

基于 EMD-SVM 的江水浊度预测方法研究

王军栋, 齐维贵

(哈尔滨工业大学电气工程及自动化学院, 黑龙江哈尔滨 150001)

摘 要: 针对江水浊度序列宽频、非线性、非平稳的特点, 将经验模态分解 (EMD) 和支持向量机 (SVM) 回归方法引入浊度预测领域, 建立了基于 EMD-SVM 的浊度预测模型. 通过 EMD 分解, 将原始非平稳的浊度序列分解为若干固有模态分量 (IMF), 根据各 IMF 序列的特点, 选择不同的参数对各 IMF 序列进行预测, 最后合成原始序列的预测值. 将该方法应用于实际浊度预测, 并与径向基神经网络 (RBF) 预测及单独支持向量机回归预测结果进行比较, 仿真结果表明该方法预测精度有明显提高.

关键词: 浊度; 预测; 经验模态分解; 支持向量

中图分类号: TN911.7

文献标识码: A

文章编号: 0372-2112 (2009) 10-2130-04

Prediction of River Water Turbidity Based on EMD-SVM

WANG Jun-dong, QI Wei-gui

(School of Electrical Engineering and Automation, Harbin Institute of Technology, Harbin, Heilongjiang 150001, China)

Abstract: Due to the nonlinear and nonstationary characteristics of river water turbidity, a novel intelligent forecasting method based on empirical mode decomposition (EMD) and support vector machines (SVMs) is proposed. The intrinsic mode functions (IMFs) are adaptively extracted via EMD from a time series of turbidity according to the intrinsic characteristic time scales. Then tendencies of these IMFs are forecasted with SVMs respectively, in which the kernel functions are appropriately chosen with these different fluctuations of IMFs. Finally these forecasting results are combined to output the ultimate forecasting result. The proposed model is applied to a water turbidity tendency forecasting example, and the simulation results show that the forecasting performance of the hybrid model outperforms SVMs and RBF ahead forecasting.

Key words: turbidity; prediction; empirical mode decomposition; support vector

1 引言

松花江水浊度受地表径流、温度以及人类活动等的影响, 波动明显, 在不同的月份有着很大的变化, 表现出非平稳、非线性的特点. 对其进行分析和预测, 对于河流生态评价、航运安全以及以江河水为原水的饮用水生产具有重要的指导意义.

国内外在浊度序列分析方面的研究文献较少, 通常都是综合考虑各种水质参数而对浊度进行预测, 采用较多的是人工神经网络等非线性模型方法^[1,2]. 这种模型结构复杂, 要求原始数据丰富, 在实际操作中实现较为困难. 此外, 对于江水浊度这一具有宽带频谱的小样本混沌时间序列, 采用单一的预测方法, 将会把原始浊度序列中的各种不同特征信息同质化, 势必影响其预测精度. 采用经验模态分解 (Empirical Mode Decomposition, EMD) 将浊度序列分解后分别预测, 再进行合成将可能提高其预测精度. 不同于小波变换, 在对信号进行经验

模态分解时不需要先验基底, 每一个固有模态函数 (Intrinsic Mode Function, IMF) 包含的频率成分不仅与采样频率有关, 并且还随着信号本身的变化而变化, 具有自适应性. 能够把局部时间内含有的多个模态的非线性、非平稳信号分解成若干个彼此间影响甚微的基本模态分量, 这些分量具有不同的尺度, 从而简化系统间特征信息的干涉或耦合^[3]. 支持向量机 (Support Vector Machines, SVM) 是建立在统计学习理论上的一种机器学习方法, 是目前针对小样本统计估计和预测学习的较好方法^[4], 对统计学习理论的发展起到巨大推动作用并得到广泛应用^[5~8]. SVM 有良好的泛化能力, 并解决了模型选择与欠学习、过学习问题及非线性问题, 避免了局部最优解, 克服了“维数灾难”, 且人为设定参数少, 便于使用, 已成功应用于许多分类、识别和回归问题^[5,6,8].

根据江水浊度序列的特点, 结合 EMD 和 SVM 两种方法的不同功能, 本文提出了基于 EMD-SVM 模型的预测方法, 用于江水浊度的预测.

2 基本理论

2.1 经验模态分解(EMD)

假设任一信号都是由若干固有模态函数 IMF 组成的,任何时候,一个信号都可以包含多个固有模态信号.固有模态信号是满足以下两个条件的信号:

(1) 整个数据中,零点与极点数相等或至多相差 1;

(2) 信号上任意一点,由局部极大值点确定的包络线和由局部极小值点确定的包络线的均值均为 0,即信号关于时间轴局部对称.

对任一信号 $s(t)$, 首先确定出 $s(t)$ 上的所有极值点,然后将所有极大值点和所有极小值点分别用一条曲线连接起来,使两条曲线间包含所有的信号数据.将这两条曲线分别作为 $s(t)$ 的上、下包络线.若上、下包络线的平均值记作 m , $s(t)$ 与 m 的差记作 h , 则:

$$s(t) - m = h \quad (1)$$

将 h 视为新的 $s(t)$, 重复以上操作,直到当 h 满足一定的条件(如 h 变化足够小)时,记

$$c_1 = h \quad (2)$$

将 c_1 视为一个 IMF, 再作

$$s(t) - c_1 = r \quad (3)$$

将 r 视为新的 $s(t)$, 重复以上过程,依次得第二个 IMF c_2 , 第三个 IMF c_3 , ... 当 $cn(n=N)$ 或 r 满足给定的终止条件(如分解出的 IMF 或残余函数 r 足够小或 r 成为单调函数)时,筛选过程终止,得分解式:

$$s(t) = \sum_{i=1}^n c_i + r \quad (4)$$

其中, r 称为残余函数,代表信号的平均趋势.

2.2 支持向量机(SVM)

对于给定的非线性样本数据 $\{(x_i, y_i) | i = 1, 2, \dots, k\}$, (其中 $x_i \in R_n$ 为样本输入, $y_i \in R_n$ 为样本输出), 利用非线性映射 (\cdot) 将训练数据集非线性映射到一个高维特征空间(Hilbert 空间), 将在输入空间中的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.

设函数形式为:

$$f(x) = \sum_{i=1}^n (x_i) + b \quad (5)$$

寻找结构风险最小化函数的 $f(x)$, 使下式最小:

$$R_{reg} = \frac{1}{2} \sum_{i=1}^n (x_i)^2 + R_{emp} \quad (6)$$

式中: R_{reg} 表示结构化风险; $R_{emp}(\cdot) = \frac{1}{n} \sum_{i=1}^n L(y_i, f(x_i, \cdot))$ 反映了训练误差即经验风险; $\sum_{i=1}^n (x_i)^2 = \|\cdot\|^2$ 反映了模型的复杂度; λ 决定着经验风险和正则化部分之间的平衡,称为平衡系数或惩罚因子.通过引入松弛变量 ξ_i

和 ξ_i^* , 式(7)转化为求解约束最优问题:

$$\min_{b, \xi, \xi^*} J = \frac{1}{2} \sum_{i=1}^n (x_i - \xi_i - \xi_i^*)^2 + \sum_{i=1}^n (\xi_i + \xi_i^*) \quad (7)$$

$$s. t. \begin{cases} y_i - \sum_{i=1}^n (x_i) - b + \xi_i \\ \sum_{i=1}^n (x_i) + b - y_i + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases}$$

该问题实际上成为二次规划问题,通过构造 Lagrange 函数对 b, ξ, ξ^* 求最小化,得到回归决策函数为:

$$f(x) = \sum_{i=1}^n (\xi_i - \xi_i^*) (x_i \cdot x) + b \quad (8)$$

由于在高维空间的内积运算相当复杂,根据泛函理论,一种满足 Mercer 条件的核函数对应某一变换空间中的内积,即:

$$K(x_i, x) = (x_i) \cdot (x) \quad (9)$$

通过这种核函数技术,可以将低维空间的非线性运算转化为高维空间的线性运算,大大简化运算量,从而得到支持向量机回归方程的估计式:

$$f(x) = \sum_{i=1}^n (\xi_i - \xi_i^*) K(x_i, x) + b \quad (10)$$

3 基于 EMD-SVM 的浊度预测模型

通过支持向量机学习一个时间序列模型的最简单方法就是将时间序列的延迟样本作为支持向量机的输入样本.时间序列越复杂,则需要的过去信息就越多.经验模式分解不仅使原始信号中包含的信息通过各基本模态分量得以充分体现,而且还简化了系统间特征信息的干涉或耦合.对各基本模态分量分别进行支持向量机学习时,不仅所需要的过去信息明显减少,而且网络训练的迭代次数明显减少,大大简化了学习任务.

本文尝试将经验模态分解(EMD)和支持向量机结合(EMD-SVM),即通过 EMD 将原始浊度序列分解为若干固有模态分量,对分解得到的各序列分别采用具有支持向量机回归方法进行模拟预测,再将各序列得到的预测值合成重构得到整个浊度序列的预测值,模型如图 1.

设第 t 年第 i 天的浊度值 $L^*(t, i)$ 可由前 q 个历史浊度值进行预测,基于 SVM 的时间序列模型可表述为:

$$L^*(t, i) = \{L^*(t-1, i), L^*(t-2, i), \dots, L^*(t-q, i)\} \quad (11)$$

式中: (\cdot) 为非线性映射; q 为模型阶数.该模型为多输入、单输出的 SVM 模型.

剔除原始数据中的异常点,并将数据归一化到 $[0, 1]$, 通过相空间重构得到输入向量和输出向量.将各序列预测值反归一化即可绘制浊度预测曲线.

以平均相对误差(MRE)评价预测精度:

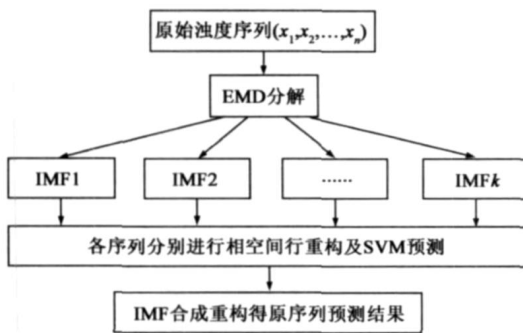


图1 预测模型示意图

$$MRE = \frac{1}{n} \sum_{i=1}^n \left| (y_i - \hat{y}_i) / y_i \right| \times 100\% \quad (12)$$

其中, y_i 为真实值, \hat{y}_i 为预测值。

(1) 原始浊度序列的 EMD 分解

采用镜像法对数据端点进行延拓,以减弱端点效应,将原始浊度序列分解为一系列基本模态分量。

(2) 模型的输入输出样本选取

对训练样本数据进行相空间重构,即将一维的时间序列转化为矩阵形式,获得数据间的关联信息,以尽可能大地挖掘数据的信息量。可通过嵌入窗法、GP算法和C-C算法等方法来确定嵌入维数 m , 和延迟时间。通过相空间重构,来构造预报样本 $\{x_i, y_i\}$, 其中: x_i 为 m 维向量。

用 X 和 Y 分别表示输入和输出样本,则有:

$$X = \begin{bmatrix} x_1 & x_2 & \dots & x_m \\ x_2 & x_3 & \dots & x_{m+1} \\ \dots & \dots & \ddots & \dots \\ x_{n-m} & x_{n-m+1} & \dots & x_{n-1} \end{bmatrix}, Y = \begin{bmatrix} x_{m+1} \\ x_{m+2} \\ \dots \\ x_n \end{bmatrix} \quad (13)$$

(3) 模型参数选取及核函数的确定

支持向量机参数的选择,即不敏感损失函数和误差惩罚因子。影响支持向量的数目,值越大,支持向量数目就越多,估计的函数精度越低,反之亦然;取得小,训练误差变大,系统的泛化能力变差,值越大, $\frac{1}{2}^T$ 的权重就小,同样泛化能力下降。以平均相对误差最小为寻优条件,用网格搜索法并通过试算分别对各序列的输入输出矩阵进行参数选择,得到不同序列的最优参数。

不同的核函数决定了不同特征空间的结构。目前常用的核函数有线性函数、多项式函数、径向基函数和Sigmoid函数等。本研究针对不同的输入输出对象分别采用了线性核函数和径向基核函数。

4 仿真

4.1 EMD-SVM 浊度预测实验

本研究采用以松花江水为生产原水的哈尔滨某制

水厂 2005 年全年 365 天的浊度数据构成时间序列,对其进行 EMD 分解,得到 7 个固有模态分量。

选取 180 个数据点为训练样本,15 个数据点作为测试样本。以原始数据为样本,采用 C-C 算法^[9]确定嵌入窗 w ,从而确定嵌入维 $m=10$,延迟时间 $\tau=1$ 。进行 15 天的浊度预测,得到各子序列的预测结果,如图 2。

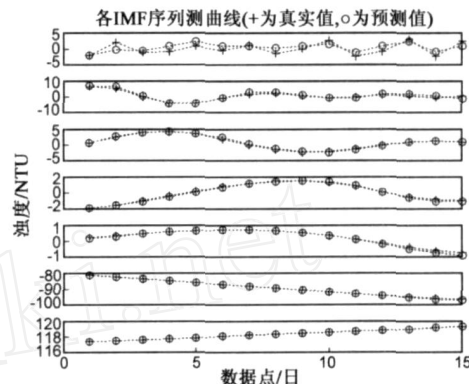


图2 各IMF预测曲线(从上至下依次是IMF1,...,IMF6,Res.)

由图 2 可以看到, SVM 对高频信号预测效果并不理想,随着 IMF 频率的降低,预测精度逐渐提高,这种趋势可以从误差上反映出来,见表 1。

表 1 各 IMF 预测误差(%)

样本	IMF1	IMF2	IMF3	IMF4	IMF5	IMF6	Res.
MRE	77.10	94.81	23.05	17.23	13.05	2.8189e-002	3.7943e-003

而另一方面,预测误差较大的序列,较之误差较小的序列其幅值较小,所以当将各 IMF 预测序列合成后,大误差序列对预测结果影响甚微。

对于样本序列 IMF2, IMF3, ..., IMF7, 当采用各种核函数时得到的预测精度基本相同,在此,本文选择径向基核函数 ($K(x, y) = e^{-\|x-y\|^2/2}$)。对于 IMF1, 采用不同的核函数得到的预测精度差别较大,故选择了在此处精度较高的线性核函数 ($K(x_i, x_j) = x_i^T x_j$)。

将各 IMF 预测序列合成重构,得到原序列的预测曲线如图 3。

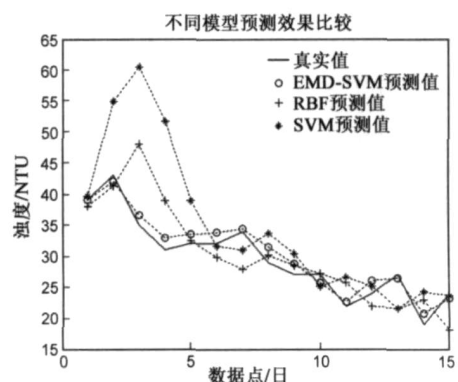


图3 不同模型浊度预测曲线比较

4.2 对比分析

分别采用 RBF 神经网络模型和单 SVM 模型对该组数据进行预测. 其中, RBF 模型径向基函数分布值选为 8, 隐层节点数选为 13; svm 模型嵌入维 $m = 10$, 延迟时间 $\tau = 1$, 核函数选为径向基核函数. 得到预测曲线如图 3.

可见, 采用 EMD-SVM 模型后, 曲线的拟合程度较之直接对序列进行 SVM 预测及用 RBF 进行预测有很大提高, 平均相对误差由 20.73 % 和 13.08 % 降至 4.65 %, 精度明显提高, 误差对比见表 2.

表 2 误差比较(%)

模型	EMD-SVM 模型	SVM 预测模型	RBF 预测模型
MRE	4.65	20.73	13.08

5 结论

本文对江水浊度预测建模进行了研究, 提出基于 EMD-SVM 的预测模型, 用 EMD 跟踪江水浊度序列的非平稳特性, 用 SVM 跟踪序列的非线性和小样本特性. 通过仿真分析得出以下结论:

(1) 采用 EMD 对原始信号进行分解, 显微式的提取了信号的原有信息, 对得到的一系列 IMF 采用相应的模型参数和核函数进行 SVM 建模预测, 仿真结果表明该模型用于江水浊度预测是有效的.

(2) 通过与 SVM 模型以及 RBF 模型进行预测对比, 表明该 EMD-SVM 预测模型的精度较高.

(3) 对不同特点的数据序列, 其模型参数以及核函数, 要有不同的选择. 对于大多数时间序列的预测, 支持向量机核函数的选择具有任意性, 但针对高频浊度信号, 线性核函数预测精度较高.

(4) 基于 EMD-SVM 的浊度预测模型, 既能体现浊度的变化趋势, 又能对水文突变进行有针对性的预测, 可用于河流水质监测, 也可作为制水厂投药控制的依据.

参考文献:

- [1] Robert J. May, et al. Application of partial mutual information variable selection to ANN forecasting of water quality in water distribution systems[J]. Environ. Model. Softw, 2008, 23(10-11): 1289-1299.
- [2] Bowden, G. J., 2003. Forecasting Water Resources Variables Using Artificial Neural Techniques[D]. Ph. D, University of Adelaide.
- [3] HUANG N E. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stastinary time series analysis[A]. Proceedings of the Royal Society of London series A-Mathematical Physical and Engineering Sciences[C]. London, 1998.
- [4] Vapnik V N. The nature of statistical learning theory[M]. New York: Springer-Verlag, 2000.
- [5] Guo G D, Li S Z. Contents based audio classification and retrieval by support vector machines[J]. IEEE Trans on Neural Network, 2003, 14(1): 209-215.
- [6] 艾玲梅, 王钰. 基于双谱分析和支持向量机的手震颤加速度信号识别[J]. 电子学报, 2008, 11: 2165-2170.
- [7] Li S Z. Contents based classification and retrieval of audio using the nearest feature line method[J]. IEEE Trans on Speech Audio Processing, 2000, 8(9): 619-625.
- [8] 赵登福, 王 蒙, 张讲社, 等. 基于支撑向量机方法的短期负荷预测[J]. 中国电机工程学报, 2002, 22(4): 26-30.

作者简介:



王军栋 男, 1974 年 12 月出生于黑龙江省哈尔滨市. 现为哈尔滨工业大学电气工程及自动化学院博士研究生. 主要研究方向为信号处理及预测控制等.
E-mail: lili161711@sina.com

齐维贵 男, 1944 年 12 月出生于辽宁黑山, 教授, 博士生导师, 研究方向为先进控制策略, 现代信号处理技术及节能控制技术等.