

融合提升小波降噪和 LSSVM 的网络流量在线预测

李明迅*, 孟相如, 袁荣坤, 温祥西, 陈新富

(空军工程大学 电讯工程学院, 西安 710077)

(*通信作者电子邮箱 coolmx@163.com)

摘要: 针对网络流量数据被噪声污染而无法进行准确建模与预测的问题, 将提升小波降噪(LWD)技术和在线最小二乘支持向量机(LSSVM)相结合, 提出了一种网络流量的集成式在线预测方法。该方法首先对采集的流量数据进行降噪, 然后采用相空间重构理论计算流量的时延、嵌入维数, 据此确定训练样本并建立在线预测模型, 对网络流量数据进行预测。实验结果表明, 该方法能有效滤除流量噪声, 实现在线预测, 提高预测精度。

关键词: 网络流量预测; 提升小波降噪; 最小二乘支持向量机; 在线算法

中图分类号: TP393.06 **文献标志码:** A

Online prediction of network traffic by integrating lifting wavelet de-noising and LSSVM

LI Ming-xun*, MENG Xiang-ru, YUAN Rong-kun, WEN Xiang-xi, CHEN Xin-fu

(Institute of Telecommunication Engineering, Air Force Engineering University, Xi'an Shaanxi 710077, China)

Abstract: Concerning the problem that the network traffic data has been polluted by noise so that accurate modeling and predicting cannot be achieved, an integrated network traffic online predicting method based on lifting wavelet de-noising and online Least Squares Support Vector Machines (LSSVM) was proposed. First, the Lifting Wavelet De-noising (LWD) was used to pre-process network traffic data, then the phase space reconstruction theory was introduced to calculate the delay time and embedded dimension. On this basis, the training samples were formed and the online LSSVM prediction model was constructed to predict the network traffic. The experimental results show that this prediction model can eliminate the noise effectively and predict the network traffic.

Key words: network traffic prediction; Lifting Wavelet De-noising (LWD); Least Squares Support Vector Machine (LSSVM); online algorithm

0 引言

随着通信网规模和业务的增加, 网络故障管理也越来越繁重与复杂。网络流量预测能为网络故障预测和健康管理提供有效依据, 使故障在尚未发生时就得到控制, 能增强网络的可生存性和减小维护成本。

研究表明网络流量存在混沌特性^[1], 网络流量的自相似现象与混沌现象之间存在着紧密联系, 某些特征量具有相同的值, 故网络流量是可进行短期预测的^[2]。传统的预测方法如基于分形自回归整合滑动平均模型(Fractional Auto-Regressive Integrated Moving Average, FARIMA)、自回归滑动平均模型(Auto-Regressive and Moving Average Model, ARMA)、指数平滑等线性方法难以对其建模^[3], 而基于神经网络的预测方法易出现过学习问题而影响其泛化性并容易陷入局部最优。以统计学习理论为基础的支持向量机(Support Vector Machine, SVM)综合了核技术和结构风险最小化原则, 较好地解决了小样本、非线性和局部极小点等实际问题, 具有较强的泛化能力^[4]。标准 SVM 预测在训练时需解二次规划问题, 其计算量较大, 为提高 SVM 的计算速度, Suykens 等^[5]提出了最小二乘支持向量机(Least Squares Support Vector Machines, LSSVM), 简化运算。现有的网络流量预测技术通

常采用的是离线学习的预测方法, 即使用固定的学习样本来建立训练模型, 但最初的训练模型对新样本的预测能力会随时间推移大幅下降, 而使用滑动窗口在线更新样本可以较好地解决这一问题。本文在学习新样本时引入迭代算法, 大大降低了运算的复杂度, 加快学习的速度。

实时采集的网络流量序列不可避免地被各种噪声所污染, 若不对其进行降噪处理, 将增加系统的复杂性, 弱化序列的自相关性, 使预测模型的准确度和泛化性大大降低。所以, 在实际应用中, 对采集的网络流量数据进行降噪是非常必要的。非线性小波变换阈值降噪是目前发展成熟的方法, 在混沌信号降噪方面有很好的应用。传统的小波变换通常使用 Mallat^[6]算法, 这种方法计算量很大, 计算复杂度高, 对存储空间要求高, 不利于应用于在线降噪。与传统小波主要从频域分析问题不同, 称之为二代小波的提升小波^[7]是在时(空)域进行变换, 无需借助傅里叶变换, 并且可以将所有的传统小波都通过提升方法构造出来。提升小波变换结构简单, 运算量小, 原位运算, 节省存储空间, 逆变换可以直接翻转实现, 以及有可逆的整数到整数的变换, 适合在线降噪。

本文将提升小波降噪(Lifting Wavelet De-noising, LWD)和最小二乘支持向量机回归相结合, 引入相空间重构理论和滑动窗口技术, 提出一种网络流量的集成式在线预测方法, 实

收稿日期: 2011-07-12; 修回日期: 2011-09-09。 基金项目: 陕西省自然科学基金资助项目(SJ08F14, 2009JQ8008)。

作者简介: 李明迅(1987-), 男, 四川成都人, 硕士研究生, 主要研究方向: 网络故障预测; 孟相如(1963-), 男, 陕西蓝田人, 教授, 博士, 主要研究方向: 宽带通信网络; 袁荣坤(1986-), 男, 陕西杨凌人, 硕士研究生, 主要研究方向: 网络可生存性; 温祥西(1984-), 男, 江苏连云港人, 博士研究生, 主要研究方向: 人工智能、网络健康管理; 陈新富(1973-), 男, 浙江金华人, 讲师, 硕士, 主要研究方向: 音频、视频与通信信号处理。

现降噪、预测双在线,对实时网络流量进行预测。

1 算法原理

1.1 提升小波在线降噪

1.1.1 算法描述

提升小波变换分为三个阶段:分裂(split),预测(predict)和更新(update)。

1) 分裂。将输入信号序列 S_j 通过懒小波变换(lazy wavelet transform),根据其序数的奇偶性分解为两部分,一个是偶数序列 e_{j-1} ,另一个是奇数序列 o_{j-1} ,即

$$Split(S_j) = (e_{j-1}, o_{j-1}) \quad (1)$$

2) 预测。由于奇偶序列之间存在一定的相关性,因此可以通过偶数序列 e_{j-1} 来预测奇数序列 o_{j-1} ,引入预测算子 $P(\cdot)$ 。定义如下:

$$d_{j-1} = o_{j-1} - P(e_{j-1}) \quad (2)$$

其中 d_{j-1} 是原始值 o_{j-1} 与预测值 $P(e_{j-1})$ 之间的误差,反映了两者之间的逼近程度,称之为细节系数,一般由原始信号 S_j 的高频部分组成。

3) 更新。为了获得原始信号 S_j 的逼近 S_{j-1} ,即近似系数,引入更新算子 $U(\cdot)$ 。定义如下:

$$S_{j-1} = e_{j-1} + U(d_{j-1}) \quad (3)$$

从频域上看,近似系数 S_{j-1} 一般由原始信号 S_j 的低频部分组成。若对 S_{j-1} 重复进行上面3步操作,就可以得到数据的多级分解。

提升小波的重构(merge)非常简单,仅仅是运算符的改变,通过如下公式即可获得:

$$\begin{cases} e_{j-1} = S_{j-1} - U(d_{j-1}) \\ o_{j-1} = d_{j-1} + P(e_{j-1}) \\ S_j = \text{merge}(e_{j-1}, o_{j-1}) \end{cases} \quad (4)$$

提升小波变换过程如图1所示。

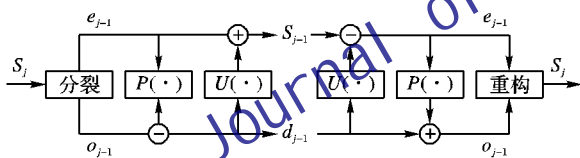


图1 提升小波变换过程

1.1.2 网络流量在线降噪

目前的小波降噪技术一般都是对测量数据进行离线批处理,但对于实时更新的网络流量序列,离线降噪方法已不再适用,因此本文引入滑动窗口法进行在线降噪。

本文降噪所采用的阈值如下:

$$\begin{cases} thr = \sigma \sqrt{2\ln(N)} \\ \sigma = \text{median}(|d(n)|)/0.6745 \end{cases} \quad (5)$$

其中: thr 为阈值, N 为细节系数序列的长度, σ 为噪声估计标准差, $\text{median}(\cdot)$ 为中值函数。

经典阈值函数是 Donoho 提出的硬阈值函数和软阈值函数^[8],由于硬阈值函数在小波域内不连续,在 $\pm thr$ 之间有间断,这在重构时会产生振荡。故本文采用软阈值函数如下:

$$\hat{d}(n) = \begin{cases} \text{sgn}(d(n))(|d(n)| - thr), & |d(n)| \geq thr \\ 0, & |d(n)| < thr \end{cases} \quad (6)$$

1.2 在线 LSSVM 预测

1.2.1 最小二乘支持向量机

LSSVM 是 SVM 的一种改进,通过引入最小二乘损失函

数和等式化约束的方法,使问题的求解变为解线性方程,避免了解二次规划问题,所需的计算资源较少,具有更快的求解速度,更适合于在线预测。

对于样本集 $S = \{x_i, y_i\}_{i=1}^l$, x_i 为 m 维输入向量, l 为样本个数, y_i 为一维输出向量。由于 x_i 与 y_i 间为非线性关系,因此将 x_i 映射到高维特征空间中, LSSVM 的基本思想是在高维空间中对样本进行线性回归:

$$y = w^T \varphi(x) + b \quad (7)$$

其中: $\varphi(x)$ 为非线性映射函数, w 为权向量, b 为偏差量。根据结构风险最小化原理,对以上问题的求解可描述如下:

$$\min_{w, b, \xi} J(w, \xi) = \frac{1}{2} \|w\|^2 + \frac{C}{2} \sum_{i=1}^l \xi_i^2 \quad (8)$$

$$\text{s. t. } y_i - \xi_i = w^T \varphi(x_i) + b, i = 1, 2, \dots, l$$

其中: C 为惩罚因子, ξ_i 为训练误差。为求解此优化问题,可引入如下 Lagrange 函数:

$$L(w, b, \xi; \alpha) = J(w, \xi) + \sum_{i=1}^l \alpha_i [y_i - \xi_i - w^T \varphi(x_i) - b] \quad (9)$$

其中 $\alpha_i (i = 1, 2, \dots, l)$ 为 Lagrange 乘子。由 Karush-Kuhn-Tucker (KKT) 条件得到如下关系式:

$$\begin{cases} \frac{\partial L}{\partial w} = 0 \rightarrow w = \sum_{i=1}^l \alpha_i \varphi(x_i) \\ \frac{\partial L}{\partial \xi_i} = 0 \rightarrow y_i - \xi_i - w^T \varphi(x_i) - b = 0 \\ \frac{\partial L}{\partial b} = 0 \rightarrow \sum_{i=1}^l \alpha_i = 0 \\ \frac{\partial L}{\partial \xi_i} = 0 \rightarrow \alpha_i = C \xi_i \end{cases} \quad (10)$$

核函数 $K_{ij} = K(x_i, x_j) = [\varphi(x_i), \varphi(x_j)]$, 式(10)的求解可形式化为

$$\begin{bmatrix} 0 & e^T \\ e & Q + C^{-1}I \end{bmatrix} \times \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix} \quad (11)$$

其中: Q 是元素为 K_{ij} 的 $l \times l$ 阶核矩阵, I 为单位矩阵, 向量 $e = [1, 1, \dots, 1]^T$, 向量 $\alpha = [\alpha_1, \alpha_2, \dots, \alpha_l]^T$, 向量 $y = [y_1, y_2, \dots, y_l]^T$ 。求解式(11)得到 α_i, b 代入式(7)中即可得到 LSSVM 的输出:

$$y = \sum_{i=1}^l \alpha_i K(x_i, x) + b \quad (12)$$

1.2.2 在线迭代学习

由于网络流量序列是一个混沌的时间序列,经过一段时间后,最初训练的模型对新样本的预测会出现较大的误差。因此,模型应随样本的更新而更新。文献[9-10]提出一种在线学习方法,采用滑动窗口更新样本,即每增加一个最新样本,就丢弃一个最旧样本,保持学习样本集大小不变;若每更新一次样本集都要重新学习一遍,会导致预测效率急剧下降,不符合网络流量预测的时效性。所以本文采用一种迭代方法进行在线学习^[11],根据新样本的特性迭代修改回归预测函数,减少学习时间。

2 网络流量集成式在线预测模型

2.1 模型框架

相空间重构是非线性时间序列分析的重要步骤,重构的质量直接影响到模型的建立和预测。其关键是时延 τ 和嵌入维数 m 的确定^[2]。对每一个确定的时间序列,都应存在一个最优的 τ 和 m 。如果 τ 太小,将不能展示系统的动力学特征,

τ 太大会使简单轨道变得复杂且会减少有效的数据点数。同样 m 太小嵌入空间无法包容动力系统的吸引子,从而无法全面展现系统的动力学特性; m 太大不仅会减少可用数据长度,增加计算工作量,而且可能会增大预测误差。本文采用互信息法求时延,假近邻法求嵌入维数,将降噪后的网络流量序列重构到低维的相空间中,然后再进行预测。

综上所述,针对含噪声的网络流量预测本文提出的网络流量集成式在线预测模型框架如图2所示。

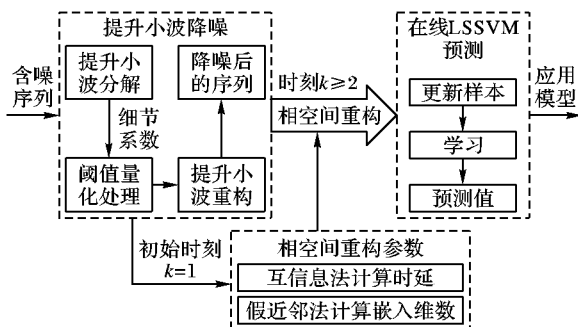


图2 预测模型框架

2.2 在线降噪预测算法流程

本文假设已获得一段平稳流量数据,取滑动窗口长度为 l 的流量数据进行降噪,计算其嵌入维 m 与时延 τ ,并在以后的相空间重构过程中均采用此参数。

为了体现本文方法的在线特性,算法流程仅对时刻 $k \geq 2$ 的情况进行描述。

- 1) 对 $k(k \geq 2)$ 时刻滑动窗口中长度为 l 的流量数据 $S_j = \{s_k, s_{k+1}, \dots, s_{k+l}\}$ 进行提升小波分解,分解为细节系数和近似系数;
- 2) 确定处理细节系数的阈值 thr ;
- 3) 对细节系数进行阈值量化处理;
- 4) 利用经过阈值量化处理后的细节系数和近似系数进行提升小波重构,得到降噪后的流量数据 \hat{S}_j ;
- 5) 依据已计算出的相空间重构参数对 \hat{S}_j 进行相空间重构,然后进行 LSSVM 迭代学习并计算出预测值 \hat{y} ;
- 6) $k+1 \rightarrow k$, 返回1)继续。

3 实验与分析

实验所采用的原始流量样本数据来自 <http://ita. ee. lbl. gov/html/contrib/LBL-TCP. html> 中的实际流量数据 LBL-tcp-3. tcp。此数据采样时间共 2 h,数据共 1 789 995 个。采用 1 s 为间隔进行重采样,得到长度为 7 199 的流量序列,并对此序列的前 3 000 个采样点进行实验。

3.1 数据在线降噪

首先对实验数据 $ns(i)$ 进行归一化,取窗口长度 $W_i = 256$,然后对归一化后的数据进行 3 层提升小波在线降噪,滑动窗口每次的计算复杂度为 $O\left[3 \times \lg \frac{256-1}{N-1}\right]$,平均降噪时间为 0.023 7 s,远远低于采样时间,满足在线要求。降噪后数据为 $ns'(i)$,如图3所示。

由于网络流量序列是一个未知的模型,无法用信噪比、均方根误差和动力学误差来评价降噪效果。故采用二维相图法来评价降噪效果,如图4所示。

由图4可以看出,降噪前的相图呈现出杂乱的伪随机性,而降噪后的相图呈现出清晰、有规律的混沌吸引子自相似结构,说明降噪后恢复了网络流量序列的混沌吸引子特性,取得

了良好的降噪效果。

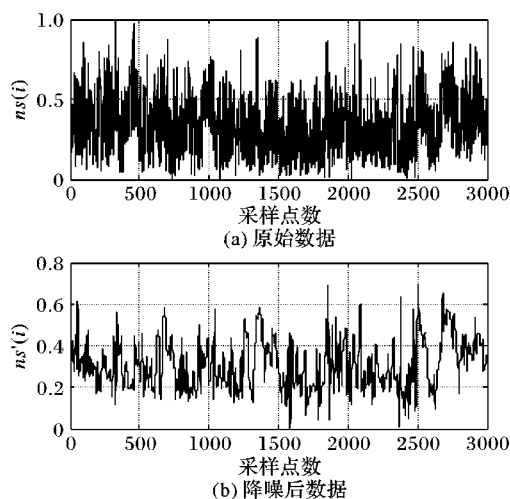


图3 提升小波降噪结果

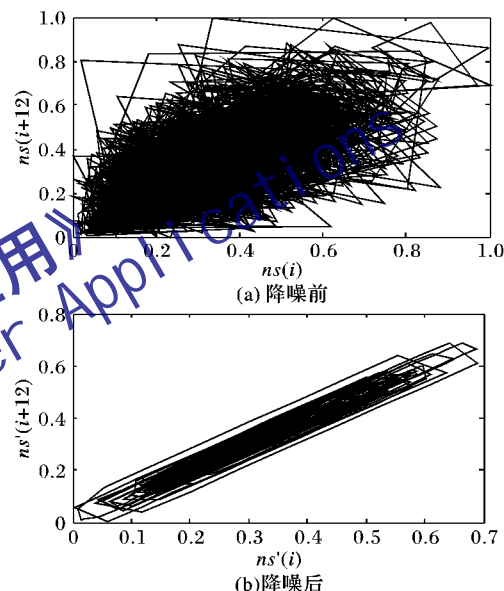


图4 网络流量序列降噪前后二维相图

3.2 实验结果

对原始数据进行在线降噪预测,对降噪后的数据进行相空间重构。选择重构参数时,采用互信息法得到的时延为 $\tau = 12$,利用假近邻法得到的嵌入维数为 $m = 4$ 。本文选择径向基核函数,参数选择 $C = 2000, \delta = 0.5$;窗口长度为 $W_i = 256$,预测范围从采样点 256 到 3 000。窗口每滑动一次迭代预测计算复杂度为 $O(256^2)$,平均每次降噪预测耗时 0.038 s,相对于 1 s 的采样间隔,满足预测的实时性。预测效果如图5所示,可以看出模型较好地预测了流量变化。

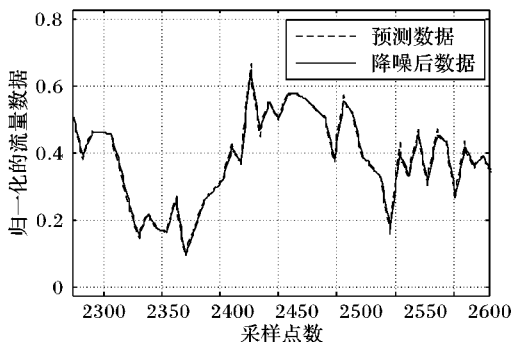


图5 在线降噪预测结果

综合实验结果分析可以得知,因子图一和积算法在精度上略高于或等于联立方程组求解法,但是在时间上完全优于

联立方程组求解法。因此,因子图一和积算法有很好的时间计算复杂度,也具有更好的实用性和可扩展性。

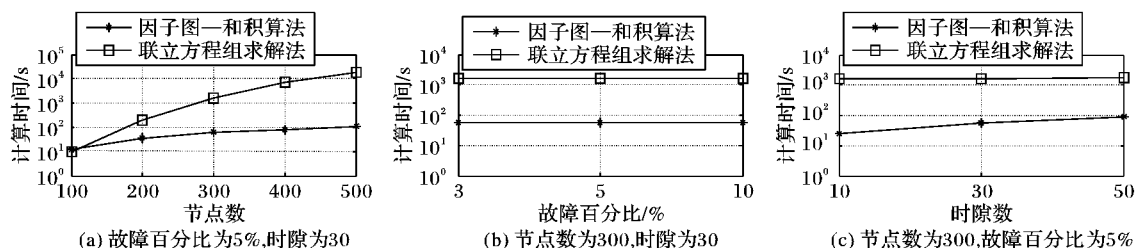


图6 不同情况下的计算时间比较

4 结语

联立方程组求解法通过联立多个线性方程求解网络内部链路的先验故障概率,其精度虽然较高,但随着网络规模增大,它的计算时间增长较快,难以适应实际的大规模网络环境。针对该算法的不足,提出一种基于最大后验准则的因子图一和积算法估计链路故障概率,该方法采用图模型形象地描述链路状态分布和路径状态分布的关系,通过消息传递机制计算链路状态分布的最大后验估计。本文方法在保证较高精度的前提下,能够显著降低计算时间。仿真实验验证了其有效性和可扩展性。

参考文献:

- [1] CASTRO R, COATES M, LIANG G, *et al.* Network tomography: Recent developments [J]. *Statistical Science*, 2004, 19(3): 499 - 517.
- [2] DUFFIELD N G, PRESTI F L, PAXSON V, *et al.* Network loss tomography using striped unicast probes [J]. *IEEE/ACM Transactions on Networking*, 2006, 14(4): 697 - 710.
- [3] CHEN AIYOU, CAO JIN, BU TIAN. Network tomography: identifiability and Fourier domain estimation [C]// IEEE INFOCOM 2007: 26th IEEE International Conference on Computer Communications. Piscataway: IEEE, 2007: 1875 - 1883.
- [4] ERIKSSON B, DASARATHY G, BARFORD P, *et al.* Toward the practical use of network tomography for Internet topology discovery [C]// IEEE INFOCOM 2010. Piscataway: IEEE, 2010: 1 - 9.
- [5] DUFFIELD N G. Network tomography of binary network performance characteristics [J]. *IEEE Transactions on Information Theory*, 2006, 52(12): 5373 - 5388.
- [6] PADMANABHAN V N, QIU L, WANG H J. Server-based inference of Internet link lossiness [C]// INFOCOM 2003: Twenty-Second Annual Joint Conference of the IEEE Computer and Communications Societies. Piscataway: IEEE, 2003, 1: 145 - 155.
- [7] NGUYEN H X, THIRAN P. The boolean solution to the congested IP link location problem: theory and practice [C]// INFOCOM 2007: 26th IEEE International Conference on Computer Communications. Piscataway: IEEE, 2007: 2117 - 2125.
- [8] GHITA D, ARGYRAKI K, THIRAN P. Network tomography on correlated links [C]// IMC '10: Proceedings of the 10th Annual Conference on Internet Measurement. New York: ACM, 2010: 1 - 10.
- [9] NGUYEN H X, THIRAN P. Network loss inference with second order statistics of end-to-end flows [C]// IMC '07: Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement. New York: ACM, 2007: 1 - 13.
- [10] GHITA D, NGUYEN H X, KURANT M, *et al.* Netscope: Practical network loss tomography [C]// IEEE INFOCOM 2010. Piscataway: IEEE, 2010: 1 - 9.
- [11] MAO YONGYI, LI BAOCHUN. A factor graph approach to link loss monitoring in wireless sensor networks [J]. *IEEE Journal on Selected Areas in Communications*, 2006, 23(4): 820 - 829.
- [12] GUO DONG, WANG XIAODONG. Bayesian inference of network loss and delay characteristics with applications to TCP performance prediction [J]. *IEEE Transactions on Signal Processing*, 2003, 51(8): 2205 - 2218.

(上接第342页)

通过对网络流量序列进行降噪,抑制了流量序列中的高频噪声,恢复了混沌吸引子的自相似结构,为相空间重构奠定基础。采用在线更新样本,使模型对新样本也有较好的预测能力。

4 结语

本文提出了融合提升小波降噪和支持向量机回归的在线网络流量预测方法。实验结果表明该方法能有效滤除实时网络流量中的噪声并进行准确预测,该方法对新样本也有较强的适应能力。下一步工作计划结合网络故障管理将本文方法应用于网络故障预测。

参考文献:

- [1] 陆锦军,王执铨.基于混沌特性的网络流量预测[J].南京航空航天大学学报,2006,38(2):217-221.
- [2] 吕金虎,陆君安,陈士华.混沌时间序列分析及其应用[M].武汉:武汉大学出版社,2002.
- [3] 姜明,吴春明,胡大民,等.网络流量预测中的时间序列模型比较研究[J].电子学报,2009,37(11):2353-2358.
- [4] HABIB T, INGLADA J, MERCIER G, *et al.* Support vector reduction in SVM algorithm for abrupt change detection in remote sensing [J]. *IEEE Geoscience and Remote Sensing Letters*, 2009, 6(3): 606 - 610.
- [5] SUYKENS J A K, VANDEWALLE J. Least squares support vector machine classifiers [J]. *Neural Processing Letters*, 1999, 9(3): 293 - 300.
- [6] MALLAT S. A theory for multiresolution signal decomposition: the wavelet representation [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1989, 11(7): 674 - 693.
- [7] SWELDEN W. The lifting scheme: A custom-design construction of biorthogonal wavelets [J]. *Applied and Computational Harmonic Analysis*, 1996, 3(2): 186 - 200.
- [8] DONOHO D L. De-noising by soft-thresholding [J]. *IEEE Transactions on Information Theory*, 1995, 41(3): 613 - 627.
- [9] 叶美盈,汪晓东,张浩然.基于在线最小二乘支持向量机回归的混沌时间序列预测[J].物理学报,2005,54(6):2568-2573.
- [10] MA J S, THEILER J, PERKINS S. Accurate online support vector regression [J]. *Neural Computation*, 2003, 15(11): 2683 - 2703.
- [11] 肖支才,王杰,王永生.基于在线LSSVM算法的变参数混沌时间序列预测[J].航空计算技术,2010,40(3):29-33.