联合 EMD 和 FSVM 的非平稳时间序列预测

龚邦明 王文波 赵 攀
(武汉科技大学理学院 武汉 430065)

摘 要 提出一种基于经验模态分解(EMD)和模糊支持向量机(FSVM)的非平稳时间序列组合预测方法。首先,利用 EMD 对非平稳时间序列进行分解,将其分解为时间尺度特征较为单一的单模态分量,降低待预测信号的非线性复杂度;然后,利用模糊支持向量机对 EMD 分解后的各固有模态函数进行预测;最后将各固有模态函数独立预测的结果进行叠加,即可得到原始序列的预测值。以带噪声的 Lorenz 系统和太阳黑子月平滑值序列为实验数据,对提出的预测方法进行了仿真分析。实验结果表明,与 BP 神经网络预测和传统的 SVM 预测方法相比,提出的方法具有更好的预测精度,而且对带有孤立点、噪声的序列信号具有较强的适应能力。

关键词 非平稳时间序列,经验模态分解,模糊支持向量机,组合预测
中图法分类号 TP181 **文献标识码** A

EMD-FSVM Prediction for Nonstationary Time Series

GONG Bang-ming WANG Wen-bo ZHAO Pan
(College of Science, Wuhan University of Science and Technology, Wuhan 430065, China)

Abstract This paper proposed a novel method to predict non-stationary time series, based on the empirical mode decomposition and fuzzy support vector machine. Firstly, uisng EMD, the non-stationary time series are decomposed into single modal components, reducing the prediction signal nonlinear complexity. Then, using the fuzzy support vector machine, each intrinsic mode function is predicted. Finally, the results predicted by each intrinsic mode function are superimposed to obtain the final forecast. Using Lorenz and sunspot month smooth value sequence with noise as the experimental data, our method was compared with BP neural network prediction and SVM prediction method by experiments. And this method has stronger adaptability to the sequence signal with isolated points and noise, and better prediction accuracy.

Keywords Non-stationary time series, Empirical mode decomposition, Fuzzy support vector machine, Combination forecast

1 引言

科学家对世界的探索永未停止,随着 20 世纪信息时代的到来,大家认知到非线性性才是世界的真谛,非线性理论及其应用技术急需得到发展完善。非线性信号处理、时间序列预测以及非线性控制都是非常重要的研究领域。信号处理、时间序列分析及其预测是非线性控制的前提,精确的预测为非线性控制所提供的价值是不可估量的。然而时间序列、机械振动信号等大都具有非平稳性和非线性性,而且现实中收集的信号因为偶然原因或实际条件的限制常带有孤立点、噪声,导致信号的真实性降低,极大地增加精确预测的难度。

1995 年 Vapnik 等学者基于有限样本统计学习理论提出支持向量机理论(Support Vector Machine,简称 SVM)[1],其在非线性领域应用中有效地解决了过学习、"维数灾难"、局部极小值等问题。为进一步提高 SVM 对带噪声和孤立点的模糊信号的适应能力,Chunfu Liu 等人将模糊理论与支持向量机结合,提出模糊支持向量机(Fuzzy Support Vector Machine,FSVM)[2-8],采用模糊隶属度的形式给模糊信号中的点赋予隶属度,使得噪声点与孤立点拥有较小隶属度,如今 FSVM 已经成功应用到分类、手写字识别以及混沌预测等领域。经验模态分解(Empirical Mode Dcomposition,EMD)根据非平稳信号的局部特性将其分解为有限个固有模态函数(Intinsic Mode Function,IMF)之和,各个 IMF 分量为平稳序列,相互之间信息影响甚微[9-11]。EMD 方法因具有自适应性和多分辨力而在数字信号处理和分析等领域得到广泛应用。非平稳序列预测中用 EMD 方法将信号分解为 IMF 序列,通过 SVM 模型分别预测各个分量的值,经过 EMD 重构过程将各平稳序列预测值相加即可得到原始非平稳信号的预测值。

FSVM 中模糊隶属度和 EMD 端点效应为两种新理论的应用带来了瑕疵。在模糊信号点隶属度设置中,FSVM 采用不同的隶属度函数将产生不同的效应,本文采用基于 Gauss 型隶属度函数方式[4]。EMD"筛分"过程中可能由于 3 次样

本文受国家自然科学基金(11201354)资助。
龚邦明(1989—),男,硕士生,主要研究方向为模式识别、小波分析及应用,E-mail:351gbm@sina.com;王文波(1978—),男,副教授,主要研究方向为多尺度分析理论、信号处理;赵 攀(1987—),男,硕士生,主要研究方向为多尺度分析以及应用。

条拟合而产生人为的影响，并且在多次分解时影响效应向数据内部扩展而“污染”整条数据，因此本文应用镜像延伸法削弱端点效应[12,13]，并提出把模糊支持向量机与经验模态分解相组合的非平稳时间序列预测方法。将带噪声的 Lorenz 系统和太阳黑子序列作为样本，采用 SVM 预测、EMD-FSVM 预测以及 BP 神经网络法进行预测比较，研究表明：EMD-FSVM 对非平稳噪声信号仍然体现出卓越的泛化能力和学习能力，预测精度较高。

2 相关理论

2.1 模糊支持向量机

模糊支持向量机（Fuzzy Support Vector Machine，SVM）是 Chunfu Liu 等人对 Vapnik 提出的 SVM 模型的改进，通过消除噪声的影响来解决传统支持向量机因噪声数据或野值的存在而导致过学习的问题[4,14]。

给定的线性可分性未知的训练集：

$$T'=\{(x_1,y_1),(x_1,y_1)\cdots(x_1,y_1)\}\in(R^n,Y)^l$$

其中，$x_i\in R^n$，$y_i\in Y=R$，$i=1\cdots\cdots l$。构造出相应的基于分类形式的模糊训练集：

$$T=\{((x_i,y_i+\varepsilon),1,\mu_{i+}),((x_i,y_i-\varepsilon),-1,\mu_{i-})\}$$

其中，$x_i\in R^n$，$y_i\in Y(=R)$，$i=1,\cdots,l$，$\mu_{i+},\mu_{i-}\in[0,1]$，$\varepsilon$ 为 SVM 回归超平面带宽。模糊支持向量机回归机问题为求目标函数(1)最小值：

$$\min_{\omega,b}\frac{1}{2}\|\omega\|^2+C\sum_{i=1}^{l}(\mu_{i+}\xi_i+\mu_{i-}\xi_i^*) \tag{1}$$

$$\text{s.t.}\begin{cases}((\omega\cdot\varphi(x_i))+b)-y_i\leqslant\varepsilon+\xi_i\\ y_i-((\omega\cdot\varphi(x_i))+b)\leqslant\varepsilon+\xi_i^*\\ (\xi_i)^*\geqslant 0,i=1\cdots l\end{cases}$$

ω 为回归超平面权值向量，b 为偏差系数，C 为惩罚参数（取常值），ε 为回归超平面带宽，ξ_i,ξ_i^* 为松弛变量（决策函数在$(\varphi(x_i),y_i)$处“损失”值 $\xi_i+\xi_i^*=0$ 当 $|y_i-((\bar{\omega}\cdot\varphi(x_i))+\bar{b})|<\varepsilon$，否则取值为 $|y_i-((\bar{\omega}\cdot\varphi(x_i))+\bar{b})-\varepsilon|$），$\mu_{i+},\mu_{i-}\in[0,1]$表示模糊训练集点在正类集和负类集中的隶属度，非线性变换 φ 使得训练集 T 在映射空间先行可分。

引入 Lagrange 函数，根据对偶理论，将式(1)所述最优化问题转化为如下问题，有：

$$\min_{\alpha_i,\alpha_i^*,\eta_i,\eta_i^*}\ \frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i^*-\alpha_i)(\alpha_j^*-\alpha_j)(\phi(x_i)^T\cdot\phi(x_j))+\varepsilon\sum_{i=1}^{l}(\alpha_i^*+\alpha_i)-\sum_{i=1}^{l}y_i(\alpha_i^*-\alpha_i) \tag{2}$$

满足下述约束条件：

$$\begin{cases}\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)=0\\ \omega=\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)\varphi(x_i)\\ \alpha_i\in[0,\mu_{i+}C],\alpha_i^*\in[0,\mu_{i-}C]\end{cases},i=1,\cdots,l$$

根据 Mercer 定理（核函数的特征[14]）定义的核函数 $K(x_i,x_j)=\varphi(x_i)^T\cdot\varphi(x_j)$，求解对偶问题(2)可得回归函数：

$$g(x)=\sum_{i=1}^{l}(\bar{\alpha}_i^*-\bar{\alpha}_i)K(x_i-x)-b \tag{3}$$

2.2 EMD 及镜像延伸

脉冲信号、振动信号等大都具有非平稳性，传统的 Fourier 分析和小波方法均体现出其固有的缺陷。1998 年 Norden E. Huang 等人提出 Hilbert-Huang 变换，经验模态分解（EMD）为其中的一种处理方法[9]。EMD 方法根据信号的局部时变特征自适应地将信号分解为有限个固有模态函数（IMF）之和，而 IMF 分量均为平稳序列，分别对各个 IMF 分量进行分析能有效改善信号序列分析的问题。

固有模态函数（IMF）要求整条数据区间中极值点数和过零点数必须相等或仅相差一个；并且局部极大值包络线与局部极小值包络线平均值为零，即上下包络线关于时间轴对称。

EMD“筛分”过程[9]将信号分解为有限个 IMF 分量和一个残量：

$$x(t)=c_1+c_2+\cdots+c_n+r_n \tag{4}$$

但是“筛分”过程中引入 3 次样条插值计算上下包络线，由于对端点值的处理方式不同使得端点处有不一样的差异性，并且对 IMF 的不断循环求解，导致差异性向数据内部扩散，可能“污染”整条数据[12,13]。端点可能是局部极大值或局部极小值，端点附近一段时间内的数据若仅有微弱的波动，则同时是极大值和极小值，当然也有可能端点处不是极值点。基于这些原因，有必要对数据进行处理，以便削弱端点效应。

若信号 $x(t)$ 长为 l，引入镜像延伸法将信号从起点复制为 x_L，其长度为 L，其中 $L\in[0.2l,+\infty]$，然后在 x_L 左端点处放置镜面翻转 x_L 使其变为右端点，并与 $x(t)$ 左端点相接。同理，将信号 $x(t)$ 以右端点作为起点（相当于镜面的作用）进行复制，得到长度为 L 的序列 x_R，其中 $L\in[0.2l,+\infty]$，然后将 x_R 左端点与 $x(t)$ 右端点相接。如此便把信号 $x(t)$ 扩展成长为 $L+l+L$ 的数据，仿真实验过程中将以 $0.2l$ 为步长对数据进行镜像延伸，直到 $L=0.2nl$ 与 $L=0.2(n-1)l$ 所获得的原始数据 IMF 分量之差小于阈值为止，并取 $L=0.2nl$ 为镜像延伸长度。

3 EMD-FSVM 预测算法

非线性时间序列预测通常依据历史数据进行相空间重构，选取适当的嵌入维数 m 和延迟时间 τ，得到相空间轨迹：

$$X=[X_1,X_2,\cdots,X_N]^T=\begin{pmatrix}x(t_1-(m-1)\tau) & \cdots & x(t_1)\\ x(t_2-(m-1)\tau) & \cdots & x(t_2)\\ \vdots & & \vdots\\ x(t_N-(m-1)\tau) & \cdots & x(t_N)\end{pmatrix}$$

其中重构后的相点数 $N=n-(m-1)\tau$。由 Takens 嵌入定理[15]可知，在重构后的相空间中存在着相应的映射 $g:R^m\to R^m$，使得

$$X_{i+p}=g(X_i) \tag{5}$$

式中，X_{i+T} 表示 X_i 经 P 步演化后的状态。

本文提出的基于 EMD-FSVM 非平稳序列预测的具体步骤如下：

(1)镜像延伸。对给定的时间信号序列$\{x(t_i),i=1,\cdots,l\}$，选取合适的 L 按照镜像延伸法将其延伸为 $L+l+L$ 长的序列$\{x(t_i),i=1,\cdots,(2L+l)\}$。

(2)EMD 分解。将信号序列$\{x(t_i),i=1,\cdots,(2L+l)\}$按照 EMD 方法进行分解，获得有限个 *IMF* 分量，并截取对应原信号 $x(t)$ 的部分，记为 $x_1(t),x_2(t),\cdots,x_n(t)$ 及残余量序

列 $x_r(t)=x_{n+1}(t)$。

(3)FSVM 训练集构造。针对序列 $x_j(t)(j=1\sim n+1)$选择适当的嵌入维数 m 和延迟时间 τ，进行相空间重构，构造训练样本：

$$X=\begin{bmatrix} x(t_1-(m-1)\tau) & \cdots & x(t_1) \\ x(t_2-(m-1)\tau) & \cdots & x(t_2) \\ \vdots & & \vdots \\ x(t_N-(m-1)\tau) & \cdots & x(t_N) \end{bmatrix}=\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}$$

$$Y=\begin{bmatrix} x(t_1+1) \\ x(t_2+1) \\ \vdots \\ x(t_N+1) \end{bmatrix}=\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

(4)训练集模糊化。选取回归超平面带宽 ε，根据 Gauss 型隶属度函数对训练集前 N'个进行模糊化处理得到$\{(X', Y'+\varepsilon e), +1, \mu_+\}$与$\{(X', Y'-\varepsilon e), -1, \mu_-\}$，将 X 中后 $N-N'$个作为预测集。

(5)FSVM 时间序列预测。将已经模糊化处理的 $2N'$训练集点输入支持向量预测模型进行训练，构造支持向量机预测函数 $y_i=x(t_i+1)=g(x_i)$。然后将原始数据集中后 $N-N'$个点根据预测函数进行预测，得到预测集 Y_j。

(6)重复步骤(3)－(5)求得所有 IMF 分量预测值，并按照 EMD 重构方式求出 $Y_1+\cdots+Y_{n+1}$，得出原始时间序列 $x(t)$的预测值，并与原始数据进行精度评价。

4 实验分析

为了检验本文提出的基于经验模态分解模糊支持向量机非平稳时间序列的预测效果，采用带噪声的 Lorenz 系统和太阳黑子月平滑值序列作为实验数据进行实验分析，并且应用绝对误差对预测结果进行局部评判，同时采用均方误差(MSE)与相关系数进行全局精度评价。

4.1 Lorenz 系统预测

大气环流对天气预报有着重大的指导意义，20 世纪 60 年代 Lorenz 等专家经过长期的研究发现大气流动的数学模型——Lorenz 系统，其具有高度的非线性性，并且在一定情况下呈现出混沌特征。其诱导方程组如下：

$$dx/dt=\sigma(y-x)$$

$$dy/dt=\rho x-y-xz$$

$$dz/dt=-\beta z+xy$$

研究表明当 $\rho>24.7$ 时表现出混沌、非平稳特性，随 ρ 增大其非平稳度增大，本文选取 $\sigma=16, \rho=45.92, \beta=4$。设初值 $x_0=-5, y_0=0, z_0=5$，利用龙格-库塔方法求方差的数值解，取采样间隔 $\Delta T=0.05$，对其 x 相进行采样有采样点集 T'，共有 $N=3000$ 个点。对 T'中点添加方差为 $g\sigma=0.5$，均值 $gE=0$ 的高斯噪声，并随机选择 $N/200$ 个点为孤立点，随机改变其值为 $x(1\pm1/20)$，得到含有噪声和孤立点的集合 T，选择 T 中前 2500 个点作为 FSVM-EMD 预测训练集$\{L_Train\}$，后 500 个点为效验集$\{L_Test\}$，Lorenz 系统和含有噪声 Lorenz 系统非平稳序列如图 1 所示。然后对 T 进行镜像延伸，并将 EMD 分解到第 4 个 IMF 分量，然后截取出实验用到的 4 个 IMF 分量片段(见图 2)，分别对其前 2500 个点进行 FSVM 训练建模，其中模型参数集为$\{C, \sigma\}$，采用交叉验证的方法选取其最优值。利用 4 个 IMF 分量的后 500 个数据作为预测测试集，将 4 组测试值重构得到原始 Lorenz 系统预测值，并进行预测精度评价。图 3 为 SVM 预测、本文方法的预测值与 Lorenz 系统的实际输出值之间的比较曲线，其绝对误差如图 4 所示。从图 4 可以看出，与 SVM 相比，本文方法的预测精度较高，误差基本小于 0.5，相对较稳定，预测值与系统实际值之间具有很高的吻合度。表 1 为 SVM 预测、BP 神经网络预测与本文方法预测均方误差与相关系数，在噪声的影响下，3 种方法精度都不是很高，但是本文方法预测的均方误差最小，且相关系数最大。

表 1 $m=3$ 时 Lorenz 系统不同方法预测结果比较

	SVM	BP 神经网络	本文方法
均方误差(MSE)	0.0391	0.0427	0.0312
相关系数(R)	0.8968	0.8896	0.9115

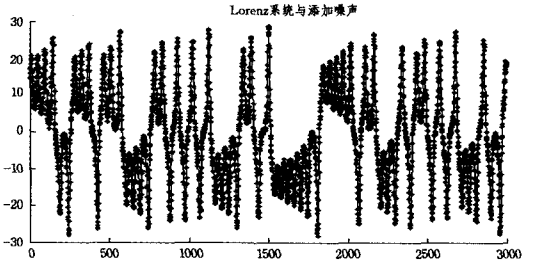

图 1 Lorenz 序列与含噪声的 Lorenz 序列

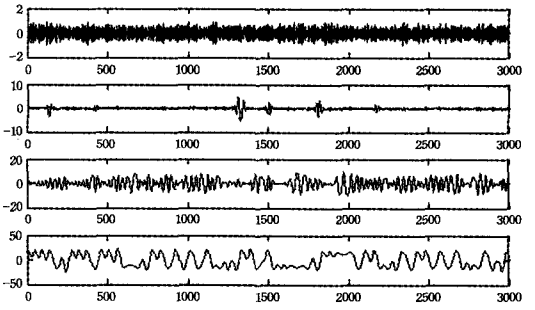

图 2 Lorenz 序列的 4 个 IMF 分量

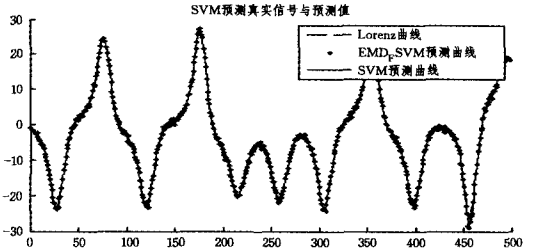

图 3 Lorenz 序列实际值和预测值比较曲线

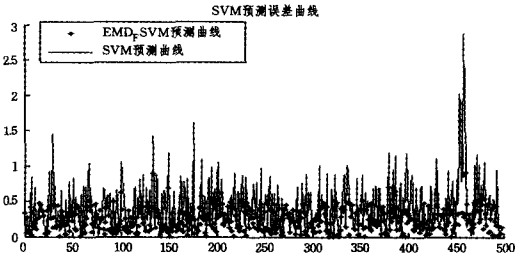

图 4 Lorenz 序列预测绝对误差

4.2 太阳黑子月平滑值预测

太阳黑子是太阳自身发生的一种变化，通常以黑斑的形

式、单独或成群地出现在太阳表面。太阳作为太阳系的主体，它的变化对其内的成员，包括人类的家园——地球，拥有不可忽视的影响。研究表明太阳黑子数的多少将影响厄尔尼诺、拉尼娜现象[16]，还与地球强震活动有关[17]，而由太阳黑子引发的风暴则直接对近地空间电磁辐射产生影响[18]。人们对太阳黑子的观测可以追溯到 1700 年，本文将选取 1751 到 2010 年的太阳黑子月平滑数作为训练集，该数据来源于太阳影响数据分析中心。

令该数据集为 T，共有 260 年即 3120 个月的数据。选择 T 中前 3000 个月(250 年)作为 FSVM-EMD 预测训练集{*ssn_Train*}，后 120 个月(10 年)为效验集{*ssn_Test*}。对 T 进行镜像延伸，将 EMD 分解到第 3 个 IMF 分量，并分别对每个分量前 3000 个数据进行 FSVM 训练建模。利用 3 个 IMF 分量的后 120 个数据作为预测测试集，将 3 组测试值进行组合重构得到原始太阳黑子月平滑值预测值，并进行预测精度评价。本文方法的预测值与太阳黑子月平滑值的实际输出值之间的比较曲线如图 5 所示，其绝对误差如图 6 所示。从图 6 可以看出，本文方法与 SVM 方法绝对误差均很小(0.01 左右)，但 SVM 在某些点处出现大幅值的波动(误差为 0.25)。表 2 为 SVM、BP 神经网络以及本文方法的均方误差和相关系数的比较，可以看出，本文方法在精度提升方面有一定优势。

表 2　$m=1$ 时太阳黑子月滑值不同方法预测结果比较

	SVM	BP 神经网络	本文方法
均方误差(MSE)	0.00220	0.00287	3.732e−04
相关系数(R)	0.9797	0.9755	0.9999

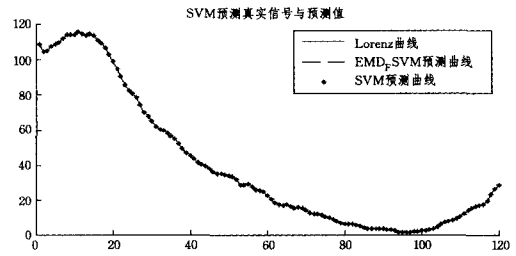

图 5　太阳黑子月滑值实际值和预测值比较曲线

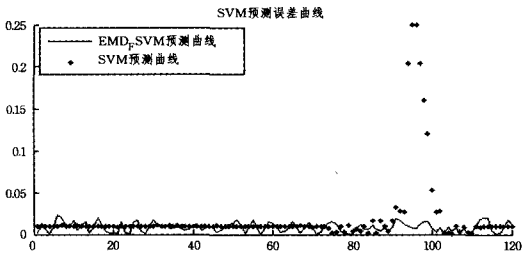

图 6　太阳黑子月滑值预测绝对误差曲线

结束语　本文将模糊支持向量机(FSVM)与经验模态分解(EMD)结合应用到非平稳时间序列的预测中。EMD 可将非平稳序列按频态分解为固有模态分量(IMF)，使单个 IMF 的非平稳性大大降低，针对单个 IMF 分量的预测可获得更高的精度。模糊支持向量机具有卓越的非线性拟合能力，并且针对训练集具有更高的自适应能力。本文提出联合 EMD 与 FSVM 的组合预测方法，首先对预测数据进行 EMD 分解，然后针对每一个分量进行 FSVM 建模训练，之后进行重构得到原始数据预测模型。以带噪声的 Lorenz 系统和太阳黑子月平滑值序列为实验数据，对本文方法的预测精度进行了仿真分析，并与经典的 SVM、BP 神经网络等预测方法进行了对比。实验结果表明：基于 EMD-FSVM 非平稳序列预测可以获得更高的预测精度，而且该方法对噪声序列预测有较强的稳健性，非常适用于非平稳时间序列的预测分析。

参 考 文 献

[1] Vapnik V. The nature of statistical learning theory[M]. Springer, 2000

[2] Chapelle, Olivier. Training a support vector machine in the primal[J]. Neural Computation, 2007, 19(5): 1155-1178

[3] 杨晓伟，郝志峰. 支持向量机算法分析与设计[M]. 北京：科学出版社，2013

[4] 阳爱民. 模糊分类模型及其集成方法[M]. 北京：科学出版社，2008

[5] 马芳芳，仝卫国，宋雨倩. 模糊支持向量机的研究与应用[J]. 电脑与信息技术，2013(1)：25-29

[6] 张永，迟忠先. 基于时间序列的模糊支持向量回归[J]. 计算机工程，2007，33(19)：47-48

[7] Sun Z, Sun Y. Fuzzy support vector machine for regression estimation[C]// IEEE International Conference on Systems, Man and Cybernetics, 2003. IEEE, 2003, 4: 3336-3341

[8] Batuwita R, Plalde V. FSVM_CIL: fuzzy support vector machines for class imbalance learning[J]. IEEE Transactions on Fuzzy Systems, 2010, 18(3): 558-571

[9] 于德介，程军圣，杨宇. Hilbert-Huang 变换在齿轮故障诊断中的应用[J]. 机械工程学报，2005，41(6)：102-107

[10] 玄兆燕，杨公训. 经验模态分解法在大气时间序列预测中的应用[J]. 自动化学报，2008，34(1)：97-101

[11] Huang N E, Shen Z, Long S R, et al. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis[J]. Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences, 1998, 454(1971): 903-995

[12] 吴炳胜，徐芮，姜金俊. 基于 EMD-SVM 镜像延拓的转子故障诊断研究[J]. 河北工程大学学报：自然科学版，ISTIC，2012，29(1)

[13] 程军圣，于德介，杨宇. Hilbert-Huang 变换端点效应问题的探讨[J]. 振动与冲击，2005，24(6)：40-42

[14] 邓乃扬，田英杰. 支持向量机——理论、算法与拓展[M]. 北京：科学出版社，2009

[15] Stark J, Broomhead D S, Davies M E, et al. Takens embedding theorems for forced and stochastic systems[J]. Nonlinear Analysis: Theory, Methods & Applications, 1997, 30(8): 5303-5314

[16] 赵佩章，陈健，赵文桐. 太阳黑子对厄尔尼诺，拉尼娜的影响[J]. 地球物理学进展，2001，16(3)：85-90

[17] 方炜，刘春，张春生. 太阳黑子与全球强震活动[J]. 高原地震，2003，15(4)：27-31

[18] 庄得新，周玉芳. 太阳黑子活动对近地空间的电磁辐射影响[J]. 北京理工大学学报，2005(1)