改进的 BP 神经网络算法在水质监测中的应用①

李 福, 郭 健

(南京理工大学 自动化学院, 南京, 210094)

摘 要: 针对一类多输入多输出系统进行辨识, 以“A simulation of the western basin of Lake Erie”为例, 通过分析河流湖泊的水质特征, 针对伊利湖湖泊水质建立数学模型, 由于该环境系统为多输入多输出系统, 文章采用了一种改进的 BP 神经网络算法, 利用 Matlab 神经网络工具箱进行数据分析, 绘出实际输出与模型输出的曲线以分析相关情况, 检验建立的模型对于系统的辨识水平, 给出传统BP网络和改进BP网络对该系统辨识的结果进行分析对比. 文章还对不同噪声层次下的数据进行分析比较, 并研究白噪声对于人工神经网络模型的影响.

关键词: 环境系统; 系统辨识; BP 神经网络; 数学建模; Matlab

Application of an Improved BP Neural Network Algorithm in Water Quality Monitoring

LI Fu, GUO Jian

(School of Automation, Nanjing University of Science and Technology, Nanjing 210094, China)

Abstract: This paper proposes an identification for a class of MIMO system. Taking “A simulation of the western basin of Lake Erie” as an example, quality characteristics of water system is analyzed and mathematical models of Lake Erie is made in this paper. An optimized BP ANN model is used for this MIMO system and the MATLAB’s NNT is used to carry on data processing. The effectiveness of system identification is inspected by the curves between models' output and actual results. The comparison between traditional and optimized BP ANN is given at the end of this paper. In this paper data collected under different noises is compared to study on the effect of white noises on ANN.

Key words: environmental systems; systemidentification; BPANN; mathematical modeling; Matlab

环境污染是当今社会的热门研究话题, 针对典型的环境系统如河流、空气、湖泊等建立模型可以有效的对污染水平进行预测以及时来预防污染.

以上环境系统一般都为多输入多输出复杂模型, 传统的最小二乘算法[1]针对单输入单输出线性系统有较好的辨识效果, 但不适用于非线性系统, 人工神经网络(ANN)模型[2]是通过模拟人的大脑神经处理信息的方式, 进行信息并行处理和非线性转换的复杂网络系统, 神经网络的优点是多输入多输出实现了数据的并行处理及自学习能力. 因此, ANN 对于多输入多输出的复杂系统辨识效果较好.

本文以“伊利湖”系统为例 , 其输入有水的温度、电导率、酸碱性、NO3 含量以及水的总硬度, 输出为氧气溶解性和水藻密度, 是一个典型的五输入——二输出系统. 文章给出了四种不同噪声层次下的数据, 每组数据(5 inputs and 2 outputs)给出 57 个采样点, 采样时间为一个月, 分别对四种模式下的系统采用改进的 BP 神经网络建立模型并分析辨识结果.

1 算法介绍

前向反馈(back propagation, BP)网络[1]是是一种反向传递并能修正误差的多层映射网络, 参数适当时, 此网络能够收敛到较小的均方差, 是目前应用最广的神经网络之一.

BP 网络是一种多层前馈神经网络, 它的名字源于在网络训练中, 调整网络权值的学习算法是反方向传播算法. 它是一种具有三层或者三层以上神经元的神经网络, 包括输入层、中间层(隐含层)和输出层, 上

① 收稿时间:2015-01-22;收到修改稿时间:2015-03-12

下层之间实现全连接, 而同一层神经元之间无连接. 输入神经元与隐含层神经元之间的连接值是网络的权值, 其意思是两个神经元之间的连接强度. 隐含层或输出层任一神经元将前一层所有神经元传来的信息进行整合, 还会在整合过的信息中添加一个阈值, 这主要是模仿生物学中的神经元必须到达一定的阈值才会触发的原理, 然后将整合过的信息作为该层神经元输入. BP 网络的训练过程[3]由正向传播与误差反向传播组成. 对于正向传播, 输入样本要从输入层传入, 经过隐含层处理, 之后传向输出层, 如果输入层的实际输出与期望输出不符, 则将进入误差的反向传播阶段. 所谓误差反传是对输出误差通过某种方式经过隐含层逐层反传至输入层, 并将误差分配给各层中的所有单元, 权值的不断调整就是网络的训练过程. BP 算法的核心是数学中的'负梯度下降'理论, 即 BP 网络的误差调整方向总是沿误差下降最快的方向进行.

而传统的 BP 网络存在着收敛速度慢, 网络易陷于局部极小, 学习过程常常发生振荡等缺陷, 故本文在传统 BP 网络的基础上提出了一些改进, 包括以下两个方面:

(1)初始权值和阈值的优化. 初始权值和阈值选取的随机性对网络的预测精度有着很大的影响, 文章利用遗传算法[4]的搜索能力,采用实数编码方式对初始权值和阈值进行优化.

(2)权值调整方法的改进. 传统 BP 网络中,权值的调节只考虑按照某时刻的负梯度方向修正权值,并没有考虑到以前积累的经验,本文在原有的基础上增加一个附加动量[5],不仅考虑误差在梯度上的作用,而且也考虑在误差曲面上的变化趋势的影响.

2　辨识过程

由于三层 BP 网络具有逼近任意函数的能力, 所以本文针对"伊利湖"系统设计一个三层神经网络, 包括输入层, 输出层以及一个隐层, 具体辨识步骤如下:

1)样本的建立及分类

根据采样数据建立一个 57*29 矩阵, 共 57 个采样点. 其中, 第一列为样本序列号, 故有效采样数据为一个 57*28 的矩阵. 每个采样点包含 7 个数据, 其中 5 个为输入, 2 个为输出, 按照噪声水平分别为 0, 10%, 20%, 30%每个数据对应 4 种情况, 共 7*4=28 列数据.

2)数据预处理

由于神经网络隐含层激励函数作用范围的限制以及数据自身特点对神经网络模型适应程度的影响, 首先根据式(1)分别对样本的输入、输出数据进行规格化处理,

$$\tilde{x}_k = \frac{x_k - x_{min}}{x_{max} - x_{min}} \tag{1}$$

其中, x_k 为规格化前的变量; x_{max} 和 x_{min} 分别为 x 的最大值和最小值; $\tilde{x}_k$ 为规格化后的变量.

MATLAB 中可直接调用归一化函数 premnmx[5]对数据进行预处理.

3)神经网络模型的建立

设网络的输入为 x, 第一层 BP 网络的输出为 $a1_i\ (i = 1, 2, \ldots, s1)$, 第二层 BP 网络的输出为 $a2_j\ (j = 1, 2, \ldots, s2)$. 第三层 BP 网络的输出为 a_3, 则三层 BP 网络的数学模型[2]为:

$$\begin{cases} a1_i = f_1\left(w1_i x - \theta 1_i\right) \\ a2_j = f_2\left(\sum_{k=0}^{s_1} w2_{kj} a1_k - \theta 1_j\right) \\ a_3 = f_3\left(\sum_{k=0}^{s_2} w3_k a2_k - \theta 3\right) \\ i = 1, 2, \ldots, s1 \\ j = 1, 2, \ldots, s2 \end{cases} \tag{2}$$

其中, w 为各层神经元的权值, f 为变换函数, θ 为各层神经元的阈值.

4)隐层节点数的确定

网络的输入与输出节点数是由实际问题的维数决定的, 与网络性能无关. 而隐含层节点数 l 的设计就非常重要了, 目前还没有统一的规范来解决这个问题. 一般可以利用经验公式来确定:

$$l = \sqrt{m + n} + a \tag{3}$$

其中, m、n 分别为输入节点数目与输出节点数目; a 为 1–10 之间的常数.

隐层节点数可以根据式(3)得出一个初始值, 然后利用逐步增长或逐步修剪法[6]最终确定神经元的个数. 逐步增长是从一个较简单的网络开始, 若训练结果不符合要求, 则逐步增加隐含层神经元个数直到合适为止. 逐步修剪则从一个较复杂的网络开始逐步删除隐含层神经元. 本次辨识研究经过逐步修剪法, 最终取隐含层神经元节点数为 8.

5)初始权值和阈值的优化.

遗传算法的基本思想[4]及步骤如图 1 所示.

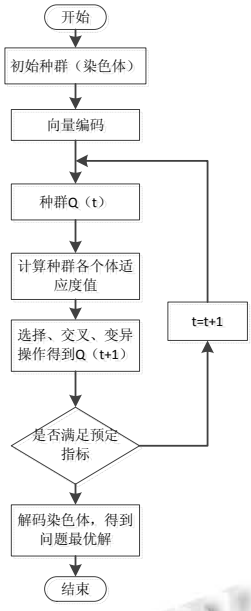

图 1 遗传算法流程图

由于遗传算法具有不需要其他辅助信息、内在启发式随机搜索最优值等优点, 文章采用实数编码的遗传算法对网络权值和阈值的初始值进行优化, 调用 Matlab 自带的遗传算法工具箱[7]来实现, 设置编码长度为 66, 适应度函数为:

$$g=\frac{1}{\frac{1}{M}\sum_{p=1}^{M}(Y_p-\tilde{Y}_p)^2} \tag{4}$$

式中 M 为训练样本数目, Y_p 为期望输出, $\tilde{Y}_p$ 为实际输出, 可表示为:

$$\tilde{Y}_p=w_p^2[1+e^{-T}]^{-1}+\theta_P^2 \tag{5}$$

其中,

$$T=\sum_{j=1}^{n}w_p^1\frac{(xp,j-aj)}{(Xp,1-a1)^2+\ldots+(Xp,n-an)^2}+\theta_p^1,$$

$Xp=(Xp,1\ldots Xp,n)$ 为输入样本, 第 p 个样本隐含层与输入层、输出层连接权值分别为 w_p^1、w_p^2, 神经元阈值为 θ_p^1、θ_p^2. 采用轮盘赌方式选择最优值, 群体规模取 100; 交叉概率取 0.7; 变异概率取 0.09; 代沟 0.9; 终止代数 200. 运行结果如图 2、3 所示:

6)学习速率的选定

学习速率[4]参数 net.trainParam.Lr 不能选择的太大, 否则会出现算法不收敛的情况; 也不能太小, 否则会使训练过程时间太长. 一般选择为 0.01–0.1 之间的值, 再根据训练过程中梯度变化和均方误差变化值来确定.

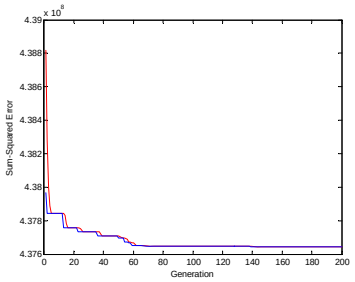

图 2 均方误差曲线

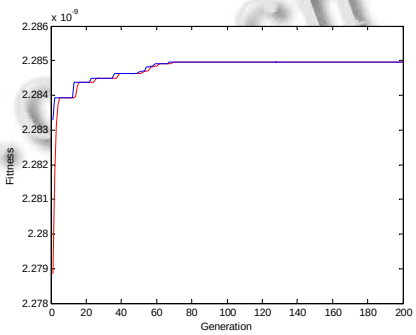

图 3 适应度曲线

7)网络的训练

传统的 BP 算法权值调节为负梯度调节:

$$\Delta w=-\gamma\frac{\partial E}{\partial w} \tag{6}$$

式(6)只考虑按照某时刻的负梯度方向修正权值, 并没有考虑到以前积累的经验,从而在学习过程中发生振荡,收敛缓慢, 本文在原有的基础上增加了一个附加动量:

$$\Delta w(t+1)=\gamma\left(-\frac{\partial E}{\partial w}\right)\Big|_{w=w(t)}+\alpha\Delta w(t) \tag{7}$$

式(6)、(7)中, γ 为学习率, E 为均方误差, $\alpha\in[0,1]$ 为动量因子[8], 一般取 0.95. 动量因子将最后一次权值变化的影响不断传递, 促使权值的调节向着误差曲面底部的平均方向变化, 有助于使误差从局部最小值跳出, 减小了学习过程的振荡趋势, 从而使收敛性得到改善.

设置网络训练采用最大训练轮回为 50000, 学习速率为 0.05, 当均方误差低于 0.001 时自动停止训练.

3 辨识结果分析

本针对四种不同噪声情况下的数据采用改进的 BP 神经网络进行辨识,结果如下:

(1)无白噪声情况辨识结果

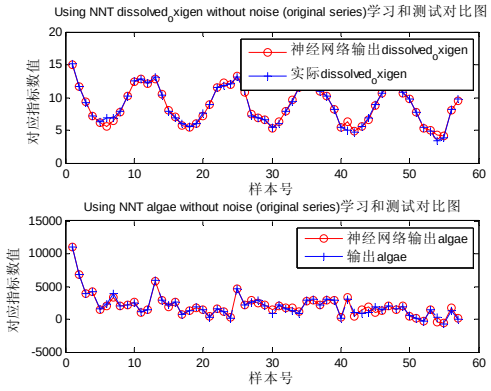

图 4　无噪声情况

(2)10%白噪声情况辨识结果

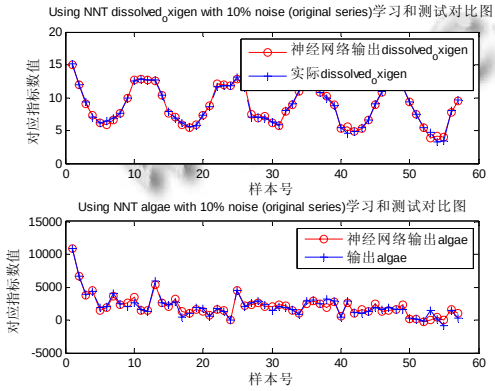

图 5　含 10%白噪声情况

(3)20%白噪声情况辨识结果

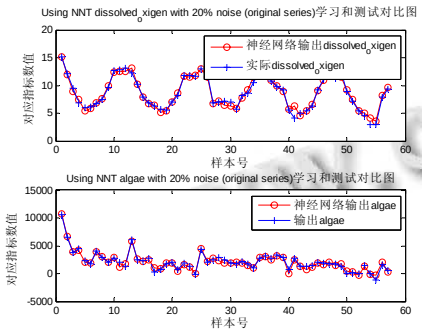

图 6　含 20%噪声情况

(4)30%白噪声情况辨识结果

由图 4–图 7 可知, 改进的 BP 神经网络建立的模型输出与实际输出的拟合水平都能达到 0.99 以上, 模型能较好的反映实际系统的性能.

此外, 对比不同噪声等级下的辨识结果可以看出, ANN 算法对于白噪声有较强的抗干扰能力, 对于 30%白噪声的情况, 该算法的拟合结果也非常好, 即白噪声的存在对训练结果几乎无影响.

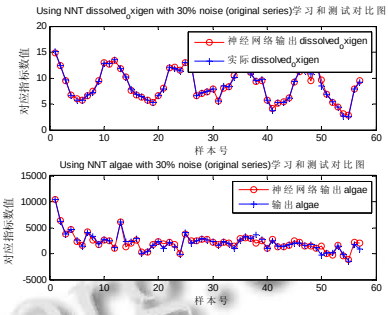

图 7　含 30%噪声情况

(5)传统算法与改进算法效果对比

下面以无白噪声情况给出两种算法运行效果, 训练情况如图 8 所示((a)为传统算法, (b)为改进算法).

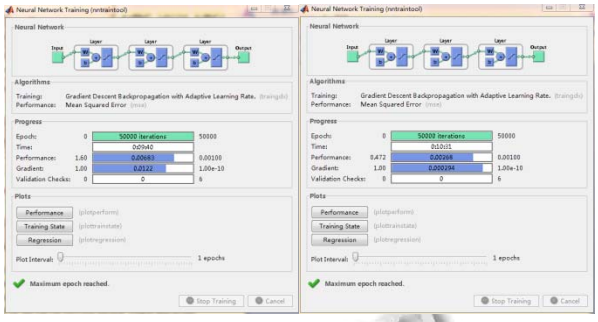

图 8　训练过程, 左(a)右(b)

收敛情况如图 9 所示.

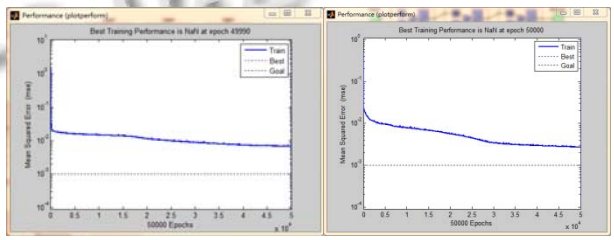

图 9　收敛情况, 左(a)右(b)

数据拟合情况如图 10 所示.

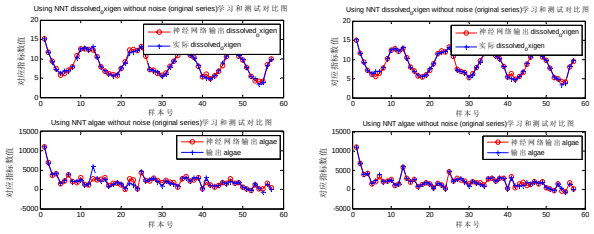

图 10　数据拟合情况, 左(a)右(b)

表 1 给出两者训练结果的对比.

表 1 传统算法与改进算法训练结果对比

	传统算法	改进算法
训练数据(组)	57	57
训练时间(分)	10:31	09:40
训练步数	49990	50000
拟合程度	0.99344	0.984

从表 1 可以看出, 对于 57 组数据的训练改进 BP 神经网络在保持拟合精度的情况下速度提高了约 1 分钟, 且在最大训练步数达到之前就完成了训练.

4 算法评价与总结

传统的 BP 网络对本系统的辨识已有较好的效果, 但同时也存在一些缺点, 如收敛速度慢, 网络易陷于局部极小, 学习过程常常发生振荡. 针对这些缺点, 文章从两方面进行改进: (1)结合遗传算法的全局搜索能力优化初始权值和阈值,为进一步进行网络的训练奠定基础; (2)增加动量因子考虑均方误差变化的经验值, 加快网络收敛速度. 实际测试显示, 改进后的 BP 神经网络弥补了传统算法的缺陷, 加快了训练速度, 而且对针对文章中的水质系统建模效果很好.

从仿真结果可以看出, 文章研究的改进 BP 神经网络算法对于多输入多输出复杂系统辨识程度很高, 该方法不仅使用于文中的水质监测, 对于其他非线性环境系统也适用.

参考文献

1 李鹏波,胡德文等.系统辨识基础.北京:中国水利水电出版社, 2006.

2 杨承志,孙棣华,张长胜.系统辨识与自适应控制.重庆:重庆大学出版社,2003.

3 Moore H. MATLAB for Engineers(Second Edition). Beijing: Publishing House of Electronics Industry, 2011.

4 周明,孙树栋.遗传算法原理及应用.北京:国防工业出版社, 1999:32–60.

5 张德丰等.MATLAB 程序设计与综合应用.北京:清华大学出版社,2012.

6 王淑玲,李振涛,邢棉.一种优化神经网络结构的遗传禁忌算法.计算机应用,2007,27(6):1426–1429.

7 雷英杰,张善文,李续武,周创明.MATLAB 遗传算法工具箱及应用.西安:西安电子科技大学出版社,2004:2–4

8 郭伟,张昭昭.熵在 BP 神经网络修剪算法中的应用.信息与控制,2009,38(5):633–636.