

基于小波分解和最小二乘支持向量机的 COD 预测

刘 伟, 李春青, 张艳芬

(天津工业大学 计算机科学与软件学院, 天津 300387)

摘 要: 膜生物反应器(MBR)的化学需氧量(COD),其值大小反映污水处理系统是否良好运行,对整个系统正常运行有重要意义. 我们针对 COD 的非线性特征,引入基于小波分析和最小二乘向量机(LSSVM)的预测模型(WLSSVM). 利用 Mallat 塔式分解算法获取趋势项与随机项,然后利用最小二乘支持向量机对时间序列分别预测,最后将各尺度下的分量整合作为预测值. 试验结果表明,该模型具有较高的精度,是科学可行的.

关键词: 膜生物反应器;时间序列;最小二乘支持向量机;小波分解

中图分类号: TP181

文献标识码: A

文章编号: 1000-7180(2013)04-0026-04

Prediction of COD Time Series Based on Wavelet and LS-SVM

LIU Wei, LI Chun-qing, ZHANG Yan-fen

(Dept of Computer Science and Software, Tianjin Polytechnic University, Tianjin 300387)

Abstract: MBR (Membrane Bio-Reactor)'s COD (chemical oxygen demand) have important effects on entire system. We design our wavelet transformation and LSSVM (Least Squares Support Vector Machine) method based on COD's nonlinear characteristic. First, we use Mallat deconstruction and reconstruction arithmetic to get trend and random time series, and then we use LSSVM to predict separately. Finally we combine every time series to get final result. Through simulation, our method show higher precision.

Key words: MBR; time series; LSSVM; wavelet decomposition

1 引言

当今世界水资源匮乏,据估计到 2025 年,全球百分之八十的人口将处于缺水地区^[1]. 水资源的合理利用以及循环利用,将产生巨大市场价值,同时也将产生深远的社会意义. 膜生物反应器(Membrane Bio-Reactor, MBR)是一种由膜分离技术与生物处理技术有机结合的新型污水处理系统^[2]. COD 其中文名称为化学需氧量,其值大小能反应水中微生物的情况,该值越大说明出水中微生物越多,水质越差.

对 COD (chemical oxygen demand) 的预测可以更好地了解系统运行的状况,有利于提前为 MBR 系统可能发生的情况做好准备^[3-4]. 当前对膜生物反应器系统的仿真,大多集中于自回归、神经网络、支持向量机等方法. 自回归方法由于自身原理限制

对于非线性问题预测效果较差;神经网络以及衍生方法,存在过学习的现象,泛化能力差;支持向量机虽然效果较好,但数据量较大时,运算时间长不适合实时运行. 小波预测方法,常应用于金融领域^[5],在污水处理系统至今没有应用. 我们通过观察发现 COD 的时间序列与金融时间序列具有很多相似的地方. 本文结合以上现象,提出基于小波变换和最小二乘支持向量机的模型 WLSSVM,应用于污水处理系统,具有良好的精度.

2 WLSSVM 模型

本论文在小波和最小二乘支持向量机预测分析时间序列的基础上提出 WLSSVM(Wavelet Least Squares Support Vector Machine)模型.

采用 WLSSVM 模型的首要原因是 COD 时间

收稿日期: 2012-07-18; 修回日期: 2012-08-25

基金项目: 国家自然科学基金项目(40808130)

序列具有波动性,引入小波分解作为预处理手段,可使时间序列更加平滑,预测值更加准确;另外,各尺度中不同的时间序列通过不同参数的最小二乘支持向量机进行 NARX 非线性回归,最后引入带有权值的最小二乘法将输出值整合得到预测值.该模型结构如图 1.

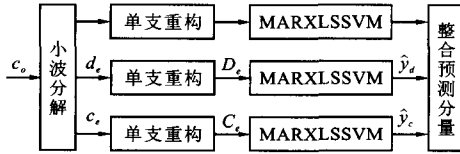


图1 WLSSVM模型过程

WLSSVM模型主要包括以下三部分:小波分解与单支重构、基于NARX(nonlinear autoregressive with exogenous inputs)模型的LSSVM预测及预测分量整合.

2.1 Mallat 塔式小波变换

小波变换与傅里叶变换相比,小波变换是时间(空间)频率的局部化分析,它通过伸缩平移运算对信号(函数)逐步进行多尺度细化,最终达到高频处时间细分,低频处频率细分,能自动适应时频信号分析的要求,从而可聚焦到信号的任意细节,解决了Fourier变换的困难问题,成为继Fourier变换以来在科学方法上的重大突破.小波变换称为“数学显微镜”^[6-7].

小波变换分为离散小波变换与连续小波变换.离散小波变换公式如下

$$WT_f(a, b) = \langle f(t), \psi_{a,b}(t) \rangle \\ = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{+\infty} f(t) \psi^* \left(\frac{t-b}{a} \right) dt \quad (1)$$

其中, $f(t)$ 是输入信号, a 是缩放, b 是平移, $\psi(t)$ 是母小波. 离散小波以 2 的平方作为缩放与平移的基数. 离散小波变换具有零冗余, 高效, 携带足够可供重构的信息等特点. 因此离散小波变换被广泛应用于数据压缩与特征提取. 我们使 $a_0 = 2, b_0 = 1$, 如果 $\psi(t)$ 如下:

$$\psi_{j,k}(t) = 2^{-j/2} \psi(2^{-j}t - k) \quad (2)$$

离散小波变换可由以下公式表示

$$C_{j,k} = \int_{-\infty}^{+\infty} f(t) \psi_{j,k}^*(t) dt = \langle f, \psi_{j,k} \rangle \quad (3)$$

一个多分辨率分析包含一系列连续的向量空间 V_j . 更准确地说, 子空间 V_j 满足下式:

$$V_0 = V_1 \oplus W_1, V_1 = V_2 \oplus W_2, \dots, V_j = V_{j+1} \oplus W_{j+1}, \dots$$

j 是整数. 当 j 越小, 向量空间越大. 当 $j=4$,

得到如图 2 的向量空间.

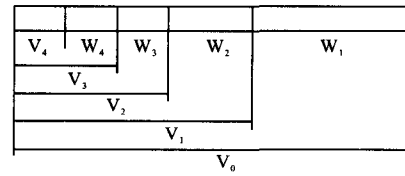


图2 $j=4$ 得到的小波分解

通过 $L^2(R) = \bigcup_{j \in \mathbb{Z}} V_j$ 和 $V_{j-1} = V_j \oplus W_j$, 得到

$$L^2(R) = \bigoplus_{j=-\infty}^{\infty} W_j \quad (4)$$

基于多分辨率分析理论, Mallat 塔式算法可以快速计算离散小波变换. 首先, 我们把信号分别通过高通滤波器与低通滤波器进行分解. 输出序列的长度减半. 经过高通滤波器得到的序列包含整个信号的细节部分, 经过低通滤波器得到的序列包含趋势部分. 把趋势部分作为下一次的输入, 如此循环. 流程如图 3.

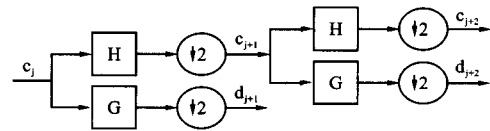


图3 Mallat 塔式小波分解算法

公式如下:

$$c_{j,k} = \sum_m h(m-2k) c_{j-1,m} \quad (5)$$

$$d_{j,k} = \sum_m g(m-2k) c_{j-1,m} \quad (6)$$

重构流程如图 4 所示.

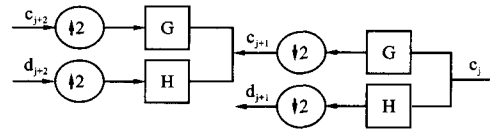


图4 Mallat 塔式小波重构算法

重构公式如下:

$$c_{j-1,m} = \sum_k h(m-2k) c_{j,k} + \sum_k g(m-2k) d_{j,k} \quad (7)$$

2.2 最小二乘支持向量机

SVM 是基于统计学理论的新型机器学习方法. SVM 的核心思想是将低维下线性不可分的问题, 转化为高维, 进而求得低维下的非线性方程. SVM 由于引入结构风险最小化, 使得 SVM 能够处理高度非线性的分类和回归等问题. 它克服了人工神经网络训练时间长、泛化能力差、易陷入局部极小等缺点^[8-9]. 最小二乘支持向量机由 Suykens J. A. K 提出. 是对 SVM 的改进, 将 SVM 中二次规划的约束条件从不等式转为等式, 通过线性求解的方式, 简

化了 SVM 的运算复杂度,并具有唯一全局最优解. 由于其求解过程类似于最小二乘法,故取名最小二乘支持向量机. 原理如下:

对于训练样本 $(x_1, y_1) \cdots (x_l, y_l), x \in R^n$ 为 n 维输入向量, $y \in R$ 为输出向量,求一个最优的 $y(x)$ 表示 x 到 y 的映射关系. 设:

$$y(x) = w^T \varphi(x) + b \quad (8)$$

式中: $\varphi(x): R^n \rightarrow R^k$ 为输入空间到高维特征空间的非线性映射,这种非线性映射可将输入空间的非线性回归问题变为高维特征空间中的线性回归问题. LSSVM 回归算法可表示为求解如下约束优化问题:

最小化泛函:

$$J(w, e) = \frac{1}{2} w^T w + \frac{1}{2} \gamma \sum_{k=1}^N e_k^2 \quad (9)$$

约束条件为:

$$y_k = w^T \varphi(x) + b + e_k, \quad k=1, \dots, N \quad (10)$$

式中, $w \in R^k$ 为权重向量, $e^k \in R$ 为误差变量, b 为偏置量, γ 为调整参数. 根据式(9)可定义 Lagrange 函数:

$$L(w, b, e, \alpha) = J(w, e) - \sum_{k=1}^N \alpha_k \{w^T \varphi(x) + b + e_k - y_k\} \quad (11)$$

式中, α_k 为拉格朗日乘子, $\alpha_k \in R$. 计算 $L(w, b, e, \alpha)$ 对 w, b, e, α 的偏导并置零,可得式(11)的最优条件如下:

$$\begin{cases} \frac{\partial L}{\partial w} = 0 \rightarrow w = \sum_{k=1}^N \alpha_k \varphi(x) \\ \frac{\partial L}{\partial b} = 0 \rightarrow \sum_{k=1}^N \alpha_k = 0 \\ \frac{\partial L}{\partial e_k} = 0 \rightarrow \alpha_k = \gamma e_k \\ \frac{\partial L}{\partial \alpha_k} = 0 \rightarrow w^T \varphi(x) + b + e_k - y_k = 0 \end{cases} \quad k=1, \dots, N \quad (12)$$

通过式(12)则可以求出 w, e (用 α, b 表示). 由于支持向量机采用了核函数方法,在上述非线性变换中只需在原空间中利用满足 Mercer 条件的核函数进行内积运算即可实现^[5],而无需知道非线性映射的具体形式. 设核函数 $K(x_i, x_j) = \varphi(x_i) \varphi(x_j)$, 则求解的优化问题变为求解线性方程:

$$\begin{bmatrix} 0 & 1 & \cdots & 1 \\ 1 & K(x_1, x_1) + 1/\gamma & \cdots & K(x_1, x_N) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & K(x_N, x_1) & \cdots & K(x_N, x_N) + 1/\gamma \end{bmatrix} \begin{bmatrix} b \\ \alpha_1 \\ \vdots \\ \alpha_N \end{bmatrix} = \begin{bmatrix} 0 \\ y_1 \\ \vdots \\ y_N \end{bmatrix} \quad (13)$$

$$= \begin{bmatrix} 0 \\ y_1 \\ \vdots \\ y_N \end{bmatrix}, \quad i, j=1, \dots, N \quad (13)$$

从式(13)中求解出 α 和 b ,即可得到训练样本集的拟合函数:

$$y(x) = \sum_{k=1}^N \alpha_k K(x_i, x_j) + b \quad (14)$$

LSSVM 在回归过程中均采用等式约束,求解优化问题最终转为求解线性方程,计算过程得到了极大的简化,因此求解过程比采用不等式约束的 SVM 要快.

2.3 WLSSVM 模型流程

- (1) 确定小波分解的层数,用 Mallat 小波算法对时间序列进行分解与重构;
- (2) 将时间序列变换成 NARX 矩阵;
- (3) 用原理分析或暴力方法确定 LSSVM 的核函数;
- (4) 用网格搜索法确定 LSSVM 非线性回归的最佳参数 σ^2, γ ;
- (5) 将趋势项与随机项整合,得到预测值;
- (6) 比对预测值与真实值,返回第一步进行调整.

3 实例分析

本文所选实例为石家庄市某污水处理厂三个月来的工程数据,选取出水数据作为试验数据,共 82 组,其中选取前 77 组作为训练数据,后 5 组作为测试数据. TN(含氮量)与 TP(含磷量)作为输入量, COD(需氧量)作为预测量.

3.1 小波分解

经过一层小波分解后得到趋势时间序列与原时间序列降低采样之后对比如图 5.

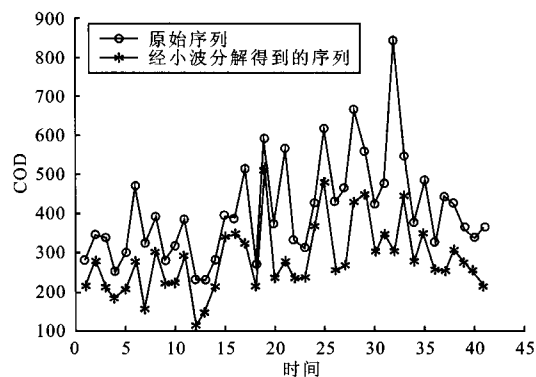


图 5 原始序列与小波分解对比图

3.2 对比分析

本实验采用一层 Mallat 塔式分解与单支重构,最小二乘支持向量机采用 RBF 核函数,趋势序列参数为 $\gamma=4.22862, \sigma^2=11.0247$; 随机序列参数为 $\gamma=57.34555, \sigma^2=235.45786$; cod 的预测值与真实值对比如图 6 所示。

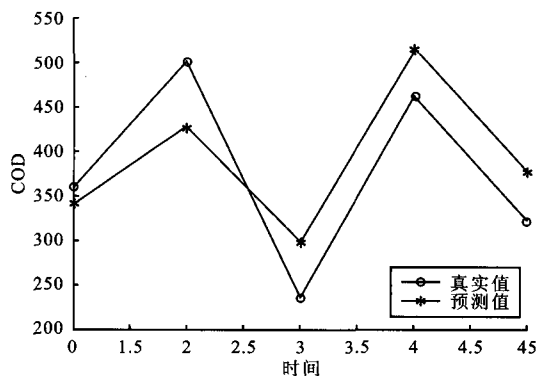


图 6 真实值与预测值对比

从表 1 可以看出,本文提出的 WLSSVM 的预测结果要好于 LSSVM 方法。虽然利用 WLSSVM 方法对 COD 预测的结果令人满意,但预测值与真实值相比差距仍然较大,主要是因为影响膜生物反应系统出水的需氧量因素众多,要对其作出更为接近的预测,必须加入更多因素的考量。在加入更多学习样本之后,预测的精度将会提高。

表 1 两种方法对比

	LSSVM	WLSSVM
MAE(平均绝对误差)	52.2164	50.6712
MAPE(平均相对误差)	0.144042	0.138588

4 结束语

基于 MBR 中需氧量时间序列的波动性的特点,建立了基于小波的 LSSVM 预测模型,并利用该模型对我国石家庄某污水处理厂的实测数据进行验证。与原 LSSVM 模型的预测数据的比较表明该模型预测精度高,泛化能力强,具有较高的可靠度,用于 MBR 出水需氧量预测具有一定优势。考虑到本模型步骤多,相关参数多,用于以上实例的参数未必是最佳的,今后可在参数调优以及参数相互影响上做进一步研究。

参考文献:

- [1] 石虹,浅谈全球水资源态势和中国水资源环境问题[J]. 水土保持研究,2002,9(1):145—150.
- [2] Bouju H, Buttiglieri G, Malpei F. Are MBRs really more efficient in removing pharmaceuticals Comparison of a full scale conventional activated sludge process and a MBR pilot plant [C] // 5th IWA Conference on Membranes for Water and Wastewater Treatment. Beijing: IWA — MTC 2009 Organising Committee, 2009.
- [3] 上海市政工程设计研究总院. CECS 246:2008 给水排水工程顶管技术规程[M]. 北京:中国计划出版社, 2008.
- [4] Qin J J, MH O, Tao G, et al. Osmotic membrane bioreactor: effect of intermittent air scouring on membrane Fouling[C] // 5th IWA Conference on Membranes for Water and Wastewater Treatment. Beijing: IWA — MTC 2009 Organising Committee, 2009.
- [5] Reynaud O, Starck J L, Murtagh F. Predictin based on a multiscale decomposition[J]. International Journal of Wavelets, Multiresolution and Information Processing, 2003,1(2):217—232.
- [6] 邸继征,小波分析原理[M]. 北京:科学出版社,2010.
- [7] 周伟. 基于 MATLAB 的小波分析应用[M]. 西安:西安电子科技大学出版社, 2010.
- [8] Suykens J A K, Vandewalle J. Least squares support vector machine classifiers [J]. Neural Processing Letters, 1999, 9(3): 293—3.
- [9] Li Fangfang, Zhao Yingkai, Jiang Zhibing. The prediction of oil quality based on least squares support vector machines and daubechies wavelet and mallat algorithm[C] // Proceedings of the Sixth International Conference on Intelligent Systems Design and Applications (ISDA'06). China:Jinan, 2006.

作者简介:



刘伟 男,硕士研究生. 研究方向为 MBR 模拟仿真。