟合误差加权修正权重, 其拟合预测精度进一步得到了提高。然而, 加权修正系数的增大在降低模型经验风险的同时, 会扩大回归拟合函数的复杂度, 降低模型的泛化性能。所以当加权修正权重过大时, 将会出现模型的过拟合现象。在实际应用中, C 的取值建议为 4。

5 结 论

本文研究了最小二乘支持向量回归模型中, 时序峰值样本的拟合预测误差偏大的问题, 提出了一种适用于时序峰值预测的最小二乘支持向量回归模型。该模型主要是通过样本分布密度的倒数加权样本经验风险损失函数, 解决因稀疏分布引起的峰值样本拟合预测误差偏大的问题; 同时采用模型输出期望幅值加权, 进一步降低峰值样本的拟合预测误差。研究表明, 该模型在保持样本整体预测精度的同时, 能够显著提高时序峰值预测精度。本文给出的峰值误差修正算法还可以应用于咸潮预报、污染预测等注重峰值预报需要的相关领域, 具有一定的普适性。

需要指出的是, 随着样本维数的增加, 分布密度加权因子呈几何级数增大, 会降低模型的泛化推广能力。本文思想中密度权因子的作用主要是消除样本分布对模型拟合预测精度的影响, 所以当样本具有高维特征时, 采用相对分布密度加权也能达到算法的目的, 同时可以避免上述问题。

参考文献(References)

- [1] Suykens J A K. Least squares support vector machines[R]. Leuven: KULeuven, 2002: 10-84.
- [2] Tony Van Gestel, Johan A K Suykens, Dirk-Emma Baestaens, et al. Financial time series prediction using least squares support vector machines within the evidence framework[J]. Neural Networks, 2001, 12(4): 809-821.
- [3] Li Yuancheng, Fang Tingjian, Yu Erkeng. Short-term electrical load forecasting using least squares support vector machines[C]. PowerCon, 2002: 230-233.

- [4] Marcelo Espinoza, Suykens J A K, Bart De Moor. Load forecasting using fixed-size least squares support vector machines[J]. Lecture Notes in Computer Science, 2005, 3512: 488-527.
- [5] Adamowski J F. Peak daily water demand forecast modeling using artificial neural networks[J]. J of Water Resources Planning and Management, 2008, 134(2): 119-128.
- [6] Enrico Pisoni, Marcello Farina, Claudio carnevale, et al. Forecasting peak air pollution levels using NARX models[J]. Engineering Applications of Artificial Intelligence, 2009, 22(4-5): 593-602.
- [7] Hyndman R J, Shu Fan. Density forecasting for long-term peak electricity demand[J]. IEEE Trans on Power Systems, 2010, 25(2): 1142-1153.
- [8] 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(1-4): 85-105.
- [9] Cui Wentong, Yan Xuefeng. Adaptive weighted least square support vector machine regression integrated with outlier detection and its application in QSAR[J]. Chemometrics and Intelligent Laboratory Systems, 2009, 98(2): 130-135.
- [10] Quan Tingwei, Liu Xiaomao, Liu Qian. Weighted least squares support vector machine local region method for nonlinear time series prediction[J]. Applied Soft Computing, 2010, 10(2): 562-566.
- [11] Silverman B W. Density estimation for statistics and data analysis[M]. London: Chapman & Hall, 1986: 7-21.
- [12] Gavin J Bowden, Graeme C Dandy, Holger R Maier. Input determination for neural network models in water resources applications, Part I: Background and methodology[J]. J of Hydrology, 2005, 301(1/2/3/4): 75-92.
- [13] Caro Lucas M M, Araabi M B, Elaheh Kamaliha. Fuzzy descriptor systems and spectral analysis for chaotic time series prediction[J]. Neural Computer & Application, 2009, 18(8): 991-1004.
- [14] http://www.nyiso.com/public/markets_operations/market_data/load_data/index.jsp[EB/OL]. 2011.
- [15] Yuan Conggui, Zhang Xinzhen, Xu Shuqiong. Partial mutual information for input selection of time series prediction[C]. Proc of 2011CCDC. Mianyang, 2011: 2010-2014.

(上接第1744页)

- [9] Kim Y H, Lewis F L. Neural network output feedback control of robot manipulators[J]. IEEE Trans on Robotics and Automation, 1999, 15(2): 301-309.
- [10] Sun Fuchun, Sun Zengqi, Woo Pengyung. Neural network-based adaptive controller design of robotic manipulators with an observer[J]. IEEE Trans on Neural Network, 2001, 12(1): 54-67.
- [11] Travis Dierks, Sarangapani Jagannathan. Neural network output feedback control of robot formations[J]. IEEE Trans on Systems, Man, and Cybernetics: B, 2010, 40(2): 383-399.
- [12] Wang C, Hill D J. Learning from neural control[J]. IEEE Trans on Neural Network, 2006, 17(1): 130-146.
- [13] Wang C, Hill D J. Deterministic learning theory for identification[M]. Boca Raton: CRC Press, 2009: 17-36.
- [14] Ge S S, Hang C C, Zhang Tao. Adaptive neural network control of nonlinear systems by state and output feedback[J]. IEEE Trans on Systems, Man, and Cybernetics: B, 1999, 29(6): 818-828.
- [15] Tee K P, Ge S S. Control of fully actuated ocean surface vessels using a class of feedforward approximators[J]. IEEE Trans on Control System Technology, 2006, 14(4): 750-756.
- [16] Kurdila A J, Narcowich F J, Ward J D. Persistency of excitation in identification using radial basis function approximants[J]. SIAM J of Control and Optimization, 1995, 33(2): 625-641.
- [17] Liu Tengfei, Wang Cong, David J Hill. Learning from neural control of nonlinear systems in normal form[J]. System Control Letter, 2009, 58(9): 633-638.
- [18] Loría A, Panteley E. Uniform exponential stability of linear time-varying systems: Revisited[J]. System Control Letter, 2002, 47(1): 13-24.
- [19] Khalil H K. Nonlinear systems[M]. 3rd ed. Upper Sanddle River: Prentice-Hall, 2002: 167-196.
- [20] Huang S J, Lee J S. A stable self-organizing fuzzy controller for robotic motion control[J]. IEEE Trans on Industry Electron, 2000, 47(2): 421-428.