旱涝灾害的遗传-神经网络集成预测方法研究*

A Study on Genetic Algorithms-Neural Network Ensemble Forecasting Methods of Drought and Water-logging Disasters

吴建生
WU Jian-sheng

（广西柳州师范高等专科学校数学与计算机科学系，广西柳州 545004）
（Department of Mathematics and Computer Science，Liuzhou Teacher School，Liuzhou，Guangxi，545004，China）

摘要：利用遗传算法的全局搜索能力同时进化设计三层 BP 神经网络的结构和连接权，并以进化后的网络结构和连接权作为新的神经网络结构和初始连接权，再进行新一轮附加动量的 BP 神经网络训练，把训练后的结果简单平均集成，以此建立旱涝灾害的遗传-神经网络集成预测新方法。应用该方法对广西桂林 6 月（主汛期 1995～2005 年）的降水量进行实例预测的结果表明，该方法的收敛速度快，预报精度高，易于操作，是一种具有较高应用价值的预测方法。

关键词：旱涝灾害 预测遗传算法 神经网络集成
中图法分类号：TP183 **文献标识码**：A **文章编号**：1005-9164(2006)03-0203-04

Abstract：Evolved the neural network architecture and connection weights by using global research ability of genetic algorithms，new neural network architecture and beginning start connection weights be made of the evolution network structure and the connection，and train again the traditional back propagation by training samples and ensemble results by mean，this method be established the forecast new model of drought and water-logging. The application example is build with monthly mean rainfall of Guilin of Guangxi during 1995 to 2005. The calculation result express that our method of forecast can improves convergence speed and forecast accuracy. It is a useful model for forecasting.

Key words：drought and water-logging，genetic algorithms，neural network ensemble

旱涝灾害的气候预测问题是减灾防灾的重要研究课题。随着全球气候模式和区域气候模式的发展，旱涝灾害的气候动力学方法研究有了很大进展。20 世纪 90 年代以来，以神经网络方法为代表的非线性人工智能预报建模方法，已经应用在大气学科和气候分析等领域[1～4]。但是，由于神经网络方法缺乏严密理论体系指导，其应用效果完全取决于使用者的经验。在实际应用中，加上缺乏问题的先验知识，研究人员往往要经过大量费力、耗时的实验摸索，才能确定合适的网络模型和各种参数的设置，其效果也完全取决于使用者的经验，即使采用同样的方法解决同样的问题。由于操作者不同，其结果也可能大相径庭，这样会导致在应用中出现过拟合问题，影响网络的泛化能力，极大限制神经网络在实际旱涝灾害的天气业务中的应用[5,6]。

神经网络集成是用有限个神经网络对同一个问题进行学习，集成在某输入示例下的输出由构成集成的各神经网络在该示例下的输出共同决定[7,8]。该方法可以显著地提高神经网络系统的泛化能力，即使是缺乏经验的普通工程技术人员也可以从中受益，被视

收稿日期：2005-12-19
修回日期：2006-03-29

作者简介：吴建生(1974-)，陕西咸阳人，硕士，讲师，主要从事神经网络应用及智能优化研究。

*广西科学研究与技术发展计划项目（桂攻关：0592005-2A），广西教育厅项目（200508234）。

为一种非常有效的工程化神经计算方法[9]。目前神经网络集成技术已经被成功地应用到很多领域中，如光学字符识别、人脸识别、地震分类、医学等领域[10~13]。

遗传算法是从自然进化思想和理论发展而来的一种全局性搜索算法。是进化计算的一种。近年来利用遗传算法提高神经网络的泛化性能是一个十分活跃的研究领域[14,15]。本文针对神经网络在实际天气业务应用中的问题，提出利用遗传算法的全局搜索能力同时进化设计三层 BP 神经网络的结构和连接权，并以进化后的网络结构和连接权作为新的神经网络结构和初始连接权，再进行新一轮附加动量的 BP 神经网络训练，把训练后的结果采用简单平均集成，研究新的旱涝灾害预报方法。用该方法对广西桂林主汛期(6 月)的降水量进行实例分析，计算结果表明该方法的收敛速度和预报精度较高、容易操作，是一种具有较高应用价值的预测方法。

1 遗传-神经网络集成方法

遗传-神经网络集成方法的基本思想是先利用遗传算法全局性搜索的特点，寻找合适的神经网络初始连接权和网络结构，使得神经网络的训练对其初始连接权不再异常敏感，再对其进行附加动量的 BP 神经网络训练，把训练后的结果采取简单加权平均方法集成。

遗传-神经网络的进化问题数学描述如下：

$$\begin{cases} \min E(w,v,\theta,\gamma)=\dfrac{1}{N_1}\sum_{k=1}^{N_1}\sum_{t=1}^{n}[y_k(t)-\hat{y}_k(t)]^2<\varepsilon_1, \\ \hat{y}_k(t)=\sum_{j=1}^{p}v_{jk}\cdot f[\sum_{i=1}^{m}x\cdot w_{ij}+\theta_j]+\gamma_t, \quad (1) \\ f(x)=\dfrac{1}{1+e^{-x}}, \\ \text{s.t } w\in R^{m\times p},v\in R^{p\times n},\theta\in R^{p},\gamma\in R^{n}, \end{cases}$$

其中，x 为训练样本，$\hat{y}_k(t)$ 网络的实际输出，$y_k(t)$ 网络的期望输出，w_{ij} 为输入层 i 节点到输出层 j 节点的权值，v_{jk} 为隐层 j 节点到输出层 k 节点的权值，θ_j 为隐层 j 节点处的阀值，r_t 输出 t 节点处的阀值，$f(x)$ 为激活函数。

利用遗传 - 神经网络集成方法求解上述进化问题，先定义适度函数为：

$$F(w,v,\theta,\gamma)=\frac{1}{1+\min E(w,v,\theta,\gamma)}. \quad (2)$$

具体实现步骤如下：

(Ⅰ) 利用训练样本训练三层 BP 神经网络使其满足(1)式，把其连接权中的最大值和最小值分别记为 $u_{\max}$ 和 $u_{\min}$，以该区间$[u_{\min}-\delta_1,u_{\max}+\delta_2]$(其中 $\delta_{1,2}$ 为调节参数) 作为连接权基本解空间。

(Ⅱ) 对基本解空间进行编码，其中编码生成的码串由控制码和权重系数码两部分组成。控制码主要是控制隐节点的个数，它是由 0 ~ 1 组成的串，其中 0 表示无连接，1 表示有连接，串长 l_1 可由输入节点个数的 0.5 ~ 1.5 倍来确定。而权重系数码主要是控制网络的连接权，采用浮点数编码，串长 $l_2=m\times l_1+l_1+l_1\times n+n$(其中 m 为输入节点的个数，n 为输出节点个数)。编码按一定的顺序级联成一个长串，每个串对应一组网络结构和连接权。以 3 个输入节点为例，则隐层节最多有 6 个点。详见图 1。

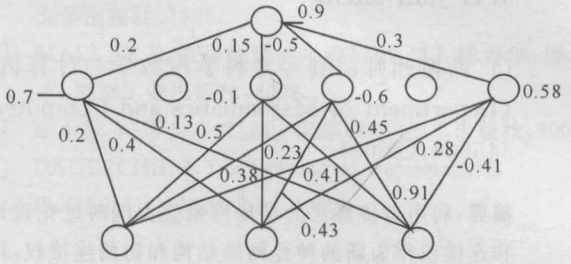

图 1 网络结构及其连接权系数

Fig. 1 The neural network and connection weights

按照图 1 标注的权值和阈值，以及网络的连接情况，可以给出对应的控制编码串和权重系数编码串。

控制码串：1 0 1 1 0 1

神经元输入层至隐层的权重矩阵：

$$\begin{bmatrix} 0.20 & 0.50 & 0.38 & 0.45 \\ 0.40 & 0.23 & 0.43 & 0.28 \\ 0.13 & 0.41 & 0.91 & -0.41 \end{bmatrix},$$

把它按一定的顺序展开：0.20 0.50 0.38 0.45 0.40 0.23 0.43 0.28 0.13 0.41 0.91 −0.41 基因串。

隐层神经元的阀值：0.70 −0.10 −0.60 0.58。

隐层到输出层的权重：0.20 0.15 −0.50 0.30。

输出神经元的阀值：0.9。

则控制码和权重码级联成的整个码串为：1 0 1 1 0 1 0.20 0.50 0.38 0.45 0.40 0.23 0.43 0.28 0.13 0.41 0.91 −0.41 0.70 −0.10 −0.60 0.58 0.20 0.15 −0.50 0.30 0.9。

(Ⅲ) 初始群体由 L 个个体构成，每个个体由两部分组成，第一部分是串长为 l_1 的 0~1 串；第二部分是区间$[u_{\min}-\delta_1,u_{\max}+\delta_2]$上的 l_2 个均匀分布随机数。

(Ⅳ) 计算群体中每个个体的适应度，由控制码得到网络的隐节点个数，由权重系数码得到网络的连接权，输入训练样本，按照(2)式计算每个个体的适应度。

(Ⅴ)保留群体中适应度最高的个体不参与交叉和变异运算，直接将其复制到下一代。群体中的其它个体，采用轮盘选择法进行选择。

(Ⅵ)对于控制码的交叉和变异采用基本遗传算法中的方法，在变异运算时，当某个神经元被变异运算删除时，相应的有关权重系数编码被置为0，而当变异运算增加某个神经元时，则随机初始化有关权重系数编码。连接权采用浮点数编码，交叉算子和变异算子如下。

以 p_c 的概率对选择后的个体进行交叉。设在第 i 个体和第 $i+1$ 个体之间进行交叉，交叉算子为：

$$\begin{cases} X_i^{t+1} = c_i \cdot X_i^t + (1 - c_i) \cdot X_{i+1}^t, \\ X_{i+1}^{t+1} = (1 - c_i) \cdot X_i^t + c_i \cdot X_{i+1}^t。 \end{cases} \tag{3}$$

式中 X_i^t, X_{i+1}^t 是一对交叉前的个体，X_i^{t+1}, X_{i+1}^{t+1} 是交叉后的个体，c_i 是区间[0,1]的均匀分布的随机数。

以 p_m 的概率对交叉后的个体进行变异。设对第 i 个体进行变异，变异算子为：

$$X_i^{t+1} = X_i^t + c_i。 \tag{4}$$

式中 X_i^t 是变异前的个体，X_i^{t+1} 是变异后的个体，c_i 是区间$[u_{\min} - \delta_1 - X_i^t, u_{\max} + \delta_2 + X_i^t]$上的均匀分布随机数。这样可以保证变异后的个体仍在搜索区间内。

(Ⅶ)生成新一代群体。反复进行4～6次，每进行一次，群体就进化一代，直到适应度满足要求或者达到总的进化代数(总的进化代数 K)。

(Ⅷ)把进化后的最后一代 L 个体全部解码，得到 L 组神经网络的连接权和网络结构，以其作为新的神经网络初始连接权和网络结构，再次利用训练样本进入新的附加动量的BP神经网络训练。

(Ⅸ)用 L 个训练后的神经网络同时对检测样本预报，然后把 L 个预测结果简单平均作为最终预测结果。

总之，遗传-神经网络集成方法可以归纳为，先通过训练样本遗传进化得到一组神经网络连接权和网络结构，进一步把其作为新的神经网络连接权的初始值和网络结构，再进行附加动量BP神经网络训练，以训练后的一组神经网络同时进行检测样本预测，把所有预测值取简单平均作为最终预测输出。

2 建模前的数据清洗

气象资料在收集过程中受许多人为因素影响，数据不可避免地包含有噪声，由此所建立的预测模型会失真，预测结果会出现偏差。为了提高预测的准确率，需要尽量有效地减少样本序列中噪声的影响。我们采用奇异谱分析(Singular Spectrum Analysis，SSA)方法[16]对原始降水序列重构，并用均生函数(Mean Generating Function，MGF)方法[17]对重构序列构造均生函数延拓矩阵，以其作为自变量，原始降水序列作为因变量，将自变量利用偏最小二乘(Partial Least-Squares Regression，PLS)方法[18]进行处理，提取对因变量影响强的成分作为神经网络的输入因子，原始序列作为输出因子，建立基于遗传算法的进化神经网络集成预测模型。

3 应用实例

利用SSA-MGF方法对广西桂林6月份1995～2005年降雨量45个原始降水序列重构，选取时间主成分方差积累贡献率达到75%的值，得到重构序列。重构序列和原始序列的相关系数达到0.8711，重构结果如图2所示。

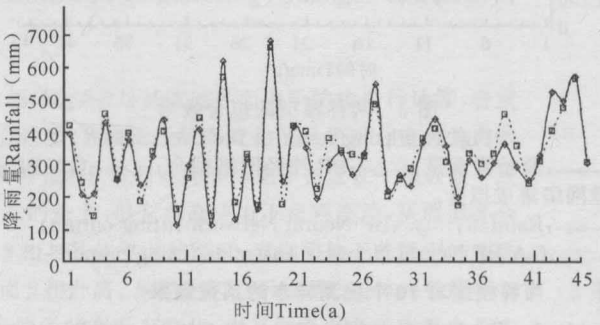

图2 原始数据和重构数据

Fig. 2 Actual data and reconstruction data

—○—：奇异谱计算值；···□···：原始降雨量。

—○—：Singular spectrum rainfall；···□···：Actual rainfall

从图2可以看出，通过数据重构，有效提取了原始序列中的主要趋势成分和震荡周期成分，并且有效地降低了原始序列中的噪声。再由重构序列生成均生函数延拓矩阵作为自变量矩阵，原始降水序列为因变量，将自变量利用偏最小二乘回归处理，提取对因变量影响强的成分，在交叉检验有效时，共提取到7个综合变量 $F_1, F_2, \cdots, F_7$，以其作为神经网络的输入因子，因变量作为神经网络的输出。

建立传统的BP神经网络模型和遗传-神经网络集成模型，分别对桂林1995～2005年主汛期(6月份)45个降水样本拟合和对10个样本预报。

两种模型对45个样本拟合的各种统计指标结果见表1，拟合效果见图3。从表1和图3的结果可以看出，遗传-神经网络集成模型的4种指标均比BP神经网络模型的好，其拟合的精度稍高BP神经网络模型。

表2是两种模型对广西桂林1995～2005年主汛期(6月份)10个降水量的预报结果。从表2的结果可以看出，在建模样本相同，预报因子相同的条件下，遗

传-神经网络集成模型对10个样本的预报精度明显优于BP神经网络模型，而且预报结果稳定。

表1 两种模型对45个样本拟合的统计评价指标

Table 1 The fitting evaluate index about 45 samples of two prediction models

模型 Model	*MAPE*(%)	*MSE*	*MAE*	*PR*
BP	11.15	37.93	30.79	0.841
GA-BP	4.21	14.89	11.88	0.964

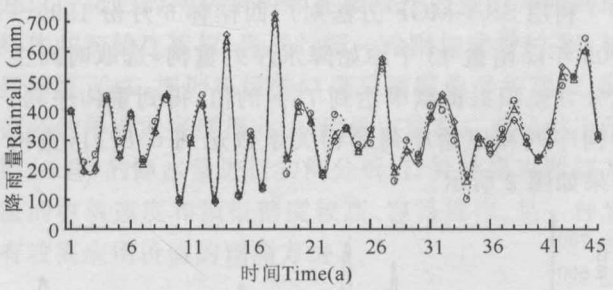

图3 两种模型的拟合效果

Fig. 3 Fitting result of the two model

—◇—:原始降雨量;···○···:BP神经网络拟合;--△--:GA-BP神经网络集成拟合。

—◇—:Rainfall;···○···:BP Neural Network fitting output;

--△--:GA-BP Neural Network ensemble fitting output

表2 两种模型对10个检测样本的预报结果

Table 2 Prediction results about 10 testing samples of two prediction models

年份 Year	实况 Actual data (mm)	BP神经网络模型 BP model			遗传-神经网络集成模型 GA-BP ensemble model		
		预报 Prediction	绝对误差 Absolute error	相对误差 Relative error (%)	预报 Prediction	绝对误差 Absolute error	相对误差 Relative error (%)
1996	288.7	229.25	59.45	20.59	255.27	33.42	11.58
1997	212.8	149.81	62.99	29.60	253.03	40.23	18.91
1998	766.7	687.10	79.60	10.38	800.49	33.79	4.41
1999	237.0	233.48	3.52	1.49	299.31	62.31	26.29
2000	432.4	256.00	176.40	40.80	442.19	9.79	2.26
2001	228.7	65.52	163.18	71.35	232.93	4.23	1.85
2002	758.0	663.07	94.93	12.52	728.32	29.68	3.92
2003	303.7	244.29	59.41	19.56	238.22	65.48	21.56
2004	301.7	269.40	32.30	10.71	277.66	24.04	7.97
2005	644.8	554.61	90.19	13.99	595.80	49.00	7.60
平均值 Average			82.20	23.10		14.91	10.63

4 结束语

本文利用SSA-MGF方法对原降水序列重构并延拓，以延拓矩阵作为自变量，原序列作为因变量，再利用PLS方法提取对系统解释最强的综合变量作为神经网络的输入因子，原始降水序列作为输出因子，建立基于遗传算法进化的BP神经网络集成预测模型。通过对广西的桂林主汛期降水量的实例计算对比表明，该方法具有：

（Ⅰ）利用SSA－MGF方法对原始数据降噪和重构，并利用PLS处理，提取对系统解释性最强的综合变量，克服了变量之间的多重相关性，提高模型精度和可靠性；又对神经网络的输入矩阵降维，使得网络结构规模变小，增强网络的稳定性。

（Ⅱ）利用遗传算法的全局搜索能力同时进化设计三层BP神经网络的结构和连接权，以进化后的网络结构和连接权作为新的神经网络结构和初始连接权，再进行新一轮附加动量的BP神经网络训练，这种方法避免了一般神经网络依靠经验确定网络结构和初始连接权困难，克服了由于神经网络初始权的随机性和网络结构确定过程中所带来的网络振荡，以及网络极易陷入局部解问题。

（Ⅲ）采用神经网络-遗传算法-神经网络混合的训练方法，有效结合了神经网络局部调节能力强和遗传算法全局优化的能力，进一步把训练后神经网络采用简单平均集成，来决定最终的预测输出，极大提高系统的泛化能力。在建模样本和预报因子相同的条件下，该预测模型的预报精度明显优于传统BP模型，而且预报结果稳定，在实际预报中易于操作，具有一定的普遍适应性。

（Ⅳ）遗传算法优化神经网络结构和初始连接权过程中，由于采用神经网络-遗传算法-神经网络并集成，使得计算量很大，需要计算机运行的时间较长，如何改进交叉、变异操作和设置有效的计算参数，加快计算速度，以及神经神经网络集成中，如何选择集成个体，提高神经网络泛化能力，将是我们下一步要做的工作。

参考文献：

[1] DEAN, ANDREW R, BRIAN H FIEDLER. Forecasting warm-season burn-off low clouds at the San Francisco international airport using linear regression and a neural network[J]. Appl Meteor, 2002, 41(6): 629-639.

[2] HSIEH WILLIAM W. Nonlinear canonical correlation analysis of the tropical Pacific climate variability using [J]. Neural Network Approach, 2001, 14(12): 2528-2539.

[3] 胡江林，涂松柏，冯光柳. 基于人工神经网络的暴雨预报方法探索[J]. 热带气象学报，2003，19(4)：422-428.

[4] 吴建生，金龙，农吉夫. 遗传算法BP神经网络的预报研究和应用[J]. 数学的实践和认识，2005，35(1)：83-88.

[5] JIN LONG, JU WEIMIN, MIAO QILONG. Study on Ann-based multi-step prediction model of short-term climatic variation [J]. Advances Atmosphere Sciences, 2000, 17(1): 157-164.

（下转第211页 Continue on page 211）

响[J]. 轻合金加工技术,2002,30(2):1-5.
[2] EDWARDS G A,STILLER K,DUNLOP G L,et al. The precipition sequence in Al-Mg-Si alloys[J]. Acta Mater,1998,46(11):3893-3901.
[3] ANDERS G F,RAGNVALD H. Bonding in MgSi and Al-Mg-Si compounds relevant to Al-Mg-Si alloys[J]. Physical Review,2003,B67:224106-224117.
[4] DERLET P M,ANDERSEN S J,MARIOARA C D,et al. A first-principles study of the -phase in Al-Mg-Si alloys[J]. Journal of Physics:Condense Matter,2002,14:4011-4024.
[5] ANDERSEN S J,MARIOARA C D,Frøseth A,et al. Crystal structure of the orthorhombic U2-Al4Mg4Si4 precipitate in the Al-Mg-Si alloy system and its relation to theand phases[J]. Materials Science and Engineering,2005,A390:127-138.
[6] RAVI C,WOLVERTON C. First-principles study of crystal structure and stability of Al-Mg-Si-(Cu) precipitates[J]. Acta Materialia,2004,52:4213-4227.
[7] MARIOARA C D,ANDERSEN S J,JANSEN J,et al. Atomic model for GP-zones in A6082 Al-Mg-Si system [J]. Acta Mater,2001,49:321-328.
[8] MSTSUDA K,GAMADA H,FUJII K,et al. High-resolution electron microscopy on the structure of GP zones in an Al-1.6% Mg_2Si alloy [J]. Metal Mater Trans,1998,29A:1161-1167.
[9] 王莉,蒋大鸣. 时效对6063铝合金力学性能及阻尼特性的影响[J]. 轻合金加工技术,2003,31(12):35-37.
[10] 高英俊,李云雯,王太成,等. Al-Mg-Si合金强化作用的键分析[J]. 轻金属,2005,2:55-57.
[11] 张瑞林. 固体与分子经验电子理论[M]. 长春:吉林科学技术出版社,1993:29-73.
[12] 科瓦索夫 ΦH,弗里德良捷尔 И H. 工业铝合金[M]. 韩秉诚,蒋香泉,译. 北京:冶金工业出版社,1987:50-64.
[13] 刘志林,李志林,刘伟东. 界面电子结构与界面性能[M]. 北京:科学出版社,2002:30-184.
[14] GAO YINGJUN,HUANG CHUANGGAO,et al. Atomic bonding and properties of Al-Mg-Sc alloy[J]. Materials Trans,2005,46(6):1123-1127.
[15] GAO YINGJUN. Electron structure and interface energy of GP Zone in Al-Zn alloy[J]. Mater Sci Forum,2005,475-479:3131-3135.
[16] GAO YINGJUN. Atomic bonding and properties of Al-Mg-Zr-Sc alloy[J]. Trans Nonferrous Met Soc China,2004,14(5):922-927.
[17] ZHANG J,FAN Z,WANG Y Q,et al. Microstructural development of Al-15wt% Mg_2Si in situ composite with mischmetal addition[J]. Mater Sci & Eng,2000,A281:109-116.
[18] 蒙多尔福 L F. 铝合金的组织与性能[M]. 北京:冶金工业出版社,1984:504-514.
[19] 刘伟东,刘志林,屈华,等. 高合金化钛合金拉伸延性的价电子理论分析[J]. 金属学报,2002,38(10):1037-1041.
[20] 波特 D A,伊斯特林 K E. 金属和合金中的相变[M]. 李长海,余永宁,译. 北京:冶金工业出版社,1988:143-188.
[21] 赖祖涵. 金属的晶体缺陷与力学性质[M]. 北京:冶金工业出版社,1988:217-242.

(责任编辑:邓大玉)

(上接第206页 Continue from page 206)

[6] 金龙,况雪源,黄海洪,等. 人工神经网络预报模型过拟和研究[J]. 气象学报,2004,62(1):62-69.
[7] HANSEN L K,SALAMON P. Neural network ensembles [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,1990,12(10):993-1001.
[8] SOLLICH P,KROGH A. Learning with Ensembles:How Over-fitting can be useful[C]//Touretzky D S,Mozer M C, Hasselmo M E. Advances in Neural Information Processing Systems 8. MA:the MIT Press,1996:190-196.
[9] 周志华,陈世福. 神经网络集成[J]. 计算机学报,2002,25(1):1-8.
[10] GUTTA S,WECHSLER H. Face recognition using hybrid classifier systems: Proceedings of the ICNN 1996-IEEE International conference network [C]. Washington DC,1996:1017-1022.
[11] MAO J. A case study on bagging boosting and basic ensembles of neural networks for OCR: Proceedings of the IJCNN 1998-IEEE International Joint conference on neural networks[C]. Anchorage,Alaska,1998,3:1828-1833.
[12] SOLLICH P,INTRATOR N. Classification of seismic signals by integrating ensembles of neural networks[J]. IEEE Transactions Signal Processing, 1998, 46(5):1194-1021.
[13] LI NING,ZHOU HUAJIE,LING JINJIANG,et al. Spiculated lesion detection in digital mammogram based on Artificial Neural Network Ensemble: Advances in Neural Networks ISNN[C]. Springer Press,2005(Ⅲ):790-795.
[14] GALLANT P J,AITKEN J M. Genetic Algorithm Design of complexity-controlled time-series predictors: Proceedings of the 2003 IEEE XIII Workshop on Neural Networks for Signal Processing[C]. Toulouse, 2003:569-574.
[15] TIAN L,NOORE A. Evolutionary neural network modeling for software cumulative failure time prediction [J]. Reliability Engineering and System Safety, 2005, 87:45-51.
[16] VAUTARD R. SSA:A toolkit for noisy chaotic signals [J]. Physical D,1992,58:95-126.
[17] 魏凤英,曹鸿兴. 长期预测的数学模型及应用[M]. 北京:气象出版社,1990.
[18] 王惠文. 偏最小二乘回归方法及其应用[M]. 北京:国防工业出版社,1999.

(责任编辑:邓大玉 凌汉恩)