DOI: 10.3969/j.issn.1005-8982.2017.23.026
文章编号： 1005-8982(2017)23-0124-03

基于改进 BP 神经网络模型的肺结核发病率预测*

徐学琴[1],张知鹜[2],王瑾瑾[1],闫国立[1],裴兰英[1],孙春阳[1],刘晓蕙[1]
(1.河南中医药大学 基础医学院,河南 郑州 450046;2.郑州大学
第一附属医院 血液科,河南 郑州 450052)

摘要:目的 建立用于肺结核发病率预测的人工神经网络模型,预测肺结核疫情发生发展趋势,为肺结核的预防和控制提供理论依据。**方法** 选取肺结核 2000～2014 年发病率数据,采用改进的误差反向传播(BP)神经网络算法建立预测模型。其中以 2000～2013 年的发病率数据作为训练样本,以 2014 年的发病率数据来检验模型的有效性。并对 2015～2019 年肺结核的发病率进行预测。**结果** 采用单隐层神经网络模型,输入层节点数为 3,隐含层节点数为 7,输出层节点数为 1。建立的肺结核发病率预测模型在仿真预测样本处的平均相对误差为 0.7597%,在检验样本处的相对误差为 0.2649%。经预测,2015～2019 年肺结核的发病率分别为 69.33/10 万、71.16/10 万、64.49/10 万、62.41/10 万和 72.78/10 万。**结论** 采用改进的 BP 神经网络算法建立的肺结核发病预测模型具有较高的预测精度及较低的预测相对误差,为肺结核疫情预测提供一种新的预测模型。

关键词: 肺结核;误差反向传播;神经网络;模型;预测
中图分类号: R181.2 **文献标识码:** B

结核病是一种慢性传染性疾病,由结核分枝杆菌感染引起,可侵害脑、肠、骨骼及淋巴结等,以肺部结核最为常见。我国是全球 22 个结核病高负担国家之一,结核病人数量占全球 12%,居世界第 2 位[1-2]。我国第五次结核病流行病学抽样调查结果显示,2010 年全国 15 岁及以上人群中,活动性肺结核患病人数约 499 万,患病率为 459/10 万,涂阳患病率为 66/10 万[3]。肺结核是我国传染病防治法中规定的乙类传染病,其发病数一直居于甲乙类法定报告传染病的第 1、2 位[4]。因此,肺结核的预防和控制是公共卫生工作中非常重要的一部分。

在传染病的预防控制工作中,疾病的预警、预测起着关键作用,能够早期发现疾病的发生发展趋势。通过建立肺结核疫情的预测模型,对其未来发生、发展趋势进行科学的预测,将为制定肺结核的预防和控制的策略和措施,以及有效预防肺结核提供重要的参考依据。在众多预测模型中,人工神经网络模型越来越多地用于传染病的预测和监测工作中,其中误差反向传播(back error propagation,BP)神经网络模型是应用最普遍、最成熟的人工神经网络模型之一[5-6]。本研究便采用误差反向传播神经网络模型建立肺结核年发病率预测模型,并对我国肺结核未来 5 年的发病率进行预测,为制定肺结核预防和控制措施提供理论依据。

1 资料与方法

1.1 BP 神经网络模型的基本原理

BP 神经网络模型是多层前馈网络,它包括 1 个输入层、1 个或多个中间层和 1 个输出层,每层包含若干个节点。该模型的特点是相邻层节点之间为完全连接方式,同层节点之间却彼此独立。训练过程为:输入层接收外源性样本信号,然后传递到中间层,经中间层逐层处理之后由输出层输出信号,此即信正向传播;若输出信号与期望值之间存在误差,则该误差进入反向传播,并通过调整各单元的权重,使误差逐渐下降,此即误差的反向传播。即在模型训练过程中,最终的目标是通过反复地调整权重和偏差,最终使网络的输出值和实际值之间的整体误差降到期

收稿日期:2016-06-06
*基金项目:河南省软科学研究重点项目(No:102400440002);河南中医学院科研苗圃工程项目(No:MP2014-07)

望精度[7]。

因常规的 BP 神经网络算法存在收敛速度较慢、易陷入极端最小值的缺陷，人们在传统算法的基础上做改进来解决该问题。本研究采用 Levenberg-Marquardt 数值优化算法来改进网络模型，该算法能够加快收敛的速度、提高训练的效率及预测的精确性[8-9]。

1.2　数据收集

我国 2000～2014 年肺结核发病率数据来自于国家卫生和计划生育委员会发布的历年卫生统计年鉴及我国法定报告传染病监测系统，人口资料来自于国家统计局。

1.3　样本数据的准备

由于原始发病率数据大小不一，不利于模型的建立及训练。为了提高网络模型的效率和泛化能力，需要将历年发病率数据进行处理。归一化处理可防止较大的输入值覆盖较小的输入值、降低网络误差及加快训练速度[10]。归一化方法有多种，本研究采用的是用每个发病率除以某个数值，能使该发病率变为 0～1 之间的数据，本研究中将该数值定为 120/10 万。

1.4　BP 神经网络模型的基本参数

建立 BP 神经网络模型需先确定基本参数，即输入层节点的数目、中间层的层数及其节点数、各层节点的传递函数。根据 BP 神经网络的定理，对于任何在闭区间内的一个连续函数都可以用单隐层（1 个中间层）的 BP 神经网络逼近，因此 1 个 3 层 BP 网络已具有很强的非线性映射能力[11]。本研究便采用 3 层网络模型，其中包括 3 个输入层节点数，1 个输出层节点数，输出层传递函数采用对数型 Sigmoid 函数。根据 Kolmogorov 定理，中间层节点数应为 2×3+1=7，中间层传递函数采用正切型 Sigmoid 函数。本研究将预测的期望精度定为 0.0001。

1.5　BP 神经网络模型的训练及检验

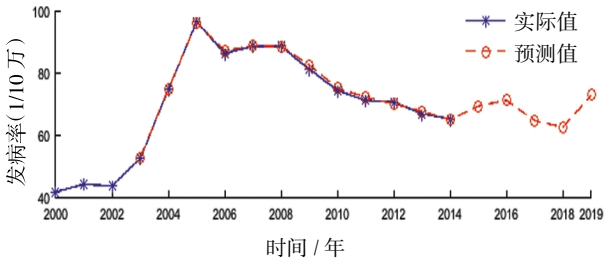
若采用神经网络模型进行肺结核发病率的预测，需先利用已有样本数据对模型进行训练。训练过程中，通过不断调整各单元的权重，使误差逐渐减小，当减小到预先设定的期望精度时，训练便完成。此时，实际发病率和预测发病率的值非常接近。

训练完成的模型便可用来进行发病率的预测。通常采用新陈代谢预测法，即按照事先设定的参数，以连续 3 年的肺结核发病率数据组成输入序列，输出第 4 年的发病率预测值。输入序列中每加入新的 1 个发病率数据，将舍弃最前面的 1 个数据。本研究中，以 2000～2002 年的数据预测 2003 年数据，以 2001～2003 年数据预测 2004 年数据，依次进行。2014 年发病率数据用来检验模型，若检验合格，即可用来预测未来 5 年（2015 年～2019 年）肺结核的发病率。预测所得的数据还需反归一化处理，即各预测值乘以 120/10 万。以上工作均在 Matlab 7.0 软件中实现。

2　结果

2.1　仿真预测及模型的检验情况

经计算，在仿真预测样本处的平均相对误差为 0.7597%。用 2014 年数据检验该模型，结果显示：2014 年肺结核的预测发病率为 64.85/10 万，而实际发病率为 65.02/10 万，其相对误差为 0.2649%，实际值与预测值吻合度较高。各年发病率的真实值和预测值均非常接近。见附表和附图。



附图　肺结核发病率的真实值与预测值的曲线比较

附表　肺结核实际发病率、预测发病率（反归一化值）及相对误差

年份	实际发病率（1/10 万）	预测发病率（1/10 万）	相对误差 /%
2000	41.68	-	-
2001	44.06	-	-
2002	43.58	-	-
2003	52.36	52.74	0.7234
2004	74.64	74.80	0.2104
2005	96.31	96.11	0.2079
2006	86.23	87.10	1.0088
2007	88.55	88.61	0.0692
2008	88.52	88.50	0.0181
2009	81.09	82.45	1.6813
2010	74.27	75.00	0.9832
2011	71.09	72.14	1.4744
2012	70.62	70.03	0.8391
2013	66.80	67.56	1.1405
2014	65.02	64.85	0.2649

2.2 对未来 5 年肺结核发病率的预测

利用该模型，通过新陈代谢法预测，得到我国肺结核 2015～2019 年发病率（反归一化值）分别为 69.33/10 万、71.16/10 万、64.49/10 万、62.41/10 万及 72.78/10 万。

3 讨论

肺结核病是长期危害人类健康的慢性传染性疾病，因此，肺结核的监测与预测有着非常重要的流行病学意义[12]。在肺结核发病率的预测中，目前常用的模型有灰色模型、自回归积分移动平均模型、马尔科夫链模型及 BP 神经网络模型等。灰色动态模型对样本容量和概率分布没有严格要求，模型简单，预测效果好，适合于样本量较小、流行因素较稳定的疾病的短期预测[13]。马尔科夫链模型可进行随机波动较强序列的预测，揭示序列在不同状态间转移的内在规律，但预测结果的准确性会受到不同状态划分的影响[14]。相比之下，BP 神经网络模型因其能应对复杂的大规模数据资料而被广泛应用于传染病的预测。该模型的应用无需具备先验知识，也无需资料满足正态性、线性及独立性等条件。它能学习和存贮大量的输入 - 输出模式映射关系，通过数据训练、人工智能、机器学习能描述许多复杂的非线性相关性[15]。

肺结核发生及流行的影响因素很多，如肺结核接触史、吸烟、频繁的人际交往、粉尘环境暴露、疫苗免疫、温度及湿度等[16-17]。该因素和结核病的发生与流行之间往往是一种非常复杂的非线性关系。而 BP 神经网络模型恰好具有很强的解决非线性问题的能力，能根据已学会的知识和处理问题的经验对复杂问题做出合理的判断，对未来疾病的发生做出较为精确的预测[18]。因此，BP 神经网络模型很适合于肺结核发病率的预测。因常规的 BP 神经网络模型存在训练时间长、易陷入局部极端最小值的缺点，为克服该缺点，本文所采用 Levenberg-Marquardt 算法来改进 BP 神经网络算法。

本研究以我国历年肺结核发病率作为样本数据，应用改进的 BP 神经网络算法建立肺结核的预测模型。该模型在仿真预测样本处的平均相对误差为 0.7597%，在检验样本点的相对误差为 0.2649%，真实值和预测值吻合度非常高，具有良好的预测精度，适于进行肺结核发病率的预测。

经该模型对我国 2015～2019 年肺结核发病率进行预测，结果显示，我国肺结核的发病率在经历连续 8 年（2007～2014 年）的持续下降后，将在 2015 年及 2016 年呈现连续上升趋势，2017～2018 年发病率将有所下降，但 2019 年发病率将会再次升高。因此，对肺结核的预防和监测工作还需进一步加强。

参 考 文 献：

[1] 山珂. 肺结核发病空间聚集分布及影响因素研究[D]. 济南: 山东大学, 2014.

[2] 汪学智, 王琳, 夏凡, 等. 军队“三位一体”结核病防治模式成效初探[J]. 中华疾病控制杂志, 2013, 6: 545-547.

[3] 王黎霞, 成诗明, 陈明亭, 等. 2010 年全国第五次结核病流行病学抽样调查报告[J]. 中国防痨杂志, 2012, 8: 485-508.

[4] 金瑾, 景睿. 山东省 2013 年学生肺结核疫情特征分析[J]. 中华流行病学杂志, 2015, 36(8): 871-874.

[5] 董选军, 贾伟娜. ARIMA 时间序列和 BP 神经网络在传染病预测中的比较[J]. 现代实用医学, 2010, 22(2): 142-143.

[6] 严文娟, 张晶, 胡广芹, 等. BP 神经网络用于肝炎患者舌诊近红外光谱的研究[J]. 光谱学与光谱分析, 2010, 30(10): 2628-2631.

[7] WANG Y M, LI J, GU J Z, et al. Artificial neural networks for infectious diarrhea prediction using meteorological factors in Shanghai[J]. Applied Soft Computing, 2015, 35: 280-290.

[8] SHAHROKH A, JAMAL S, PEYMAN A, et al. A new hybrid artificial neural networks for rainfall-runoff process modeling[J]. Neurocomputing, 2013, 121: 470-480.

[9] HUANG H X, LI J C, XIAO C L. A proposed iteration optimization approach integrating backpropagation neural network with genetic algorithm[J]. Expert Systems with Applications, 2015, 42(1): 146-155.

[10] PHOSSEINZADEH T. Multilayer perceptron with different training algorithms for stream flow forecasting[J]. Neural Comput, 2014, 24(3-4): 695-703.

[11] 郑立华, 李民赞, 潘娈, 等. 基于近红外光谱技术的土壤参数 BP 神经网络预测[J]. 光谱学与光谱分析, 2008, 28(5): 1160-1164.

[12] 李云, 穆卫明, 陆建方. 肺结核发病率的灰色模型预测及时间趋势分析[J]. 中华疾病控制杂志, 2011, 15(1): 87-88.

[13] 尹志英, 方春福. 传染病预警预测方法探讨[J]. 中国卫生统计, 2010, 27(2): 218-220.

[14] 易静, 胡代玉, 杨德香, 等. 三种预测模型在肺结核发病预测中的应用[J]. 中国全科医学, 2012, 15(13): 1495-1497.

[15] ALMEIDA J S. Predictive non-linear modeling of complex data by artificial neural networks[J]. Current Opinion in Biotechnology, 2002, 13(1): 72-76.

[16] 靳成娟, 杜建, 杨怀盛, 等. 中国人群肺结核发病危险因素的荟萃分析[J]. 军事医学, 2014, 38(5): 355-359.

[17] 胡婧媛, 蒋梦姣, 景元书, 等. 基于神经网络的气象条件对泸州市肺结核发病率预测[J]. 科技通报, 2013, 5: 19-23.

[18] 李异凡, 彭健, 张阳德. 神经网络方法在医学中的应用[J]. 中国现代医学杂志, 2003, 13(13): 8-11.

（李科 编辑）