

· 基础理论与方法 ·

神经网络在麻疹预测预警中的应用

余滨 丁春 魏善波 陈邦华 刘普林 罗同勇 王家刚 潘志伟 陆君安

【导读】探讨神经网络在麻疹发病数预警中的应用,以武汉市1986年1月至2006年8月麻疹病例数为依据,分别以月和周为单位,应用神经网络建立模型进行预测。结果表明,在用两层反向传播(BP)神经网络建立时间序列动态模型时,当 $p=9$ 时,网络的收敛速度可以接受,相关系数达到0.85,接近1,按月预测较准确;用概率神经网络(PNN)按周分类预测较好。数据积累较多时,用两层BP神经网络预测作早期预警较为可行;对于某些疾病因数据的积累时间较短,利用时间序列精确预测不能得到令人满意的结果时,可以用PNN预测。该方法有望在疾病暴发早期预警系统的建立中得到应用。

【关键词】 麻疹;神经网络;预警

Early warning on measles through the neural networks YU Bin¹, DING Chun², WEI Shan-bo³, CHEN Bang-hua¹, LIU Pu-lin¹, LUO Tong-yong¹, WANG Jia-gang³, PAN Zhi-wei³, LU Jun-an⁴. 1 Wuhan Center for Disease Control and Prevention, Wuhan 430015, China; 2 Nanjing Marine Radar Institute; 3 Wuhan Public Health Bureau; 4 School of Mathematics and Statistics, Wuhan University

Corresponding author: DING Chun, Email: tingnt@126.com

This work was supported by a grant from the Ministry of Health China and WHO (No. WP/2006/CHN/CSR/1.5/001).

【Introduction】 To discuss the effects on early warning of measles, using the neural networks. Based on the available data through monthly and weekly reports on measles from January 1986 to August 2006 in Wuhan city. The model was developed using the neural networks to predict and analyze the prevalence and incidence of measles. When the dynamic time series model was established with back propagation (BP) networks consisting of two layers, if p was assigned as 9, the convergence speed was acceptable and the correlation coefficient was equal to 0.85. It was more acceptable for monthly forecasting the specific value, but better for weekly forecasting the classification under probabilistic neural networks (PNN). When data was big enough to serve the purpose, it seemed more feasible for early warning using the two-layer BP networks. However, when data was not enough, then PNN could be used for the purpose of prediction. This method seemed feasible to be used in the system for early warning.

【Key words】 Measles; Neural networks; Early warning

麻疹是由麻疹病毒引起的一种急性呼吸道传染病,传染性极强。近年来,国内部分省份麻疹疫情明显回升,尤其是原来一些发病率较低的省份,发病率大幅度上升。采取合适的方法分析,及时预警,并开展一系列防控措施可有效降低麻疹流行强度^[1]。人工神经网络作为一种新技术已在流行病学预测^[2-4]和疾病诊断^[5]方面得到广泛应用。本研究利用两层反向传播(BP)神经网络和概率神经网络(PNN)模

型,结合实际数据预测武汉市麻疹发病数,并分别评价两个模型的预测效果。

基本原理

分别采用BP和PNN神经网络,建立时间序列动态预测模型^[6]。

1. BP神经网络模型:该模型可以看成是从输入到输出的高度非线性映射。图1为具有一个隐层的BP神经网络模型^[7]。BP网络的每一连线连接着2个神经元,并附有1个权值 W ,权值表明了上一层神经元对下一层神经元的影响或连接强度。每一神经元的计算输出又是下一层所有神经元的输入,再用于计算其输出。BP网所采用的传递函数一般是sigmoid型函数,例如: $f(x) = 1/[1 + \exp(-x)]$ 。BP

DOI: 10.3760/cma.j.issn.0254-6450.2011.01.017

基金项目:卫生部与世界卫生组织卫生技术合作项目(WP/2006/CHN/CSR/1.5/001)

作者单位:430015 武汉市疾病预防控制中心(余滨、陈邦华、刘普林、罗同勇);南京船舶雷达研究所(丁春);武汉市卫生局(魏善波、王家刚、潘志伟);武汉大学数学与统计学院(陆君安)

通信作者:丁春, Email: tingnt@126.com

网络的学习过程包括正向传播过程(the forward phase)和反向传播过程(the back-ward phase)两部分。当给定网络的一个输入模式 X 时,它由输入层单元传到隐层单元,经隐层单元逐层处理后再送到输出层单元。由输出层单元处理后,产生一个输出模式 O ,称为前(正)向传播。如果输出响应与期望输出模式有误差,不满足要求,那么就将误差信号沿原来的连接通路从输出层到输入层逐层传递,并修正各层的连接权值,直到误差信号最小,该过程称为反向传播。BP算法(back-propagation algorithm)为网络的学习提供了有力方法。

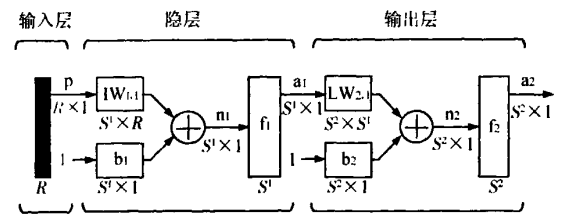


图1 两层BP神经网络模型

2. PNN模型:该模型常用于模式分类(图2)^[8]。其第一层是径向基网络层,神经元数目等于输入样本数,其权值等于 Q 个输入样本构成矩阵的转置; b 是阈值向量。第二层是竞争层,神经元数 K 等于分类模式数,其权值为目标向量 T ,无阈值向量。对于每个输入样本 p ,它在训练样本中选择与样本 p 最相似样本的模式作为网络输出,即对样本 p 最可能的模式分类结果。

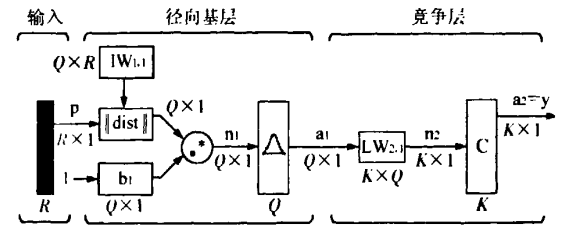


图2 PNN模型

实例分析

1. 资料来源:麻疹月分析资料来自武汉市1986年1月至2006年8月(共248个月)疫情数据库,病例包括实验室确诊病例和临床诊断病例。周分析资料源于《中国疾病预防控制系统》中武汉市2004—2006年网络疫情资料。资料以分月病例统计。

2. 两层BP神经网络建立的时间序列动态模型及评价: $x(t)$ 为1986年1月至2006年8月(共248个月)各月武汉市麻疹患者数。在用BP网络预测时,用 $x(t), x(t+1), \dots, x(t+p-1)$ 预测 $x(t+p)$,其中

p 是一个正整数,这样得到 $248-p$ 个样本。 p 依次从3到30取值,步长为3。在学习过程中采用贝叶斯归一化方法(自适应,泛化)。由于当输入向量各个分量在-1到1之间取值时,贝叶斯归一化方法效果比较好,所以要对 $248-p$ 个样本进行预处理。预处理包括数据标准化和主成分分析等。在数据预处理之后,还必须把样本划分为训练集、验证集和测试集。训练集用来训练网络,验证集是在神经网络的训练过程中,起提前终止的作用,测试集用来测试所建立的BP网络性能,对训练过程无影响。采用贝叶斯归一化方法和提前终止的技术都是为了防止网络训练过度,从而提高网络的泛化能力。但是,贝叶斯归一化方法只适合中等规模的网络,且收敛速度也较慢。

本研究采用两层BP神经网络,其中只有一个隐层。隐层和输出层的激活函数分别取为双曲正切函数(tansig)和纯线性函数(purelin)。输入层有一个向量,该向量长度由主成分分析的结果决定。训练集、验证集和测试集中样本个数之比为2:1:1,并且均匀选取,使得各个样本都有一定的代表性,其中训练集最好要包含那些特殊的样本。这要求原始数据尽可能包含所有的情况。

为了评价BP网络的性能,采用误差的平方和(SSE)作为评价标准;为了检验预测的效果如何,引入回归分析,即考察实际观测值和预测值之间的相关系数(COR)。其中,SSE越小说明该网络越好,COR越接近于1说明预测效果越好。对于每个 p ,分别进行了5次试验,都表明该结构稳定,可以重复(表1)。

表1 不同 p 时训练集、验证集和测试集的SSE及其COR

p	SSE				COR
	训练集	验证集	测试集	合计	
3	33.01	32.81	27.10	92.92	0.79
6	43.83	12.22	44.21	100.26	0.77
9	10.72	25.42	29.64	65.78	0.85
12	26.10	42.15	18.08	86.33	0.81
15	23.65	44.05	26.57	67.70	0.78
18	23.76	18.04	36.01	77.81	0.82
21	24.06	24.79	40.78	89.63	0.79
24	21.56	36.37	19.92	77.85	0.80
27	12.83	32.40	23.92	69.15	0.83
30	19.54	18.87	35.00	73.14	0.81

从表1可见, p 对训练集、验证集、测试集的SSE和COR有一定影响;当 p 较小时,训练集的SSE较大,说明网络性能较差,且此时COR也较小;当 $p>9$

时,随 p 的增大,对于训练集的SSE和COR有一定的影响,但无简单的线性变化规律。权衡各方面因素,认为 $p=9$ 较好。此时,网络的收敛速度可以接受,并且总SSE也不是太大,COR达到0.85。

经验表明,麻疹暴发的周期大致是5年,但从图3可见在1986—2006年中有一年例外,按照5年一个周期,暴发期应该在2006年到来,但是高峰期却提前于2005年。

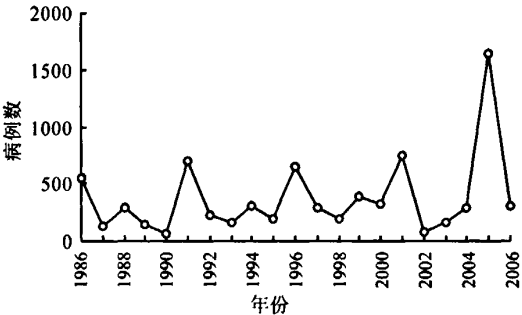


图3 1986—2006年武汉市麻疹发病数趋势

图4为 $p=9$ 时按月预测的效果。可见对于按月预测,其结果较好,不仅准确预测了平凡点,也准确预报1991、1996和2001年麻疹患者数出现较多的几个月,特别是准确预报了2005年上半年麻疹患者突然增加。这得益于训练样本选取得较为合适。

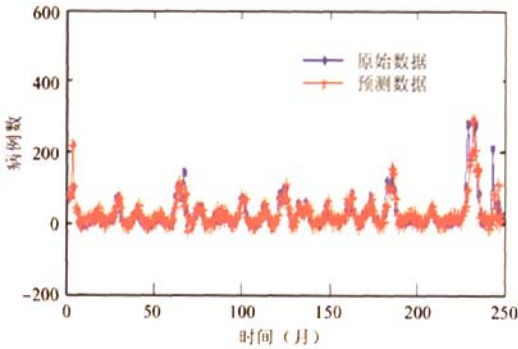


图4 p 取9时的按月预测武汉市麻疹每月患病例数

3. PNN网络建立的时间序列动态模型及评价:
图5是2004—2006年武汉市麻疹逐周报告数。对延迟 $\tau=1, 2, \dots, 4$, 分别对门限集 $S_1 = \{\bar{x}\}$, $S_2 = \left\{\bar{x}, \bar{x} + \frac{s}{2}\right\}$ 和 $S_3 = \left\{\bar{x} - \frac{s}{2}, \bar{x}, \bar{x} + \frac{s}{2}\right\}$ 进行预测。其中 $\bar{x}$ 和 s 分别是麻疹序列的平均数和标准差。门限集 S_1 将麻疹发病数分为高和低两类;门限集 S_2 将麻疹发病数分为高、中、低三类;门限集 S_3 将麻疹发病数分为高、次高、中、低四类。

图6、图7分别为对麻疹序列应用G-P算法^[8]所

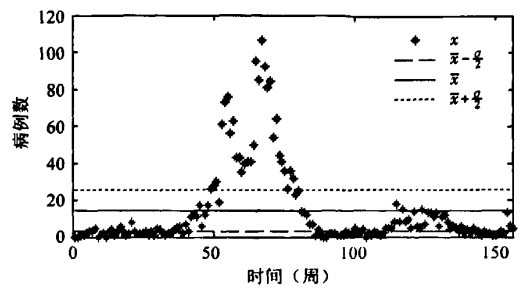


图5 2004—2006年武汉市麻疹逐周报告病例数

得的 $\ln C(r) - \ln(r)$ 和 $d-m$ 关系图。可以看出,麻疹序列在嵌入维数 $m=7$ 后关联维数 d 不再随嵌入维数 m 的增大而增加,这时的关联维数 $d=3.3628$ 。根据嵌入定理,麻疹序列可以重构成8维状态空间向量序列。将重构相空间中全部样本的前90%的样本作为训练集,用来确定PNN网络中的各个参数,后10%的样本作为测试集,用来测试的模型的准确性(表2)。

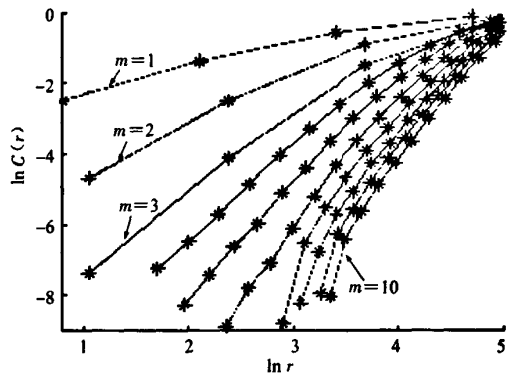


图6 $\ln C(r) - \ln(r)$

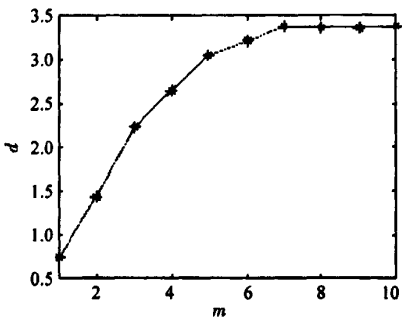


图7 $d-m$ 关系图

通过比较可以看出:①对麻疹序列分别按两类和三类预测的预测率都很高,按四类预测的预测率较低。可见,分类数对预测率的高低有很大影响。②不管是按几类预测,延迟取 $\tau=4$ 时的预测率均最小,这与由麻疹序列本身的特征有关。

表2 应用PNN预测得到训练集和测试集的预测率

项目 延迟	门限集 S_1				门限集 S_2				门限集 S_3			
	1	2	3	4	1	2	3	4	1	2	3	4
训练集	0.95	0.96	0.95	0.90	0.99	0.96	0.95	0.92	0.85	0.84	0.87	0.88
测试集	1.00	1.00	1.00	0.91	1.00	1.00	1.00	0.91	0.78	0.85	0.61	0.33

讨 论

1. 在部分时间窗口内,麻疹发病数很少(<10)。由于麻疹发病数采用人工收集、上报和统计,在麻疹发病数较少时,这种数据统计的方式对于周发病数的影响(包括延时、漏报、虚报、误报等)较大,而对月发病数的影响不大。

将麻疹月发病数看作一个非线性的时间序列,对于重构像空间中的样本,可以使用BP网络实现任意维之间的映射。另外,由于BP网络天然的鲁棒性(robustness),附加于统计数据上的微小误差对于网络性能和预测结果影响不大。在合适地选取训练集和验证集后,两层BP网络建立的时间序列动态模型经过训练集的训练,可以很好地“记忆”麻疹月发病数序列内在的非线性关系。

当 p 取9时,训练集、验证集和测试集中分别有119、59和59个样本,此时测试集的SSE为29.64, COR为0.85,说明本研究提出的对于重构像空间中的样本,利用BP神经网络进行麻疹月发病数预报的方法是可行。但该方法必须建立在观测数据记录较多的情况下,且必须覆盖所有典型的情况,否则可能出现错误预报。目前,我国已对新生儿实行了接种麻疹疫苗,这将改变麻疹暴发的内在规律,在实际使用过程中,还应当根据实际情况及时更新训练集和验证集,使其具有一定代表性,重新训练BP网络,达到较好的预测效果。

2. 对某些疾病因数据积累时间较短,或数据统计过程中有误差,精确预测不能得到令人满意的结果,可以用PNN做分类预测。实验表明,当按两类或三类进行预测时,测试集的预测准确率较高;按四类预测时,测试集的预测准确率稍低。这是由于两类和三类的分类法比较粗,且延迟取1,2,或3时,测试集的预测准确率都是1,当延迟取4时,预测准确率都是0.91。而按四类预测时,由于数据本身的误差,以及模式特征选取等各种因素的影响,当延迟取2时,预测集的预测准确率最高达到0.85。当延迟和分类数合理时,PNN模型对于这类问题的分类预测是有效的,可以在疾病暴发早期建立预警系统中应用。在使用PNN建模的过程中,应注意:①PNN的预测率虽然较高,但只是对分类做预测,并不能预测

出具体的数目。②由于PNN径向基层的神经元数等于输入样本的个数,所以当样本数量较大时需要大量存储空间,计算所需要的时间也较长。③当学习模式样本本身杂乱无章,没有明显的分类特征时,网络对输入、输出的响应呈现振荡现象,从而不能正确分类。当各类模式的特征相近时,也会出现同样情况。④尽管需要在关联维数不随嵌入维数的增大而变化时确定重构像空间的维数,但是关联维数在多大范围内变化可以认为是“不变化”,这需要根据经验来设定一个门限,这就具有不少随意性,给该模型带来潜在隐患,需要改进估计重构像空间维数的算法。

参 考 文 献

- [1] Wei XQ, Lan X, Huang XH, et al. Study on the effect of measles control under early warning system. Prev Med Trib, 2008, 12(14): 1208-1210. (in Chinese)
魏小庆, 蓝毅, 黄星辉, 等. 预警机制对麻疹防控的效果研究. 预防医学论坛, 2008, 12(14): 1208-1210.
- [2] Qu B, Huang DS, Guo HQ, et al. The model of back-propagation neural network about meteorological factors and typhoid fever, paratyphoid fever in drought area. Chin J Health Stat, 2004, 21(3): 165-167. (in Chinese)
曲波, 黄德生, 郭海强, 等. 干旱地区伤寒副伤寒与气象因素关系的BP神经网络模型. 中国卫生统计, 2004, 21(3): 165-167.
- [3] Qu B, Huang DS, Guo HQ, et al. The model of back-propagation neural network about meteorological factors and epidemic cerebrospinal meningitis. J Chin Med Uni, 2006, 35(2): 158-159. (in Chinese)
曲波, 黄德生, 郭海强, 等. 流行性脑脊髓膜炎与气象因素关系的BP神经网络模型研究. 中国医科大学学报, 2006, 35(2): 158-159.
- [4] Yan WR, Shi LY, Zhang HJ, et al. Introduction on a forecasting model for infectious disease incidence rate based on radial basis function network. Chin J Epidemiol, 2007, 28(12): 1219-1222. (in Chinese)
严薇荣, 施侶元, 张惠娟, 等. 径向基函数神经网络在传染病发病率预测中的应用. 中华流行病学杂志, 2007, 28(12): 1219-1222.
- [5] Jiang Y, Zhou ZH, Xie Q, et al. Application of neural network ensemble in lung cancer cell identification. J Nanjing Uni (Natural Sciences), 2001, 37(5): 529-533. (in Chinese)
姜远, 周志华, 谢琪, 等. 神经网络集成在肺癌细胞识别中的应用. 南京大学学报(自然科学), 2001, 37(5): 529-533.
- [6] Kantz H, Schreiber T. Nonlinear time series analysis. Cambridge University Press, 1997.
- [7] Demuth H, Beale M. Neural network toolbox users' guide—neural network toolbox for use with Matlab. The Math Works, Inc., 1998.
- [8] Lv JH, Lu JA, Chen SH. The analyse and application on the chaos time-series. Wuhan: Wuhan University Press, 2002. (in Chinese)
吕金虎, 陆君安, 陈士华. 混沌时间序列分析及其应用. 武汉: 武汉大学出版社, 2002.

(收稿日期: 2010-07-09)

(本文编辑: 张林东)