基于BP神经网络的研究生教育发展规模预测

陆芳　陶芳芳　王萍　尹平[Δ]

（华中科技大学同济医学院公共卫生学院流行病学与卫生统计学系，Δ 通讯作者）

随着社会经济的发展，研究生教育发展问题越来越受到人们的关注。要促进研究生教育的健康可持续发展，必须结合社会发展过程中对高级人才的需求实际和国家的整体发展建设规划做好研究生教育发展规模的科学预测。常用的统计学预测方法较多，如线性回归模型、回归自回归混合模型等，这些方法简单、便于操作。但由于我国研究生教育正处于由过去的计划逐步向市场需求调节过度的转轨时期，不确定性的政策和社会因素也较多，这些因素相互作用，往往构成一个非线性系统，导致基于这些常规方法的预测结果与实际的偏差较大、精度不高，难以得到满意的结果。BP 神经网络是介于灰箱和黑箱之间的系统，对非典型数据有着良好的适应性，且在处理缺失值和非线性问题时有着明显的优越性。因此，在预测上更具有普遍适应性。本文通过建立 BP 神经网络模型，借助 Matlab7.0 软件对我国的研究生教育发展规模进行模拟，以期获得更加科学的预测结果。

1　BP 神经网络

BP（Backpropagation反向传播）模型是一种用于前向多层神经网络的误差反向传播学习算法，由美国加州大学的D.E.Rumelhart和J.L.Mcclelland等人在研究并行分布式信息处理方法，探索人类认识微结构的过程中于 1986 年提出。BP神经网络采用的是并行网络结构，包括输入层、隐含层和输出层，经作用函数后，再把隐节点的输出信号传递到输出节点，最后给出输出结果。该算法的学习过程由信息的前向传播和误差的反向传播组成。在前向传播的过程中，输入信息从输入层经隐含层逐层处理，并传向输出层。第一层神经元的状态只影响下一层神经元的状态。如果在输出层得不到期望的输出结果，则转入反向传播，将误差信号(目标值与网络输出之差)沿原来的连接通道返回，通过修改各层神经元权值，使得误差均方最小。神经网络理论已经证明BP网络具有强大的非线性映射能力和泛化功能，任一连续函数或映射均可采用三层网络加以实现[1]，其网络结构见图 1。

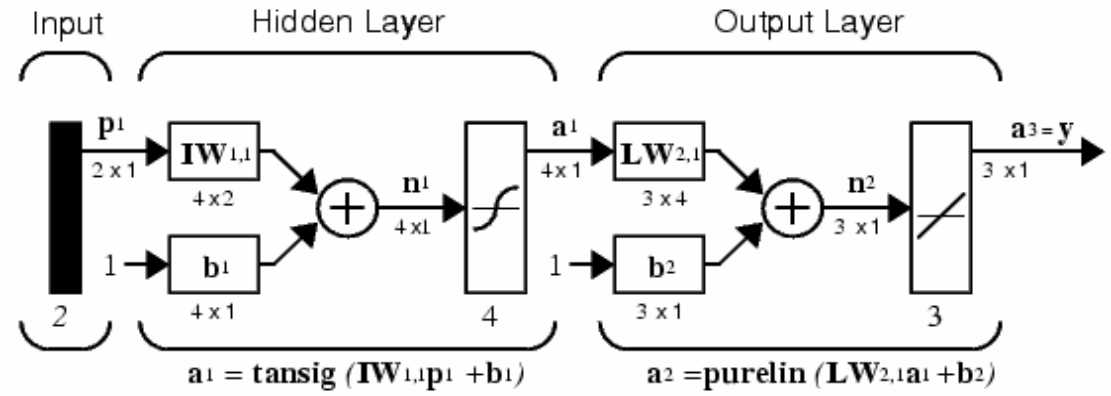

图 1　BP 神经网络结构

2　基于 BP 网络的研究生教育规模预测模型的建立

2.1 对样本数据的预处理

以 1991～2002 年全国全日制在校研究生数来衡量研究生教育规模（非全日制研究生数尚无完整的统计资料），对其产生影响的因素也是多方面的，这里，我们考虑了全国人均 GDP、投入人年、投入经费、导师数、财政预算内教育支出及国家财政科技拨款等 6 个因素。

为了缩小网络输入值（p）和目标值（t），对样本数据进行标准化处理。在Matlab中，可用prestd函数来完成，即[pn,meanp,stdp,tn,meant,stdt] = prestd(p,t);。这一过程标准化输入值和目标值，以使它们具有零均值和统一的标准差[2]。

2.2 网络结构的确定

神经网络理论定理Kolmogorov定理已经证明，经充分学习的三层BP网络可以逼近任何函数，因此选择三层BP网络，即只有一个隐含层的BP网络，如图 1 所示。该网络输入层的节点数由输入向量的维数决定，输入向量的维数是 6，所以输入层节点数确定为 6 个。输出层节点数由输出向量的维数决定，这里输出节点数为 1。隐层节点数的选择在所有BP网络中目前还没有理论上的指导，过多的网络节点会增加训练网络的时间，也会使网络的泛化能力减弱，网络的预测能力下降。但是网络节点过少则不能反映后续值与前驱值的相关关系，建模不充分。隐含层节点数可参考下式：$m=(p+q)^{0.5}+a$，其中m为隐层节点数，p为输入层节点数，q为输出层节点数，a为 1-10 之间的常数[3]。经反复训练，隐含层节点数定为 9。这样就形成了一个 6-9-1 神经网络。

2.3 学习算法的选择

基本 BP 算法采用梯度下降法使得误差均方（mse）趋向最小，直至达到误差要求。但在实际应用中，存在收敛速度慢、局部极值等缺点。Matlab7.0 神经网络工具箱中提供了十

多种快速学习算法，一类是采用启发式学习方法，如引入动量因子的 traingdm 算法、变速率学习算法 traingda、“弹性”学习算法 trainrp；另一类采用数值优化方法，如共轭梯度学习算法 traincgf 等、Quasi-Newton 算法 trainbgf 等、Levenberg-Marguardt 算法 trainlm。其中 Levenberg-Marguardt 数值优化算法适用于中小型网络，并且学习速率最快，所以本文选择 trainlm 算法。

2.4 Matlab7.0 中 BP 网络的学习、训练与模拟

（1）建立网络　net=newff(minmax(pn),[9,1],{'tansig','purelin'},'trainlm');newff()为建立 BP 神经网络的函数，minmax(pn)表示样本数据经预处理后的网络输入 pn 的取值范围，[9，1]表示隐层节点数是 9，输出层节点数是 1，{'tansig','purelin'}表示隐含层中的神经元采用 tansig 转换函数，输出层采用 purelin 函数，'trainlm'表示选择的学习算法。

（2）权重和阈值初始化　net=init(net);给各连接权重 IW{1,1}、LW{2,1}及阈值 b{1}、b{2}赋予（-1,+1）间的随机值。

（3）学习　[net,tr]=train(net,pn,tn);tn 为目标向量，根据网络学习误差逆传递算法，利用阻尼最小二乘算法迭代，由前一次训练得到的网络权重及阈值训练得到新的网络权重及阈值。

（4）模拟　an= sim(net,pn)；a= poststd(an,meant,stdt)；根据训练好的网络及输入向量进行模拟网络输出，由于事先对样本进行了标准化处理，所以还需用 poststd()函数将网络输出结果转换为原始数据的格式。

（5）预测方法　由于 1991 年-2010 年这 20 年中对研究生教育规模影响的因素是不断变化的，网络连接权重和阈值需不断校正，所以采取实时学习训练及仿真，来预测 2003 年-2010 年研究生教育规模。先采用 1991 年-2002 年共 12 年的统计资料作为一个学习样本，进行学习训练、仿真，预测后两年（2004 年、2005 年）的研究生教育规模，并不断将新的预测资料增加到学习模式中，增加后两年资料的同时剔除最早的两年资料。

3　几种模型预测结果的比较

对 1991 年至 2002 年研究生教育规模及其影响因素的数据，考虑存在共线性问题，先进行主成分分析后再分别利用线性回归模型、回归自回归混合模型进行预测[4]，并与BP神经网络模型的预测结果进行比较，见表 1。

表 1　不同模型预测结果比较

年份	实际值	线性回归模型		回归、自回归模型		BP 神经网络	
	（万人）	预测值	相对误差%	预测值	相对误差%	预测值	相对误差%
1991	8.7873	4.7869	-45.52	6.857	10.02	8.7834	-0.04
1992	9.3975	4.9861	-46.94	7.831	16.66	9.4011	0.04
1993	10.6405	9.3868	-11.78	15.1	41.91	10.6573	0.16
1994	12.7651	12.9038	1.09	14.858	16.39	12.7616	-0.03
1995	14.5148	16.3579	12.70	17.325	19.36	14.5338	0.13
1996	16.2035	19.4163	19.83	18.716	15.50	16.1982	-0.03
1997	17.5629	21.1501	20.42	18.152	3.35	17.5157	-0.27
1998	19.8356	24.5986	24.01	22.329	12.57	19.8852	0.25
1999	23.2563	27.5857	18.62	23.312	0.24	23.2335	-0.1
2000	30.0437	31.5409	4.98	27.52	8.40	30.0471	0.01
2001	39.2364	37.0396	-5.60	37.536	4.33	39.2402	0.01
2002	50.0873	42.5782	-14.99	44.511	11.13	50.0860	0.00
2003	-	48.1573		54.506		58.1611	
2004	-	54.5155		59.465		66.3745	
2005	-	61.7941		65.045		73.2746	
2006	-	70.1603		71.339		77.6843	
2007	-	79.8136		78.459		82.5487	
2008	-	90.991		86.538		88.2482	
2009	-	103.9752		95.734		97.8141	
2010	-	119.1032		106.234		108.4143	
平均	-	-	16.89	-	13.32	-	0.09

4　讨论与思考

⑴对 1991～2002 年全国全日制在校研究生的 12 年数据资料采用不同的模型进行拟合和预测，从结果来看 BP 神经网络模型的预测效果最好，其次是回归、自回归混合模型预测法，而线性回归模型的预测效果最差，这可能主要是线性回归模型没有考虑时间趋势的滞后效应所致。

(2)与线性回归模型相比，BP 网络的参数确定过程是一个反复的学习过程，数据的强共线性在进行线性回归时常会引起麻烦，而在神经网络模型中不会带来多大问题。BP 神经网络模型给出的结果是权重和阈值，模型中因变量和自变量关系的解释不像线性回归或时间序列分析那样直观。在一个多层网络中，很难说当其它自变量不变时，某一自变量会引起因变量多大变化。其原因是：在线性模型中，各自变量的作用是可以相互分离的，而在 BP 网络模型中，某个自变量对因变量的影响，不但取决于此自变量变化的大小，还依赖于其它自变量的取值。用 BP 网络所失去的是模型解释的直观性，所得到的是更精确的预测结果。

(3)利用 BP 神经网络，虽然可解决传统处理方法所不能处理的问题，但在实际应用中，对如何选择和确定一个合适的神经网络结构没有确切的理论指导，需反复实验以确定一个合适的网络结构。同时，由于 BP 神经网络训练过程中的权系数和阈值是随机产生的，所以每次训练、预测的结果都不同，即预报结果极不稳定。需反复训练，当多次输出结果在一定误差范围内时才取用。

参考文献

1. 杨行峻.人工神经网络与盲信号处理.北京：清华大学出版社，2003.
2. Preprocessing and Postprocessing, Full Product Family Help of MATLAB7.0
3. 邹广宇，王宏峰，汪定伟. 基于神经元网络模型的城市用水量预测. 信息与控制，2004,33(3):364-368.
4. 孙振球.医学统计学. 北京：人民卫生出版社，2002.