# 基于神经网络的预测模型中输入变量的选择

杨奎河[1,2] 王宝树[1] 赵玲玲[2]

(西安电子科技大学计算机学院 西安710071)[1] (河北科技大学信息学院 石家庄050054)[2]

## Input Variables Selection of Forecasting Model Based on Neural Network

YANG Kui-He[1,2] WANG Bao-Shu[1] ZHAO Ling-Ling[2]

(Computer School, Xidian University, Xi'an 710071)[1]

(Information School, Hebei University of Science and Technology, Shijiazhuang 050054)[2]

**Abstract** It is important to select input variables when the neural network forecasting model is proposed. In this paper, by using the autocorrelation function on input variables sets selection for neural network forecasting model, a systemic and scientific method for input variables sets selection is put forward. FFT is adopted to accomplish the speediness calculation, which enhances the maneuverability of this approach. A forecasting example is given, whose result indicates that the method is effective.

**Keywords** Neural network, Input variables, Forecasting

影响预测的因素和数据非常多，如何从大量的影响因素和数据中选择出对期望输出影响较大的一些因素，组成一个有效输入变量集，成为神经网络预测方法首先要面对的问题。现在对神经网络预测模型中输入变量的选择尚未提出一种比较系统的方法，一般都根据设计者的经验选取[1,2]。

用相空间嵌入法[3]来确定神经网络的输入变量，能够在历史数据序列中寻找对预报时刻影响最大的数据，但直接将选择结果用于预测时效果比较差。OLS 法[4]对输入变量进行正交变化，可求出各因素的单独贡献，但该方法不适合处理随时间连续变化的数据序列，而且计算较复杂。笔者将自相关函数的概念应用于神经网络预测中的输入变量选择，从历史数据中选择与期望输出相关度较大的数据集合作为输入变量集，并通过采用 FFT 来实现对数据自相关函数的快速计算，增加了该方法的可操作性。

## 1 自相关函数的快速计算

### 1.1 历史数据的自相关函数

建立预测的神经网络数学模型先要从历史记录中找出系统变化的规律与特性。设历史数据为确定性实信号 $x(t)$，为了计算方便，把信号取成离散形式，即 $x(n)$，然后用数字的方法对它们进行计算。广义平稳随机实信号自相关函数的定义为：

$$R(m)=E\{X(n)X(n+m)\} \tag{1}$$

如果 $X(n)$是各态遍历的，则上式的集总平均可以由单一样本 $x(n)$的时间平均来实现：

$$R(m)=\lim_{N\to\infty}\frac{1}{2N+1}\sum_{n=-N}^{N}x(n)x(n+m) \tag{2}$$

在实际的应用中，所遇到的信号都是因果性的，即当 $n<0$时，$x(n)=0$，这样，

$$R(m)=\lim_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N-1}x(n)x(n+m) \tag{3}$$

实际计算中能得到的只是 $x(n)$的 $N$ 个观察值 $x_N(0)$，$x_N(1)$，…，$x_N(N-1)$，对 $n\geqslant N$ 的值 $x(n)$只能假设为零。现在的任务是如何由这 $N$ 个观察值来估计出 $x(n)$的自相关函数 $R(m)$。

在式(3)中，由于观察值的点数 $N$ 为有限值，则用下式可求 $R(m)$的估计值 $\hat{R}(m)$：

$$\hat{R}(m)=\frac{1}{N}\sum_{n=0}^{N-1}x_N(n)x_N(n+m) \tag{4}$$

由于 $x(n)$只有 $N$ 个观察值，因此，对每一个固定的延迟 $|m|$，可以利用的数据只有 $N-1-|m|$个，且在0～$N-1$的范围内，$x_N(n)=x(n)$，所以在实际计算 $\hat{R}(m)$时，式(4)变为：

$$\hat{R}(m)=\frac{1}{N}\sum_{n=0}^{N-1-|m|}x(n)x(n+m) \tag{5}$$

对于一个固定的延迟 $|m|$，当 $N$ 足够大时，$\hat{R}(m)$是 $R(m)$的一致估计。在利用式(5)估计 $\hat{R}(m)$时，如果 $N$ 和 $m$ 都比较大，则需要的乘法次数太多，因而其应用受到了限制。采用快速傅立叶变换(FFT)可实现对 $\hat{R}(m)$的快速计算。

### 1.2 用 FFT 对自相关函数的快速计算

通过式(4)对 $\hat{R}(m)$作傅立叶变换，得：

$$\sum_{m=-(N-1)}^{N-1}\hat{R}(m)e^{-j\omega m}=\frac{1}{N}\sum_{m=-(N-1)}^{N-1}\sum_{n=0}^{N-1}x_N(n)x_N(n+m)e^{-j\omega m}$$

$$=\frac{1}{N}\sum_{n=0}^{N-1}x_N(n)\sum_{m=-(N-1)}^{N-1}x_N(n+m)e^{-j\omega m} \tag{6}$$

两个长度为 $N$ 的序列的线性卷积，其结果是一长度为 $(2N-1)$的序列。为了能用离散傅立叶变换(DTF)来计算线性卷积，需要把这两个序列的长度扩充到 $(2N-1)$。利用 DTF 来计算线性相关时，同样也是如此。为此，现把 $x_N(n)$补 $N$ 个零，得到 $x_{2N}(n)$，即：当 $0\leqslant n\leqslant N-1$时，$x_{2N}(n)=x_N(n)$；当 $N\leqslant n\leqslant 2N-1$时，$x_{2N}(n)=0$。记 $x_{2N}(n)$的傅立叶变换是 $X_{2N}(e^{j\omega})$，则：

$$\sum_{m=-(N-1)}^{N-1}\hat{R}(m)e^{-j\omega m}=\frac{1}{N}\sum_{n=0}^{2N-1}x_{2N}(n)\sum_{m=-(N-1)}^{2N-1}x_{2N}(n+m)e^{-j\omega m} \tag{7}$$

令 $l=n+m$，由于 $x_{2N}(n+m)=x_{2N}(l)$的取值区间是0～$2N-1$，因此 $l$ 的变化范围是0～$2N-1$，这样，

$$\sum_{m=-(N-1)}^{N-1}\hat{R}(m)e^{-j\omega m}=\frac{1}{N}\sum_{n=0}^{2N-1}x_{2N}(n)e^{j\omega n}\sum_{l=0}^{2N-1}x_{2N}(l)e^{-j\omega l} \tag{8}$$

杨奎河 副教授，博士研究生，主要研究方向为人工神经网络，计算机信息处理与智能控制。王宝树 教授，博士生导师，主要研究方向为人工神经网络，多传感器数据融合。

即：$\sum_{m=-(N-1)}^{N-1}\hat{R}(m)e^{-j\omega m}=\frac{1}{N}|X_{2N}(e^{j\omega})|^2$ (9)

式(9)中$|X_{2N}(e^{j\omega})|^2$是有限长信号$x_{2N}(n)$的能量谱，除以$N$后即为其功率谱。这说明自相关函数$\hat{R}(m)$和$x_{2N}(n)$的功率谱是一对傅立叶变换。式(9)中$X_{2N}(e^{j\omega})$可用FFT快速计算。由此可以得出用FFT计算自相关函数的一般步骤：

(1)对$x_N(n)$补$N$个零，得$x_{2N}(n)$，求$x_{2N}(n)$的频谱得$X_{2N}(k)$，$k=0,1,\cdots,2N-1$。

(2)求$X_{2N}(k)$的幅平方，然后除以$N$，得$\frac{1}{N}|X_{2N}(k)|^2$。

(3)对$\frac{1}{N}|X_{2N}(k)|^2$作逆变换，得$\hat{R}_0(m)$。

$\hat{R}0(m)$并不简单地等于$\hat{R}(m)$，而是等于将$\hat{R}(m)$中$-(N-1)\leqslant m\leqslant 0$的部分向右平移$2N$点后形成的新序列。

设历史数据序列为$x(i)(i=0,1,\cdots,N-1)$，由上述函数自相关理论的分析可知，只要历史数据序列$N$取得足够长，就可估算出序列$x(i)$的自相关函数$\hat{R}(m)$。在计算过程中，为了使求出的相关系数更具有可比性，把$\hat{R}(m)$按下式作归一化处理：

$$\hat{\rho}(m)=\hat{R}(m)/\hat{R}(0) \quad (10)$$

## 2 神经网络输入节点的选取

实际工作中很多地方都要进行预测，如电力负荷、城市供热负荷、城市用水量等，本文以电力负荷神经网络预测模型为例讨论输入变量的选择，采用的方法同样适用于其他神经网络预测模型。取石家庄地区电网2000年7月6日至8月5日各小时整点负荷值，共24×31=744个数据，即公式中的$N=744$，算出8月5日24点与它前面8天24×8=192个时段的自相关函数曲线，如图1所示。可以验证该自相关函数曲线对负荷各个时段具有代表性。

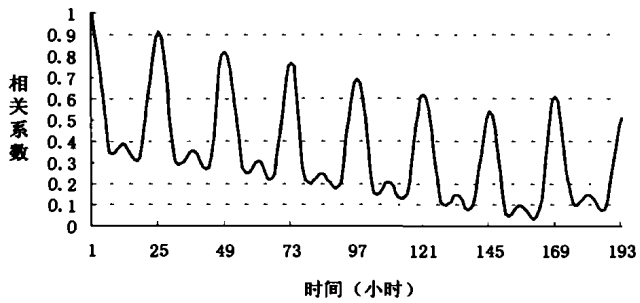


图1 负荷的自相关函数曲线

若假设当前时刻为$t$，记当前负荷值为$x(t)$，记$x(t-k)$表示在$t$时刻之前$k$个时段的负荷值，且记$\rho(t-k)$表示负荷$x(t)$与$x(t-k)$的相关系数，则从图1可以得出结论：在负荷序列中，当前负荷与以前其他各个时段负荷的相关程度是不同的。图1中负荷$x(t)$与$x(t-k)$相关系数大于0.760的点分别为$\rho(t-1)=0.911$、$\rho(t-2)=0.823$、$\rho(t-22)=0.795$、$\rho(t-23)=0.889$、$\rho(t-24)=0.912$、$\rho(t-25)=0.887$、$\rho(t-26)=0.786$、$\rho(t-47)=0.803$、$\rho(t-48)=0.819$、$\rho(t-49)=0.799$、$\rho(t-72)=0.768$。因此，当前负荷点与上述历史负荷点相关性较大，完全可以通过相关系数的显著性校验，可以考虑用上述点的数据作为输入变量。

在从历史负荷数据中选取神经网络输入变量时，考虑是对未来某日负荷进行预测，此时$x(t-1)$与$x(t-2)$不存在，因此可选用$x(t-22)$、$x(t-23)$、$x(t-24)$、$x(t-25)$、$x(t-26)$、$x(t-47)$、$x(t-48)$、$x(t-49)$和$x(t-72)$相关性较大的九点作为神经网络的负荷输入节点。对于未来23时及24时两个时段，$x(t-22)$、$x(t-23)$不存在，选取上一天的值代替。虽然还可选取其他负荷点作为输入节点，但其相关系数要小于上述各点，且会大量增加输入节点的数量。实践证明，选用上述负荷点作为输入节点已可以满足预测精度的要求。

综上所述，进行负荷预测的数学模型可用下式表示：

$$x(t)=f(x(t-22),x(t-23),x(t-24),x(t-25),x(t-26),x(t-47),x(t-48),x(t-49),x(t-72)) \quad (11)$$

其中$x(t)$为$t$时刻的预测值，$x(t-k)$为$t$时刻前$k$个时刻的负荷值。

电力负荷是一个很复杂的非线性系统，有许多因素都会对负荷的变化产生影响。本文主要讨论负荷的历史数据对预测结果的影响，其他因素的影响同样可以采取这种方法进行分析。根据对负荷的历史数据分析可知，负荷变化明显地以24小时为小周期变化，为了反映这个特征，分别选用24个前馈神经网络来作为未来一日不同时段负荷的预测模型。从图1可以看出，负荷以7天为大周期变化的影响对于当前要预测的负荷点相关性已相对较小，可不予考虑。

## 3 计算实例

为了验证上述选取输入节点的有效性，采用石家庄地区电网2000年11月1日至11月20日的数据训练神经网络，对11月21日作预测，算例分别采用两种不同的输入节点：方法一根据负荷影响"远大近小"的原则[1,2]，选择负荷输入节点为$x(t-24)$、$x(t-25)$、$x(t-26)$、$x(t-27)$、$x(t-28)$、$x(t-29)$、$x(t-30)$、$x(t-31)$、$x(t-32)$，方法二为本文提出的选择方法。

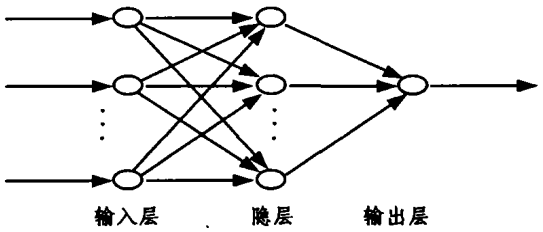


图2 神经网络的结构

表1 两种不同方法的预测结果 （负荷值单位：MW）

| 时间 | 实际值 | 方法一 | | 方法二 | |
|---|---|---|---|---|---|
| | | 预测值 | 误差(%) | 预测值 | 误差(%) |
| 2:00 | 1280.0 | 1249.7 | −2.37 | 1296.1 | 1.26 |
| 6:00 | 1390.3 | 1341.8 | −3.49 | 1414.8 | 1.76 |
| 10:00 | 1649.3 | 1693.2 | 2.66 | 1596.4 | −3.21 |
| 14:00 | 1590.9 | 1649.0 | 3.65 | 1635.1 | 2.78 |
| 18:00 | 1868.8 | 1893.7 | 1.33 | 1857.2 | −0.62 |
| 22:00 | 1609.9 | 1564.2 | −2.84 | 1634.5 | 1.53 |

选用含有一个隐层的三层前馈神经网络作为其预测模型，如图2所示，结构为9×18×1，即输入层有9个节点，隐含层有18个节点，在此每个神经网络只预测一个整点负荷，所以只有一个输出节点。隐层和输出层的函数都采用Sigmoid函数，训练算法都采用BP算法。

2000年11月21日六个时段的负荷预测结果如表1所示。

方法一和方法二的平均绝对误差分别为2.72%和1.86%，方法二的最大误差和绝对平均误差明显小于方法一。

**结束语** 影响预测的因素非常多，适当将这些因素考虑进去可以增加预测的精度；对神经网络的结构和BP算法进行改进同样可以进一步增加预测的精度。笔者在本文重点讨

（下转第143页）

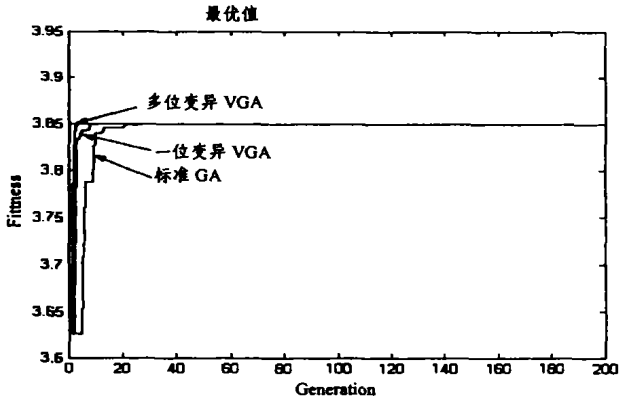


图2

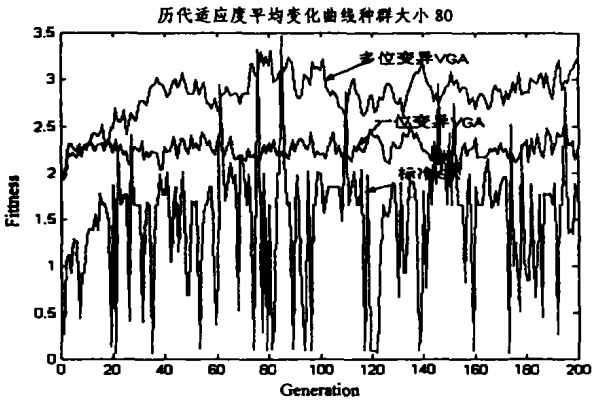


图3

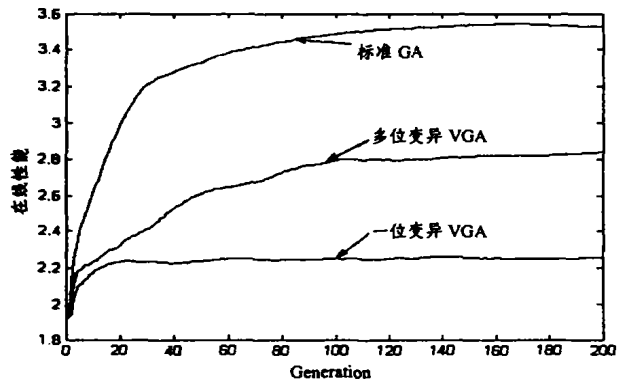


图4

三种遗传算法的历代进化的个体适应度平均值变化曲线如图3所示。

各组实验所得优化值的最优值反映了算法所能达到的最大精度，平均值反映了算法的效率。自适应多位变异遗传算法具有全局收敛性且收敛速度最快。

### 4.2 算法的在线特性和离线特性[6]

在线特性和离线特性。在线特性算法的动态性能；离线特性反映了算法的收敛性能。对上述函数进行测试，结果如图4和图5所示。

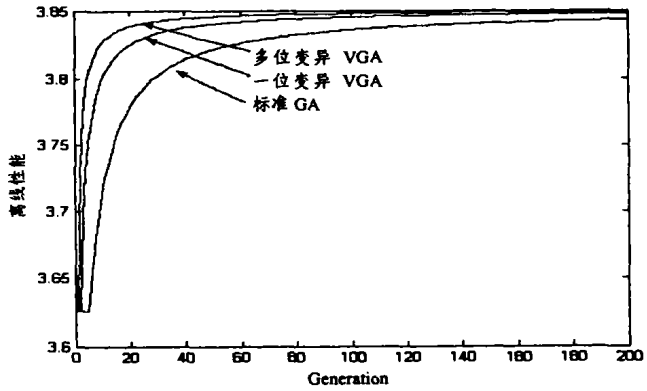


图5

从图中可以看出，自适应多位变异遗传算法的在线特性介于标准遗传算法和自适应一位变异遗传算法之间；而离线特性要优于自适应一位变异遗传算法和标准遗传算法。

**结论** 遗传算法的控制参数对算法本身的性能有重要的影响。基本遗传算法的控制参数在进化过程中保持不变，在多峰值函数优化时收敛速度慢。本文实现的自适应多位变异遗传算法充分考虑了遗传算子在不同进化时期的作用，改善了算法的性能，并且在多峰值函数优化求解中显示了优良的特性。

## 参 考 文 献


1 John H. Adaptation in Nature and Aritificial Systems . The University of Michigan Press，1975

2 王小平，曹立明. 遗传算法--理论、应用与软件实现. 西安交通大学出版社，2002，1(1)

3 周激流，丁晶，金菊良. 一种新型遗传算法及其在暴雨强度公式参数优化中的应用研究. 四川大学学报，2000，37(4)

4 Binary and Real-Valued Simulation Evolution for Matlab Copyright (C)1996 C. R. Houck, J. A. Joines, M. G. Kay

5 飞思科技产品研发中心. MATLAB 6.5辅助优化计算与设计. 电子科技出版社，2003，1(1)

6 De Jong K A. Analysis of the Behavior of a Class of Genetic Adaptive Systems [D]. University of Michigan，1975


(上接第140页)

论了历史数据对预测结果的影响。将自相关函数的概念应用于神经网络预测模型中的输入变量集选择，有明确的理论依据，通过该方法可以得到更准确的输入变量集，FFT 的采用增加了该方法的实用性和可操作性，预测误差更小，算例也验证了该方法的有效性。本文所提出的方法同样适用于城市供热负荷、城市用水量等神经网络预测模型中输入变量集的选择。

## 参 考 文 献


1 Vila J P, Wagner V, Neveu P. Bayesian Nonlinear Model Selection and Neural Networks: A Conjugate Prior Approach. IEEE Trans on neural networks, 2000,11(2):265～278

2 Matsui T, Iizaka T. Peak Load Forecasting Using Analyzable Structured Neural Network [A]. IEEE PES 2001 Winter Meeting, Columbus, Ohio USA, 2001

3 Drezga I, Rahman S. Input Variable Selection for ANN-Based Short-Term Load Forecasting. IEEE Trans on Power Systems, 1998, 13(4):1238～1244

4 Mastorocostas P A, Theocharis J B, Bakirtzis A G. Fuzzy Modeling for Short Term Load Forecasting Using the Orthogonal Least Squares Method. IEEE Trans on Power Systems, 1999,14(1):29～36

5 荣辉. 基于前馈神经网络的电力系统负荷预测研究:[学位论文]. 北京:清华大学，1998

6 Yuan J L, Fine T L. Neural Network Design for Small Training Sets of High Dimension. IEEE Trans on Neural networks, 1998, 9(2):266～280