中华流行病学杂志2009年9月第30卷第9期 Chin J Epidemiol, September 2009, Vol.30, No.9

·基础理论与方法·

单纯ARIMA模型和ARIMA-GRNN组合模型在猩红热发病率中的预测效果比较

朱玉 夏结来 王静

【导读】 探讨单纯求和自回归滑动平均(ARIMA)模型和求和自回归滑动平均模型与广义回归神经网络(GRNN)组合模型在猩红热发病率研究中的应用。该研究对某市2000—2006年猩红热月发病率资料建立ARIMA模型,然后将其拟合值作为GRNN的输入,实际值作为网络的输出训练网络,然后比较两个模型的效果。结果表明,单纯ARIMA模型和组合模型的平均误差率(MER)分别为31.6%、28.7%;决定系数(R^2)分别为0.801、0.872。组合模型的效果要优于单纯ARIMA模型,可以用于发病率的拟合与预测。

【关键词】 猩红热; 自回归滑动平均模型; 广义回归神经网络

Comparison of predictive effect between the single auto regressive integrated moving average (ARIMA) model and the ARIMA-generalized regression neural network (GRNN) combination model on the incidence of scarlet fever

ZHU Yu, *XIA Jie-lai*, *WANG Jing*. *Department of Epidemiology and Health Statistics, College of Public Health, Anhui Medical University, Hefei* 230032, China

Corresponding author: *WANG Jing*, Email: jwang2006@126.com

【**Introduction**】 Application of the 'single auto regressive integrated moving average (ARIMA) model' and the 'ARIMA-generalized regression neural network (GRNN) combination model' in the research of the incidence of scarlet fever. Establish the auto regressive integrated moving average model based on the data of the monthly incidence on scarlet fever of one city, from 2000 to 2006. The fitting values of the ARIMA model was used as input of the GRNN, and the actual values were used as output of the GRNN. After training the GRNN, the effect of the single ARIMA model and the ARIMA-GRNN combination model was then compared. The mean error rate (MER) of the single ARIMA model and the ARIMA-GRNN combination model were 31.6%, 28.7% respectively and the determination coefficient (R^2) of the two models were 0.801, 0.872 respectively. The fitting efficacy of the ARIMA-GRNN combination model was better than the single ARIMA, which had practical value in the research on time series data such as the incidence of scarlet fever.

【**Key words**】 Scarlet fever; Auto regressive integrated moving average model; Generalized regression neural network

近年来猩红热发病虽有大幅度下降,但由于缺乏特异预防措施,控制其流行的任务依然很艰巨[1]。在疾病的预防控制中,对发病率的预测有着重要的作用,而发病率的预测方法有多种,本研究以某市疾病预防控制中心提供的猩红热月发病率资料进行发病率拟合与预测,首先采用求和自回归滑动平均(ARIMA)模型进行拟合,由于ARIMA模型的拟合值与实际值有着高度的相关性,所以可以作为一组训练样本建立广义回归神经网络(GRNN),建立ARIMA-GRNN组合模型,探讨单纯ARIMA模型与组合模型两种方法的可行性,为猩红热的监测和防治提供科学依据。

基本原理

1. ARIMA模型[2-4]:该模型是用于描述非平稳资料的一种方法,由自回归AR(p)、差分I(d)和滑动平均MA(q)三个部分组成,模型中p表示模型的自回归阶数、d表示非平稳资料转化成平稳资料的差分阶数、q表示模型移动平均阶数,当时间序列资料

DOI:10.3760/cma.j.issn.0254-6450.2009.09.025

基金项目:安徽省教育厅人文社科重点项目(2009sk192zd);安徽医科大学学术技术带头人科研资助

作者单位:230032 合肥,安徽医科大学公共卫生学院流行病与卫生统计学系(朱玉、王静);第四军医大学公共卫生学院卫生统计学教研室(夏结来)

通信作者:王静, Email:jwang2006@126.com

含有季节性变动趋势时建立ARIMA乘积模型。ARIMA(p,d,q)表达式：

$$\varphi(B)\nabla^d X_{(t)}=\theta(B)a_{(t)}$$

$$\varphi(B)=1-\varphi_1 B-\varphi_2 B^2-\cdots-\varphi_p B^p$$

$$\theta(B)=1-\theta_1 B-\theta_2 B^2-\cdots-\theta_q B^q$$

其中∇表示差分算子，B表示后移算子，$a_{(t)}$属于白噪声。

(1)模型的识别：ARIMA建模的前提条件是要求时间序列是平稳的。如果一个时间序列的概率分布不随时间变化，称为严格的平稳时间序列；如果时间序列的一、二阶距存在，并在任意时刻均值是常数、协方差为时间间隔的函数，称为宽平稳时间序列，一般所谓的平稳是指宽平稳。对于非平稳的时间序列资料，可以通过差分和数据变换实现序列的平稳化。通过观察时间序列资料的序列图(sequence chart)及自相关系数(autoeorrelation function, ACF)和偏自相关系数(partial autocorrelation function, PACF)初步判断序列的平稳性，其中还可以根据ACF、PACF初步确定p、d、q值。通常先对序列的季节性成分进行识别，再识别非季节成分。

(2)模型的参数估计：模型的阶数确定后，要对模型的参数进行估计，参数估计的方法主要有矩估计法、最小二乘估计法和极大似然估计法等。

(3)模型的诊断：建模后，要对模型进行诊断以选择最佳模型，对模型进行诊断的指标有决定系数(R^2)、赤池信息准则(AIC)、舒瓦茨(BIC)及舒瓦茨贝叶斯准则(SBC)，还可以根据残差是否为白噪声序列等。

2. GRNN[5-7]：该网络是径向基函数神经网络的一个分支，含有输入层、中间层和输出层结构，输入层到隐含层采用径向基函数变换，而从隐含层到输出层采用特殊线性变换，其权函数是规范化点积权函数，是一种新颖有效的前馈式神经网络模型(图1)。在训练过程中只需改变光滑因子，来调整各单元的传递函数，以获得最佳回归估计结果。人为调节的参数只有一个光滑因子，网络的学习全部依赖样本，使网络最大限度地避免主观因素对预测结果的影响。该网络在逼近能力、分类能力和学习速度上比BP神经网络模型有较强的优势，最后又收敛于样本量聚集较多的优化回归面，有很好的外推能力；当样本量很少时，其预测效果也较好，并能处理不稳定的数据。

GRNN初始化就是对训练样本的学习过程，学习样本确定，则相应的网络结构和各神经元之间的连接权值也随之确定，所以网络的训练实际上只是确定光滑因子的过程。与传统的误差反向传播算法不同，GRNN的学习算法在训练过程中无需调整神经元之间的连接权值，而是改变光滑因子，从而调整各单元的传递函数，获得最佳回归估计结果。光滑因子σ对网络的预测性能影响较大，当光滑因子越小，网络对样本的逼近性能就越强；而光滑因子越大，网络对样本数据的逼近过程就越平滑。

(1)学习样本的选择：在建立GRNN神经网络之前要选择合适的输入与输出样本，即学习样本。学习样本一般需事先确定，如不够明确，再进行一番筛选。输入样本必须是对输出样本影响大且能够检测的样本，另外输入样本互不相关或相关性小，这就是选择输入样本的两点原则。输出样本代表网络要实现的功能，其选择相对简单。

(2)数据处理：需要对学习样本进行尺度变换也称归一化处理，使输入输出样本控制在[0,1]或[-1,1]之间，数据已位于其区间内，无需再进行归一化处理。另外在学习样本中随机选取一个或是两个样本为待估点，用来确定网络的光滑因子。

3. ARIMA-GRNN组合模型：建立组合模型时[8]，首先是要根据原始数据信息建立ARIMA模型，利用最优的ARIMA模型，计算每个观察对象的

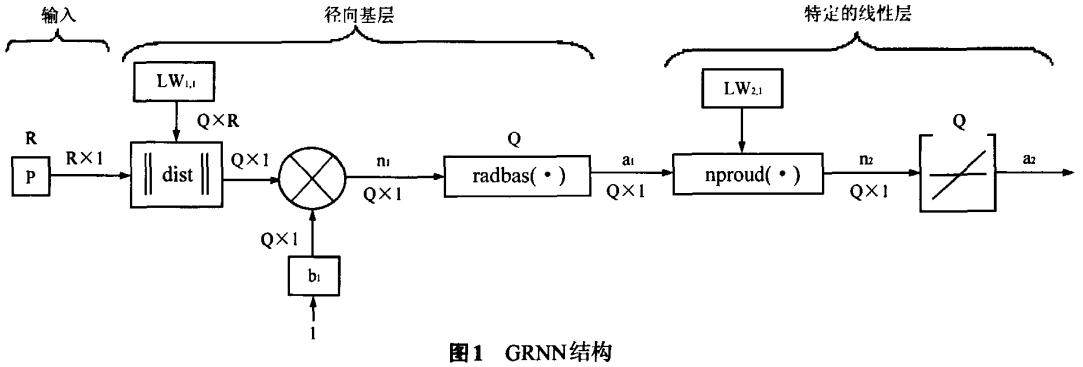

图1 GRNN结构

拟合值，很显然每个观察对象的拟合值与实际值之间存在密切的相关关系。然后将每个观察对象的拟合值作为输入样本，将实际值作为输出样本，建立一维输入、一维输出的GRNN，通过GRNN不断的训练，可以模拟、归纳出输入变量和输出变量之间的关系。当GRNN训练完成后，可以用于发病率的预测。ARIMA-GRNN组合模型相当于应用GRNN的优势对ARIMA模型值进行校正，使结果更符合观察对象的实际值。

实例分析

1. 资料来源：某市2000—2006年猩红热月发病人数与总人口数是由该市疾病预防控制中心提供，这样可以精确计算猩红热的发病率，保证了资料的可靠性。

2. ARIMA模型的建立过程及结果：

（1）模型的识别：绘制猩红热月发病率的序列图（图2）。可见发病率2000—2004年是一个平稳的状态，2004—2006年有波动，可能含有异常值，总体上看资料不平稳，并含有季节趋势。

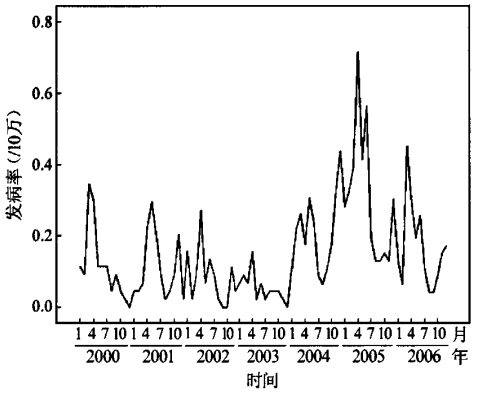

图2　猩红热月发病率的序列图

绘制猩红热月发病率的ACF、PACF图。从图3可见前3阶的自相关函数值有意义，随后自相关函数没有下降为0，反而在11～13阶表现出有意义，说明序列不平稳或是季节性；从图4可见序列存在1阶自回归，即为AR(1)；再考虑分别采用不同阶数的差分及季节差分。所以初步定下ARIMA(1,0,0)、ARIMA(0,1,1)、ARIMA(1,1,1)、ARIMA(1,0,0)×(1,0,0)$_{12}$、ARIMA(0,1,1)×(1,0,0)$_{12}$、ARIMA(1,1,1)×(1,0,0)$_{12}$ 6种模型进行发病率的拟合，并比较模型的效果。

（2）模型的参数估计：模型的阶数确定后，要对模型的参数进行估计，用SPSS 17.0统计分析软件分

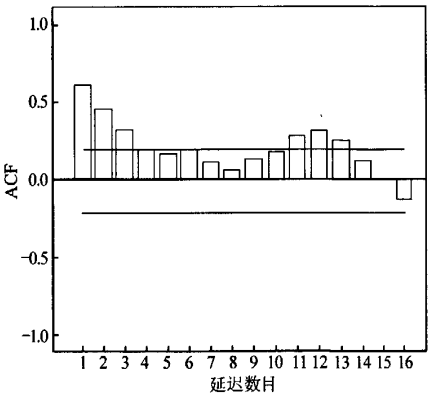

图3　猩红热月发病率的ACF图

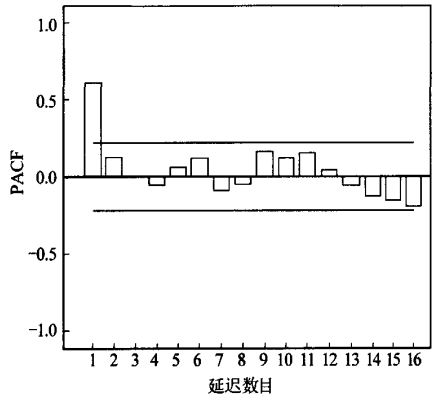

图4　猩红热月发病率的PACF图

别拟合上述6种模型，计算6种模型的R^2、BIC值（表1）。综合考虑了各项评价指标后最终决定ARIMA(1,0,0)×(1,0,0)$_{12}$模型为最优模型。参数的估计值如表2。

表1　各个模型的R^2、BIC值

指标	ARIMA (1,0,0)	ARIMA (0,1,1)	ARIMA (1,1,1)	ARIMA (1,0,0)×(1,0,0)$_{12}$	ARIMA (0,1,1)×(1,0,0)$_{12}$	ARIMA (1,1,1)×(1,0,0)$_{12}$
R^2	0.644	0.692	0.629	0.801	0.598	0.690
BIC	-4.717	-4.581	-4.594	-4.904	-4.514	-4.575

表2　ARIMA(1,0,0)×(1,0,0)$_{12}$模型的参数估计

参数	ARIMA(1,0,0)×(1,0,0)$_{12}$		
	β	t值	P值
常数	0.097	4.598	0.000
AR1	0.339	3.505	0.001
SAR1	0.536	5.280	0.000

（3）模型的诊断：ARIMA(1,0,0)×(1,0,0)$_{12}$模型R^2为0.801、BIC为-4.904；残差的ACF、PACF与Box-Ljung Q为0.408（P=0.408），可知残差属于白噪声，用该模型回代2000—2006年月发病率的拟合值与实际值如图5，可见用模型拟合的值与预测值

基本吻合。说明建立的模型是可以用来预测猩红热的月发病率。

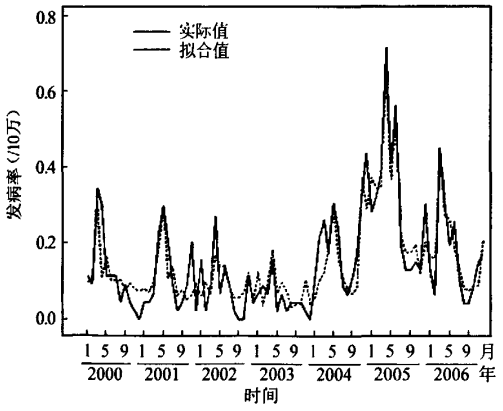

图5 ARIMA模型的拟合值与实际值的比较

3. 组合模型的建立过程及结果：

（1）学习样本的选择：因为用ARIMA模型回代的每个拟合值与其实际值之间存在密切相关关系，因此本文以ARIMA模型对2000—2006年月发病率的拟合值作为输入，实际值作为输出为学习样本，来训练GRNN神经网络。

（2）数据处理：由于样本数据已位于区间[0,1]，无需再进行归一化处理。从学习样本中随机选取两个样本为待估点，用来确定网络的光滑因子。

（3）网络的建立与训练：利用Matlab 7.0软件中的神经网络工具箱编程构建猩红热月发病率的GRNN模型。由于光滑因子对GRNN的性能有很大的影响，所以在学习样本中随机选取两个样本作为待估点，来确定最优光滑因子。通过对光滑因子的不同取值进行多次尝试来确定最优值。光滑因子从0.01开始取值，每次增加一个单位量0.01到0.1，分别对待估点进行预测，计算待估点预测值与实测值误差序列的RMSE值，将不同光滑因子对应的RMSE值绘图6。

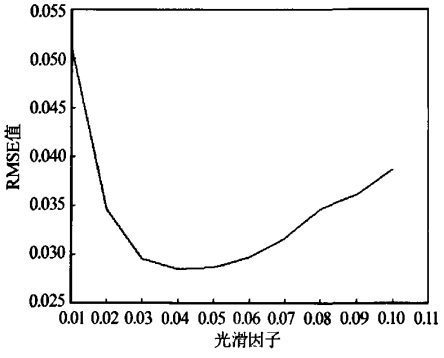

图6 不同光滑因子对应的RMSE值

经过对输出结果的检查发现，光滑因子越小，网络对样本的逼近性能就越强；光滑因子越大，网络对样本数据的逼近过程就越平滑。从图6可见当光滑因子为0.04时，待估点的RMSE值达到最小为0.0285，所以确定光滑因子为0.04。确定了光滑因子后，用训练好的GRNN拟合往年发病率，并与实际值比较，结果如图7。训练网络的程序：

确定光滑因子的程序：

```
P=[0.09718 0.10329 0.31215 0.10794……0.08868];
T=[0.11519 0.09215 0.34557 0.29950……0.15065];
P_daigudian=[0.07478 0.20882];
T_daigudian=[0.04304 0.17217];
for spread=0.01:0.01:0.1;
net=newgrnn (P,T,spread);
a=sim(net,P_daigudian);
error=a-T_daigudian
end;
```

用网络预测的程序；

```
P_yuce=[0.09718  0.10329  0.31215  0.10794……
        0.20882];
T_yuce=sim(net,P_yuce);
```

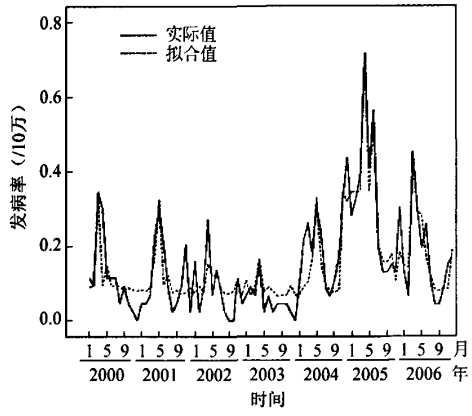

图7 组合模型的拟合值与实际值的比较

4. 精度评价：对于发病率的拟合已结束，用平均误差率（MER）及R^2评价单纯的ARIMA模型与组合模型的效果，评价结果如表3。

MER=平均误差绝对值/实际值的均值

$R^2=(SS_{实}-SS_{误})/SS_{实}$

其中SS为离均差平方和。

表3 ARIMA模型与组合模型的效果比较

指标	ARIMA模型	组合模型
MER(%)	31.6	28.7
R^2	0.801	0.872

从表3可以看出组合模型的效果要优于单纯的ARIMA模型，提示组合模型可以用于发病率的拟合与预测，且模型的效果很好。

讨　论

用于发病率拟合与预测的方法有很多，对猩红热发病率，沈艳辉等[9]对北京市猩红热的发病率采用单纯ARIMA模型进行拟合与预测。但是ARIMA模型非线性映射性能弱，传染病的发病影响因素众多，相互作用又复杂，通常具有一定非线性特征的时间序列，单纯应用ARIMA模型进行拟合与预测时，其预测精度不尽人意。为有效地利用各种预测模型的优点，Bates和Granger[10]1969年首次提出了组合预测的理论和方法，将不同的预测方法进行组合，以求产生较好的预测效果。组合预测综合利用各种单一预测方法所提供的信息，以适当的加权平均形式得出组合预测模型，其主要优点是可以对独立、有价值信息的模型进行合理的综合，求出一个最大程度利用信息的协调解，从而提高预测精度、可靠性和抗风险性。

吴伟等[11]用ARIMA与GRNN组合模型对肾综合征出血热发病率进行研究。总体上讲组合模型要好于单纯一个模型，确定各个单纯模型的组合权重是组合模型的核心问题，这样能有效提高预测精度，GRNN具有函数逼近能力强、学习速度快、对小样本预测准确、受主观因素影响的少和预测结果稳定等特点，所以通过GRNN确定组合权重来进行组合预测。

本文结果表明组合模型的效果好于单纯的ARIMA模型，说明组合模型可以用于传染病发病率的拟合与预测，其预测精度高于单纯的ARIMA模型。但组合模型的效果要直接受到每个单一模型的影响[12]，在本研究中对猩红热的月发病率建立单纯的ARIMA模型时，发现发病率序列中存在异常点，这必然影响了ARIMA模型的拟合，这也是影响组合模型效果的一个原因。再者，判断哪个模型最适合猩红热发病率的预测，尚需要扩大地区范围、扩大样本含量进一步探讨。

参 考 文 献

[1] Gidaris D, Zafeiriou D, Mavridis P, et al. Scarlet fever and hepatitis: a case report. Hippokratia, 2008, 12(3): 186-187.

[2] 王振龙，胡永宏. 应用时间序列分析. 北京：科学出版社，2005.

[3] Ubeyli ED, Guler I. Spectral analysis of internal carotid arterial Doppler signals using FFT, AR, MA, and ARMA methods. Comput Biol Med, 2004, 34: 293-306.

[4] Stadnytska T, Braun S, Werner J. Comparison of automated procedures for ARMA model identification. Behav Res Methods, 2008, 40(1): 250-262.

[5] 钟珞，饶文碧，邹承明. 人工神经网络及其融合应用技术. 北京：科学出版社，2007.

[6] 董长虹. Matlab神经网络与应用. 北京：国防工业出版社，2007.

[7] Specht DF. A general regression neural network. IEEE Trans Neural Networks, 1991, 2(6): 568-576.

[8] 严薇荣，徐勇，杨小兵，等. 基于ARIMA-GRNN组合模型的传染病发病率预测. 中国卫生统计，2008，25(1)：82-83.

[9] 沈艳辉，江初，敦哲，等. 北京市城区1957－2004年猩红热流行趋势及预测. 现代预防医学，2008，35(7)：1224-1226.

[10] Bates JM, Granger CWJ. The combination of forecasts. Operational Research Quarterly, 1969, 20(4): 451-468.

[11] 吴伟，郭军巧，周宝森. GRNN组合预测模型对辽宁省及部分地区肾综合征出血热发病率的预测研究. 中国媒介生物学及控制杂志，2008，19(1)：44-48.

[12] Flores BE, White EM. A framework for the combination of forecasts. J Ac Market Sci, 1988, 16(3): 95-103.

（收稿日期：2009-04-25）

（本文编辑：张林东）