文章编号:2095-1116(2014)06-0572-05

·流行病学·

利用遗传算法优化的 ARIMA-BP 组合模型预测手足口病发病趋势

吴文博[1],李虹艾[1],万鹏程[2],袁秀琴[1]
(1.南华大学公共卫生学院,湖南 衡阳 421001;2.南华大学经济管理学院)

摘 要: **目的** 探讨 ARIMA 模型及遗传算法优化的 ARIMA-BP 神经网络组合模型在传染病预测与预警中的应用,为相关部门制定防治措施提供参考。 **方法** 选择某市 2009~2013 年的手足口病发病数作为研究对象,首先建立 ARIMA 模型,得到的拟合值作为神经网络输入值,真实观测值作为输出值,带入通过遗传算法优化的 BP 神经网络中训练,并比较两种模型的预测精度。 **结果** 对建立的 ARIMA(1,0,0)(1,1,0)$_{12}$模型预测的相对误差为 39.89%,决定系数为 0.786,经统计检验残差为白噪声序列;组合模型预测的相对误差为 26.25%,决定系数为 0.852。 **结论** 该组合模型的预测精度高于 ARIMA 模型,且对于极值的预测较为精确,可以为其他传染病的预测及建立统计预警提供参考。

关键词: ARIMA; 遗传算法; BP 神经网络; 组合模型; 手足口病
中图分类号:R181.2 **文献标识码:**A

Predicting the Incidence Trend of Hand-Feet-Mouth Diease by Using the GA Optimized ARIMA-BP Combination Modeling

WU Wenbo, LI Hongai, WAN Pengchen, et al
(*School of Public Health, University of South China, Hengyang, Hunan* 421001, *China*)

Abstract: **Objective** To discuss the protential usage of ARIMA modeling and the model combinated with BP neural network and Gentic Algorithm, particularly, in the field of disease prediction. **Methods** The ARIMA and combination model are established based on the monthly incidence rate of HFMD in Hengyang, comparing the accuracy of the models. **Results** The relative error of the following models are 26.25% &39.89%, the infinitive coefficents are 0.786&0.852. **Conclusion** The combined modeling has a better predicating effect on the HFMD, which can be a reference to other infectious diseases.

Key words: ARIMA; Gentic Algorithm; BP neural network; combination model; HFMD

近年来,手足口病的发病数逐年增多,作为丙类传染病中为我省重点监测的疾病之一,因其发病率高、目前尚无疫苗进行免疫接种、并发症危害大,故其防控形势较为严峻。在传染病的防控中,通过建立相对准确的统计预测模型,从而建立预警监测机制,为制定防控政策和卫生资源配置提供依据,这是传染病防控的重点及难点之一[1]。本文基于某市近年来手足口病的发病资料,先后建立差分自回归移动平均模型(ARIMA)、遗传算法(Gentic Algorithm,GA)优化的 ARIMA-BP 神经网络模型,比较两种模型的预测准确性,并探讨组合模型在预测方面的优越性。

1 材料与方法

1.1 资料来源

数据资料来源于中国疾病预防控制信息系统(传染病报告信息管理系统),按发病日期检索某市 2009 年 1 月 1 日~2013 年 12 月 31 日的手足口病

收稿日期:2014-04-21
基金项目:南华大学学生科研重点课题立项.
作者简介:吴文博,10 级本科预防医学专业在读,研究方向:公共卫生与预防医学,E-mail:beanwuwenbo@163.com. 通讯作者袁秀琴,硕士,教授,硕士研究生导师,研究方向:疾病监测与防控,E-mail:wtjysh@126.com.

月发病数。

1.2 方法

将数据导入 SPSS18.0 中,检查有无缺失数据。以 2009 年 1 月 1 日 ~2012 年 12 月的发病数据作为模型拟合值,预测 2013 年手足口病月发病数,将预测值与实际值进行比较,以相对偏差的大小衡量模型的精确性。在建模方法上,首先建立 ARIMA 模型,将 ARIMA 模型预测值作为 BP 神经网络的输入值,真实值作为输出值,同时应用遗传算法优化 BP 神经网络结构,不断寻找预测值与真实值的关系,从而调整 ARIMA 模型的预测精度。比较这两种模型的相对偏差,并对模型预测精度进行评价。

1.3 原理

1.3.1 ARIMA 模型 ARIMA 模型的基本思想是将预测对象随时间推移而形成的数据序列视为一个随机序列,用一定的数学模型来近似描述这个序列[2]。该模型由美国数学家 Box 与英国统计学家 Jenkins 提出,在预测具有季节周期性的时间序列中,以乘积季节性模型最为常用,记 ARIMA(p,d,q)$(P,D,Q)_n$,其中 p,q 为非季节性模型的自回归项及移动平均项数,d 为时间序列平稳化时所做的差分次数;P,Q 为季节性模型的自回归及移动平均项,D 为季节差分的阶数。其建模过程见图 1。

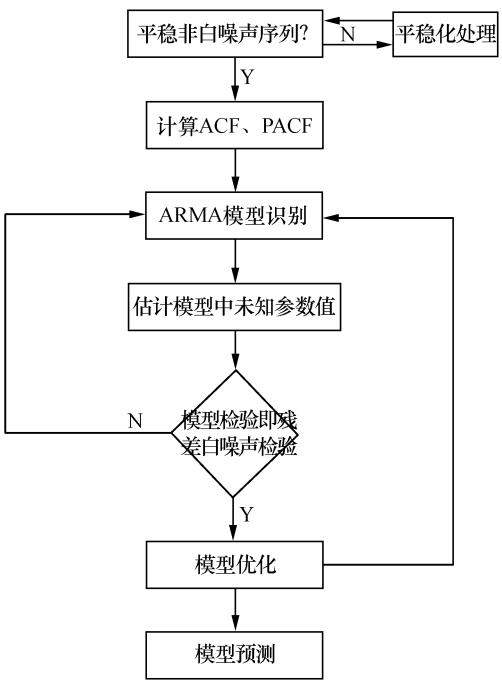

图 1 ARIMA 的建模过程

1.3.2 BP 神经网络及遗传算法 BP 神经网络是一种多层前馈神经网络,该网络的主要特点是信号前向传递,误差反向传播。在前向传递中,输入信号从输入层经隐含层逐层处理,直至输出层。每一层的神经元状态只影响下一层神经元状态[3]。如果输出层得不到期望输出,则转入反向传播,根据预测误差调整网络权值和阈值,从而使 BP 神经网络预测输出不断逼近期望输出。BP 神经网络的拓扑结构如图 2 所示。根据输入向量 X,输入层和隐含层间连接权值 ω_{ij},以及隐含层阈值 a,可以计算隐含层输出向量 H,即:

$$H_j = f\left(\sum_{i=1}^{n} \omega_{ij} X_i - a_j\right), \text{其中 } j = 1,2,3\cdots,l.$$

式中,l 为隐含层节点数;f 为隐含层激励函数,本模型中选用 $f(x) = \dfrac{1}{1+e^{-x}}$

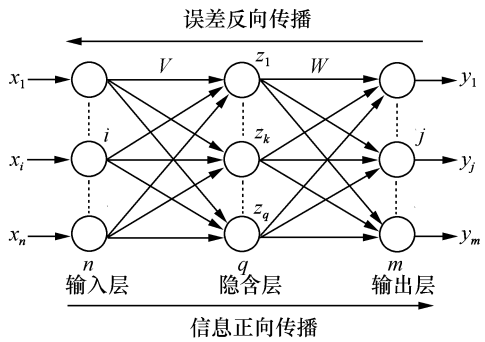

图 2 BP 神经网络拓补结构

但 BP 神经网络容易受数据极值的影响,从而导致预测精度的下降,同时由于各层权重的权值是主观经验确定的,导致 BP 神经网络的主观倾向性过大,为此引入 GA 算法。

遗传算法(Genetic Algorithm,GA)是一种进化算法,其基本原理是仿效生物界中的“物竞天择、适者生存”的演化法则。遗传算法的做法是把问题参数编码为染色体,再利用迭代的方式进行选择、交叉以及变异等运算来交换种群中染色体的信息,最终生成符合优化目标的染色体[4]。遗传算法所优化的内容是 BP 神经网络的初始权值和阈值,神经网络的权值和阈值一般是通过随机初始化为[-0.5,0.5]区间的随机数,这个初始化参数对网络训练的影响很大,引入遗传算法就是为了优化出最佳的初始权值和阈值,以提高研究的预测精度。

1.4 软件平台

ARIMA 模型构建采用 SPSS18.0,神经网络及遗传算法编程采用 MATLAB 7.0,操作系统为 Windows XP。

2 结　果

2.1 ARIMA 模型的建立

(1)模型的平稳化识别

首先定义估计区间,依据 2009 年 1 月 ~ 2012 年 12 月的发病数拟合模型,并绘制序列图,见图 3。从序列图中可见,手足口病发病数序列存在着明显的季节性周期波动规律,在每一年 5 ~ 8 月份,出现发病高峰;报告数在 2012 年明显增多。总的来看,手足口病发病数呈现逐年增多的趋势,提示该序列是一个非平稳的序列,需对其进行差分,和对数转化,以达到序列平稳化的目的。

(2)季节性 ARIMA 模型的建立:通过做出自相关函数(ACF)、偏自相关函数(PACF)图,见图 4。依据"截尾性"估计季节性模型参数;同时在残差中识别非季节性模型,确定若干个备选模型后,利用 Bayesian 信息准则(BIC)最小原则,确定最后的模型,模型 ARIMA(1,0,0)(1,1,0)$_{12}$的 BIC 最小,为 13.002。并对该模型进行诊断、对模型参数进行估计,相关统计量见表 1,表 2。同时利用 Ljung-Box 方法检验残差白噪声,得到 L-B 统计量为 16.847,P > 0.05,可以认为残差为白噪声序列。综上,选用 ARIMA(1,0,0)(1,1,0)$_{12}$模型进行预测。

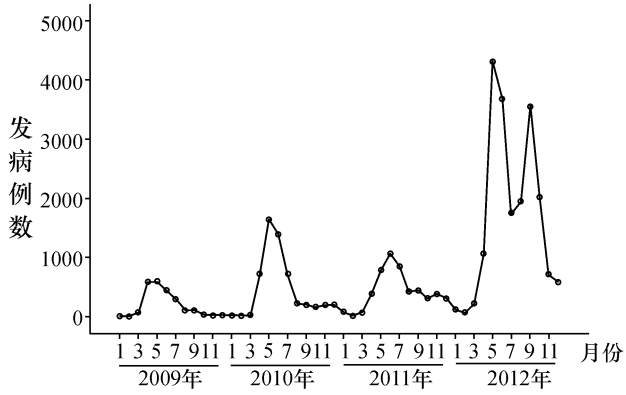

图 3　手足口病发病数序列图

表 1　预测模型的相关检验统计量

	Df	ARIMA(1,0,0)(1,1,0)$_{12}$
R^2	16	0.839
MAE	16	369.109
SMAE	16	367.143
BIC	16	13.002
B-J	16	7.839

表 2　模型的参数估计

	估计值	标准误	t 值	P 值
常数项	3.875	0.997	2.075	0.038
自回归滞后 1	0.943	0.146	3.875	0.011
季节性差分	1			

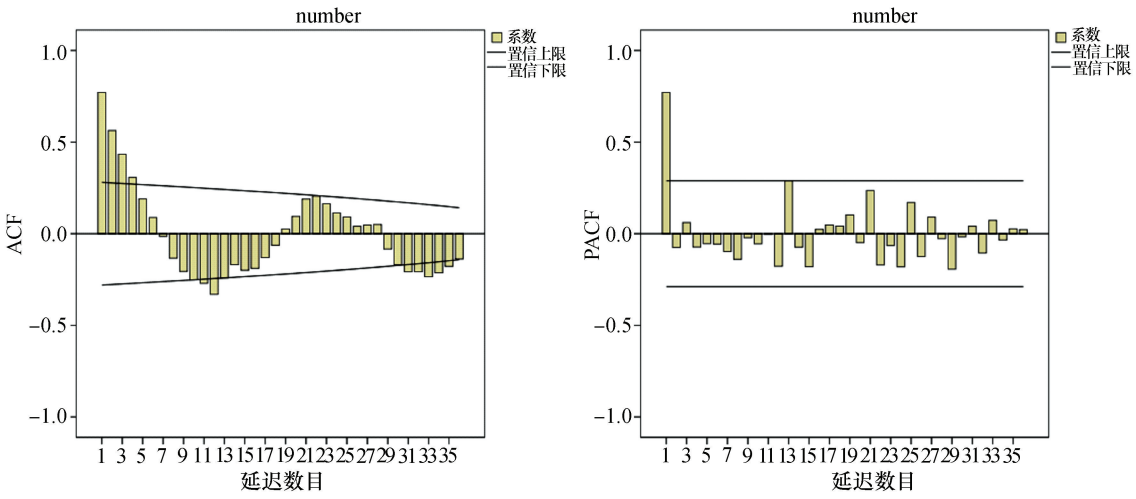

图 4　手足口病序列的 ACF、PACF 图

2.2 ARIMA-BP 组合模型的建立与遗传算法的优化

(1)将 ARIMA 模型预测得到的预测值、2013 年 1 月～12 月的手足口病实际发病数数据进行归一化处理,使其集中在[-0.5,0.5]之间;

(2)学习样本的选择

输入变量:根据最优的 ARIMA 模型得到的预测值 $\hat{y}_i$;

输出变量:2013 年 1 月～12 月的实际观测值 y_i。

(3)网络初始化,利用遗传算法进化 50 代寻找最佳初始权值和阈值。确定遗传算法参数的程序代码如下:

```
%% 定义遗传算法参数
NIND = 5;            % 个体数目
MAXGEN = 50;         % 最大遗传代数
PRECI = 10;          % 变量的二进制位数
GGAP = 0.95;         % 代沟
px = 0.7;            % 交叉概率
pm = 0.01;           % 变异概率
trace = zeros( N + 1, MAXGEN);
        % 寻优结果的初始值
FieldD = [ repmat ( PRECI, 1, N); repmat ([ -0.5;0.5], 1, N); repmat ([1;0;1;1], 1, N)];   % 区域描述器
```

(4)依据遗传算法寻找的最佳阈值与连接权值,计算隐含层与输出层;

(5)计算误差,不断迭代更新阈值与权值,直至误差保持在稳定水平时输出最后结果。迭代进化过程中误差变化见图 5。

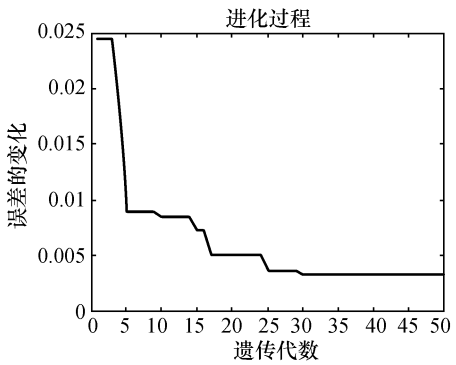

图 5 遗传算法进化过程中误差的变化

(6)进化完成后,使用最优的网络参数,将测试数据代入 BP 神经网络进行仿真,通过优化后的网络结构进行学习,不断寻找 $\hat{y}_i$ 与 y_i 之间的关系,从而达到优化 ARIMA 预测精度的最终目的。

利用 BP 神经网络进行预测的程序如下:

```
net. iw{1,1} = reshape( w1, hiddennum, inputnum);
net. lw{2,1} = reshape( w2, outputnum, hiddennum);
net. b{1} = reshape( B1, hiddennum, 1);
net. b{2} = reshape( B2, outputnum, 1);
%% 训练网络
net = train( net, inputn, outputn);
% BP 网络预测
% 预测数据归一化
inputn_test = P_test. /80;
% 网络预测输出
an = sim( net, inputn_test);
```

2.3 ARIMA 模型与组合模型的预测精度比较

利用上述模型得到 2013 年各月某市手足口病的预测发病数。预测精度通过统计量相对偏差 RSE 衡量,RSE 反映点预测值 $\hat{y}_i$ 与真实值 y_i 的误差,即:$RSE = \frac{y_i - \hat{y}_i}{y_i}$。结果见表 3。

表 3 ARIMA 模型与组合模型预测 2013 年各月份发病数的结果比较

月份	ARIMA 模型		组合模型	
	预测值	RSE	预测值	RSE
1	178	0.45	117	-0.05
2	53	-0.29	124	0.66
3	289	0.29	187	-0.16
4	1000	-0.07	993	-0.07
5	2516	-0.42	4091	-0.05
6	4546	0.23	3435	-0.07
7	2283	0.30	1394	-0.21
8	1101	-0.44	1042	-0.47
9	1944	-0.45	2535	-0.29
10	1410	-0.30	480	-0.76
11	1468	1.06	820	0.15
12	601	0.02	457	-0.22

由表 4 可见,组合模型的平均预测精度高于 ARIMA 模型,且对于部分极端值(尤其是发病高峰 5～8 月)的预测较为准确。对模型整体的评价指标见表 4。其中决定系数 R^2 用来反映模型的拟合效果,即:

$$R^2 = \frac{SS_i - SS_o}{SS_o}$$

SS_i 表示预测值的离均差平方和,SS_o 表示实际值得离均差平方和。

表 4 两种模型的综合评价

	RSE(%)	R^2
ARIMA 模型	39.89	0.786
组合模型	26.25	0.852

3 讨 论

传染病的发病预测是当前传染病疾病预防与控制的难点[5]。用于发病预测的方法有很多,实际工作中通常的做法是定性预测法,即基于日常的疾病监测数据、传染病的发病特点,进行趋势外推。这种做法主观因素较大,没有充分利用监测数据,难以保证预测的准确性。对于组合模型,国内外相关的研究相对较少,如:Gamer 分解定律、对不同模型给予一定的权重[6],但权重的设定仅能依靠经验的判断,导致预测的主观性增加,影响了预测的精度。有关基于 ARIMA 的组合模型预测,朱玉等[7]利用 ARIMA-GRNN 对猩红热的发病进行拟合;章勤等[8]则利用 BP 神经网络对矽肺的发病情况,进行了预测,取得了较好的效果。虽然上述采用神经网络的组合预测方法,避免了对各分模型权重大小选取的讨论,但是神经网络结构的阈值以及连接权值,均需要依靠经验反复设定,在实际利用中较为繁琐。本次研究的创新点在于使用了遗传算法,通过进化代数的迭代,优化了神经网络参数,使得预测结果更加科学可信。

从预测结果上看,利用遗传算法优化的 ARIMA-BP 组合模型实现了对手足口病的发病趋势的预测,相比于 ARIMA 模型,该组合模型对极端值的预测效果较好,此外,从整体上看,手足口病的发病数年年攀升,需引起足够重视,由于手足口病致病病原体种类繁多,难以进行免疫接种干预,故应在每年的疾病高发时期,建立统计预警,加大对托幼机构、学校的消毒、卫生检查以及卫生宣教;广大医疗机构应加强监测筛检机制,动员医疗资源做到早发现、早诊断、早治疗。

由于纳入本次建模的数据样本量并不大,组合模型的效果变化并不是非常明显,但当数据量增大时,采用本法的组合模型,将会依旧保持良好的预测精度。此外,介于本模型的非线性映射能力较强、不受极端数值的影响[9],基于 ARIMA 的组合模型尚可以用于金融、人口、水文以及环境保护领域,但仍需进一步的研究与讨论。

参考文献:

[1] 罗静,杨舒,张强,等. 时间序列 ARIMA 在艾滋病疫情预测中的应用[J]. 重庆医学, 2012, 41(13): 1255-1257.

[2] 张晋昕,医学时间序列分析及其预测应用相关问题的研究[D]. 西安:第四军医大学博士研究生毕业论文, 2000:23-25.

[3] SimonHaykin. 神经网络原理[M]. 北京:机械工程出版社,2004:154-167.

[4] 史峰,王小川,郁磊,等. MATLAB 神经网络 30 个案例分析[M]. 北京:北京航天航空大学出版社,2010: 21-25.

[5] 牟瑾,谢旭,李媛,等. 将 ARIMA 模型应用于深圳市 1980~2007 重点法定传染病预测分析[J]. 预防医学论坛,2009,15(11):1051-1055.

[6] Erxu Pi, Mantri, Sai Ming Ngai. BP-ANN for fitting the temperature-germination model and its application in predicting sowing time and region for bermudagrass[J]. PLoS One 2013,11(8):e82413.

[7] 朱玉,夏结来,王静. 单纯 ARIMA 模型和 GRNN-ARIMA 组合模型在猩红热发病率中的研究[J]. 中华流行病学杂志,2009,30(9):964-966.

[8] 章勤,田晶,孙傲冰,等. 基于 BP 神经网络的矽肺病组合模型预测研究[J]. 计算机科学, 2009, 36(4): 265-269.

[9] 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.

(此文编辑:蒋湘莲)