信息替换的均生函数主分量多步预测*

李祚泳　张辉军

（成都气象学院，610041）

提　要

提出随着时间的推移，用新信息取代旧信息的“限定记忆”的时间序列数据处理方法。且仅对均生函数外延矩阵的前 L 阶方阵作主分量分析，用数目较少又包含主要信息量的主分量因子对时间序列建模。该方案用于四川省28个县市的年平均气温的多步预测值与实况值的相对误差均在±4%以内，表明该方案用于气温多步预测是有效的。

关键词： 信息量　均生函数　主分量　气温预测

引　言

时间序列的自回归模型(AR)、自回归滑动平均模型(ARMA)、灰色GM(1,1)模型和方差分析周期叠加外推法，都是常用的时序分析统计模型[1,2]。最近，曹鸿兴等提出了均生函数概念，并通过对均生函数 Gram-Schmidt 正交化处理，建立了时序变量的长期多步预测模型[3]。

本文在时间序列均生函数主分量预测建模的基础上[4]，提出在建立时间序列多步预测模型过程中，随着时间的推移，不断用新信息取代旧信息的数据处理方法，并且认为仅对已包含大部分信息的均生函数外延矩阵的前 L 阶方阵进行主分量分析即可。用这种方法建立的新旧信息替换的均生函数主分量模型具有多步预测能力。

1　均值生成函数及外延矩阵概念

时间序列

$$x(t)=\{x(1),x(2),\cdots,x(K)\} \quad (1)$$

的均值生成函数定义为[3]

$$\bar{x}_l(i)=\frac{1}{n_l}\sum_{j=0}^{n_l-1}x(i+jl),$$

$$i=1,2,\cdots,l,l\leqslant[\frac{K}{2}] \quad (2)$$

式中，n_l 为满足 $n_l\leqslant[\frac{K}{2}]$ 的最大整数，$[\frac{K}{2}]$ 表示对 $\frac{K}{2}$ 取整数。对 $\bar{x}_l(i)$ 按下式进行周期外延

$$\bar{x}_l(i)=\bar{x}_l(i),t\equiv i(\mathrm{mod}(l)),$$

$$t=1,2,\cdots,K \quad (3)$$

并构造外延矩阵

$$\underset{K\times L}{\bar{F}}=\begin{bmatrix}\bar{x}_1 & \bar{x}_2(1) & \bar{x}_3(1) & \cdots & \bar{x}_l(1)\\ \bar{x}_1 & \bar{x}_2(2) & \bar{x}_3(2) & \cdots & \bar{x}_l(2)\\ \bar{x}_1 & \bar{x}_2(1) & \bar{x}_3(3) & \cdots & \bar{x}_l(3)\\ \bar{x}_1 & \bar{x}_2(2) & \bar{x}_3(1) & \cdots & \bar{x}_l(4)\\ \cdots & & & & \\ \bar{x}_1 & \bar{x}_2(i_2) & \bar{x}_3(i_3) & \cdots & \bar{x}_l(i_l)\end{bmatrix} \quad (4)$$

其中 $\bar{x}_2(i_2)$ 表示取 $\bar{x}_2(1)$、$\bar{x}_2(2)$ 之一，$\bar{x}_3(i_3)$ 表示取 $\bar{x}_3(1)$、$\bar{x}_3(2)$、$\bar{x}_3(3)$ 之一，余类推。

2　建模方案

2.1　建模基本思想

为了使建立的时序变量预测模型具有多步预测能力，随着时间的推移，模型应不断地获得新信息，但又不会对计算机的容量和运

* 本文属国家自然科学基金项目。

算速度提出太高要求，可采用新信息取代旧信息的方法使建立的每一步预测模型的时序长度保持不变。为此，采用序列长度为 K 的“限定记忆法”。首先用时序变量 $\{x(1),x(2),\cdots,x(K)\}$ 建模，作 $(K+1)$ 时刻的变量 $\hat{x}(K+1)$ 值预测。其次，删去序列的第 1 个值 $x(1)$，把 $(K+1)$ 时刻的预测值 $\hat{x}(K+1)$ 补充到序列的最末，再用 $\{x(2),x(3),\cdots,x(K),x(K+1)\}$ 序列建模，作 $(K+2)$ 时刻的变量值 $\hat{x}(K+2)$ 预测。如此相继进行下去，就可实现时序变量新旧信息替换的多步预测。

2.2 建模过程

2.2.1 对长度为 K 的时间序列按式(2)生成均生函数 $\bar{x}_l(i)$，$L=[\frac{K}{2}]$，并由式(3)得外延矩阵(4) $\underset{K\times L}{\tilde{F}}=\{\bar{x}_l(i)\}$

2.2.2 外延矩阵 $\tilde{F}$ 的前 L 行和 L 列组成方阵 $\underset{L\times L}{\tilde{X}}=\{\bar{x}_l(i)\}$ 中，已经包含时间序列生成的所有均生函数的全部信息。因此，只需对 $\underset{L\times L}{\tilde{X}}$ 进行主分量分析即可。其作法是按协方差公式[5]

$$s_{ij}=\frac{1}{n}\sum_{k=1}^{n}(x_{ik}-\bar{x}_i)(x_{jk}-\bar{x}_j),$$
$$i=1,2,\cdots,L;j=1,2,\cdots,L \qquad (5)$$

计算协方差阵 $S=\{s_{ij}\}_{L\times L}$，并计算 S 按从大到小顺序排列的特征值 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_L\geqslant 0$ 和相应的特征向量组成的矩阵 $\underset{L\times L}{C}$。由 L 个特征向量可将 L 阶均生函数线性组合成 L 个主分量矩阵 $\underset{L\times L}{V}$。其中前面 H 个主分量相应特征值之和占全部特征值总和的比值满足

$$\rho=\sum_{i=1}^{H}\lambda_i/\sum_{i=1}^{L}\lambda_i\geqslant 0.85 \qquad (6)$$

时，用前面 H 个主分量已包含原 L 个均生函数(变量)的大部分信息。因此，只需用这 H 个相互独立的主分量 $v_1,v_2,\cdots,v_H$ 作为自变量，对一维时序变量 $\{x(i)\}=\{x(1),x(2),\cdots,x(K)\}$ 建立线性模型

$$\underset{K\times 1}{X}=\underset{K\times H}{V}\underset{H\times 1}{\Phi'}+\underset{K\times 1}{\varepsilon'} \qquad (7)$$

式中，V 为主分量，Φ' 为用主分量作自变量对时序变量 $\{x(i)\}$ 建模的回归系数，ε' 为误差项，Φ' 由下式计算

$$\underset{H\times 1}{\Phi'}=(\underset{H\times K}{V^T}\underset{K\times H}{V})^{-1}\underset{H\times K}{V^T}\underset{K\times 1}{X} \qquad (8)$$

“T”和“-1”分别表示矩阵转置和求逆。若时序变量 $\{x(i)\}$ 直接用均生函数 $\bar{x}_l(i)$ 作自变量建模，则应有

$$\underset{K\times 1}{X}=\underset{K\times L}{\tilde{F}}\underset{L\times 1}{\Phi}+\underset{K\times 1}{\varepsilon} \qquad (9)$$

式中，$\tilde{F}$ 为均生外延矩阵，Φ 为用均生函数建模的回归系数，ε 为误差项。因为

$$\underset{K\times H}{V}=\underset{K\times L}{\tilde{F}}\underset{L\times H}{C^T}$$

所以

$$\underset{K\times L}{\tilde{F}}=\underset{K\times H}{V}\underset{H\times L}{C} \qquad (10)$$

式中，C 是 H 个特征向量组成的正交矩阵，它满足

$$CC^T=C^TC=I$$

式(10)代入式(9)，误差项 ε' 和 ε 若忽略不计，则由式(7)和式(9)，得回归系数 Φ 和 Φ' 之间关系

$$\underset{L\times 1}{\Phi}=\underset{L\times H}{C^T}\underset{H\times 1}{\Phi'} \qquad (11)$$

Φ 求出后，由式(9)知，时序变量某时刻 K 的值

$$x(K)=\sum_{j=1}^{L}\Phi(j)\tilde{f}_{k,j}$$

因而 $K+1$ 时刻的预测值为

$$\hat{x}(K+1)=\sum_{j=1}^{L}\Phi(j)\tilde{f}_{K+1,j} \qquad (12)$$

式中，$\tilde{f}_{K+1,j}$ 为均生外延矩阵 $\underset{(K+1)\times L}{\tilde{F}}$ 的第 $K+1$ 行第 j 列的元素。可见，要预测时序变量 $K+1$ 时刻的值，需把均生函数外延为 $\underset{(K+1)\times L}{\tilde{F}}$ 矩阵。

2.3 气温变量信息替换的建模实例

表 1 列出了成都市 1959 年 5 月—1989 年 4 月按水文年* 计算的年平均气温资料。用 1984 年 4 月前 25 年的历史资料建模，用 1984 年 5 月—1989 年 4 月 5 年气温资料作模型的试报检验。其中，$K=25$，$L=[\frac{K}{2}]=$

* 前一年 5 月至次年 4 月。

表1 成都市1959年5月—1989年4月年平均气温/℃

	0	1	2	3	4	5	6	7	8	9
1960	15.9	16.5	16.2	16.5	16.3	16.2	15.6	15.7	15.9	16.0
1970	16.2	15.9	16.6	16.0	16.3	15.7	15.2	16.3	16.5	15.9
1980	16.4	15.6	15.5	15.9	15.6	15.8	16.2	15.7	15.9	15.9

12。由式(2)生成均生函数。对均生函数的外延矩阵 $\underset{26\times12}{\tilde{F}}$ 的前 L 阶方阵 $\underset{12\times12}{\bar{X}}$ 进行主分量分析，选取满足式(6)的前 $H=4$ 个主分量为

$$\underset{4\times1}{\Phi'} = (0.009550 \quad -0.599950 \quad 0.337419 \quad -0.241970)^T$$

由式(11)得到用均生函数作自变量的回归系数

$$\underset{12\times1}{\Phi'} = (0.000595 \quad -0.002868 \quad 0.114420 \quad 0.039073 \quad -0.008516 \quad 0.244406$$
$$-0.074301 \quad -0.283199 \quad 0.271372 \quad 0.009814 \quad 0.169349 \quad 0.519521)^T$$

外延矩阵 $\underset{26\times12}{\tilde{F}}$ 的第26行元素为

$$\underset{26\times j}{\tilde{f}} = (16.02 \quad 16.02 \quad 15.91 \quad 16.10 \quad 16.08 \quad 16.02$$
$$16.23 \quad 16.27 \quad 15.45 \quad 15.95 \quad 16.40 \quad 16.25)$$

由式(12)得1985年($K+1=26$)的年平均气温预测值

$$\hat{x}(26) = \Phi_{1,1}\tilde{f}_{26,1} + \Phi_{2,1}\tilde{f}_{26,2} + \cdots + \Phi_{12,1}\tilde{f}_{26,12} = 15.95(℃)$$

删去气温序列的第1个数据15.9，并在末尾数据15.6后补充上预测值15.95，得下一步建模序列

$$x(t) = (16.5, 16.2, \cdots, 15.6, 15.95).$$

重复第一步预测步骤，得到1986年平均气温预测值，余类推。成都市1985—1989各年年平均气温预测值如表2所示。表3给出了用该模型对前25年的拟合值及其与实况值的相对误差。拟合和试报的效果均较好。该模型用于四川省20个县市1985—1989年的年平均气温试报。试报值与实况值的相对误差均在±4%以内。相对误差定义为

$$\delta = \frac{\text{预测值} - \text{实况值}}{\text{实况值}}$$

表2 成都市1985—1989年年平均气温预测值及相对误差

年份	1985	1986	1987	1988	1989
预测值/℃	15.95	16.09	16.12	15.72	16.19
实况值/℃	15.8	16.2	15.7	15.9	15.9
δ/%	0.6	−0.7	2.7	1.1	1.8

自变量建立气温 $x(t)$ 的线性预测模型，由式(8)计算回归系数

表3 成都市1959—1984年间年平均气温拟合值、实况值及相对误差

年份	1960	1961	1962	1963	1964	1965	1966
实况值/℃	15.9	16.5	16.2	16.5	16.3	16.2	15.6
拟合值/℃	16.2	16.2	16.4	16.8	15.7	16.1	16.0
δ/%	0.2	−0.2	0.2	2.0	−4.0	−0.2	3.0
年份	1967	1968	1969	1970	1971	1972	1973
实况值/℃	15.7	15.9	16.0	16.2	15.9	16.6	16.0
拟合值/℃	15.8	16.2	15.9	15.8	16.1	16.3	16.2
δ/%	0.8	2.0	−1.0	−2.0	0.2	−2.0	1.0
年份	1974	1975	1976	1977	1978	1979	1980
实况值/℃	16.3	15.7	15.2	16.3	16.5	15.9	16.4
拟合值/℃	16.4	16.0	15.7	16.1	16.0	15.9	16.1
δ/%	0.6	0.2	2.0	−2.0	−2.0	0.0	−2.0
年份	1981	1982	1983	1984			
实况值/℃	15.6	15.5	15.9	15.6			
拟合值/℃	16.0	15.1	16.2	16.2			
δ/%	3.0	−3.0	2.0	4.0			

3 结论

3.1 信息替换的均生函数主分量多步预测与文献[4]的均生函数主分量时序预测建模有以下两点不同：(1)该模型不对时间序列均生函数外延矩阵 $\tilde{F}_{K\times L}$ 进行主分量分析，而只对其前 L 阶方阵 $\tilde{F}_{L\times L}$ 进行主分量分析。这样可对时间序列用个数更小又不丢失主要信息的主分量建模，减少了计算量。(2)该模型随着时间的推移，不断用新信息取代旧信息的"限定记忆法"建模，既不会增大运算量，又保持了模型的有效性和合理性。

3.2 用新信息取代旧信息建模，随着预测步数增加，势必在建模序列中引入越积越多的误差，降低可靠性，因此，一般说来，预测步长应限制在原时序长度的 1/6—1/5。我们用于四川省 20 个地市的气温建模数值试验中，都只作了 5 步预测。

3.3 该模型适用于较平稳的时间序列建模，对于含有异常值的时间序列的异常值预测结果误差较大。因此，若将该模型用于异常值预测，还有待进一步探索。

致谢：曹鸿兴研究员对该文提供了宝贵文献，并给予了支持和帮助，在此致以衷心谢意。

参考文献

1 Akaike, H., A new look at the statistical identification model. IEEE Trans. Auto. Control, 1974, 19: 716—723.

2 魏凤英，曹鸿兴. 长期预测的数学模型及其应用. 北京：气象出版社，1990：29—36，91—114.

3 曹鸿兴，魏凤英. 基于均值生成函数的时间序列分析. 数值计算与计算机应用，1991，12(2)：82—89.

4 魏凤英，曹鸿兴. 建立长期预测模型的新方案及其应用. 科学通报，1990，(10)：777—780.

5 施能. 气象统计预报中的多元分析方法. 北京：气象出版社，1992：182—226.

Multi-Steps Forecast Model with Principal Component of Mean Generating Function by Replacement of Information

Li Zuoyong　　Zhang Huijun

(Chengdu Institute of Meteorology, 610041)

Abstract

A method of data analysis of temproal sequences, based on the replacement of new-old information of restrictive memory with lapse of time is developed. This appraoch is based on the modelling of the principal component analysis for the prior L square matrix of period extrapolation matrix of mean generating function only, with a few principal components inclouding main information of temporal sequence. The scheme is applied to the multi-step forecast of annual mean temperatures of 28 cities in Sichuan, and the relative errors between forecast and real values are not more than ±4%, and it is shown that the scheme is effective for the multi-step forecast of temperature.

Key Words: information quantity　mean generating function　principal component analysis　temperature forecast