复杂系统数据挖掘的多尺度混合算法*

康 卓[1], 黄竞伟[2], 李 艳[1], 康立山[3+]

[1](武汉大学 计算中心,湖北 武汉 430072)

[2](武汉大学 计算机学院,湖北 武汉 430072)

[3](武汉大学 软件工程国家重点实验室,湖北 武汉 430072)

A Multi-Scale Mixed Algorithm for Data Mining of Complex System

KANG Zhuo[1], HUANG Jing-Wei[2], LI Yan[1], KANG Li-Shan[3+]

[1](Computation Center, Wuhan University, Wuhan 430072, China)

[2](School of Computer Science, Wuhan University, Wuhan 430072, China)

[3](State Key Laboratory of Software Engineering, Wuhan University, Wuhan 430072, China)

+ Corresponding author: E-mail: kang@whu.edu.cn

Received 2002-09-07; Accepted 2003-01-20

Kang Z, Huang JW, Li Y, Kang LS. A multi-scale mixed algorithm for data mining of complex system. *Journal of Software*, 2003,14(7):1229~1237.

http://www.jos.org.cn/1000-9825/14/1229.htm

Abstract: Any complex system must be controlled by some basic laws, including macroscopic level submicroscopic level and microscopic level laws. How to discover its necessity-laws from these contingency – phenomena (observed data) is the most important task of data mining (DM) and KDD, and it is the goal of this paper too. Based on the evolutionary computation and natural fractal, a multi-scale dynamic prediction system is proposed, which models the macro-behavior of the system by ordinary differential equations while models the micro-behavior of the system by natural fractals. The financial data such as the stock market data of Jun'an stock price and the scientific observed data such as rainfall data of Wuhan in flood season are used as the test data for simulated test of analysis and prediction. The experimental results show that this system fits the data very well, and the simulated prediction is good too, even for modeling the time series with large undulating.

Key words: evolutionary computation; complex system; time series prediction; multi-scale dynamic prediction system

摘 要: 任何复杂系统都要受到某些基本规律的约束,包括宏观、中观与微观的多层次规律的约束.怎样从一个系统的这些偶然现象(观测数据)中找出它的必然规律,是知识发现(KDD)与数据挖掘(DM)的首要任务,也是研究目标.建立了一个基于演化计算与自然分形相结合的多尺度的动态预测系统.它以微分方程描述系统的宏观行为,以自然分形刻画系统的微观行为.同时,以股票市场数据(君安证券股票数据)和科学观测数据(武汉汛期

* Supported by the National Natural Science Foundation of China under Grant Nos.40275034, 70071042, 60133010 (国家自然科学基金)

第一作者简介: 康卓(1970－),男,湖北武汉人,博士生,副教授,主要研究领域为演化算法,智能计算.

雨量数据)为例,进行了分析与预测模拟.数值实验表明,该系统的描述(拟合)性能优越,即使是对起伏波动很大的时间序列,也能拟合得很好,预测效果也较好.

关键词: 演化计算;复杂系统;时间序列预测;多尺度动态预测系统

中图法分类号: TP18 **文献标识码**: A

复杂系统的观测与实验数据,如气象数据、太阳黑子数据、化学实验数据与金融数据等,是该系统的某个侧面的数量关系的反映(样本).与其他系统,如物理、工程等系统数据相比,它具有更大的随机性与偶然性.但是,任何复杂系统都要受到某些基本规律的约束,包括宏观、中观与微观的多层次规律的约束.如中国股票市场数据既要受到全球性经济波动的影响,也要受到国内宏观经济调控的影响,还要受到许多局部的错综复杂的个人与集团经济行为的影响.股票市场是否可以预测是一个有争议的问题,但即使是认为股票市场不可预测的人,他们仍然在进行股票数据的分析与预测[1,2],以决策自己的投资意向,正如 Lorenz 在气象预报建模时提出的一个简化了的常微分方程组模型:

$$\begin{cases}\dfrac{\mathrm{d}x}{\mathrm{d}t}=a(y-x)\\ \dfrac{\mathrm{d}y}{\mathrm{d}t}=bx-y-xz.\\ \dfrac{\mathrm{d}z}{\mathrm{d}t}=cz+xy\end{cases}$$

该模型中仅含 xz 与 xy 两个非线性项,且不含 t 的显式,看起来形式相当简单,但当其参数 a,b,c 取某些值时(如 a=10,b=28,c=8/3),系统的演化行为却十分复杂,出现混沌(chaos)现象,从而引起了气象是否可以长期预测的争论.尽管如此,人们仍在进行长程气象预测.对股市数据,除了许多人在采用经典的时间序列分析方法以外,也有人采用演化算法进行金融数据处理[1~4].

我们假定复杂系统数据是受到宏观、中观与微观多层次规律(因素)约束的,故在分析与预测时相应地也采用多层次、多尺度的建模思想与方法.本文以连续动力学模型(常微分方程)来描述复杂系统的宏观行为,以自然分形(一类自然离散小波)为模型来刻画复杂系统的微观行为,建立了一个多层次、多尺度的复杂数据分析与预测系统.该系统为复杂时间序列的实时分析与预测提供了一个客观的、有力的工具.我们以金融数据和科学观测数据为例进行了分析与预测模拟.数值实验表明,该模型的描述(拟合)性能优越,即使是对于产生混沌行为的复杂系统(混沌时间序列),也能拟合得很好,该模型的预测效果有时也比较好.

本文第 1 节介绍宏观动力系统预测模型,第 2 节研究微观自然分形预测模型及其与宏观动力模型的结合,第 3 节是一些金融数据与科学观测数据实例的数值实验结果.

1 时间序列的宏观动力学模型

复杂时间序列大都具有多层次与多尺度特征.不失一般性,本文假设它们具有宏观与微观两种层次、多种尺度.为了对它进行宏观描述,人们采用了多种时间序列的预处理方法.由于各自的目的不同,处理方法也有区别.Iba 等人[1]作股市数据预测的目标是提供一个有效法则:“何时”和处理“多少”股票,即什么时候抛出或买进多少股票.汛期雨量预报的目的则是预测今后是否会有洪涝发生.本文强调的是从复杂系统数据中发现其演化规律,包括它的宏观规律(宏观动力学模型)与微观规律(微观动力学模型).

1.1 原始数据的分解

为了从复杂系统数据中发掘它的宏观规律,首先就要对原始数据进行预处理,文献[1]介绍了 7 种预处理方法.本文对原始数据进行预处理的目的是将原始数据分解为两部分:光滑部分与粗糙(非光滑)部分,并假设光滑部分的演化行为是由宏观因素控制的,粗糙部分的演化行为是由微观因素控制的.分解以后的数据分别采用不同的数学模型来描述,光滑数据用微分方程来描述,粗糙部分用自然分形(自然离散小波)来刻画,从而发现两类“知识”或模型(model).本文采用积分分解方法将连续的原始数据 $x(t)$分解为解光滑部分 $\bar{x}(t)$ 与粗糙部分 $\tilde{x}(t)$:

$$x(t)=\bar{x}(t)+\tilde{x}(t),\quad t_0\le t\le T. \tag{1}$$

其中

$$\bar{x}(t)=\begin{cases}\dfrac{1}{t}\displaystyle\int_{t_0}^{t}x(\xi)\mathrm{d}\xi, & \text{当 } t_0<t<l \text{ 时},\bar{x}(t_0)=x(t_0)\\ \dfrac{1}{l}\displaystyle\int_{t-l}^{t}x(\xi)\mathrm{d}\xi, & \text{当}l\le t \text{ 时}\end{cases}, \tag{2}$$

称为 $x(t)$的光滑部分,l 称为光滑参数.l 的大小选取一般与 T 成正比.l 越大,数据越光滑.

$$\tilde{x}(x)=x(t)-\bar{x}(t) \tag{3}$$

表示 $x(t)$的粗糙部分.

对于 $x(t)$离散数据 $x(t_i)$,可相应地分解如下: $x(t_i)=\bar{x}(t_i)+\tilde{x}(t_i),i=0,1,...,m=T/\Delta t$. 其光滑部分为

$$\bar{x}(t_i)=\begin{cases}\dfrac{1}{i+1}\displaystyle\sum_{j=0}^{i}x(t_j), & \text{当 } i<l\text{时}\\ \dfrac{1}{l+1}\displaystyle\sum_{j=i-l}^{i}x(t_j), & \text{当 } l\le i\le m\text{时}\end{cases}, \tag{4}$$

其粗糙部分为

$$\tilde{x}(t_i)=x(t_i)-\bar{x}(t_i). \tag{5}$$

注意:式(4)和式(5)是与光滑参数 l 的选择有关的,l 越大,时间序列 $\{\bar{x}(t_i)\}$ 越光滑,l 的大小一般与 m 成正比.

1.2 高阶微分方程建模问题

这一节重点讨论光滑数据 $\{\bar{x}(t_i)\}_{i=0}^{m}$ 的建模与预测问题.由于它描述着动态系统的宏观行为,决定着系统的总体趋势,所以是复杂系统数据处理的重要部分.又由于它是该系统数据的光滑部分,故可假设 $\bar{x}(t)$ 是充分光滑的,即设 $\bar{x}(t)\in C^n[t_0,T],1\le n\le 4$.

动力系统 $\bar{x}(t)$ 的建模问题:寻求一个 n 阶常微分方程的初值问题:

$$\begin{cases}x^{(n)}(t)=f(t,x(t),x'(t),x''(t),...,x^{(n-1)}(t))\\ x^{(i)}(t)\big|_{t=t_0}=x^{(i)},i=0,1,...,n-1\end{cases}, \tag{6}$$

使得它的解 $x^*(t)$在$t=t_i,i=0,1,...,m$ 时与序列 $\{\bar{x}(t_i)\}$ 之均方误差

$$\left\|x^*-\bar{x}\right\|\equiv\frac{1}{m}\sqrt{\sum_{i=0}^{m}(x^*(t_i)-\bar{x}(t_i))^2} \tag{7}$$

尽可能地小.

换言之,在模型空间 F 中寻求一个模型 f,使得 $\min\limits_{f\in F}\left\|x^*-\bar{x}\right\|$.

1.3 一阶微分方程组的建模问题

为了建模方便起见,我们先将问题(6)化为一阶常微分方程组的初值问题.作函数代换:

$$y_1(t)=x(t),\quad y_{i+1}(t)=x^{(i)}(t),\quad i=1,2,...,n-1, \tag{8}$$

则问题(6)化为一阶常微分方程组:

$$\begin{cases}\dfrac{\mathrm{d}y_j(t)}{\mathrm{d}t}=y_{j+1}(t),\ j=1,2,...,n-1\\ \dfrac{\mathrm{d}y_n(t)}{\mathrm{d}t}=f(t,y_1,y_2,...,y_n)\end{cases} \tag{9}$$

与初始条件:

$$y_{j+1}(t_0)=\frac{\mathrm{d}^j x(t)}{\mathrm{d}t^j}\bigg|t=t_0,j=0,1,...,n-1. \tag{10}$$

记 $x(t_i)$为 x_i.应用差分公式: $\Delta^{i+1}x_k=\Delta^i x_k-\Delta^i x_{k-1},i=0,1,2,3.$

记 $t_{s+i}=t_i+s\Delta t, 0\le s\le 4$.

Newton-Gregory 向前插值多项式为

$$x(t_{s+i})=P_4(t_{s+i})+\text{error}=x_i+s\,\Delta x_i+\frac{s(s-1)}{2!}\Delta^2x_i+\frac{s(s-1)(s-2)}{3!}\Delta^3x_i+\frac{s(s-1)(s-2)(s-3)}{4!}\Delta^4x_i+\text{error} \tag{11}$$

其中

$$\text{error}=\frac{s(s-1)...(s-4)}{5!}(\Delta t)^5x^{(5)}(\xi), t_i\le\xi\le t_{i+4}. \tag{12}$$

我们有近似公式:

$$x^{(j)}(t_{s+i})=\frac{\mathrm{d}^j}{\mathrm{d}s^j}P_4(t_{s+i}), j=1,2,3. \tag{13}$$

应用式(8)与式(13)就能得到下述数据(矩阵 Y):

$$Y=\begin{bmatrix} y_1(t_0) & y_2(t_0) & y_3(t_0) & y_4(t_0)\\ y_1(t_1) & y_2(t_1) & y_3(t_1) & y_4(t_1)\\ \vdots & \vdots & \vdots & \vdots\\ y_1(t_m) & y_2(t_m) & y_3(t_m) & y_4(t_m)\end{bmatrix}. \tag{14}$$

这样一来,建模问题(6),(7)就转换成下述建模问题:给定数据 Y,寻求模型(9),(10),使得

$$\min_{f\in F}\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}(y_j(t_i)-y_j^*(t_i))^2}, \tag{15}$$

即

$$\min_{f\in F}\left\|y-y^*\right\|, \tag{16}$$

其中 F 是模型空间,$n\le 4$, $y^*(t)$ 是问题(9),(10)的解.

1.4 一阶微分方程组建模的演化算法

演化算法是用计算机模拟大自然的演化过程,特别是生物演化过程来求解复杂问题的一类计算方法,具有自适应、自组织、自学习以及内在并行性等智能特征.它用染色体来表示问题的可行解.首先随机地在解空间(基因型空间)产生一组染色体,然后执行遗传操作在解空间中进行搜索,最后根据达尔文的自然淘汰法则从这些染色体中选择适应值最好的那些染色体(解)作为下一代,这样一代一代地演化下去,最终求得问题最好的解.

遗传程序设计(一种自动程序设计方法)是一种特殊的演化算法,它用树来表示染色体.如四阶微分方程 $y^{(4)}(t)=y'''+4y'\cdot y''-t\cdot e^y$ 就可以通过式(8)和式(9)表示成如图 1 所示的树结构.

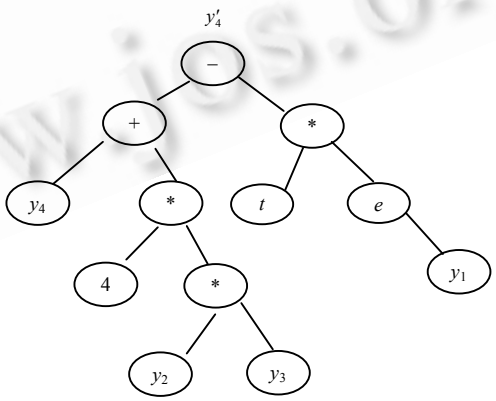

Fig.1 The tree structure of ordinary differential equation

图 1 常微分方程的树结构

对应于不同的 $t=t_0,t_1,\ldots,t_m$.树的叶节点的内容取自数据集 $D=\{R,Y\}$,即其取值或者属于 R,或者可从原始数据(矩阵 Y)(式(14))中找到它们的相应值.树的非叶节点的内容取自运算集(函数集)$O=\{+,-,\times,\exp\}$,其中+,-,×为二元运算,exp 为一元运算.若给定建模问题(9)与(10)的数据集 $D=\{R,y_1,y_2,y_3,\ldots,y_n\}$ 和运算集

O={+,−,×,/,exp,ln,sin, cos,…}以及树的高度 H,问题的模型空间 F 就被定义了.模型空间的复杂度是由 $\varphi(|O|,n,H)$ 来确定的,这里,φ 是指数函数[5,6].本文限定 $n\le4$,$|O|\le8$,$H\le7$.

当给定了一个树表达式,即一个微分方程组模型(9)以后,依初始条件(14)(即矩阵 Y 的第 1 行)即可用数值方法,如 Runge-Kutta 方法计算出问题(9)、问题(10)的解 $y^*(t)$在相继时刻 $t_1,t_2,\dots,t_m$ 的值矩阵 Y^*来,从而可以计算出此模型的拟合误差:$\|Y-Y^*\|$ 以评估该模型的好坏.当最佳模型(9)确定之后,就可以用矩阵 Y 的最后一行作为初始条件,用同样的数值方法计算出解 $y^*(t)$在时刻 $t=t_{m+1},t_{m+2},\dots,t_{m+q}$ 上的值来,对 $\bar{x}(t)$ 进行预测.

高阶微分方程建模的演化算法过程可简单地描述如下:

PROCEDURE 1

begin

Initialize population $P(0)=\{p_1(0),p_2(0),\dots,p_N(0)\}$; (随机产生 N 个树)

t:=0;

Evaluate the fitness of $p_i(t)$,i=1,2,…,N;

while not terminated **do**

begin

$P_c(t)$=crossover{$P(t)$};

$P_m(t)$=mutation{$P_c(t)$};

Evaluate $P_m(t)$;

$P(t+1)$=selection{$P_m(t),P(t)$};

t:=t+1;

end

Prediction $P(t)$;

end

建模过程的详细描述请参见文献[6].

2 时间序列的微观自然分形模型

对时间序列 $\{x(t_i)\}_{i=0}^m$ 的粗糙部分 $\{\tilde{x}(t_i)\}_{i=0}^m$,

$$\tilde{x}(t_i)=x(t_i)-\bar{x}(t_i),i=0,1,\dots,m \tag{17}$$

建立多尺度微观自然分形模型.

2.1 自然基小波的构造

记时间序列 $\{x(t_i)\}_{i=0}^m$ 的平均值为

$$\bar{x}=\sum_{i=0}^{m}\tilde{x}(t_i)\Big/(m+1), \tag{18}$$

方差为

$$\sigma^2=\sum_{i=0}^{m}(\tilde{x}(t_i)-\bar{x})^2\Big/(n-1). \tag{19}$$

为了搜索序列(17)的一个尺度为 l 的自然基小波,我们将序列 $\{\tilde{x}(t_i)\}_0^m$ 分成 l 组(由行至列地顺序排列成下面的矩阵,其中每列为一组,共 l 组),记 $\tilde{x}(t_i)=x_i$.

	1	2	…	S+1	…	l
1	x_0	x_1	…	x_S	…	x_{l-1}
2	x_1	x_{1+1}	…	x_{1+S}	…	x_{2l-1}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
k	$x_{(k-1)l}$	$x_{(k-1)l+1}$	…	$x_{(k-1)l+S}$		
各列的平均值	$\bar{x}_1^l$	$\bar{x}_2^l$	…	$\bar{x}_{S+1}^l$	…	$\bar{x}_1^l$

$$\bar{x}_i^l=\left(\sum_{j=1}^{k_i^*}x_{(j-1)l+i-1}\right)\Big/k_i^*, \tag{20}$$

其中
$$k_i^*=\begin{cases}k, & \text{当 } i\ge S+1\text{时}\\ k-1, & \text{当 } i>S+1\text{时}\end{cases}. \tag{21}$$

当 $S=l-1$ 时,则 $k_i^*=k$,这时,$m+1=kl$,即$(m+1)/l=k$.

在 x-t 平面上过点 $(\bar{x}_1^l,t_i),i=1,2,...,l$ 连成的折线:

$$x^l(t)=\begin{cases}\left[(t-t_i)\bar{x}_{i+1}^l-(t-t_{i+1})\bar{x}_i^l\right]\Big/(t_{i+1}-t_i), & \text{当 } t_i\le t\le t_{i+1},0\le i\le m-1\\ 0, & \text{当 } t<t_1 \text{ 或 } t>t_l\end{cases}. \tag{22}$$

显然,函数 $x^l(t)$以$[t_1,t_l]$为局部紧支集,称它是尺度为 l 的基小波.

为了检验函数 $x^l(t)$是否为时间序列(17)的自然基小波,我们引进 3 种形式的离差.

(1) 总离差:

$$S_1^2=\sum_{i=1}^{m}(x_i-\bar{x})^2, \tag{23}$$

其自由度为 m.

(2) 组(列)离差:

$$S_2^2=\sum_{i=1}^{l}\sum_{j=1}^{k_i^*}(x_{(j-1)l+i-1}-\bar{x}_i^l)^2, \tag{24}$$

其自由度为 $m-l+1$.

(3) 组间离差:

$$S_3^2=\sum_{i=1}^{l}k_i^*(\bar{x}_i^l-\bar{x})^2, \tag{25}$$

其自由度为 $l-1$.

容易检验:

$$S_1^2=S_2^2+S_3^2. \tag{26}$$

在一定条件下可以证明方差比

$$E_l=\left[S_3^2/(l-1)\right]\Big/\left[S_2^2/(m-l+1)\right] \tag{27}$$

是服从自由度为$(l-1,m-l+1)$的 F-分布.对不同的信度α及自由度$(l-1,m-l+1)$可以查 F 表得到 $F_\alpha(l-1,m-l+1)$.

当 $E_l\ge F_\alpha(l-1,m-l+1)$时,就认为在信度$\alpha$下序列(17)存在尺度为 l 的自然基小波(22);

当 $E_l<F_\alpha(l-1,m-l+1)$,就认为在信度α下序列(17)不存在尺度为 l 的自然基小波(22).

作为基小波,一般要求

$$\int_{-\infty}^{\infty}x^l(t)\mathrm{d}t=0. \tag{28}$$

由于我们构造 $x^l(t)$时采用分段线性函数使序列(22)连续化,它一般不能严格地满足条件(28),但采用拓广其局部紧支集的办法可使它满足条件(28).由于我们在以后应用自然基小波时,只在等距时间步长上采样,故可假定它满足条件(28),而无须具体研究其局部紧支集的拓广定义方法.

2.2 微观自然分形模型

为了建立时间序列的粗糙部分 $\tilde{x}(t)$ 的数学模型,我们从微观上(尺度为 l)构造它的多尺度自然分形模型的过程如下:

PROCEDURE 2

begin

initialize $\bar{x}:=\tilde{x}$; where $\tilde{x}=\{\tilde{x}(t_i)\}_{i=0}^m$

$x^*:=0$; where $x^*=\{x^*(t_i)\}_{i=0}^{m+q}$

for l=2, L, **do**

using $\{\bar{x}(t_i)\}_{i=0}^{m}$ calculate $\bar{x}^l=\{\bar{x}_1^l,\bar{x}_2^l,...,\bar{x}_l^l\}$;

if $E_l \geq F_\alpha(l-1,m-l+1)$ **then**

$x^*:=x^*+x^l$; where $x^l(i)=\bar{x}_{j+1}^l,\ j\equiv i \pmod l$

end for

$$\varepsilon:=\sum_{i=o}^{m}\frac{\tilde{x}(i)-x^*(i)}{m+1};$$

for i=0, $m+q$ **do**

$x^*(i):=x^*(i)+\varepsilon$;

end for

end

注 1. 计算结果 $\{x^*(t_i)\}_{i=0}^{m}$ 为 $\{\tilde{x}(t_i)\}_{i=0}^{m}$ 的拟合部分; $\{x^*(i)\}_{i=m+1}^{q}$ 为 $\tilde{x}(t)$ 的预测部分.

ε为平均拟合误差,并作为校正量(随机误差)加以消除.

注 2. 过程的第 1 部分是依次检测时间序列 $\{\tilde{x}(t_i)\}$ 是否存在尺度不超过 L 的自然基小波并作小波展开到 $[t_0,t_{m+q}]$,以分解时间序列 $\{\tilde{x}(t_i)\}=\{\bar{x}(i)\}+\{x^l(i)\}$ 及合成时间序列 $\{x^*(i)\}_0^{m+q}$ 以预测 $\tilde{x}(t)$ 在 $t\in[t_{m+1},t_{m+q}]$ 上的值.

注 3. 过程的第 2 部分是估算 $\{\tilde{x}(t_i)\}$ 的随机误差,并在拟合与预测时加以消除.

注 4. L 是基小波的尺度的上界,一般 $L\leq(m+1)/3$,对于一些特殊问题,如 m 较小时,L 可以放大到 $L\leq(m+1)/2$. 更详细的讨论将在下一节结合数值实例进行.

2.3 多层次多尺度动态建模与预测系统

调用高阶微分方程的宏观模型(9)、(10)的建模过程 PROCEDURE 1 与自然分形的微观模型的建模过程 PROCEDURE 2 就可以建立复杂时间序列的自然分形动态预测过程了.调用过程 PROCEDURE 1 建立的微分方程(9),以矩阵 Y(14)的最后一行数据 $y_1(t_m),y_2(t_m),\ldots,y_4(t_m)$为初始值($y_i(t)$在 $t=t_m$ 时刻的值),以Δt 为步长,以建模时采用的数值方法(如 Runge-Kutta 方法或改进的欧拉方法)计算 q 步,得到 $x(t)$的光滑部分 $\bar{x}(t)$ 在 $t_{m+1},t_{m+2},\ldots,t_{m+q}$ 时的预测值:$y_1(t_{m+1}),y_1(t_{m+2}),\ldots,y_1(t_{m+q})$,再调用过程 PROCEDURE 2 得到 $x(t)$的粗糙部分 $\tilde{x}(t)$ 在 $t_{m+1},t_{m+2},\ldots,t_{m+q}$ 时的预测值 $x^*(m+1),x^*(m+2),\ldots,x^*(m+q)$,将它们迭加起来,就得到了所需要的预测值:

$$y_1(t_i)+x^*(i)=x(t_i),i=m+1,m+2,\ldots,m+q.$$

这一过程可描述如下:

PROCEDURE 3

begin

Decompose data x [0,m] into $\bar{x}[o,m]$ and $\tilde{x}[o,m]$;

Call PROCEDURE 1 to get $y_1[0,m+q]$;

Call PROCEDURE 2 to get $x^*[0,m+q]$;

for i =0, m **do**

$x^*(i):=y_1(i)+x^*(i)$;

$e(i):=x(i)-x^*(i)$;

end for

for $i=m+1$, $m+q$ **do**

$x^*(i):=y_1(i)+x^*(i)$;

end for

print x^*,e

end

注 1. 过程的第 1 步:将原始时间序列 $x(t)$分解为光滑部分 $\bar{x}(t)$ 与粗糙部分 $\tilde{x}(t)$.

注 2. 过程的第 2 步调用过程 PROCEDURE 1 处理光滑数据 $\bar{x}(t)$ 得到高阶微分方程模型及解的值 $y_1(t_0)$, $y_1(t_1),\ldots,y_1(t_m),y_1(t_{m+1}),\ldots,y_1(t_{m+q})$.其前 m+1 个数值为 $\bar{x}(t)$ 的拟合值,后 q 个数值为 $\bar{x}(t)$ 的预测值.

注 3. 过程的第 3 步调用过程 PROCEDURE 2 处理粗糙数据 $\tilde{x}(t)$ 得到自然分形模型及其解 $x^*(0),x^*(1),\ldots,x^*(m)x^*(m+1),\ldots,x^*(m+q)$,前 m+1 个数值为 $\tilde{x}(t)$ 的拟合值,后 q 个数值为 $\tilde{x}(t)$ 在 $t=t_{m+1},t_{m+2},\ldots,t_{m+q}$ 时的预测值.

注 4. 过程的第 4 步合成光滑部分的拟合值与粗糙部分的拟合值,得到 $x(t)$的拟合时间序列 $\{x^*(i)\}_0^m$,并计算出拟合误差 e.

注 5. 过程的第 5 步合成光滑部分的预测值与粗糙部分的预测值,得到 $x(t)$在 $t=t_{m+1},t_{m+2},\ldots,t_{m+q}$ 时的预测值.其中预测长度 q 可由用户根据其需要而定.

3 数值实验研究

这一节研究多层次多尺度动态预测系统在复杂系统的建模与预测中的应用.这些系统数据是股票市场数据(君安证券股票数据)和科学观测数据(武汉汛期雨量数据).

3.1 君安证券股票数据建模与预测

对君安证券股票数据中 692 天的数据进行建模与预测,结果如图 2 所示.图中横坐标表示时间,以天为单位.该数据从深圳市君安证券公司获得.取 692 天数据中前面的 682 天的数据作为历史数据(即训练数据)来建模并预测后 10 天的数据.其中在第 1 阶段宏观建模过程中的参数设置为:光滑参数 l=10,m=58,q=10,t_0 =0,Δt=0.01(代表一年),N=100,n=2(二阶微分方程),运算集 O={+,−,*,/,sin,cos,ln,exp};在第 2 阶段微观自然分形建模过程中的参数设置为 m=682,L=m/3,q=10,α=0.025.得到的宏观微分方程模型为 $\frac{d^2x}{dt^2}=180654.8125-105.646973\cdot t-\frac{dx}{dt}\Big/t$.

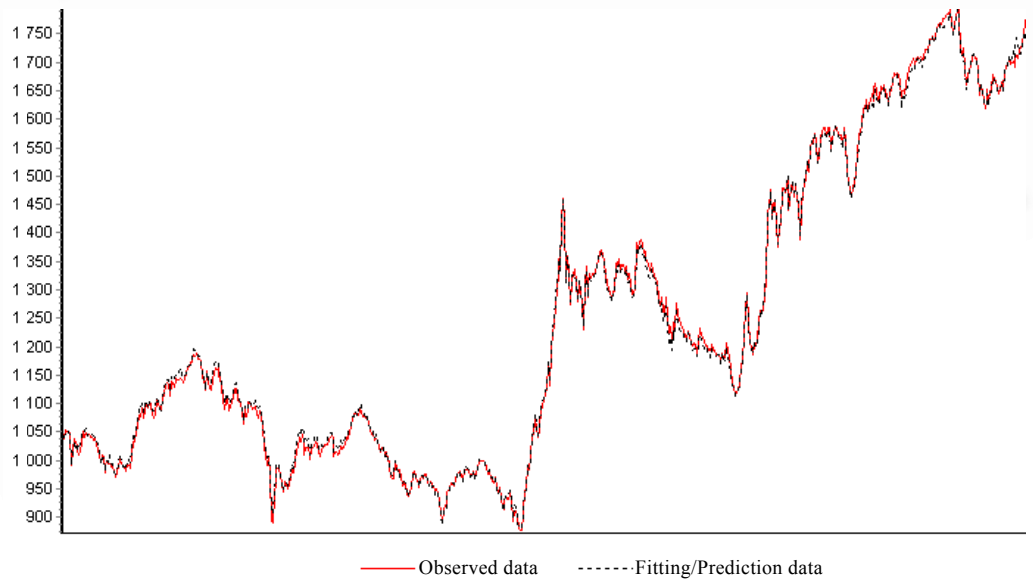

Fig.2 The fitting and prediction curves for Jun'an Stock data
图 2 君安证券股票数据的拟合与预测曲线

3.2 武汉汛期雨量数据建模与预报

对 1946 年~2002 年中的武汉雨量数据进行建模与预测结果如图 3 所示.图中横坐标表示时间,以年为单位.该数据从武汉暴雨研究所获得.应用前 56 年的历史数据来建模并预测 2003 年的汛期(5 月~9 月)的降水量.其中,在第 1 阶段宏观建模的参数设置为 l=6,m=54,q=1,t_0=0,Δt=0.01(代表一年),N=100,n=1(一阶微分方程),运算集 O 同上例;在第 2 阶段微观自然分形的参数设置为 m=54,L=m/3,q=1, α =0.1.得到的宏观微分方程模型为

$$\frac{\mathrm{d}x}{\mathrm{d}t}=64860.453125\cdot\sin\left(\frac{11878.343750}{t}\right)+(t-0.467132)/\ln(|x|)+\frac{x}{1826.261108},$$

2003 年的汛期(5 月~9 月)的降水量的预测结果为 7990.6.

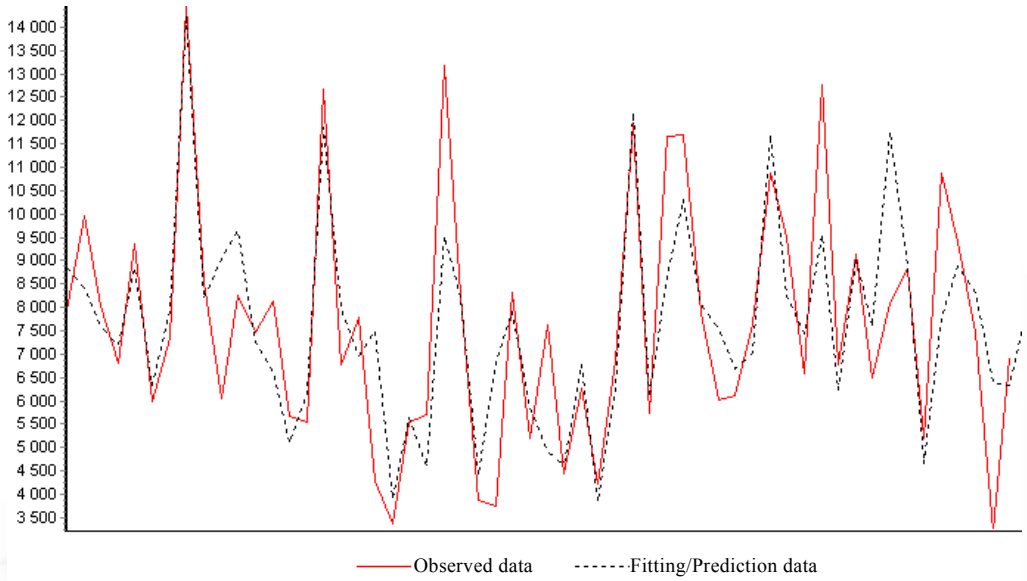

Fig.3 The fitting and prediction curves for rainfall data of Wuhan in flood season

图 3 武汉汛期雨量数据的拟合与预测曲线

从上面两个例子可以看出,应用新的动态预测系统建立的模型可以很好地描述(拟合)复杂系统(时间序列),也可以用它得到较好的预测结果.

References:

[1] Iba H, Sasaki T. Using genetic programming to predict financial data. In: Angeline PJ, ed. Proceedings of the Congress on Evolutionary Computation. Piscataway: IEEE Press, 1999. 244~251.

[2] Hafner C, Frohlich J. Generalized function analysis using hybrid evolutionary algorithms. In: Angeline PJ, ed. Proceedings of the Congress on Evolutionary Computation. Piscataway: IEEE Press, 1999. 287~294.

[3] Yoshihara I, Numata M, Sugawara K, Yamada S, Abe K. Time series prediction model building with BP-like parameter optimization. In: Angeline PJ, ed. Proceedings of the Congress on Evolutionary Computation. Piscataway: IEEE Press, 1999. 295~301.

[4] Ferreira AR, da Silva. Evolving best-basis representations. In: Angeline PJ, ed. Proceedings of the Congress on Evolutionary Computation, Vol 1. Piscataway: IEEE Press, 1999. 302~309.

[5] Kang LS, Li Y, Chen YP. A tentative research on complexity of automatic programming. Wuhan University Journal of Natural Sciences, 2001,6(1-2):59~62.

[6] Cao HQ, Kang LS, Chen YP. Evolutionary modeling of systems of ordinary differential equations with genetic programming. Genetic Programming and Evolvable Machines, 2000,1(4):309~337.