中国农业大学学报 2006,11(4):99-104
Journal of China Agricultural University

GM(1,N)灰色系统与BP神经网络方法的粮食产量预测比较研究

苏 博[1] 刘 鲁[1] 杨方廷[2]
(1. 北京航空航天大学 经济管理学院,北京 100083;2. 北京航天二院国家仿真中心,北京 100854)

摘 要 基于国家粮食安全预警系统的开发项目,针对我国粮食年产量预测中精度差和波动大的问题,分析了逐步回归、BP神经网络和GM(1,N)灰色系统3种常用预测方法的预测能力。根据能够计量和具有农学意义2个原则,选择了粮食作物播种面积、化肥施用量、粮食作物有效灌溉面积等12个重要的粮食年产量影响因子,用上述3种方法构建预测模型。在建模样本相同的情况下,结果显示,BP神经网络方法5年期拟合平均相对误差为1.44%,连续5年逐年预测平均相对误差可达到2.89%,这2个性能均优于其他2种方法,可以较好地应用于粮食安全预警系统,笔者最后探讨了对BP神经网络进一步优化的方法。

关键词 逐步回归;GM(1,N)灰色系统;BP神经网络;粮食产量预测
中图分类号 S114 **文章编号** 1007-4333(2006)04-0099-06 **文献标识码** A

Comparison and research of grain production forecasting with methods of GM(1,N) gray system and BPNN

Su Bo[1], Liu Lu[1], Yang Fangting[2]
(1. School of Economics & Management, Beijing University of Aeronautics and Astronautics University, Beijing 100083, China;
2. China Aerospace Science & Industry Corp. Simulation Center, Beijing 100854, China)

Abstract Based on the project of National Grain Warning System, aiming at predicting the grain output of China, this paper has compared and analyzed forecasting performances of three methods, namely step regression, BP neural network and GM(1,N) gray system. According to the principle of calculable and having agricultural significance, we chose twelve important effecting factors, and established respective forecasting model with the above three methods. Results showed that the average error of the method of BPNN was 1.44% and its average forecast error on five years could reach 2.89%, which is better than the other two methods in performances. It can be used in the project of National Grain Warning System. Finally the paper lists feasible methods to optimize the BPNN farther.

Key words step regression; GM(1,N) gray system; BP neural network; forecast of grain production

粮食产量预测是复杂的农学和统计学问题,受政策、自然环境、资源投入等多因素的影响。

国内外的相关研究中,不少学者构建了许多很有价值的理论假说和预测模型,主要有4类:气候生产力模型、遥感技术预测模型、投入产出模型、多元回归和因子分析模型[1-5]。这些模型和方法从不同角度对粮食产量预测进行了研究。如气候生产力模型,将粮食产量分离成经济技术产量和气象产量,两者综合建立粮食总产量预测模型,这种模型首次引入了气象和气候因子,但需要大量气候和气象数据,预测精度较低并且不能实现连续多年的长期预测;投入产出模型应用投入产出占用技术及系统科学方法实现对粮食产量、粮食进口量及自给率等多方面的综合分析,但多用于长期趋势分析,短期预测精度较低;多元回归和因子分析模型能够综合分析多方面影响因子的作用。比较有代表性的是黄继[illegible]等建

收稿日期:2005-10-13
基金项目:国家自然科学基金资助项目(70371004);博士点基金项目(20040006023)
作者简介:苏博,博士研究生,E-mail:bob_su@buaa.edu.cn;刘鲁,教授,博士生导师,主要从事系统工程和管理科学与工程研究,E-mail:liulu@buaa.edu.cn

立的中国农业政策分析和预测模型(简称 CAP-SIM),该模型分析了各种政策和外界冲击对中国各种农产品的生产、消费、价格、市场和贸易的影响以及未来中国农产品供给、需求、贸易和市场价格变动趋势,建立了农作物和畜产品生产的关联模型以实现产量预测。

上述模型多数采用传统的统计预测技术,如定性与推断技术、时间序列统计模型和因果关系方法。而近年来,随着智能技术发展,灰色系统理论预测、与专家系统相结合的预测、模糊推理预测、人工神经网络预测等新技术解决了传统方法的很多缺陷,得到广泛应用。近些年,虽然也出现了使用粗糙集理论[6]或灰色系统分析方法[7-10]等新算法构建的各种粮食产量的组合预测模型,但多数仅局限于算法的变换,采用的影响因子少,缺乏对粮食供需、市场、存储、政策、进出口贸易等多方面影响因子的系统分析和应用,而且大多存在诸如算法落后、适应性差、精度低等固有缺陷。因此,笔者旨在通过比较分析,结合国家粮食安全预警系统项目,选择一种稳定、先进的预测方法进行实际应用,分析和预测中国的粮食产量。

1 预测因子的选择和处理

影响粮食产量的因素除了无法确定的因素外,通常分为自然环境、气象、社会经济 3 类,结合国家粮食安全预警系统项目,将社会经济因素进一步分为粮食供给、需求、市场、存储、政策和进出口贸易 6 大类影响因素系统。根据能够计量及具有农学意义的原则,结合农业专家的意见,通过前期大量的影响因子分析,笔者选取了 1978—2003 年的粮食总产量(Y)为输出因子,初步选取粮食作物播种面积(x_1)、化肥施用量(x_2)、粮食作物有效灌溉面积(x_3)、受灾面积(x_4)、农村用电量(x_5)、农业机械总动力(x_6)、第一产业从业人员(x_7)、基本建设支出(x_8)、农业科技三项费用(x_9)、总农业支出(x_{10})、农村居民家庭平均纯收入(x_{11})、农村居民家庭经营平均纯收入(x_{12})12 个因子作为输入因子构筑模型,进一步的因子筛选将在具体建模过程中进行。原始数据来源于《中国统计年鉴》(2004)[11]。

为消除量纲影响,采用 Z-Score 变换法对采集的样本数据进行变换,即先求出变量数据的均值和标准差 S,然后用变量值减去其均值,再除以变量的标准差,即得标准化后的数值,用公式表示为:$X' = (X - \overline{X})/S$。

2 GM(1,N)灰色系统预测建模

2.1 灰色系统建模原理与方法

灰色系统(gray system)理论和方法是近年来广泛应用的一种预测算法。它的关联度分析方法是灰色系统分析、预测、决策的基础,可以为因素判别、优势分析和预测精度检验等提供依据。灰色系统建模是以灰色过程概念为基础,通过关联度分析,理清系统中各因素间的主要关系,找出影响最大的因素。最后将模型预测值作一次累减还原,用以对系统进行预测[12]。常用的灰色模型有以下几种:

GM(1,1)对应于一阶微分方程,进行一次累加,变量数为 1,常用残差模型;

GM(2,1)对应于二阶微分方程,进行二次累加,变量数为 1;

GM(1,N)对应于一阶微分方程,进行一次累加,变量数为 N。

这里主要采用 GM(1,N)模型,灰色系统建模的主要计算方法与步骤如下:

设 m 个输入因子的时间序列为$\{X_1^{(0)}(t)\}$,$\{X_2^{(0)}(t)\}$,…,$\{X_m^{(0)}(t)\}(t=1,2,\cdots N)$,$m$ 为各序列的长度,设定输出因子的时间序列为$\{X_0^{(0)}(t)\}(t=1,2,\cdots N)$。

1)求输入因子与输出因子的关联度。

灰色系统建模的核心就是求输入因子与输出因子的关联度。其计算方法与步骤具体如下:

①原始数据标准化变换。

②计算关联系数。经数据变换的输出数列记为$\{X_0(t)\}$,输入数列记为$\{X_i(t)\}$,则在时刻 $t=k$ 时输入序列$\{X_0(k)\}$与输出序列$\{X_i(k)\}$的关联系数 $L_{0i}(k)$ 可由下式计算

$$L_{0i}(k)=\frac{\Delta_{\min}+\Delta_{\max}}{\Delta_{0i}(k)+\rho\Delta_{\max}} \qquad (1)$$

式中:$\Delta_{0i}(k)=|x_0(k)-x_i(k)|$ $(1\leq i\leq m)$;$\Delta_{\max}$ 和 $\Delta_{\min}$ 分别表示所有比较序列各个时刻绝对差中的最大值与最小值。ρ 为分辨系数,其意义是削弱最大绝对差数值太大引起的失真,提高关联系数之间的差异显著性。

③求关联度。

$$r_{0i}=\frac{1}{N}\sum_{k=1}^{N}L_{0i}(k) \qquad (2)$$

r_{0i} 为输入序列 i 与输出序列 0 的关联度,N 为比较

序列的长度。

④排关联序。

⑤列出关联矩阵。

2) 筛选输入因子,因为 GM(1,N)预测数据量要求不能太大,一般不超过 10 个输入输出因子,因此需要根据灰色系统关联度分析的结果进行数据筛选。

3) 建立灰色数列 GM(1, N)模型分析。

2.2　数据分析结果

运用 DPS 软件建立灰色数列 GM(1, N)模型,数据分析结果如下:

输出因子 Y 与各输入因子的关联系数分别为: $G(1,1)=0.844\ 34$, $G(1,2)=0.658\ 93$, $G(1,3)=0.939\ 05$, $G(1,4)=0.888\ 70$, $G(1,5)=0.736\ 15$, $G(1,6)=0.924\ 66$, $G(1,7)=0.914\ 44$, $G(1,8)=0.809\ 65$, $G(1,9)=0.856\ 71$, $G(1,10)=0.883\ 00$, $G(1,11)=0.888\ 68$, $G(1,12)=0.951\ 56$。

关联序为: $x_{12}>x_3>x_6>x_7>x_4>x_{11}>x_{10}>x_9>x_1>x_8>x_5>x_2$。

关联矩阵为: 0.844 34, 0.658 93, 0.939 05, 0.888 70, 0.736 15, 0.924 66, 0.914 44, 0.809 65, 0.856 71, 0.883 00, 0.888 68, 0.951 56。

因为 GM(1,N)预测数据量要求不能太大,因此,筛选 x_3、x_4、x_6、x_7、x_{10}、x_{11}和 x_{12} 7 个因子作为输入项构建 GM(1,N)模型,运算结果如下。

系数向量: $a=1.174\ 04$, $b_3=1.016\ 50$, $b_4=0.101\ 90$, $b_6=-0.153\ 20$, $b_7=-0.101\ 74$, $b_{10}=11.791\ 77$, $b_{11}=-43.430\ 62$, $b_{12}=63.390\ 88$。

系统动态环节及其传递函数:

$$\frac{Y}{x_3}=\frac{0.865\ 8}{1+0.851\ 8s},\quad \frac{Y}{x_4}=\frac{0.086\ 8}{1+0.851\ s},$$

$$\frac{Y}{x_6}=\frac{-0.130\ 5}{1+0.851\ 8s},\quad \frac{Y}{x_7}=\frac{-0.086\ 7}{1+0.851\ 8s},$$

$$\frac{Y}{x_{10}}=\frac{10.043\ 8}{1+0.851\ 8s},\quad \frac{Y}{x_{11}}=\frac{-36.992\ 5}{1+0.851\ 8s},$$

$$\frac{Y}{x_{12}}=\frac{53.993\ 9}{1+0.851\ 8s}。$$

3　BP 神经网络产量预测

3.1　BP 神经网络原理与方法

BP 神经网络(back-propagation neutral network)通常是指基于误差反向传播算法(BP 算法)采用有导师训练方式的多层前向神经网络[13],由 D. E. Rumelhart 及其研究小组在 1986 年设计。由于它可以实现输入和输出的任意非线性映射,具有高度非线性和很强的自适应学习能力,因此被广泛应用于函数逼近、模式识别、经济预测等领域。

与传统的回归分析法(一般统计方法)相比,人工神经网络提供的数据较好,分类错误的次数很少。但 BP 网络也具有:1)学习过程收敛慢;2)容易陷入局部极小;3)鲁棒性不好,网络性能差等缺点。

BP 网络的学习训练过程由网络输入信号正向传播和误差信号反向传播两部分组成。在正向传播中,输入信息从输入层经隐含层逐层计算传向输出层,输出层的各神经元输出对应输入模式的网络响应;如果输出层得不到期望输出,则误差转入反向传播,按减小期望输出与实际输出的误差原则,从输出层经到中间各层,最后回到输入,层层修正各个连接权值。随着这种误差逆传播训练不断进行,网络对输入模式响应的正确率也不断提高,如此循环直到误差信号达到允许的范围之内或训练次数达到预先设计的次数为止。模型结构如图 1 所示。

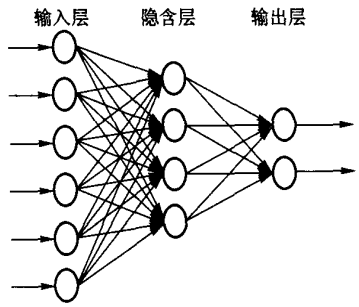

图 1　BP 神经网络结构

Fig. 1　Structure of BP neural network

算法的基本计算方法和公式[14]如下:

BP 网络学习规则是对网络权值和阈值的修正要沿着表现函数下降最快的方向——负梯度方向。

$$x_{k+1}=x_k-a_kg_k \tag{3}$$

式中: x_k 为当前的权值和阈值矩阵; g_k 为当前表现函数的梯度; a_k 为学习速率。

对 3 层 BP 网络,输入节点 x_i,隐层节点 y_j,输出节点 z_l。输入节点与隐层节点间的网络权值为 w_{ji},隐层节点与输出节点间的网络权值为 v_{lj}。当输出节点的期望值为 t_l 时,模型的计算公式如下:

隐层节点的输出

$$y_j=f\left(\sum_i w_{ji}w_i-\theta_j\right) \tag{4}$$

输出节点的计算输出

$$z_l=f\left(\sum_j v_{lj}y_j-\theta_l\right) \tag{5}$$

输出节点误差

$$E = \frac{1}{2}\sum_l (t_j - z_l)^2 = \frac{1}{2}\sum_l \left(t_l - f\left(\sum_j v_{lj} f\left(\sum_i w_{ji} x_i - \theta_j \right) - \theta_l \right) \right)^2 \quad (6)$$

3.2 建模步骤

BP神经网络本身没有从预测因子中进行筛选的功能,因此首先采用逐步回归方法从12个预测因子中筛选建模因子,并建立逐步回归预测模型,再以所得的建模因子建立BP神经网络模型,并进一步对检测样本进行预测,通过对预测结果的对比分析,讨论模型的优劣。

在回归系数及回归模型非常显著时,得到7个因子的回归方程:

$$Y = 32\,010 + 7.061\,x_2 - 0.056\,49\,x_3 - 0.071\,98\,x_4 - 5.616\,x_5 + 588.2\,x_9 - 16.77\,x_{11} + 26.86\,x_{12} \quad (7)$$

使用MATLAB软件编程实现的BP算法步骤[14]如图2。

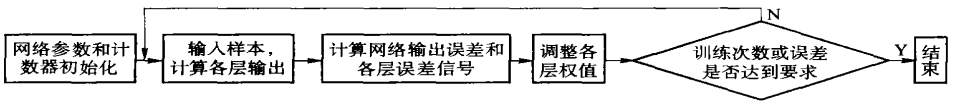

图2 BP算法流程

Fig. 2 Flow chat of BP arithmetic

其中网络参数的设定值如下:

1)网络节点 运用逐步回归的结果选定7个输入因子,因此输入节点数为7个,隐层节点数设为5,输出层节点数为1;

2)初始权值的确定 在程序中,设计了一个随机发生器程序,产生一组-0.5～+0.5的随机数,作为网络的初始权值;

3)最小训练速率 训练速率在不导致振荡前提下,越大越好。规定最小训练速率为0.9;

4)动态参数 取0.7;

5)允许误差 取0.000 01;

6)迭代次数 取1 000次;

7) Sigmoid参数 该参数调整神经元激励函数形式,取0.9;

8)数据转换 标准化转换。

建立预测模型,其网络训练过程和预测曲线如图3和4所示,对独立样本的预测结果及误差分析见表1。由图3可以看出,在所选参数下,经过近500步迭代,模型收敛,模型收敛性能较好,没有陷入局部极小。

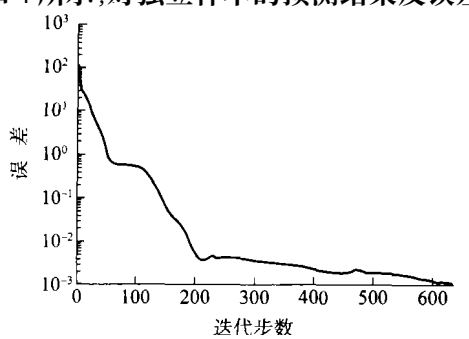

图3 BP算法训练曲线

Fig. 3 Curve of BP training

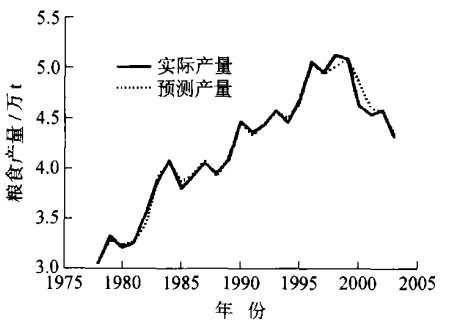

图4 BP算法模型预测曲线

Fig. 4 Forecasting curve of BP model

4 3种方法结果比较分析

首先分析各种方法的拟合精度。其中,GM(1,N)运用相关分析方法筛选出 x_3、x_4、x_6、x_7、x_{10}、x_{11}和 x_{12}7个因子作为输入项构建GM(1,N)模型;BP神经网络模型采用逐步回归方法中筛选出 x_2、x_3、x_4、x_5、x_9、x_{11}和 x_{12} 7个建模因子,建立回归方程后再运用神经网络方法对样本进行训练。x_3、x_4、x_{11}、x_{12}为2种建模方法最终筛选出的共有建模因子。

从表1可以看出,运用逐步回归方法筛选输入因子的BP网络模型整体拟合能力最优,运用相关分析方法筛选输入因子的GM(1,N)方法整体拟合

表 1　7 个预报因子的 3 种预报模型对 1998—2003 年的独立样本 1 次拟合结果

Table 1　Three forecasting models of seven factors, curve fitting result of independen sample from 1999 to 2003　万 t

方　法	年　份						平均绝对误差	平均相对误差/ %
	1998	1999	2000	2001	2002	2003		
实际值	51 229.5	50 838.6	46 217.5	45 263.7	45 705.8	43 069.9		
逐步回归方法	51 137	49 719	46 978	47 035	44 090	43 049		
绝对误差	92.5	119.6	760.5	1 771.3	1 615.8	20.9	563.39	
相对误差/ %	0.18	2.20	1.65	3.91	3.54	0.05		1.92
GM(1,N)方法	50 876.0	49 109.3	47 366.7	46 083.7	43 274.5	42 745.6		
绝对误差	353.5	1 729.3	1 149.0	820.0	2 431.3	324.3	1 134.57	
相对误差/ %	0.69	3.40	2.49	1.81	5.32	0.75	2.41	
BP 神经网络方法训练值	50 038.66	50 045.15	46 507.05	45 918.88	45 142.93	43 066.67		
绝对误差	1 190.84	793.45	289.55	655.18	562.87	3.23	582.52	
相对误差/ %	2.32	1.56	0.63	1.45	1.23	0.01		1.44

能力最差。

为达到实际预测的要求,分别采用逐步回归、GM(1,N)和 BP 神经网络方法进行逐年预测。即用 1978—1998 年 147 个样本作训练样本建模,预测 1999 年;再用 1978—1999 年 264 个样本作训练样本建模,预测 2000 年;依此类推,直到用 1978—2002 年 300 个样本作训练样本建模,对 2003 年进行预测。在此预测建模过程中,为达到最高的精度,对神经网络隐节点进行了改变,其余参数取值与前面相同,模型对独立样本的预测结果见表 2。

表 2　7 个预报因子的 3 种预报模型对 1999—2003 年的独立样本逐年预报结果

Table 2　Three forecasting model of seven factors, forecasting result of independent sample from 1999 to 2003　万 t

方　法	年　份					平均绝对误差	平均相对误差/ %
	1978—1998 预测 1999	1978—1999 预测 2000	1978—2000 预测 2001	1978—2001 预测 2002	1978—2002 预测 2003		
实际值	50 838.6	46 217.5	45 263.7	45 705.8	43 069.9		
逐步回归方法	59 586.49	47 420.08	45 711.67	40 155.63	43 004.87		
绝对误差	8 747.89	1 202.58	447.97	5 550.17	65.03	3 202.73	
相对误差/ %	17.21	2.60	0.99	12.14	0.15		6.618
GM(1,N)方法	51 043.30	50 102.23	44 640.49	43 497.55	43 039.56		
绝对误差	204.70	3 884.73	623.21	2 208.25	30.34	1 390.25	
相对误差/ %	0.40	8.41	1.38	4.83	0.07		3.018
BP 神经网络方法	50 826.84	49 586.76	48 003.46	45 596.57	43 434.07		
绝对误差	11.76	3 369.26	2 739.76	109.23	364.17	1 318.84	
相对误差/ %	0.02	7.29	6.05	0.24	0.85		2.89
隐节点个数	7	4	3	4	5		

从表 2 的结果可以看出,对 5 年独立样本作逐步预测时,运用逐步回归方法筛选输入因子的 BP 网络模型整体的平均预测误差最小,仅在 2000 年、2001 年粮食产量出现突然下跌时误差较大,但在 2002 年预测时就迅速适应调整过来,显示出其良好的适应性。GM(1,N)模型的综合预测表现好于逐步回归模型。而逐步回归模型预测误差最大,从中也可以看出逐步回归方法在增加已知信息时预测能力反而有所下降。

由以上的预测结果对比分析看出,分别在 2 组不同因子的条件下,无论是拟合精度还是逐年预测 5 个独立样本,BP 神经网络模型比逐步回归预测模型和 GM(1,N)模型的预测精度都高,而且预测结果稳定。运用相关分析方法筛选输入因子的 GM(1,N)模型虽然预测精度优于逐步回归模型,但预测结果波动大,而且拟合精度最差。从结果也可以看出 BP 网络模型方法平均预测精度只达到 2.89 %,仍有待于提高。

5 结论及存在问题

针对我国的粮食产量预测问题,分别将 BP 神经网络和 GM(1,N)灰色系统应用于国家粮食安全预警系统中,对比分析了逐步回归、BP 神经网络和 GM(1,N)灰色系统 3 种方法的预测能力。通过对比分析,证明了神经网络方法无论拟合还是预测性能均优于其他 2 种方法。但是方法本身仍存在一些问题,有待继续优化,主要有:

1)如何克服由于神经网络初始权值的随机性和网络结构确定过程中所带来的网络振荡,以及局部解问题,并有效提高网络的泛化能力。

2)BP 神经网络方法需要事先进行逐步回归筛选最优影响因子,其自身没有因子筛选的能力。

针对第 1 个问题,下一步考虑利用遗传算法全局性搜索的特点,寻找最合适的网络连接权和网络结构,优化神经网络的连接权和网络结构,会较好地克服以上问题,并且有效提高神经网络的泛化性能。

针对第 2 个问题,考虑使用自组织建模(GMDH)方法改进模型,运用它的自组织性能进行最优影响因子的筛选,从而去除逐步回归筛选因子这一环节,但其效果如何需要进一步比较研究。

参 考 文 献

[1] 王建林,王宪彬,太华杰. 中国粮食总产量预测方法研究[J]. 气象学报,2000,58(6):738-744

[2] 吴炳方. 中国农情遥感监测研究[J]. 中国科学院院刊,2004,19(4):202-206

[3] 陈锡康,郭菊娥. 中国粮食生产发展预测及其保证程度分析[J]. 自然资源学报,1996,11(3):197-202

[4] 黄季[illegible],李宁辉. 中国农业政策分析和预测模型——CAPSIM[J]. 南京农业大学学报(社会科学版),2003,3(2):30-41

[5] 毕守东,王冬平. 安徽省粮食产量的最优加权组合预测[J]. 预测,2000,3:70-72

[6] 肖智,郑大霞. 基于粗糙集的组合预测方法在粮食产量预测中的应用[J]. 统计与决策,2005,08

[7] 晏路明. 区域粮食总产量预测的灰色动态模型群[J]. 热带地理,2000,20(1):53-57

[8] 周介铭,彭文甫. 影响四川省粮食生产因素的灰色分析与粮食产量预测[J]. 四川师范大学学报(自然科学版),2005.03

[9] 吴玉鸣,李建霞. 非线性神经网络模型及在粮食生产预测中的应用[J]. 河南师范大学学报(自然科学版),2002,11

[10] 吴玉鸣,李建霞,徐建华. 中国粮食多因子灰色关联神经网络预测研究[J]. 华中师范大学学报(自然科学版),2002,36(4):419-423

[11] 国家统计局. 中国统计年鉴[M]. 北京:中国统计出版社,2004

[12] 唐启义. 实用统计分析及其 DPS 数据处理系统[M]. 北京:科学出版社,2002:10

[13] 李晓峰,徐玖平,王萌清,等. BP 人工神经网络自适应学习算法的建立与应用[J]. 系统工程理论与实践,2004,1(5):1-8

[14] 董长虹. Matlab 神经网络与应用[M]. 北京:国防工业出版社,2005:1