一种新型时间序列多分辨预测模型研究

彭喜元,王 军,彭 宇
(哈尔滨工业大学自动化测试与控制系,黑龙江哈尔滨 150080)

摘 要: 使用单一模型实现复杂时间序列预测一直是一个研究热点和难点问题.本文采用经验模式分解方法首先将复杂时间序列分解为一系列本征模式函数之和,然后对各个本征模式进行径向基神经网络预测建模,在此基础上,通过各个分量预测结果的等权求和得出综合预测结果.此外,各 RBF 网络核函数的最优参数对数值与各本征模式分量呈近似线性关系,利用该线性关系可以减少交叉验证求参数的次数,从而降低计算负担.仿真结果表明分解域多 RBF 网络预测模型对复杂时间序列预测性能好于单一的 RBF 网络预测模型.

关键词: 经验模式分解;径向基神经网络;交叉验证方法;时间序列预测
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2007) 11-2146-04

A Novel Multi-Scale Predictor for Complex Time Series

PENG Xi-yuan, WANG Jun, PENG Yu
(*Department of Auto-test and Control, Harbin Institute of Technology, Harbin, Heilongjiang* 150080, *China*)

Abstract: The task of complex time series predicting is hard to be accomplished with only one single predicting model. In this paper, a complex time series is decomposed into a series of intrinsic mode functions and a residue signal. Then a RBF network is constructed for an intrinsic mode function or the residual signal. Finally output of every predicting model is integrated into one output with equal weighted. As the sifting process of EMD is an approximate frequency dividing process, the relationship between the logarithm of optimal parameter of every RBF network and its corresponding Intrinsic Mode Function is also approximate linear. This relationship can be utilized to alleviate the computing burden of the model selecting with cross validation method. Experimental results showed that the proposed method outperformed the single RBF network in the task of predicting complex time series.

Key words: empirical mode decomposition; radial-basis function network; cross validation method; time series prediction

1 引言

时间序列数据在一些新的数据库应用以及数据挖掘等领域中日趋重要,而且这些数据往往是非平稳、非线性随机时间序列,如气象数据、太阳黑子数据、激光数据、股票价格数据、网络流量数据、电力需求数据等.许多时间序列的复杂行为都可以用内在的非线性动力系统加以解释,但从历史数据正确识别非线性动力系统是个难题,究其根源在于这些非线性时间序列所蕴涵的动力系统具有耗散性和初始敏感性等特点.目前已经有很多进行时间序列预测的方法.最为常用的预测方法为线性模型方法,如 AR、ARMA、ARX、ARMAX 等,其优点是模型简单和容易识别,但很难应用于现实中复杂时间序列的处理[1,2].非线性模型可以克服线性模型的缺点,如 MLP 神经网络、RBF 神经网络、模糊逻辑模型、双线性自回归模型 BAR、支持向量回归模型等[3,4].但正如文献[5]指出的那样,一个变化异常复杂的非线性、非平稳随机信号难以使用单一的模型进行有效的预测.因此,本文提出使用经验模式分解方法将复杂时间序列分为一系列比较简单的模式,并分别采用 RBF 神经网络对各简单模式分量进行预测建模,然后将这些分量预测结果集成为综合预测结果.而且利用各个分量参数之间的关系可减轻正交验证求参数的次数,从而大大减轻计算负担.利用 5 个真实复杂时间序列进行预测实验结果显示,本方法具有很好的预测性能.

2 理论基础

2.1 经验模式分解

现有的数据处理方法要么是针对线性非平稳过程的处理方法,如小波变换、Wagner-Ville 分布和短时 Fourier 变换,要么是针对非线性平稳过程的处理方法,如相空间表示法和时间延迟嵌入方法等.小波和短时

收稿日期:2006-12-01;修回日期:2007-02-15

Fourier 变换等方法是基于 Fourier 谱分析的,通过可调时-频窗函数来进行非平稳信号的时-频分解,在处理非平稳信号时可以达到一定的效果,但使用谐波来表示非线性信号时会发生能量向高频泄露,产生虚假频谱成分.相空间表示法和时间延迟嵌入方法在高噪声情况下将无法重构与真实物理过程微分同胚的动力系统[12].所以上述这些方法很难处理现实世界中非线性、非平稳随机信号,人们迫切需求一种新型处理方法.

1998 年,Norden E. Huang 提出一种适用于分析和处理非线性、非平稳随机信号的新方法——HHT 变换(Hilbert Huang Transform, HHT)[6,8].HHT 变换由两步组成:

第一步将任意信号分解为若干本征模式分量(Intrinsic Mode Function, IMF)和一个余项,该步骤称为经验模式分解法(Empirical Mode Decomposition, EMD);

第二步对每个本征模式或余项进行希尔伯特谱分析(Hilbert spectral analysis, HAS).

其中,IMF 分量必须满足下面两个条件:

(1)在整个数据序列中,极值点的数目与过零点的数目必须相等或至多相差一个;

(2)数据序列极大值点确定的上包络和极小值点所确定的下包络关于时间轴对称.

EMD 方法步骤如下,假设在任何信号都由不同的 IMF 组成,每个 IMF 可以是线性的,也可以是非线性的,这样任何一个信号就可以分解为有限个 IMF 之和,则 IMF 可以按以下方法"筛分"(Sifting):

(1)确定信号 $x(t)$的所有局部极值点,将所有极大值点用三次样条拟合形成上包络线,将所有极小值点用三次样条拟合形成下包络线,这两条包络线包络了所有的信号数据;

(2)将两条包络线的均值记为 m_1,求出

$$y_1(t) = x(t) - m_1 \tag{1}$$

(3)判断 $y_1(t)$是否满足 IMF 条件,若 $y_1(t)$不满足 IMF 条件,则将 $y_1(t)$作为原始数据,重复执行步骤(1)、(2),直到 $y_1(t)$满足 IMF 条件,记 $y_1(t) = c_1(t)$,则 $c_1(t)$为信号 $x(t)$的第一个 IMF 分量,它代表信号 $x(t)$中最高频率的分量;

(4)将 c_1 从 $x(t)$中分离出来,即得到一个去掉高频分量的差值信号 $r_1(t)$,即有

$$r_1(t) = x(t) - c_1(t) \tag{2}$$

(5)将 $r_1(t)$作为原始数据,重复步骤(1)、(2)、(3),得到第二个 IMF 分量 $c_2(t)$,重复 n 次,得到 n 个 IMF 分量.

经过一系列分解后,时间序列 $x(t), t = 1,2,\cdots,N$ 可表示成 n 个本征模式分量 $c_i(t)$)和一个余项 $r_n(t)$之和,即:

$$x(t) = \sum_{i=1}^{n} c_i(t) + r_n(t) \tag{3}$$

其中 c_1 到 c_n 的频率从大到小排列,c_1 所含频率最高,c_n 所含频率最低,而 $r_n(t)$为一个单调序列.

2.2 RBF 网络预测建模

文献[7]指出人工神经网络具有优良的非线性时间序列预测性能,且与支持向量回归预测模型进行比较的结果表明:人工神经网络预测模型在许多应用中优于支持向量回归预测模型.

RBF 神经网络是一种全局最优逼近能力的前向神经网络模型,其结构简单、训练简洁且学习收敛速度快,能够逼近任意非线性函数.此外,RBF 神经网络训练不需要求解大规模二次规划问题,在训练速度上优于支持向量机等.因此,本文采用 RBF 神经网络对各个 IMF 进行预测建.RBF 网络是一个包括输入层、隐含层和输出层的多输入单输出系统.隐含层利用径向基函数实现输入向量到高维特征空间的映射.径向基函数有多种形式,通常选择高斯函数.输出层可实现输出权值的线性组合.RBF 神经网络的学习算法主要有:凭借经验选取、有监督的选择中心、无监督的选择中心、正交最小二乘法 OLS 等.本文采用 Matlab 工具箱中的径向基神经网络设计函数进行 RBF 神经网络设计.该径向基网络设计函数为:

$$net = newrbe(P, T, SPREAD) \tag{4}$$

式中:*net* 为返回的 RBF 网络;*newrbe* 为 RBF 网络设计函数名;P 为输入向量;T 为目标输出向量;*SPREAD* 为径向基函数的分布系数,网络越大,网络输出越平滑,网络泛化能力也越强.

3 新型多分辨预测模型及参数选择方法

3.1 新型多分辨预测模型

本文推荐的多分辨预测模型先使用 EMD 算法对非线性非平稳随机序列进行多分辨分解为一系列对称性好的非线性序列(IMF 分量)之和,然后对各个子序列分别使用合适的 RBF 网络进行预测,最后将各个预测进行集成.

基于 EMD 分解的多 RBF 网络综合预测模型的基本工作流程如下:

(1)使用经验模式分解将非线性非平稳时间序列分解进行 n 阶分解,即分解出从高频到较低频的 n 个分量;

(2)使用 RBF 神经网络对各个分量进行预测建模.每个 RBF 网络采用与各 IMF 分量相对应的近似最优的 SPREAD 参数;

(3)对各个分量的的预测结果进行等权求和集成

为综合预测结果.

3.2 参数选择方法

获取良好预测效果的关键是选择合适的 RBF 基函数的离散度参数,该参数大小决定了 RBF 神经网络的 VC 维大小.RBF 基函数为形如 $f_{bef}(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{\|x-x_0\|}{\sigma^2}}$ 的函数,其中 σ 为离散度参数,x_0 为均值.在 Matlab 工具箱中的径向基神经网络设计函数的 SPREAD 参数为离散度参数的倒数.若 SPREAD 取值太大,则 RBF 网络对高频成份拟合效果好,但是容易造成过拟合而不能获得好的预测效果.若 SPREAD 取值太小,则 RBF 网络对低频成分拟合较好,但不能充分反映序列的细节变化,造成欠拟合,同样不能取得好的预测效果.所以,如果使用单一 RBF 网络进行全局预测建模型,则径向基函数的分布系数 SPREAD 的选取需要平衡对各个频率成分的预测性能以期达到最优预测效果.实践表明 RBF 神经网络对对称性比较好的非线性非平稳随机序列预测效果比较好,但当非线性时间序列序列对称性很差(如网络流量数据)时预测性能很差.

RBF 网络的 SPREAD 参数选择通常采用交叉验证和网格搜索(grid searching)相结合的方法进行最优参数搜索.但是由于分解分量较多,如果对各个分量的 RBF 网络参数均进行最优参数搜索将会增加很大的计算负担.针对这一缺点,本文试图找到各个 RBF 网络最优参数之间的相互关系.

从 EMD 分解出发,可以发现 EMD 分解方法实现对信号的分解为近似分频分解,这可以推断出用来拟合不同 IMF 分量的所需函数集的容量(即 VC 维大小)也以指数形式递减,而在使用 RBF 网络进行拟合时其核函数参数决定了拟合函数集的容量大小,从而可以做出如下推断:RBF 网络最优 Spread 参数对数值与 $IMF_1 \sim IMF_n$ 的曲线应该近似为线性关系.例如,采用太阳黑子数据进行实验,使用 10 阶交叉验证和网格搜索方法进行各 RBF 网络的 SPREAD 参数搜索.如图 1 所示为太阳黑子数据 EMD 分解域 IMF 分量与各 RBF 网络的 log(Spread)关系曲线,该曲线近似为线性关系,证实了上述推断的正确性.由该近似线性关系可知,只需要对两个分量进行交叉验证和网格搜索取得两个参数,然后通过线性关系求取其他分量的参数,这样可以大大减轻由于分解造成的参数选择的计算负担.

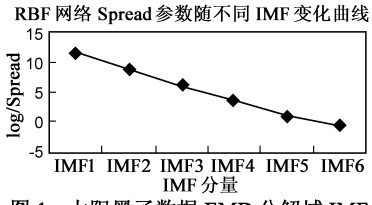

图 1 太阳黑子数据 EMD 分解域 IMF 分量与 RBF 模型 Spread 参数关系的曲线图

4 实验仿真

为了验证新型预测模型的有效性,本文采用 5 个实际的非线性、非平稳随机时间序列进行实验,并与单一分辨率下的 RBF 神经网络预测模型做对比实验.5 个实际时间序列为:sunspot database[9],Santa Fe Laser Time Series[9],Darwin Sea Level Pressure[9],Poland Electric Demand time series[9],Ethernet Packet 时间序列数据[10].

因为以太网数据包时间序列的非线性性和非对称性特点突出,此外因篇幅所限,所以在此仅给出以太网数据包时间序列在使用两种不同预测模型预测的结果显示,如图 2 所示,其中图 2(a)给出了使用单一 RBF 网络进行预测结果,图 2(b)给出了在 EMD 分解域下多 RBF 网络集成预测模型预测结果.图 2 中横坐标表示时间,纵坐标表示预测序列值或真实序列值.从图 2 可以发现 EMD 分解域下多 RBF 网络集成预测模型的预测值比单一预测模型的预测值更接近于真实序列值.

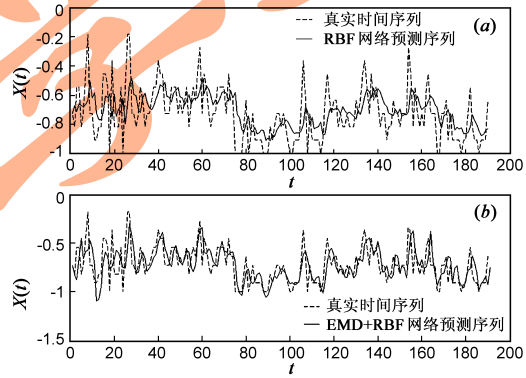

图 2 以太网数据包时间序列数据使用两种不同方法预测的预测结果

为了全面考察某一预测方法的预测性能,本文采用两种性能评价标准[11]:平均绝对误差(Mean absolute error, MAE)和规范化均方根误差(Normalized root mean square error, NRMSE).MAE 的定义式为:

$$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}|x(i)-\hat{x}(i)| \tag{5}$$

NRMSE 的定义式为:

$$\mathrm{NRMSE}=\frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}[x(i)-\hat{x}(i)]^2}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}[x(i)-\bar{x}]^2}} \tag{6}$$

其中 n 为预测集数据个数,$x(i)$为真实值,$\hat{x}(i)$为预测值,$\bar{x}$ 为序列均值.

表 1 给出了 5 个非线性非平稳时间序列预测误差.从表 1 可知,使用 EMD 分解域多 RBF 网络集成预测模型进行预测的平均绝对误差和规范化均方误差均小于使用单一 RBF 网络预测模型进行预测的平均绝对误差和规范化均方误差,所以 EMD 分解域多 RBF 网络模型

的预测性能优于单一 RBF 网络预测模型,其原因是在各个本征模式下的频率成分或波形变化与原始信号相比更简单和对称,从而更容易预测.

表 1　不同非线性时间序列预测的 MAE 和 NRMSE 误差

数据集	MAE		NRMSE	
	RBF	EMD + RBF	RBF	EMD + RBF
Ethernet Packet Series	0.1277	0.0826	0.8001	0.3197
Poland Electric Demand	0.0378	0.0220	0.0085	0.0029
Darwin Sea Level Pressures	0.1253	0.1131	0.1418	0.1369
Santa Fe Laser time series	0.0630	0.0377	0.0322	0.0184
Sunspot time series	0.1350	0.1009	0.1473	0.0827

5　结论

本文提出了一种基于 EMD 分解的复杂时间序列多 RBF 网络集成预测模型.该模型通过使用经验模式分解将非线性非平稳随机信号分解为一系列变化较简单的本征模式和一个余项之和,然后分别对各个本征模式进行 RBF 神经网络预测建模.此外,本文利用各个 RBF 网络参数之间的近似线性关系减少参数选择中交叉验证的次数,从而减轻由于多分量建模造成的计算负担.使用 5 个实际的非线性非平稳时间序列完成的有效性验证结果表明,基于 EMD 分解的多 RBF 网络集成预测模型是一种适合于非线性、非平稳时间序列预测方法,具有广阔的应用前景.因为新模型中的各个 RBF 网络的训练彼此独立,非常适合并行计算,所以下一步工作将集中于该模型的并行计算研究.此外,该模型的增量算法研究也是未来的研究方向之一.

参考文献:

[1] A Lendasse, D Francois, V Wertz, M Verleysen. Vector quantization: A weighted version for time-series forecasting[J]. Future Generation Computer Systems, 2005, 21(7): 1056 – 1067.

[2] L Ljung. System Identification Theory for User[M]. Prentice-Hall, 1987. 1 – 115.

[3] M Versace, R Bhatt, O Hinds, M Shiffer. Predicting the exchange traded fund DIA with a combination of genetic algorithms and neural networks[J]. Expert Systems with Applications, 2004, 27(3): 417 – 425.

[4] S Mukherjee, E Osuma, F Girosi. Nonlinear prediction of chaotic time series using support vector machines[A]. Proceeding of the IEEE Workshop on Neural Networks for Signal Processing [C]. Ameliz Island: IEEE, 1997: 511 – 520.

[5] Y Mitani, K Tsutsumoto, N Kagawa. Time series prediction of acoustic signals using neural network model and wavelet shrinkage[A]. Proceedings of the Tenth International Congress on Sound and Vibration [C]. Stockholm, Sweden: IIAV, 2003. 4189 – 4196.

[6] Norden E Huang, Man-Li Wu, Wendong Qu, Steven R Long, Samuel S P Shen. Applications of hilbert-huang transform to non-stationary financial time series analysis [J]. Applied Stochastic Models in Business and Industry, 2003, 19(3): 245 – 268.

[7] H Ince, T B Trafalis. Kernel principal component analysis and support vector machines for stock price prediction[A]. Proceeding of the IEEE International Joint Conference on Neural Networks[C]. Budapest, Hungary, 2004. 2053 – 2058.

[8] N Huang, N O Attoh-Okine. The Hilbert-Huang Transform in Engineering[M]. Taylor & Francis, 2005.

[9] Time Series Prediction group[EB/OL]. http://www.cis.hut.fi/projects/tsp/? page = Timeseries.

[10] Ethernet Packet[EB/OL]. http://math.bu.edu/people/murad/methods/time_series/index.html # Ethernet.

[11] D Karunasinghea, S Liongb. Chaotic time series prediction with a global model: Artificial neural network[J]. Journal of Hydrology, 2006, 323(1 – 4): 92 – 105.

[12] R Frank, N Davey, S Hunt. Time series prediction and neural network[J]. Journal of Intelligent and Robotic Systems: Theory and Applications, 2001, 31(1 – 3): 91 – 103.

作者简介:

彭喜元　男,1961 年生于内蒙古四子王旗,哈尔滨工业大学自动测试与控制系教授、博士生导师.主要研究方向为自动测试技术和智能故障诊断理论等.

王　军　男,1976 年生于浙江江山,哈尔滨工业大学自动测试与控制系博士生.主要研究方向数据挖掘、信号与信息处理和智能故障诊断理论等.E-mail: wangjunhit@chinaacc.com