

非线性自回归神经网络在肾综合征出血热流行趋势预测中的应用

吴伟 安淑一 郭军巧 关鹏 任仰武 夏玲姿 周宝森

【摘要】 目的 探讨非线性自回归(NAR)神经网络拟合及预测我国HFRS流行趋势的应用。方法 使用2004—2013年全国HFRS月报告发病数序列建立ARIMA模型和NAR神经网络模型,预测2014年HFRS月发病数,并比较两模型的拟合和预测效果。结果 对于拟合集,ARIMA模型的平均绝对误差(MAE)、均方误差平方根(RMSE)和平均绝对误差百分比(MAPE)分别为148.058、272.077和12.678%,NAR神经网络分别为119.436、186.671和11.778%;对于预测集,ARIMA模型的MAE、RMSE和MAPE分别为189.088、221.133和21.296%,NAR神经网络分别为119.733、151.329和11.431%。结论 NAR神经网络对于全国HFRS流行趋势拟合及预测效果优于传统的ARIMA模型,具有良好推广应用价值。

【关键词】 肾综合征出血热;非线性自回归神经网络;预测

Application of nonlinear autoregressive neural network in predicting incidence tendency of hemorrhagic fever with renal syndrome Wu Wei¹, An Shuyi², Guo Junqiao², Guan Peng¹, Ren Yangwu¹, Xia Lingzi¹, Zhou Baosen¹. 1 Department of Epidemiology, School of Public Health, China Medical University, Shenyang 110122, China; 2 Liaoning Provincial Center for Disease Control and Prevention

Corresponding author: Zhou Baosen, Email: bszhou@mail.cmu.edu.cn

This work was supported by grants from the National Natural Science Foundation of China (No. 81202254, No. 30771860).

【Abstract】 Objective To explore the prospect of nonlinear autoregressive neural network in fitting and predicting the incidence tendency of hemorrhagic fever with renal syndrome (HFRS), in the mainland of China. **Methods** Monthly reported case series of HFRS in China from 2004 to 2013 were used to build both ARIMA and NAR neural network models, in order to predict the monthly incidence of HFRS in China in 2014. Fitness and prediction on the effects of these two models were compared. **Results** For the Fitting dataset, MAE, RMSE and MAPE of the ARIMA model were 148.058, 272.077 and 12.678% respectively, while the MAE, RMSE and MAPE of NAR neural network appeared as 119.436, 186.671 and 11.778% respectively. For the Predicting dataset, MAE, RMSE and MAPE of the ARIMA model appeared as 189.088, 221.133 and 21.296%, while the MAE, RMSE and MAPE of the NAR neural network as 119.733, 151.329 and 11.431% respectively. **Conclusion** The NAR neural network showed better effects in fitting and predicting the incidence tendency of HFRS than using the traditional ARIMA model, in China. NAR neural network seemed to have strong application value in the prevention and control of HFRS.

【Key words】 Hemorrhagic fever with renal syndrome; Nonlinear autoregressive neural network; Predict

HFRS的防制由于受气候环境等自然因素和城市化进程、人居改造、大型工程建设、农业生产方式等人为因素影响^[1],难以在短期内完全消除其危

害。因此,采用适当方法对其流行趋势进行预测将有助于HFRS的防控。近年来,神经网络在时间序列分析中以其良好的非线性特性、并行分布式的存储结构和较高的容错性等特点取得了较好效果^[2]。非线性自回归(nonlinear autoregressive, NAR)神经网络作为一种动态神经网络,具有较好的动态性和抗干扰能力,可以用来逼近任意的非线性动态系统。为此本研究使用传统的自回归移动平均(ARIMA)模型和NAR神经网络分别建立预测模

DOI:10.3760/cma.j.issn.0254-6450.2015.12.017

基金项目:国家自然科学基金(81202254,30771860)

作者单位:110122 沈阳,中国医科大学公共卫生学院流行病学教研室(吴伟、关鹏、任仰武、夏玲姿、周宝森);辽宁省疾病预防控制中心(安淑一、郭军巧)

通信作者:周宝森, Email: bszhou@mail.cmu.edu.cn

型,比较两者的拟合和预测效果,探讨NAR神经网络在预测我国HFRS流行趋势中的应用。

资料与方法

1. 资料数据:2004年1月至2014年12月全国HFRS报告发病数源自国家卫生和计划生育委员会网站 (<http://www.nhfp.gov.cn/zhuzhan/yqxx/lists.shtml>) 发布的法定报告传染病疫情报告。2004年1月至2013年12月HFRS月发病数作为拟合集建立模型,2014年1—12月HFRS月发病数作为预测集检验模型的预测效果。

2. ARIMA模型的建立:由于HFRS月发病数存在明显的季节性,故建立季节性ARIMA模型: $ARIMA(p, d, q) \times (P, D, Q)_s$; 式中, p 是简单模型的自回归阶数, P 是季节模型的自回归阶数, d 是简单模型的差分阶数, D 是季节模型的差分阶数, q 是简单模型的移动平均阶数, Q 是季节模型的移动平均阶数, s 为周期。使用IBM SPSS Statistics 22.0软件中的专家建模器建立季节性ARIMA模型可以自动筛选出最优的模型。

3. NAR神经网络模型的建立:NAR神经网络是一种有记忆功能的神经网络,一个典型的NAR神经网络主要由输入层、隐含层和输出层及输入和输出的延时构成。网络的输出取决于当前的输入和过去的输出。NAR神经网络模型方程为: $y(t) = f[y(t-1), \dots, y(t-d)]$; 式中, $y(t)$ 是神经网络输出, d 表示延时阶数, f 表示用神经网络实现的非线性函数。为了避免神经网络过度拟合,拟合集中的样本随机分成训练样本(80%)、检验样本(10%)和测试样本(10%)。通过反复尝试确定延时阶数和隐含层神经元个数。采用Levenberg-Marquardt算法进行训练。训练阶段使用开环(open-loop)模式,训练完成后使用closeloop函数将开环模式直接改为闭环模式(close-loop)进行时间序列的多步预测。根据误差自相关图、拟合效果误差图和拟合值与实际值的相关系数选择合适的模型。NAR神经网络的实现使用MATLAB R2014b软件中的神经网络工具箱中的时间序列建模工具。

4. 效果评价:使用平均绝对误差(MAE)、均方误差平方根(RMSE)以及平均绝对误差百分比(MAPE)3个指标评价

ARIMA模型和NAR神经网络的拟合及预测效果^[3]。

结果

1. 全国HFRS月报告发病数分布:见图1。

2. 季节性ARIMA模型:通过观察HFRS月发病数序列,存在趋势性、异方差性和季节性,为非平稳序列。进行自然对数转换以去除异方差性,并在一次普通差分的基础上再进行一次季节差分以去除趋势性和季节性。原始序列经过上述处理,变成平稳序列。SPSS软件的专家建模器最终确定的最优时间序列模型结构为 $ARIMA(0, 1, 1) \times (1, 1, 0)_{12}$, 模型的参数显著性检验结果见表1。Ljung-Box $Q = 15.053, P = 0.864$, 残差序列可以认为是白噪声。

3. NAR神经网络模型:分析中发现当隐含层神经元个数为16、延时阶数为12时,模型效果较为理想。NAR神经网络在训练3次后检验样本误差增加,训练误差自相关图显示,误差在lag为0时最大,其他情况下均未超过可信区间,说明建立的模型适宜。NAR神经网络拟合效果误差见图2。拟合值与

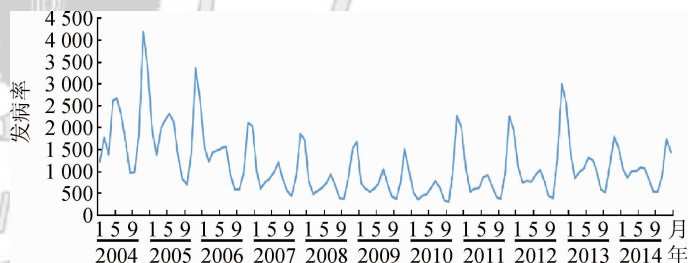


图1 2004—2014年全国HFRS的月报告发病数分布

表1 ARIMA(0,1,1)×(1,1,0)₁₂模型的参数显著性检验

模型参数	β	s_r	t 值	P 值
MA1	0.244	0.096	2.550	0.012
SAR1	-0.463	0.092	-5.027	0.000

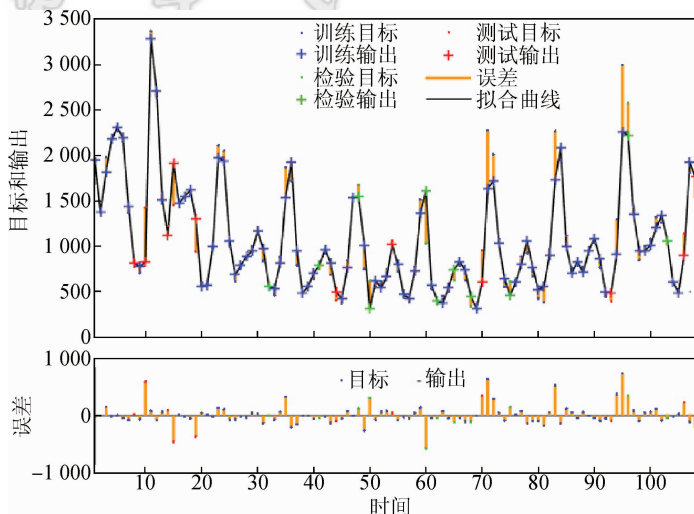


图2 NAR神经网络拟合效果误差

实际值的相关系数为0.956,较为理想。

4. 拟合和预测效果评价:2004—2014年全国HFRS月报告发病数NAR神经网络和ARIMA模型的拟合和预测值见图3,其中黑色竖线左侧部分为拟合集,右侧部分为预测集。NAR神经网络及ARIMA模型拟合和预测效果的比较见表2,NAR神经网络的MAE、RMSE和MAPE在拟合和预测阶段均小于ARIMA模型。

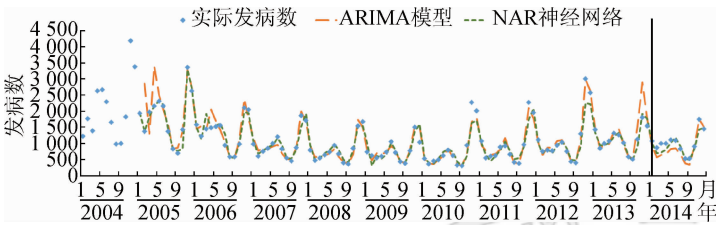


图3 2004—2014年全国HFRS月报告发病数NAR神经网络和ARIMA模型的拟合和预测值

表2 NAR神经网络及ARIMA模型拟合和预测效果的比较

方 法	拟合集			预测集		
	MAE	RMSE	MAPE(%)	MAE	RMSE	MAPE(%)
ARIMA模型	148.058	272.077	12.678	189.088	221.133	21.296
NAR神经网络	119.436	186.671	11.778	119.733	151.329	11.431

讨 论

本研究分别建立ARIMA模型和NAR神经网络拟合及预测我国HFRS流行趋势。两种方法均充分捕捉了HFRS发病季节高峰。分析拟合和预测效果评价指标,NAR神经网络的拟合和预测准确性均优于ARIMA模型。ARIMA模型作为传统的线性模型分析方法具有诸多优点^[4],但也存在明显不足,如非线性映射性能较弱,对于不规则数据序列难以确定合适的模型结构^[5]。HFRS发病受生物、环境等较多因素影响,且各影响因素之间相互作用复杂。因此HFRS发生的时间序列通常表现出一定的非线性特征。而ARIMA模型对于此类数据的拟合和预测精度往往难以令人满意。神经网络按照是否存在反馈与记忆可分为静态神经网络和动态神经网络。后者是指神经网络带有反馈与记忆功能,可将前一时刻的数据保留,使其加入到下一时刻数据的计算,使网络不仅具有动态性而且保留的系统信息也更加完整。依据动态神经网络的上述特征,非常适合用于时间序列的预测。NAR神经网络是非线性动态系统中应用较为广泛的一种神经网络,已被用于拟合和预测血吸虫病和手足口病的流行趋势^[6-7]。

在标准的NAR神经网络结构中神经网络的输

出被反馈到输入端。由于本研究在神经网络训练中期望的输出是已知的,因此在训练阶段使用开环模式,此模式下将期望输出直接反馈到输入端。这样做有两点好处。第一,可以使NAR神经网络预测结果更加准确;第二,将NAR神经网络变为单纯的前向神经网络,这样就可以直接使用静态神经网络的建模函数。训练完毕后可以使用close loop函数直接改为闭环模式以便进行多步预测。

本研究存在不足。如只使用了发病数建模以反映疾病的发展变化趋势,而影响HFRS发病的因素错综复杂,考虑到可行性的原因,其他影响因素并未被加入到模型中,这可能会在一定程度上影响到结果的准确性。因此,在实际使用过程中,在条件允许的情况下尽可能的全面收集影响HFRS的发病因素,在建模时纳入相关因素以提高模型精度。

参 考 文 献

- [1] Huang LY, Zhou H, Yin WW, et al. The current epidemic situation and surveillance regarding hemorrhagic fever with renal syndrome in China, 2010 [J]. Chin J Epidemiol, 2012, 33(7): 685-691. (in Chinese) 黄立勇,周航,殷文武,等. 2010年中国肾综合征出血热监测及疫情分析[J]. 中华流行病学杂志, 2012, 33(7): 685-691.
- [2] Zhang GP, Qi M. Neural network forecasting for seasonal and trend time series [J]. Eur J Operat Res, 2005, 160(2): 501-514.
- [3] Zhang XY, Zhang T, Liu YY, et al. Application of wavelet neural network in forecasting the incidence of hemorrhagic fever with renal syndrome [J]. Chin J Health Stat, 2012, 29(6): 864-865. (in Chinese) 张兴裕,张韬,刘元元,等. 小波神经网络在肾综合征出血热发病率预测中的应用[J]. 中国卫生统计, 2012, 29(6): 864-865.
- [4] Han Q, Su H, Wang CC, et al. Prediction on the incidence of blood and sexually transmitted diseases with models of ARIMA and GRNN [J]. Mod Prev Med, 2012, 39(6): 1337-1340. (in Chinese) 韩琴,苏虹,王忱诚,等. ARIMA模型与GRNN模型对性病发病率的预测研究[J]. 现代预防医学, 2012, 39(6): 1337-1340.
- [5] Yan WR, Xu Y, Yang XB, et al. Predict the incidence of infectious diseases based on ARIMA-GRNN combination model [J]. Chin J Health Stat, 2008, 25(1): 82-83. (in Chinese) 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[J]. 中国卫生统计, 2008, 25(1): 82-83.
- [6] Zhou L, Yu L, Wang Y, et al. A hybrid model for predicting the prevalence of schistosomiasis in humans of Qianjiang city, China [J]. PLoS One, 2014, 9(8): e104875.
- [7] Yu L, Zhou L, Tan L, et al. Application of a new hybrid model with seasonal auto-regressive integrated moving average (ARIMA) and nonlinear auto-regressive neural network (NARNN) in forecasting incidence cases of HFMD in Shenzhen, China [J]. PLoS One, 2014, 9(6): e98241.

(收稿日期:2015-03-08)

(本文编辑:张林东)