Improved Nephritis Incidence Rate Forecasting Method Based on BP Network

TIAN Fu-peng, MA Liang-liang
School of Mathematics and Computer Science, Northwest University for Nationalities, Lanzhou 730030, China
Email：tfp@xbmu.edu.cn, mll198684@126.com

Abstract: The selection of the modeling data, the data's pretreatment of nephritis incidence rate forecasting and their effects on forecasting precision were discussed. Aiming at the actual difficulties in incidence rate forecasting, a new incidence rate forecasting method in which the solar term incidence rate is used as modeling data was put forward combining ARIMA model and BP network. The practice has proved that the method will improve the precision of incidence rate forecasting effectively and has excellent robusticity. The nephritis incidence rate characteristics in each time period should be duly considered in order to carry out a more focused health prevention and treatment and to effectively reduce the hazards of nephritis.

Keywords: solar term incidence rate; ARIMA model; BP network; forecasting; incidence rate; nephritis

基于改进 BP 网络的肾炎发病率预测方法

田富鹏，马亮亮

西北民族大学 数学与计算机科学学院，兰州 中国 730030

Email：tfp@xbmu.edu.cn ,mll198684@126.com

【摘要】论述了肾炎发病率预测中建模数据的选择、预处理方法及其对预测精度的影响。针对肾炎发病率预测的实际困难，提出了一种以节气发病率作为建模数据，将 ARIMA 模型及 BP 网络相结合的发病率预测新方法。实践证明，该方法能够有效提高预测精度，并具有很好的鲁棒性。应该充分考虑肾炎在各时间段的发病率特征，以便更有重点地进行健康防治工作，有效地降低肾炎对人类的危害，保障人类的生活品质。

【关键词】节气发病率；ARIMA 模型；BP 网络；预测；发病率；肾炎

1. 引言

肾炎是一种免疫性疾病，是免疫介导的炎性反应。它是由不同的抗原微生物感染人体后，产生不同的抗体，结合成不同的免疫复合物，沉积在肾脏的不同部位，造成的病理损伤。

根据某地区肾炎发病率的特点，按照节气的划分时间采集发病率，得到节气发病率作为建模数据，从中提取趋势分量、节气周期分量，并在样本中加入日历因子，然后根据数据特点采用 ARIMA 模型和 BP 网络模型相结合的方法实现肾炎发病率预测。

2. 建立模型

2.1 建模数据的采集与处理

根据具体情况合理的选取建模数据，是方便数据预处理，提高预测精度的有效途径。选用节气发病率（按照节气时间间隔采集的肾炎发病率）作为建模数据。实践证明，采用节气发病率作为建模数据，能够有效的缩短数据采集间隔时间，进而使数据清楚地反映出该地区发病率随时间、气候、季节变化而变化的规律。

根据采集到的数据及数据特点，将发病率分离成趋势分量因子、节气周期分量因子以及白噪声，并加入日历因子，用以消除由于各节气天数不同，数据采集时间间隔差异而产生的时间噪声。步骤如下：

对于含有趋势性、节气周期规律以及日历因素的发病率模型，可以用如下乘积模型描述[1-3]：

$$y_t = x_t \square d_t \square e_t + T_t \qquad (1)$$

978-1-935068-23-5 © 2010 SciRes.

其中 y_t 为发病率预测值； x_t 为观测值周期因子；d_t 为日历因子时间噪声；e_t 为白噪声；T_t 为观测值趋势因子。时间 t 与年份 i（$i=1,2,\cdots,N$；N 为观测数据年份个数）、节气排序数 j（$j=1,2,\cdots,24$），满足关系式：

$$t=(i-1)\times 24+j \tag{2}$$

1）分离趋势变量。趋势因子反映了发病率在相当一段时间内上升或下降的趋势，属于平稳序列变量，本文采用中位移动平均法[1]，以 24 为周期，以 t 为中心，2 阶对称移动平均数字滤波来提取不含周期波动的趋势因子。设节气发病率序列为 $y_1,y_2,\cdots,y_n$，其中 n 为序列长度。n 与观测数据序列年份个数 N、节气周期数 24 以及从序列中抽取出的被预测数据个数 m 的关系是 $n=24N-m$。

$$T_t=\frac{1}{24}\left[\sum_{i=0}^{11}y_{t+i}+y_{t+i+12}\right](t=1,2,\cdots,n-12) \tag{3}$$

2）分离周期因子。设分离趋势分量后的序列为 $z_1,z_2,\cdots,z_n$，$z_t=x_t\cdot d_t\cdot e_t=y_t-T_t$，$t=1,2,\cdots,n$。为了有效减少随机因子的影响，采用移动平均 $MA_{3\times 3}$ 模型[3]分离周期因子 x_t：$x_{2+t}=MA_{3\times 3}=\frac{1}{3}\left[\frac{1}{3}(z_t+z_{t+1}+z_{t+2})+\frac{1}{3}(z_{t+1}+z_{t+2}+z_{t+3})+\frac{1}{3}(z_{t+2}+z_{t+3}+z_{t+4})\right]$ (4)

3）白噪声及日历因子。经上述计算后，就可以直接由式（4）得到白噪声与日历因子的乘积：

$$d_t\cdot e_t=(y_t-T_t)/x_t \quad t=1,2,\cdots,n \tag{5}$$

2.2 季节性乘积 ARIMA 模型

由于数据的时间序列特性及公式的特点，在数据处理过程中出现了趋势因子 T_t，周期因子 x_t 数据丢失的情况。因此，根据建模数据所具有的周期非平稳随机特性，本文采用季节性乘积 ARIMA 模型[4]求解各分量的完整序列。

季节性乘积 ARIMA(Auto-regressive Integrated Moving Average）模型即季节性累积式自回归移动平均模型，也称作双因子模型，它是由 (p,d,q) $ARIMA$ 模型与 $(P,D,Q)_S$ $ARIMA$ 模型通过两个多项式相乘而得到的，它是将 AR 模型与 MA 模型相结合，然后将这一平稳随机过程累积而得。设带有丢失值的分量序列为 $y_1,y_2,\cdots,y_t$（$t=1,2,\cdots,n-r$），r 为丢失值的个数，其模型表达式为：

$$\phi_p(B)\Phi_P(B^s)\nabla^s\nabla_d^D y_t=\theta_q(B)\Theta_Q(B^s)\varepsilon_t \tag{6}$$

其中 B 为后移算子；p 为自回归阶数；q 为移动平均阶数；d 为差分阶数；P 季节性自回归阶数；Q 为季节性移动平均阶数；D 为季节差分阶数；s 为季节参数；$\phi_p(B)=1-\phi_1B-\phi_2B^2-\cdots-\phi_pB^p$；$\theta_q(B)=1-\theta_1B-\theta_2B^2-\cdots-\theta_qB^q$；$\Phi_P(B^s)=1-\Phi_1B^s-\Phi_2B^{2s}-\cdots-\Phi_PB^{Ps}$；$\Theta_Q(B^s)=1-\Theta_1B^s-\Theta_2B^{2s}-\cdots-\Theta_QB^{Qs}$。

该模型的参数估计方法有多种，其中包括矩估计方法、逆函数法以及最小二乘法。根据建模数据特点，本文采用最小二乘法[3]对模型进行参数估计。

2.3 BP 网络及算法

人工神经网络是源于人脑神经系统的一种模型，是由许多简单且相互紧密连接的处理单元组成的计算系统，对外界输入信息进行动态响应和并行处理，具有模拟人的部分形象思维的能力。目前，在人工神经网络的实际应用中，BP 网络是人工神经网络中应用较广的一种模型，是前向网络的核心，并体现了人工神经网络最精华的部分。

BP 网络属于多层次前馈神经网络，其神经元的变换函数为 S 型函数。在网络训练过程中，它可以不依赖于输入变量和预测值之间的明确表达式，实现从输入到输出的任意的非线性映射，因而能够有效避免噪声的影响[5]。因此在实际发病率预测中，针对白噪声的影响及 ARIMA 模型本身在预测精度方面的缺陷，本文采用 BP 网络模型，将各分量作为网络模型输入，利用 BP 网络的自学习能力对数据进行训练，求得最优权值和阀值建立网络模型，进行发病率预测。

根据 Komogorol 定理，采用含有输入层、隐层和输出层的三层 BP 网络，每层分别有 n、1、m 个神经元。BP 网络的学习训练过程由正向信号传输和反向误差传输构成[2]。正向信号传输时，隐层节点输入：

$$I_i=\sum_{j=1}^{n}w_{ij}\cdot y_j+\theta_i \quad (i=1,2,\cdots;n;j=1,2,\cdots;l) \tag{7}$$

隐层节点输出：$p_i=f_1(I_i)\ (i=1,2,\cdots;n)$ (8)

输出层节点的输入：$I_k=\sum_{i=1}^{n}v_{ki}\cdot p_i+\theta_k$ $(i=1,2,\cdots,n;k=1,2,\cdots,m)$ (9)

输出层节点的输出：$y_k=f_2(I_k)$ $(k=1,2,\cdots;m)$ (10)

其中 y_j 为输入层第 j 个节点的输出；w_{ij} 为输出层节点 i 与隐含层节点 j 之间的权值；θ_j 为隐层第 j 个节点的阀值；v_{ki} 为隐层节点 k 与输出层节点 i 之间的权

值；θ_k 为输出层第 k 个节点的阀值；$f_1(x)$ 和 $f_2(x)$ 分别为隐层和输出层的递归函数。

反向误差传输时，设有 p 组学习样本 $(y_{1p},y_{2p},\cdots;y_{np};T_{1p},T_{2p},\cdots;T_{np})$，$(p=1,2,\cdots,P)$，$T_{kp}$ 为期望输出，定义误差函数

$$E=\frac{1}{2}\sum_{q=1}^{P}\sum_{k=1}^{m}\left[T_{kq}-y_{kq}\right]^2 \tag{11}$$

其中 $y_{kq}(k=1,2,\cdots,m)$ 为网络预测值。根据求得的误差可由梯度下降法得到权值和阀值的迭代计算公式

$$\sigma(k+1)=\sigma(k)+\beta\frac{\partial E}{\partial\sigma(k)} \tag{12}$$

其中 β 为学习率。

重复上述正向传输和反向误差传输运算，不断修改权值和阀值，就能得到满足精度要求的权值和阀值。

3. 实际应用

将预测模型应用到青海海西州地区 2009 年肾炎发病率预测项目中，选取该地区 2003 年至 2008 年的节气发病率历史数据共 1728 个，进行实际预测。全部肾炎发病资料取自青海海西州第一人民医院。经过核对、补漏，从而保证资料的准确和完整。运用 Excel2003 及 Eviews3.1 对 2003 年至 2009 年的发病资料进行整理分析，统计出了 2003 年至 2009 年 1 的肾炎节气发病率，共 2016 组观测数据。BP 网络输入层选用 42 个节点，隐层节点 84 个，输出层节点 1 个，传递函数均为 Sigmoid 函数。

将预测结果与 2009 年第一个季度实际节气发病率值进行比较，见图 1 和表 1。

4. 结束语

由预测结果可知，采用 ARIMA 模型对数据进行拟合并结合具有较强学习功能、联系和容错功能的 BP 网络建模，使两种方法既能够发挥各自的优势又互相补充，大大提高了网络训练速度和预测精度，在发病率预测中取得了令人满意的结果，而采用逆运算公式得到的发病率预测精度偏低，波动较大。因此，根据预测对象本身特性采集数据，并选取适当的预处理方法及预测模型，对于发病率预测结果的精确与否起着至关重要的作用。

本文中的数据是近 7 年的海西州地区肾炎发病资料，海西州既是我国西北高原地区，又是我国少数民族聚集的地区，做好该地区肾炎发病率的预测对高原少数民族地区人民的健康有重要的意义。尽管国家经济文化迅速发展，卫生条件得到改善，但是相关部门对这类疾病的预防不够重视，所以近几年来海西州地区的肾炎发病率有逐年增高的趋势，这就提醒医疗卫生部门应加强对此高原少数民族地区肾炎预防和控制的宣传工作[6-8]。

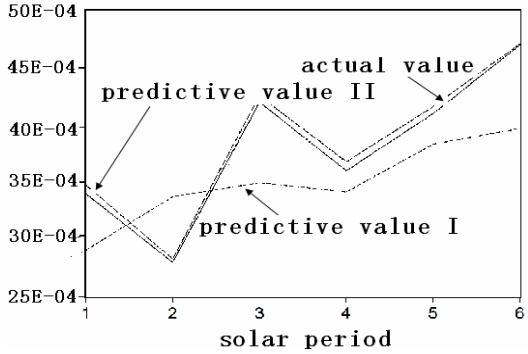

Figure 1. Forecast curve of the incidence rate on first quarter solar period of 2009

图 1. 2009 年第一季度节气发病率预测曲线

Table 1. Predictive table of the incidence rate on first quarter solar period of 2009 in Haixizhou region

表 1 海西州地区 2009 年第一季度节气发病率预测表

预测点	实际值	预测值 I	相对误差 I	预测值 II	相对误差 II
1	34	29.03253	0.1461	34.74381	0.0219
2	28	33.73233	0.2047	28.25892	0.0092
3	42	34.93135	0.1683	32.52772	0.2255
4	36	34.14227	0.0516	36.75941	0.0211
5	41	38.32187	0.0653	41.62078	0.0151
6	47	39.69769	0.1554	47.11878	0.0025

注：I 采用传统的时间序列法所求得的预测发病率；II 采用 ARIMA 模型和 BP 网络相结合的方法所求得的发病率预测值。

致谢

本文在创作过程中受到国家自然科学基金的资助，在此表示衷心的感谢。

References (参考文献)

[1] Zhang Guozhong, Huang Xiaoming, Liu Ya. Artificial Neural Net for Load Prediction in Shanghai Power Crid[J].*East China Electric Power*, 2002(2):P7-9(Ch).

[2] ZHAO Xizheng. China Electric Power Load Analysis and Prediction [M].Beijing: China Electric Power Press, 2002.

[3] XIANG Jingtian, SHI Jiuen. Methods about data processing of nonlinear system of statistical[M]. Science Press,1997.

[4] Georage E. P. Box, Gu Lan M. Time series analysis: Prediction and Control[M].Beijing: China Statistics Press,1997.

978-1-935068-23-5 © 2010 SciRes.

[5] LOU Shuntian, Shi Yang. Analysis and Design Based on MATLAB[M].Xi'an: Xi'an University of Electronic Science and Technology Press,2000.
[6] Ma Liangliang, Tian Fupeng. Research of Nephritis Morbidity Situation in Haixizhou Region Based on Season Model [J]. *Journal of Beijing Union University*, 2009, 23(3),P66-68(Ch).
[7] Ma Liangliang, Tian Fupeng. Research of Cerebral Hemorrhage Morbidity Situation in Haixizhou Region Based on PDL Model [J]. Journal of Hunan University of Arts, 2009, 21(3), P17-19(Ch).
[8] Ma Liangliang, Tian Fupeng. Research of Hypertension Morbidity Situation in Haixizhou Region Based on ADL Model [J]. Journal of Zhejiang Wanli University, 2009, 22(5),P6-9(Ch).