【论著】

BP神经网络在传染病时间序列预测中的应用及其MATLAB实现

刘天[1],姚梦雷[1],黄继贵[1],陈红缨[2],黄淑琼[2],杨雯雯[2],蔡晶[2],吴然[2]

[摘要] **目的** 利用MATLAB软件,建立基于BP神经网络的传染病时间序列预测模型。**方法** 以荆州市2005-01/2017-12和2018-01/05乙类传染病发病数作为拟合样本和预测样本,建立12-10-5三层BP神经网络模型并预测2018-01/05的逐月发病数。以最小均方差(MSE)、R^2、平均相对误差百分比(MAPE)、平均绝对误差(MAE)4个指标评价BP神经网络的拟合和预测效果。**结果** BP神经网络拟合和预测的MSE、R^2、MAPE、MAE依次分别为11 662.74,0.85,5.19%,85.87和32 729.59,0.22,12.20%,140.31。**结论** BP神经网络对传染病时间序列拟合及预测效果较好;MATLAB软件能用于BP神经网络模型的建立。

[关键词] BP神经网络;MATLAB;时间序列;预测

[中图分类号] R-05 [文献标识码] A [文章编号]1006-4028(2019)08-0812-06

Application of Back Propagation Neural Network in Prediction of Infectious Disease Time Series and Its MATLAB Implementation

LIU Tian[1], **YAO Menglei**[1], **HUANG Jigui**[1], **CHEN Hongying**[2], **HUANG Shuqiong**[2], **YANG Wenwen**[2], **CAI Jing**[2], **WU Ran**[2]

1 *Jingzhou Municipal Center for Disease Control and Prevention, Jingzhou* 434000, *Hubei Province*, *China*.

2 *Hubei Center for Disease Control and Prevention*, *Jingzhou* 430000, *Hubei Province*, *China*.

Abstract **Objective** Using MATLAB software to establish a time series forecasting model of infectious diseases based on back propagation neural network (BPNN). **Methods** The monthly reported caseload of Class B infectious diseases in Jingzhou City from January 2005 to December 2017 and January- April of 2018 were used as modeling and forecasting samples, respectively. 12-10- 5 three- layer BP neural network model was established to predict the monthly incidence of 2018- 01/05. The fitting and forecasting effects of BP neural network were evaluated based on four metrics: mean square error (MSE), R^2, mean relative percentage error (MAPE) and mean absolute error (MAE). **Results** The fitting and

基金项目:1 湖北省卫生计生委疾控专项 (项目编号:WJ2016JT-002)

2 湖北省荆州市2017年卫生科技计划项目(项目编号:2017130)

作者单位:1 荆州市疾病预防控制中心 (湖北 荆州 434000)

2 湖北省疾病预防控制中心 (武汉 430000)

作者简介:刘天(1991-),男,医师,疾病监测预警方法

通信作者:姚梦雷,E-mail:jzcrbs@163.com

forecasting effect of BPNN for MSE, R^2, MAPE and MAE were 11 662.74, 0.85, 5.19%, 85.87 and 32 729.59, 0.22, 12.20%, 140.31, respectively. **Conclusion** BP neural network has a good effect in fitting and forecasting infectious disease time series. MATLAB software can be well used for the establishment of BP neural network model.

Key words BP neural network(BPNN); MATLAB; time series; forecasting

时间序列是将某一指标在不同时间上的数值按时间先后顺序排列而成的数列[1]。对传染病时间序列进行观察、研究，找寻它发展变化的规律，并对未来准确预测，可为卫生资源合理配置提供科学依据。随着科学技术的迅猛发展和计算机的普及，ARIMA[2]、灰色模型[3]、残差自回归模型[4]等多种预测模型作为经典的预测技术，在传染病预测中得到了广泛应用。但预测模型无法处理非线性关系的局限性限制了模型的适用性。人工神经网络作为新兴的技术，能很好的处理的非线性关系，在人工智能领域取得了广泛成功。BP神经网络作为人工神经网络中最可靠、最经典的神经网络，具有学习能力强、操作简单的优点，近年来逐渐被应用于疾病监测中[5]。但目前报道多直接给出研究结果，关于如何整理数据、建立BP网络和参数设置鲜有报道。本文以荆州市2005-01/2018-05乙类传染病逐月发病数为例，旨在阐述BP神经网络的基本原理并应用MATLAB R2016a软件加以实现。

1 材料与方法

1.1 资料来源 2005-01/2018-05荆州市乙类传染病逐月发病数来源于中国疾病预防控制中心“传染病报告信息管理系统”，按现住址、发病日期统计数据。

1.2 方法 BP神经网络是一种依据误差逆向传播算法而训练的多层前馈神经网络。BP神经网络模型的拓扑结构可以分为输入层、隐含层和输出层3个层次。它是按照给定的(输入、输出)样本进行学习，按照一定的训练标准(如最小均方差)，计算网络的实际输出值与期望输出值的误差，不断进行误差反向传播，从而来调整网络的各层权重，使误差达到最小，完成学习的目的(图1)。

BP神经网络建模的基本步骤主要包括：①初始化网络，包括一些网络参数的选择和设定；②训练，训练时应尽量防止网络过度拟合；③仿真。如何初始化网络以及提高模型泛化能力对BP神经网络构建十分关键，下面介绍各种参数设置。

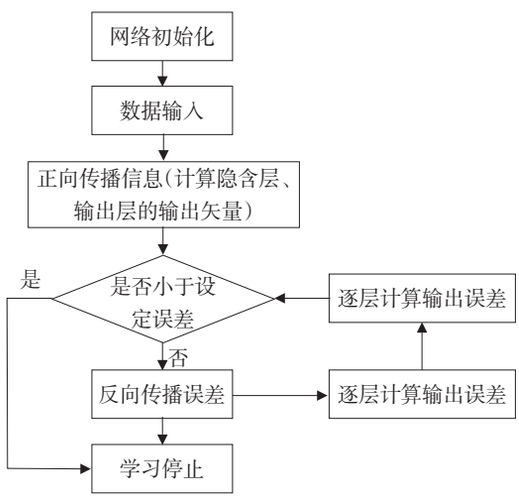

图1 BP神经网络学习流程图

1.2.1 数据归一化和反归一化 数据归一化法能消除各维度数据间数量级差别，避免因输入输出数据数量级别差别较大造成较大的网络误差，是神经网络初始化时对数据常做的处理方法。本文采用最大最小法，函数形式如下：

$$x'_k = \frac{x_k - x_{\min}}{x_{\max} - x_{\min}}$$

MATLAB归一化函数为‘mapminmax’。反归一化即指将经BP神经网络输出的数据还原，对应函数为‘reverse’。

1.2.2 传递函数的选择 常用函数包括对数S型函数、双曲正切函数和线性函数。BP神经网络的传输函数包括输入层与隐含层间、隐含层与输出层间的传输函数[6-7]。输入层与隐含层间的传输函数选择对数S型函数(‘logsig’)，隐含层与输出层间的传输函数选择线性函数(‘purelin’)。

1.2.3 神经网络层数及各层神经元个数的确定 由映射存在定理可知，任意连续函数可由一个3层感知器网络逼近[8]，本文采用3层BP神经网络。

选取t-12，t-11，……，t-1月份的12个数据作为输入；t，t+1，t+2，t+3，t+4月份的5个数据作为为输出；确定输入层、输出层神经元个数分别为12和5。隐含层神经元个数的确定通常根据经验公式$J=\sqrt{n+m}+a$选择，其中J为隐含层单元数，n、m分别为输入层、隐含层单元数，a为1-10之间常数。本文为演示方便，指定隐含层神经元个数为10。

1.2.4 初始化权值和阈值 一般对初始化权值及阈值取[-1, 1]之间的随机数，在MATLAB中由‘init’函数实现，在建立BP神经网络时自动调用。

1.2.5 训练函数的选择 BP神经网络训练算法种类较多[9]。其中Levenberg_Marquardt算法在处理函数逼近问题的能力较强，网络收敛速度最快，是BP神经网络最常用的算法。MATLAB中对应函数为‘trainlm’，本文选择‘trainlm’函数作为训练函数。

1.2.6 其他参数设置 ①选取最小均方差（MSE）作为评价指标；②期望误差最小值设为0.000 1；③最大训练步长设为1 000；④学习率设为0.05；⑤训练样本数：验证样本数：测试样本数设为10∶0∶0；⑥其余设置均为MATLAB神经网络工具箱默认设置。

1.3 评价指标 采用最小均方差（MSE）、R^2、平均相对误差百分比（MAPE）、平均绝对误差（MAE）评价神经网络的拟合和预测效果。

2 结果

2.1 数据整理 利用2005-2017年荆州市逐月乙类传染病发病数训练网络，共计140组数据；以2017年1-12月数据作为测试数据输入已建立的BP神经网络，输出2018-01/05预测值并与实际值比较，评价模型预测效果。

在Excel中录入数据，再用MATLAB调用Excel文件中的数据。基于MATLAB的神经网络建模是以矩阵的形式进行，每列为一次输入（输出）数据集，每行代表一个输入（输出）神经元数据集。输入层需建立12×140的矩阵，即在Excel中录入12行、140列数据，如A1：A12依次录入2005-01、2005-02、…、2005-12的数据；B1：B12则依次录入2005-02月、2005-03、…、2016-01的数据。输出层数据整理方法类似输入层，建立5×140的矩阵。

2.2 BP神经网络模型建立 在MATLAB R2016a中，利用神经网络工具箱中的‘feedforwardnet’、‘train’、‘sim’3个函数完成网络的创建、训练和仿真。相应的MATLAB程序如下：

```
%% 读取Excel中的数据
filename = ‘Book1.xlsx’;%数据录入在‘Book1.xlsx’文件中
sheet1 = 2;%输入数据放在第2个sheet表格中
xlRange = ‘A1:EJ12’;%12×140矩阵的输入数据
input_train=xlsread(filename,sheet1,xlRange);%读取输入数据
sheet2 = 3;%输出数据放在第3个sheet表格中
x2Range = ‘A1:EJ5’;%5×140矩阵的输出数据
output_train=xlsread(filename,sheet2,x2Range);%读取输出数据
sheet3 = 4;%测试数据放在第4个sheet表格中
x3Range = ‘A1:A12’;%12×1矩阵的测试数据
input_test=xlsread(filename,sheet3,x3Range);%读取测试数据
sheet4 = 5;%预测数据放在第5个sheet表格中
x4Range = ‘A1:A5’;%5×1矩阵的预测数据
output_test=xlsread(filename,sheet4,x4Range);%读取预测数据
%%数据归一化
[inputn, inputps]=mapminmax(input_train);%输入数据归一化
[outputn,outputps]=mapminmax(output_train);%输出数据归一化
%% BP神经网络创建
net = feedforwardnet(10, ‘trainlm’);%隐含层神经元个数为10，训练函数为trainlm
net.layers{1}.transferFcn = ‘logsig’;%设置隐含层转换函数为logsig
net.layers{2}.transferFcn = ‘purelin’;%设置输出层转换函数为purelin
net.performFcn = ‘mse’;%选择均方差（mse）作为评价指标
```

```
net.trainParam.epochs = 1 000;%最大训练步数设为1 000
net.trainParam.max_fail = 6;%最大验证失败步数设为6
net.trainParam.lr = 0.05;%学习率设为0.05
net.trainParam.goal = 0.000 1;%目标误差设为0.000 1
net.divideParam.trainRatio = 10;%训练样本占比设为100%
net.divideParam.valRatio = 0;%验证样本占比设为0%
net.divideParam.testRatio = 0;%测试样本占比设为0%
net = train(net,inputn,outputn);%开始训练
%% 计算mse
y1 = net(inputn);%查看输出拟合值
perf1 = perform(net,outputn,y1);%计算归一化后的mse
%% 查看权重值和阈值
net.iw{1};%查看权重值
net.b{1};%查看阈值
%% 预测
inputn_test = mapminmax('apply',input_test,inputps);%预测数据归一化
out = sim(net,inputn_test);%BP神经网络输出
BPoutput = mapminmax('reverse',out,outputps);%输出反归一化,得到预测值BPoutput
fit_output_train = mapminmax('reverse',y1,outputps);%得到拟合数据fit_output_train
%% 拟合效果
train_e = fit_output_train-output_train;%计算拟合值残差
train_e2 = train_e.^2;%计算拟合值残差平方
train_sse = sum(sum(train_e.^2));%计算拟合值残差平方和SSE
train_S2 = var(output_train(:));%计算实际值方差S²
train_mse = train_sse/700;%计算拟合值均方差MSE
train_r2 = 1-(train_mse/train_S2);%计算拟合R²
train_mae = mean(abs(train_e(:)));%计算拟合平均绝对误差MAE
train_ae = train_e./output_train;%计算拟合相对误差AE
train_mape = mean(abs(train_ae(:)));%计算拟合平均相对误差MAPE
%% 预测效果
test_e = BPoutput-output_test;%计算预测值残差
test_e2 = test_e.^2;%计算预测值残差平方
test_sse = sum(sum(test_e.^2));%计算预测值残差平方和SSE
test_S2 = var(output_test(:));%计算预测值方差S²
test_mse = test_sse/5;%计算预测值均方差MSE
test_r2 = 1-(test_mse/test_S2);%计算预测R²
test_mae = mean(abs(test_e(:)));%计算预测平均绝对误差MAE
test_ae = test_e./output_test;%计算预测相对误差AE
test_mape = mean(abs(test_ae(:)))%计算预测平均相对误差MAPE
save net;%保存建立的BP神经网络
```

2.3 模型拟合及预测结果 在运行上述程序中，最终得到训练实际值、训练拟合值、预测值及实际值，由此可计算出拟合和预测的MSE、R^2、MAPE、MAE，本研究经过连续10次拟合，选取预测MAPE最小者对应的网络为最优BP神经网络，其中拟合平均相对误差为5.19%，预测平均相对误差为12.20%，拟合及预测效果均较好。拟合及预测结果见表1和图2。

表1 BP神经网络拟合及预测效果

指标	拟合效果	预测效果
MAPE/%	5.19	12.20
R^2	0.85	0.22
MSE	11 662.74	32 729.59
MAE	85.87	140.31

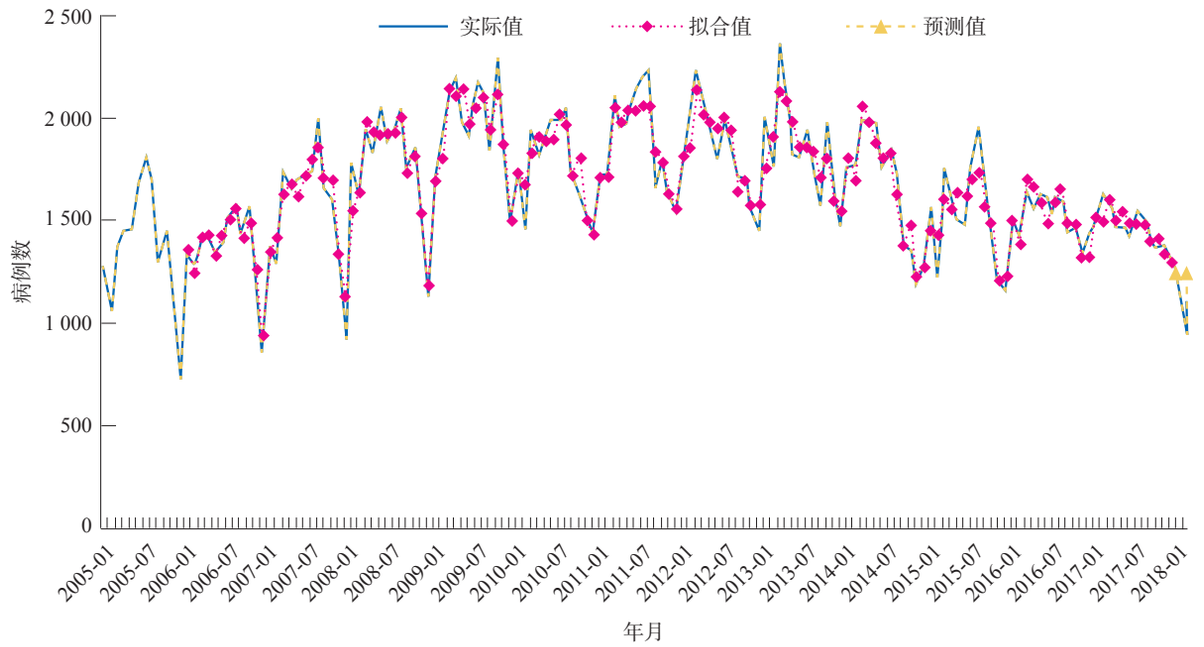

图2 BP神经网络拟合值、预测值与实际值拟合效果

3 讨论

BP神经网络作为一种非线性拟合方法，在疾病监测领域已有广泛应用，但目前BP神经网络应用于传染病时间序列预测的报道较少。BP神经网络的输入层与输出层神经元个数常需根据研究目的确定，并直接影响网络的拟合效果。对于类似于Logistic回归分析中的多因素数据的BP神经网络分析，可以充分考虑研究需求，容易确定输入层与输出层单元数[10]；而时间序列数据为单变量连续性资料，输入层与输出层单元数的确定较难从专业角度确定，目前缺乏统一标准，这也限制了BP神经网络在传染病时间序列预测预测中的应用。本研究以2005-01/2018-05荆州市乙类传染病发病数时间序列为例，考虑传染病数据一般以年为一个周期并且便于比较，尝试建立12-10-5 3层BP神经网络，拟合及预测平均相对误差分别为5.19%和12.20%，建立的模型拟合、预测精度高，可以用于预测[11]。提示BP神经网络在传染病发病数时间序列中应用效果较好。

MATLAB软件是世界三大数学软件之一，其自带的神经网络工具箱方便用户直接调用工具箱中的函数，达到建立神经网络的目的[12]。虽然MATLAB软件神经网络建模简单，但卫生人员缺乏基础编程知识，使得很多同行望而却步；另外自MATLAB 2011b版本以来，工具箱中的函数发生了变化，模型建立函数‘newff’被弃用，启用了‘feedforwardnet’函数，但目前市面教材仍以介绍‘newff’为主，极大不便于学习[13]。本文以‘feedforwardnet’函数建模，并将数据整理、模型建立过程及代码进行演示，旨在为同行理解、学习BP神经网络提供参考。

BP神经网络是神经网络中应用最广泛的模型，但它存在一定的局限性[14]。首先，BP神经网络建模参数较多，难于确定。例如隐含层神经元个数、学习率等，本文中为了演示方便[3,7]，直接指定数值。一般在实际应用中，建议采取凑试法，选取误差最小者为最优模型。其次，BP神经网络模型存在训练过度的问题，训练过度常导致网络拟合效果好，而泛化（预测）能力较差。而目前对于这一问题只能采用将样本差分为训练集、验证集和测试集，造成样本量的浪费。最后，BP神经网络可解释性很差，是一种“黑箱”操作。

综上所述，尽管BP神经网络仍然有大量问题亟需解决，但在本例中，BP神经网络的拟合及预测能力较好，提示BP神经网络可以作为传染病时间序列预测的一种方法。今后可以尝试采用凑试法调整模型参数，并将样本分割成3部分以提高模型的精度和泛化能力。

参考文献

[1]杨思凡. 时间序列建模与预测[D]. 北京：清华大学，2014：1-2.

（下转第821页）

共同参与治理的核心工作，全方位强化宣传教育、行为干预、安全套促进工作，遏制艾滋病经家庭内、非商业性行为传播，同时要全面开展安全套推广、高危人群干预工作，强化商业性性交易场所及人员监测、管控，大力打击故意传播行为，从而遏制性病艾滋病经商业性性行为传播。

参考文献

[1]中国疾病预防控制中心性病艾滋病预防控制中心.2018年第3季度全国艾滋病性病疫情[J].中国艾滋病性病，2018，24(11)：1075.

[2]张晴晴，唐作红，白永华，等.2011-2015年攀枝花市艾滋病疫情及流行趋势分析[J].预防医学情报杂志，2017，33(9)：927-930.

[3]熊馥，解瑞青，柯贤洲，等.黄石市2011—2017年艾滋病疫情及流行特征分析[J].中国热带医学，2018，18(09)：906-909.

[4]肖波，俸卫东，刘固国，等.柳州市2010-2013年艾滋病疫情分析[J].公共卫生与预防医学，2015，26(1)：91-92.

[5]赵鑫，卫晓丽，李恒新.西安市2000-2016年报告的50岁及以上HIV/AIDS病人特征分析[J].中国艾滋病性病，2017，23(12)：1101-1104.

[6]岳红林，刘亚伦，张燕，等.蒲江县50岁及以上老年人艾滋病疫情分析[J].预防医学情报杂志，2017，33(12)：1254-1258.

[7]曹妍，冉定鑫，李杨，等.2002-2017年阆中市艾滋病流行特征分析[J].现代预防医学，2018，45(19)：3484-3487.

[8]阳凯，李丽娜，彭国平，等.湖北省HIV/AIDS病例异性性接触传播特征分析[J].预防医学，2018，30(10)：997-1001.

[9]贾伯成，袁风顺，邱锋，等.2013-2017年达州市艾滋病疫情分析[J].预防医学情报杂志，2018，34(7)：992-996.

[10]中国疾病预防控制中心性病艾滋病预防控制中心.2017年12月全国艾滋病性病疫情[J].中国艾滋病性病，2018，24(2)：111.

[11]刘莉，裴晓迪，张子武，等.2009-2012年四川省艾滋病疫情估计[J].预防医学情报杂志，2014，30(9)：707-712.

[12]何静，赵西和.2013-2017年四川省绵阳市艾滋病疫情分析[J].寄生虫病与感染性疾病，2018，16(3)：119-124.

[13]余敏菊，陈聪，王美凤，等.2011-2016年内江市艾滋病疫情监测分析[J].寄生虫病与感染性疾病，2017，15(4)：211-215.

[14]黄春玲，杨林.2011-2016年遂宁市经性途径传播艾滋病疫情分析[J].预防医学情报杂志，2017，33(12)：1259-1261.

（收稿日期：2019-04-08）

（上接第816页）

[2] Zheng YL, Zhang LP, Zhang XL, *et al*. Forecast model analysis for the morbidity of tuberculosis in Xinjiang, China[J]. PLoS One, 2015, 10(3): e116832.

[3]王山.时间分布模型在甲肝流行趋势分析中的应用[D].杭州：浙江大学，2016：49-51.

[4]王永斌，柴峰，李向文，等.ARIMA模型与残差自回归模型在手足口病发病预测中的应用[J].中华疾病控制杂志，2016，20(3)：303-306.

[5] Hu H, Wang H, Wang F, *et al*. Prediction of influenza-like illness based on the improved artificial tree algorithm and artificial neural network [J]. Sci Rep, 2018, 8(1): 4895.

[6] Zhang X, Liu Y, Yang M, *et al*. Comparative study of four time series methods in forecasting typhoid fever incidence in China [J]. PLoS One, 2013, 8(5): e63116.

[7]王艳旭.基于系统聚类与BP神经网络的世界碳排放预测模型及应用研究[D].南昌：南昌大学，2016：37-38.

[8] Hagan, Martin T, Demuth, *et al*. Neural network design [M]. USA: PWS Publishing Company, 1996: 2-3.

[9]刘碧瑶.基于BP神经网络的住院费用建模研究[D].杭州：浙江大学，2006：18-19.

[10]张晓玲.基于BP神经网络的大理州艾滋病流行现状分析和疫情预测模型的研究[D].昆明：云南大学，2013：21-29.

[11] Liu L, Luan RS, Yin F, *et al*. Predicting the incidence of hand, foot and mouth disease in Sichuan province, China using the ARIMA model-CORRIGENDUM[J].Epidemiology & Infection, 2016, 144(1): 144-151.

[12]王小川.MATLAB神经网络43个案例分析[M].北京：北京航空航天大学出版社，2013：1-3.

[13]陈明.MATLAB神经网络原理与实例精解[M].北京：清华大学出版社，2013：167-172.

[14]张文彤，董伟.SPSS统计分析高级教程[M].北京：高等教育出版社，2013：346-348.

（收稿日期：2019-01-15）