·基础理论与方法·

基于 BP 神经网络的围产儿出生缺陷患病率预测

王玮 许伟 郑亚军 周宝森

【摘要】 **目的** 评价 BP 神经网络模型在围产儿出生缺陷预测中的应用价值。**方法** 选择沈阳市 1995－2005 年围产儿出生缺陷患病率数据，利用 MATLAB 6.5软件的神经网络工具箱构建 BP 神经网络模型，训练与模拟网络，预测 2006－2007 年沈阳市围产儿出生缺陷的流行趋势，并与传统的预测方法进行比较。**结果** 以1995－2003 年资料建立模型预测2004－2005 年流行水平和趋势，结果患病率回代平均误差率和 R_{NL} 分别为1.34％和0.9874，外推预测平均误差率为1.78％；以1995－2005 年资料建立模型预测2006－2007 年流行趋势，结果患病率回代平均误差率和 R_{NL} 分别为0.33％和0.9954，2006－2007 年出生缺陷患病率预测值分别为11.00‰和11.29‰。**结论** 利用 BP 神经网络进行疾病预测，不仅能获得更好的预测效果，而且对资料的类型、分布不作任何限制，是一种很好的流行病学预测方法。

【关键词】 出生缺陷；疾病预测；BP 神经网络模型；患病率

Study on a back propogation neural network-based predictive model for prevalence of birth defect

*WANG Wei**, *XU Wei*, *ZHENG Ya-jun*, *ZHOU Bao-sen*. **Department of Epidemiology*, *China Medical University*, *Shenyang* 110001, China

【**Abstract**】 **Objective** To evaluate the value of a back propogation(BP) network on prediction of birth defect and to give clues on its prevention. **Methods** Data of birth defect in Shenyang from 1995 to 2005 were used as a training set to predict the prevalence rate of birth defect. Neural network tools box of Software MATLAB 6.5 was used to train and simulate BP Artificial Neural Network. **Results** When using data of the year 1995－2003 to predict the prevalence rate of birth defect in 2004－2005, the results showed that: the fitting average error of prevalence rate was 1.34％, R_{NL} was 0.9874, and the prediction of average error was 1.78％. Using data of the year 1995－2005 to predict the prevalence rate of birth defect in 2006－2007, the results showed that: the fitting average error was 0.33％, R_{NL} was 0.9954, the prevalence rates of birth defect in 2006－2007 were 11.00‰ and 11.29‰. **Conclusion** Compared to the conventional statistics method, BP not only showed better prediction precision, but had no limit to the type or distribution of relevant data, thus providing a powerful method in epidemiological prediction.

【**Key words**】 Birth defect; Disease prediction; Back propogation network; Prevalence rate

出生缺陷包括先天畸形、智力障碍、代谢性疾病等。在一些发达国家，出生缺陷已成为围产儿死亡和婴儿死亡的主要原因[1]。目前我国每年约有 20 万～30 万肉眼可见的先天畸形儿出生，加上出生后数年才表现出来的缺陷或遗传病，出生缺陷儿童总数高达 80 万～120 万，约占年出生人口数的4％～6％[2]，严重威胁儿童的健康。由于影响出生缺陷患病率的因素很多，而这些因素又很难进行定量化描述，因此出生缺陷的患病率可认为是一个复杂的诸多因素影响的非线性系统。BP(back propogation)神经网络是一种不需要知道系统内部结构的输入输出系统的非线性映射，能够充分逼近任意复杂的非线性关系。神经网络理论中的 Kolmogorov 定理指出：给定任一连续函数 $f: U^n \rightarrow R^m$，$f(X) = Y$，式中 U 是闭单位区间，f 可以精确地用一个三层网络来实现[3]。因此，本文将沈阳市近 10 年来围产儿出生缺陷患病率所得的时间序列作为一个非线性系统进行 BP 神经网络的输入和输出，以达到对围产儿出生缺陷进行预测的目的。

作者单位：110001 沈阳，中国医科大学公共卫生学院流行病学教研室(王玮、周宝森)；沈阳市卫生局基层卫生与妇幼保健处(许伟、郑亚军)

基本原理

1. 原理[4,5]：BP 神经网络由输入层、隐含层、输出层组成，每层有若干个节点(神经元)，每个节点通

过连接权重接受来自其他节点的信息，然后通过输入输出转换函数输出信息。BP 算法的学习过程由正向传播和反向传播两个过程组成，在正向传播过程中，输入信息由输入层经隐含层传向输出层，若输出层得不到期望的输出，则转入反向传播，将误差信号沿原来的连接通路返回，修改各层节点间的连接权重值，如此反复调整网络参数，使误差函数达到极小为止。

BP 网络的时间序列预报模型：设时间序列 $X(i), i=1,2,\cdots,N$，其中的 N 为观测点的个数。预报模型可描述为：$X(t)=\Phi[X(t-1),\cdots,X(t-p)]$。式中：$\Phi(.)$为非线性作用函数；$p$ 为模型的阶数。时间序列预报模型建立的过程就是寻找 $\Phi(.)$ 的过程。已经证明：如果网络中间层神经元的特性函数具有任意阶导数，中间层可以根据需要任意设置神经元个数，那么 3 层 BP 网络模型可以任意精度逼近任何连续函数。因此，只要选取合理的神经网络结构参数，使用神经网络即可精确的反映出复杂的非线性函数 $\Phi(.)$。由时间序列 $X(i), i=1,2,\cdots,N$，可以构造 $N-p$ 个样本：

	输入层向量	输出层向量
第一个样本	$X(1),\cdots,X(p-1),X(p)$	$X(p+1)$
⋮	⋮	⋮
第 $N-p$ 个样本	$X(N-p),\cdots,X(N-2),X(N-1)$	$X(N)$

将所构造的 $N-p$ 个样本代入 BP 网络中进行学习训练，学习训练后即可得到稳定的网络结构、连接权值和阈值，这就建立了基于 BP 网络的时间序列预报模型。

2. 研究方法：利用 MATLAB 6.5神经网络工具箱进行 BP 神经网络模型构建、训练及模拟[6]，建立基于 BP 神经网络的时间序列预报模型，并采用平均误差绝对值、平均误差率及非线性相关系数 R_{NL} 指标，比较 BP 神经网络模型与时间序列 ARIMA 模型、灰色模型的拟合及预测效果。平均误差率 = 平均误差绝对值/实际值的均值；$R_{NL}=1-\sqrt{\frac{\sum(y_i-\hat{y}_i)^2}{\sum y_i^2}}$，其中 y_i 是实际值，$\hat{y}_i$ 是预测值。ARIMA 模型和 GM(1,1)分别通过 SPSS 中 ARIMA PROC 和 MATLAB 6.5编程实现[7]。

实例分析

1. 研究对象：资料来源于沈阳市围产儿出生缺陷监测资料，监测对象为 1995 年 1 月 1 日至 2005 年 12 月 31 日在沈阳市定点监测医院住院分娩的所有孕满 28 周至产后7 d的围产儿(包括活产、死胎、死产)。出生缺陷的诊断及分类均按照全国统一标准，以ICD-9为基础，按照《中国出生缺陷监测方案》及《中国出生缺陷监测手册》中有关出生缺陷的定义特征和诊断标准，由专职医生经临床体检和 B 超检查确诊并分类。

2. BP 神经网络模型构建、训练及模拟：以 1995－2003 年出生缺陷数据作为训练数据集，2004－2005 年数据作为预测样本。按 $x'=\frac{x}{\max(x)}$ 对数据进行归一化；选取 3 层网络，采用 3－5－1 体系，即 3 个输入节点、5 个隐含层节点及 1 个输出节点。隐含层的传递函数采用 Sigmoid 函数，$f(x)=\frac{1}{1+e^{-x}}$，输出层采用线性函数。网络表达形式为：net = newff(minmax(p),[3,5,1],{'logsig''logsig''pureline'},'trainlm'); net = train(net,p,t); y = sim(net,p)。

3. BP 神经网络模型与传统预测模型比较：见表 1。

表1 1995－2005 年沈阳市定点监测医院出生缺陷患病率、BP 神经网络、ARIMA 及 GM(1,1)预测值

年度	实际患病率(‰)	BP 神经网络预测值(‰)	ARIMA 模型预测值(‰)	GM(1,1)预测值(‰)
1995	6.11	–	–	–
1996	8.53	–	6.4786	7.0151
1997	6.39	–	7.8728	7.3174
1998	8.53	8.3486	7.7471	7.6327
1999	6.93	6.9312	8.3114	7.9617
2000	7.77	7.9487	8.4037	8.3048
2001	8.09	8.0859	8.6666	8.6627
2002	8.95	8.9453	8.9528	9.0360
2003	9.60	9.5966	9.3210	9.4255
2004	10.12	9.9281	9.7205	9.8317
2005	10.62	10.4476	10.1290	10.2554

由表 1 可知 BP 神经网络患病率回代平均误差率和 R_{NL} 分别为1.34％和0.9874，外推预测平均误差率为1.78％；时间序列 ARIMA 模型患病率回代平均误差率和 R_{NL} 分别为11.09％和0.8625，外推预测平均误差率为4.29％；灰色模型患病率回代平均误差率和 R_{NL} 分别为8.86％和0.8967，外推预测平均误差率为3.15％。可以看出：BP 神经网络模型平均误差率和非线性相关系数 R_{NL} 明显优于时间序列模型和灰色模型，说明 BP 神经网络模型的拟合和

预测精度均较高。

4. 用 1995 - 2005 年资料预测沈阳市 2006 - 2007 年出生缺陷的流行趋势：见表 2。

表2 1995 - 2007 年沈阳市出生缺陷患病率模型拟合和预测情况

年度	患病率(‰)		
	实际值	预测值	误差绝对值
1995	6.11	-	-
1996	8.53	-	-
1997	6.39	-	-
1998	8.53	8.5279	0.0031
1999	6.93	6.9311	0.0001
2000	7.77	7.7895	0.0235
2001	8.09	8.0838	0.0022
2002	8.95	8.8700	0.0750
2003	9.60	9.6721	0.0751
2004	10.12	10.1417	0.0206
2005	10.62	10.5879	0.0361
2006	-	10.9980	-
2007	-	11.2880	-

由表 2 可知患病率的平均误差绝对值为 0.0393，平均回代误差率为 0.33%，R_{NL} = 0.9954；2006 - 2007 年出生缺陷患病率预测值分别为 11.00‰和11.29‰。

讨 论

目前许多预测方法逐渐被用于疾病预测。传统的预测方法如时间序列模型[8]、博克斯-詹金斯(Box-Jenkins)法[9]、多元回归分析模型、弹性系数法等。前两种方法属于趋势外推法模型，需要大量的样本和典型的概率分布，且计算方法比较复杂，较少考虑疾病的影响因素；而后两种方法则是通过相关因素的分析，用数学模型来描述其关系并进行推测，其关键在于要正确地把握影响因素对事物的影响效果和其本身的发展趋势。另外，我国学者于 20 世纪 80 年代创建并发展起来的灰色系统也应用于疾病预测[10]，但此方法当数据的波动幅度较大时就不适用。Kohonen 指出：神经网络是由一些简单(通常为自适应的)的元件及其层次组织的大规模并行联接构造的网络，它致力于按照生物神经系统的同样方式处理真实世界的客观事物。生物神经系统认识事物方式的最大特点为具有学习性和自适应性，而神经网络也具有这两种特性。因此，神经网络能学习训练集输入、输出数据之间的函数关系，从而对未来的输出数据作出准确预测。神经网络模型中，最常见的为 BP 神经网络模型，已广泛应用于药学、环境科学等方面，取得了很好的效果。本文将 BP 神经网络模型应用于预测沈阳市围产儿出生缺陷的患病率，根据本研究结果可见模型的拟合误差率和外推误差率都较小，明显优于时间序列模型和灰色模型，说明模型本身预测精度很高。而且，BP 神经网络应用于疾病预测，除了拟合和预测精度很高外，尚有无需典型概率分布资料和当时间序列数据波动较大时仍然适用的优点。

本文采用沈阳市 11 年出生缺陷的监测资料，应用 BP 神经网络对沈阳市未来几年出生缺陷的患病率进行预测。按目前的发展趋势，沈阳市 2006 年出生缺陷患病率将达到 11.00‰，2007 年将达到 11.29‰，呈缓慢上升趋势。本资料监测点的选择考虑到了覆盖的人口数和监测医院的地域分布，基本能够反映沈阳市出生缺陷患病率的实际水平。但是，影响出生缺陷发生水平的因素很多，一般可以分成两类，一类是病因学因素，如有害的生物、物理、化学因素所造成的出身缺陷发病率升高等；另一类则是监测方法问题，如内脏畸形及代谢和染色体缺陷的监测与设备和技术相关，先心病检出率的上升并非发病的增加，而是其检出技术的提高所致等。因此，如何利用现有的资源控制各种致病因素的发生，并用较低的成本提高出生缺陷的诊断和识别能力应是进一步研究的方向。

参 考 文 献

[1] Rosano A, Botto LD, Botting B, et al. Infant mortality and congenital anomalies from 1950 to 1994: an international perspective. J Epidemiol Community Health, 2000, 54: 660-666.

[2] 中华人民共和国卫生部，中国残疾人联合会. 中国提高人口素质、减少出生缺陷和残疾行动计划(2002 - 2010 年). 中国生育健康杂志，2002，13：98-101.

[3] 赵振宇，徐用懋. 模糊理论和神经网络的基础与应用. 北京：清华大学出版社，1996：80.

[4] 程相君，王春宁，陈生潭. 神经网络原理及其应用. 北京：国防工业出版社，1995：24-61.

[5] 焦李成. 神经网络系统理论. 西安：西安电子科技大学出版社，1990：6-16.

[6] Guan P, Huang DS, Zhou BS. Forecasting model for the incidence of hepatitis A based on artificial neural network. World J Gastroenterol, 2004, 10: 3579-3582.

[7] 倪少凯，陈卫红. 用 Matlab 实现灰色数列模型 GM(1,1)的预测. 数理医药学杂志，2002，15：292-293.

[8] 李米英，张小平. 起伏型时间序列分析方法在流行性出血热预测中的应用. 中国卫生统计，1997，14：64.

[9] 章扬熙. 医学统计预测. 北京：中国科学技术出版社，1995：64-89.

[10] 汪爱琴，鱼敏. 灰色预测方法在疾病预测中的应用. 中华流行病学杂志，1988，9：49.

(收稿日期：2006-12-21)
(本文编辑：张林东)