BP神经网络在水华短期预测中的应用

殷高方[1]， 张玉钧[1]， 胡丽[2]， 王志刚[3]， 肖雪[1]，
石朝毅[1]， 于绍惠[1]， 段静波[1]， 刘文清[1]

(1. 中国科学院 环境光学与技术重点实验室，中国科学院 安徽光学精密机械研究所，安徽，合肥 230031；
2. 合肥学院 建筑工程系，安徽，合肥 230022；3. 扬州大学 环境科学与工程学院，江苏，扬州 225009)

摘 要：为解决影响因素多、作用关系复杂的水华预测问题，将BP神经网与水体环境因子的高频实测数据相结合，构建了巢湖水华的短期动态预测模型，该模型准确地预测了每次水华发生的时间，预测值与实际观测值相关系数可达0.608 4；在分析BP神经网络自身局限性的基础上，研究了建模过程中输入输出数据预处理、网络结构设计、训练模式选择等问题，给出了水华预测中确定环境因子和建模方案的具体方法. 该方法容易移植到其它湖库，提高了模型的实用性和通用性.

关键词：BP神经网络；水华；短期预测模型
中图分类号：X 824 **文献标志码**：A **文章编号**：1001-0645(2012)06-0655-06

Application of BP Neural Network in Algal Blooms Short-Term Forecast

YIN Gao-fang[1], ZHANG Yu-jun[1], HU Li[2], WANG Zhi-gang[3], XIAO Xue[1],
SHI Chao-yi[1], YU Shao-hui[1], DUAN Jing-bo[1], LIU Wen-qin[1]

(1. Key Laboratory of Environmental Optics & Technology, Anhui Institute of Optics and Fine Mechanics, Chinese Academy of Sciences, Hefei, Anhui 230031, China; 2. Department of Architectural Engineering, Hefei University, Hefei, Anhui 230022, China; 3. Environmental Science and Engineering, School of Yangzhou University, Yangzhou, Jiangsu 225009, China)

Abstract: Forecast of algal blooms is a difficult problem because it is influenced by many factors that are in complex relation. This paper for the first time presents a short-term forecast model of the algal blooms in Chaohu Lake based on the combination of high-frequency measured values of environmental factors in water with BP neural network technique. The model could accurately forecast the time of each bloom, and the correlation coefficient between the forecasted and observed values is up to 0.608 4. Further, the issues in the modeling process were analyzed such as the limitation of BP neural network, the input and output data preprocessing, network structure design, training mode selection and other aspects. Finally, a specific method to select environmental factors and its modeling scheme were determined. The proposed model could be easily used on other lakes showing its strong practicability and universality.

Key words: BP neural network; algal blooms; short-term forecast model

收稿日期：2011-03-31
基金项目：国家"八六三"计划项目(2007AA061502，2009AA063005)；国家重大科技专项资助项目(2009ZX07420-008-005)；安徽省自然科学基金项目(11040606M26)；合肥学院科研发展基金资助项目(12KY05ZR)；安徽光学精密机械研究所所长基金资助项目(Y03AG31144)
作者简介：殷高方(1979—)，男，博士生，E-mail：gfyin@aiofm.ac.cn.
通信作者：张玉钧(1964—)，男，博士，研究员，E-mail：yjzhang@aiofm.ac.cn.

水华的生消过程伴随着各种物理、化学和生物过程，各种无机物质和有机物质之间相互作用并互为因果[1]. 由于影响水华生消的环境因子繁多、作用关系复杂，尤其是在突发性水华暴发机制尚不明的情况下[2]，水华生态过程驱动预测方法的研究陷入了困境. 相反，基于数据驱动的预测方法基本可以不考虑水华形成的生态机制，而以建立输入输出数据之间的最优数学关系为目标的黑箱子方法，特别适用于规律不明确的系统研究[3]. 传统数据驱动模型以回归模型最为常用，近些年一些先进的智能技术被采用，如人工神经网络模型、模糊数学模型和灰色系统模型等，其中人工神经网络作为真正的多输入多输出型网络，以其强有力的适应能力、学习能力，在数据驱动预测模型中应用最广.

BP型人工神经网络，又称BP神经网络，是基于误差反向传播算法的人工神经网络. 由于BP神经网络具有高度的非线性映射能力、有很强的容错性和很快的处理速度、强有力在线自学习和自适应能力，因此它可以实现输入和输出的任意非线性映射[4]. 这些特点让BP神经网络具有强有力的函数逼近、模式识别、数据压缩能力，特别适用于影响因子多、作用关系复杂水华预测模型中，解决水华与影响因子间显现出的多元、高阶和非线性的关系问题. 然而，由于BP神经网络核心部分BP算法潜在的缺陷，导致BP神经网络存在以下不足：① BP模型存在局部极小问题，对于一些实际问题难以达到全局最优；② 普通BP算法的收敛速度很慢；③ BP模型的结构设计，特别是隐层单元数的确定缺乏理论依据；④ 神经网络的“泛化能力”不能保证，网络训练模型如何设计，网络训练何时停止，网络是否已经有效逼近样本蕴含的内在规律等问题也缺少理论指导[5]. 因此，有必要对BP算法进行改进，同时需要根据实际问题类型，深入探讨网络在应用过程中的数据预测处理、网络结构设计、训练模式选择等影响模型性能的关键问题.

1 数据来源

BP神经网络水华预测模型的建模数据来源于安徽省科技厅计划项目“浮标式多参数水质自动监测系统研制及水华预警系统研究”课题组于2009年9～11月份巢湖西半湖的浮标站点(见图1)的连续监测资料，以及距浮标站3 km处的合肥骆岗机场气象资料，数据种类包括水温(A)、pH(B)、氧化还原电位(C)、电导率(D)、氨离子(E)、硝酸根离子(F)、氯离子(G)、浊度(H)、溶解氧(I)、蓝藻叶绿素a浓度(J)、绿藻叶绿素a浓度(K)、棕藻叶绿素a浓度(L)、气温(M)、大气压(N)、能见度(O)、风速(P)、风向(Q)、光照(R)等19种水质、水文气象类参数. 数据监测周期为1 h，除去个别时间由于通信网络等原因造成的监测数据中断，最终获得的数据记录共计1 400余条.

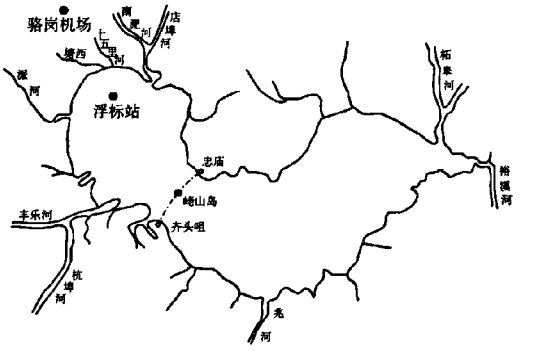

图1 巢湖外场数据监测点

Fig. 1 Map of monitoring points in Chaohu Lake

2 水华短期预测模型设计

2.1 输入输出数据预测处理

由于水体叶绿素a浓度是表征水体中藻类现存量的最直接指标，故水体中总叶绿素a浓度是模型的输出因子，通过对叶绿素a浓度的预测可以间接实现对藻类引发的水华进行预测. 在确定模型输出因子后，合理选择网络的输入因子对正确应用BP模型和保证模型预报精度非常重要[6]. 因为输入因子中，可能存在与输出因子关联弱的噪声因子，或者重叠反映系统信息的冗余因子. 无论是噪音因子还是冗余因子，都会加大分析问题的难度，增加模型的复杂性，最终影响模型预测能力. 筛选BP神经网络输入因子的基本原则是选择与输出因子相关而又彼此无关的环境因子. 通过图2相关性分析结果，剔

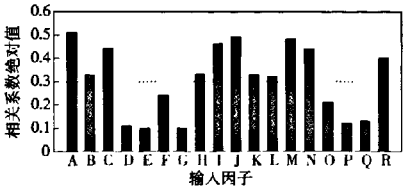

图2 输入因子与输出因子的相关性

Fig. 2 Correlation analysis between the input factors and the output factors

除与输出因子不相关(相关系数绝对值<0.3)的输入因子,再结合图 3 输入因子间的相关性分析结果,合并高度相关(相关系数绝对值>0.8)输入因子,最终压缩后的模型输入因子共 10 个:水温、pH、溶氧、浊度、气温、气压、光照、蓝藻叶绿素 a 浓度、绿藻叶绿素 a 浓度和棕藻叶绿素 a 浓度.

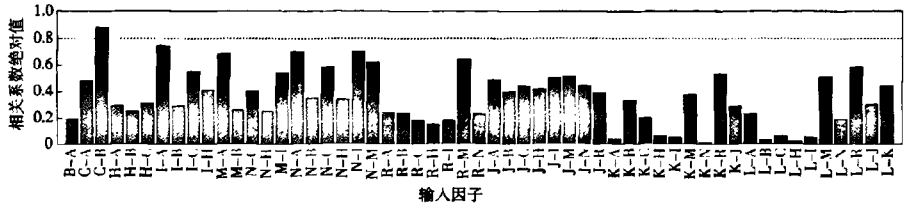

图 3　输入因子间的相关系数

Fig. 3　Correlation analysis between the input factors

2.2　网络结构

早有理论证明:3 层 BP 网络,当各层神经元均采用 S 型函数时,可满足任意复杂的非线性函数拟合逼近问题[7]. 这个存在性结论对神经网络结构的设计具有重要的指导作用. 参考该结论,确定水华预测模型为 3 层网络结构,即 1 个输入层、1 个隐含层和 1 个输出层. 然而,选取隐含层神经元数、层与层间的激活函数,虽然有规则可依,但所得结果差异较大,甚至完全不同[8]. 因此,隐含层神经元数、层与层间的激活函数仍需根据具体的研究问题而定.

2.2.1　隐含层神经元数

通过对不同隐含层神经元数的模型进行训练,比较训练样本误差的训练误差和非训练样本误差的泛化误差,从分析结果发现:随着隐含层神经元数目的增加,BP 模型的训练误差随着隐含层神经元数的增加呈下降趋势,但泛化误差却呈抛物线型(如图 4). 这意味着隐含层神经元数存在最佳值,因为判断模型是否有效逼近样本蕴含的内在规律,不是看模型对训练样本的拟合程度,而是要看模型对非训练样本的拟合能力是否接近于训练样本.

为兼顾预报精度和模型泛化能力,应选择训练误差和泛化误差接近时的隐含层神经元数据. 从训练误差和泛化误差的变化趋势图可以看出,隐含层神经元数以 7～15 个为宜,另外隐含层神经元数目通常大于等于输入因子的数量[9],所以最终将隐含层神经元个数确定为 12 个.

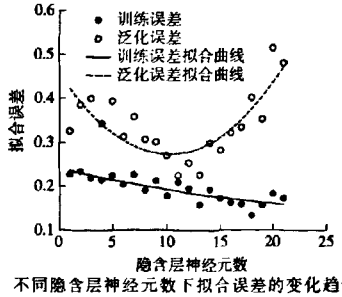

图 4　不同隐含层神经元数下拟合误差的变化趋势图

Fig. 4　Trend of fitting errors with hidden-layer neuron number

2.2.2　激活函数

BP 神经网络的激活函数反映样本输入与输出之间的对应关系,常用函数包括对数 S 型函数 logsig、正切 S 型函数 tansig、纯线性函数 purelin 3 种,通过对网络输入层-隐含层间激活函数、隐含层-输出层间激活函数的组合训练,结果表明:层间激活函数取 tansig-purelin 组合时,模型稳定收敛,训练误差和泛化误差最小,且泛化误差与训练误差最为接近;purelin-tansig、logsig-tansig 组合,模型收敛不稳定;tansig-logsig、logsig-purelin 和 purelin-logsig 组合,模型的训练和泛化误差较大(见表 1). 因此,采用 tansig-purelin 激活函数作为 BP 网络模型输入层-隐含层、隐含层-输出层间的激活函数.

表 1　不同激活函数组合条件下训练情况对比

Tab. 1　Contrast analysis of training results under different active function combinations

激活函数组合	输入层-隐含层	隐含层-输出层	收敛情况	训练误差	泛化误差
1	tansig	tansig	稳定	0.256	0.343
2	tansig	purelin	稳定	0.257	0.265
3	tansig	logsig	稳定	0.917	0.715
4	logsig	logsig	稳定	0.939	0.693
5	logsig	purelin	稳定	0.385	0.442
6	logsig	tansig	不稳定	0.276	0.209
7	purelin	purelin	稳定	0.380	0.405
8	purelin	tansig	不稳定	0.292	0.382
9	purelin	logsig	稳定	0.956	0.717

依据上述方法建立了具有输入层、隐含层和输出层的 3 层网络结构,各层节点数分别为 10,12 和

1，层间的激发函数为 tansig 和 purelin，网络拓扑结构如图 5 所示，该模型表达式为

$$y=\text{purelin}(V\cdot\text{tansig}(W\cdot x+b_1)+b_2).$$

式中：x 为输入变量，即水温、pH、溶氧、浊度、气温、气压、光照、蓝藻叶绿素 a 浓度、绿藻叶绿素 a 浓度和棕藻叶绿素 a 浓度；y 为输出变量，即未来 24 h 藻类的叶绿素 a 浓度；W 和 b_1 为输入层与隐含层间连接权重和阈值；V 和 b_2 隐含层与输出层间的连接权重和阈值.

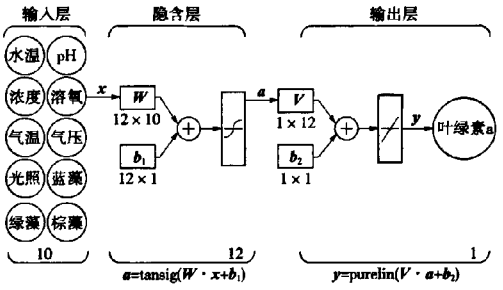

图 5 BP 网络预测模型拓扑结构

Fig. 5 Topological structure of BP neural network

2.3 训练模式

通过训练样本的训练建立合理的 BP 神经网络模型，是一个复杂而困难的过程，缺乏理论指导. 为了提高模型的训练速度和收敛的稳定性，模型训练采用了改进后的 BP 算法——带动量项批处理梯度下降算法. 该算法加入了动量因子项，可以减小学习过程中的振荡趋势，从而改善模型的收敛性，降低网络对误差曲面局部细节的敏感性，有效地抑制网络陷入局部极小[10]. 同时，为了得到合理的网络训练过程，在选择模型的训练参数时，采用枚举和交叉校验相结合的方法来判定最佳的学习速率和训练次数，即在不同训练次数和不同学习速率的条件下，对 BP 神经网络进行训练，比较训练样本的训练误差和校验样本的泛化误差来选定模型的训练参数.

2.3.1 训练次数

训练次数不足，无法提取训练样本蕴含规律，训练次数过多，又会导致过拟合现象. 判断模型的训练次数是否合理，或者说模型训练何时停止，最直接和客观的依据是校验样本的泛化误差是否和训练样本的训练误差接近. 因此，交叉校验是训练次数选择的有效途径. 图 6 给出模型训练过程中训练误差和泛化误差随训练次数的变化趋势，最初，增加训练次数，训练误差和泛化误差都在减小. 随着训练加深，训练误差仍然不断减小，而泛化误差开始呈现上升趋势. 当训练次数达到 716 次时，两者最为接近，约为 0.32，此时为最佳训练次数，交叉校验也会在此停止模型训练.

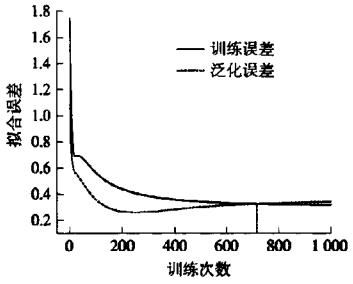

图 6 训练误差和泛化误差随训练次数的变化趋势图

Fig. 6 Trends of the training error and the generalization error with training times

2.3.2 学习速率

学习速率是网络权值和阈值的更新比率，决定着每一次循环训练中所产生的权值变化量. 对 0.001～0.300 范围内 7 个不同学习速率进行模型训练，比较训练结果以确定最佳的学习速率. 从表 2 可以看出：学习速率为 0.001 和 0.005 时，训练过程平稳收敛，且训练误差和泛化误差较小；随着学习速率的上升，拟合误差增加，训练过程逐渐变得不稳定；当学习速率达到 0.200 时，训练出现不收敛现象. 考虑到计算速度已经不是 BP 网络应用中的主要问题，为了保证 BP 网络训练的稳定性和精度，最终模型的学习速率定为 0.005.

表 2 不同学习速率条件下训练情况对比分析表

Tab. 2 Contrast analysis results of training with different learning rates

学习速率	收敛情况	训练误差	泛化误差
0.001	稳定	0.370	0.389
0.005	稳定	0.319	0.236
0.010	稳定	0.354	0.448
0.050	不稳定	0.251	0.457
0.100	不稳定	0.192	0.519
0.200	不收敛	1.077	0.582
0.300	不收敛	2.332	1.862

3 结果与分析

3.1 模型的预测结果

按 50%、25%和 25%的比例将 2009 年 9 月～2009 年 11 月巢湖实测 1 416 条样本数据，分为训练样本、校验样本和测试样本. 以水温、pH、浊度、溶

氧、气温、气压、光照、蓝藻叶绿素 a 浓度、绿藻叶绿素 a 浓度、棕藻叶绿素 a 浓度等 10 个环境参数为输入因子,依据上文讨论的网络结构、训练模型的设计方法,建立巢湖藻类叶绿素 a 浓度发展趋势的预测模型. 依据该预测模型对巢湖未来 24 h 藻类叶绿素 a 浓度进行连续 12 d(2009-10-17～2009-11-01)的预测,结果如图 7 所示,藻类叶绿素 a 浓度预测值(除 10-22,10-23)与实际观测值具有较好的一致性,相关系数为0.608 4,拟合误差为0.482 8. 同时,该模型几乎预测了所有叶绿素 a 浓度峰值发生的时间,实现了暴发性水华发生时间的准确预测. 但某些时候(如 10-22,10-23)模型的预测误差还是很明显的,从图 7 可发现,藻类叶绿素 a 浓度呈明显的昼夜变化规律,但 10-22～10-23 期间出现了异常(10-22 藻类叶绿素 a 浓度变化并无明显的昼夜变化,10-23 17:00 藻类叶绿素 a 浓度却发生了突变,由 16:00 的 1.3 μg/L 上升至 17:00 的 15.5 μg/L),这种异常可能是导致模型预测结果产生偏差的主要原因.

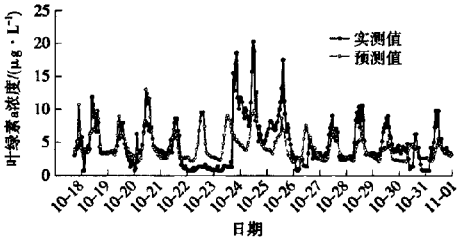

图 7　2009 年巢湖水体叶绿素 a 浓度未来 24 h 预测结果
Fig. 7　Algae chlorophyll-a concentration forecast results for the next 24 h in 2009

将模型用于未来 48,72,96 h 的水华模拟预测,从图 8～10 结果看:模型对未来 72 h 内预测效果都较为理想;随着预测周期增加,预测结果与实测值的相关度呈下降趋势,拟合误差呈上升趋势,如图 11 所示.

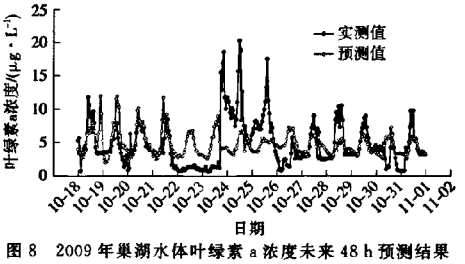

图 8　2009 年巢湖水体叶绿素 a 浓度未来 48 h 预测结果
Fig. 8　Algae chlorophyll-a concentration forecast results for the next 24 h in 2009

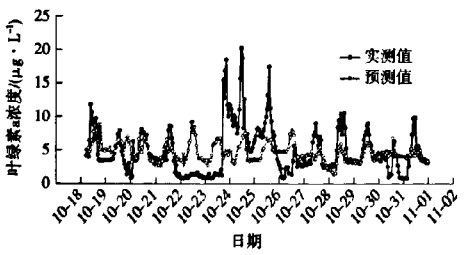

图 9　2009 年巢湖水体叶绿素 a 浓度未来 72 h 预测结果
Fig. 9　Algae chlorophyll-a concentration forecast results for the next 72 h in 2009

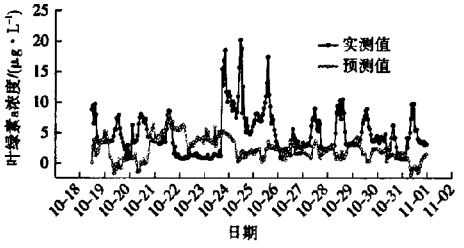

图 10　2009 年巢湖水体叶绿素 a 浓度未来 96 h 预测结果
Fig. 10　Algae chlorophyll-a concentration forecast results for the next 96 h in 2009

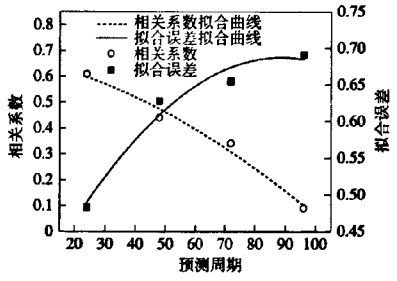

图 11　模型预测结果拟合误差与相关系数随预测周期的变化趋势
Fig. 11　Trends of correlation coefficient and fitting error with forecast period

3.2　模型的不足与改进

分析巢湖水华预测模型预测结果可知,该模型存在两个不足之处:

① 预测模型无法实现藻类叶绿素 a 浓度突变的预测:藻类叶绿素 a 浓度突变的本质是环境的异常引起的,如人工打捞蓝藻、降雨. 由于 BP 神经网络本身无法解决突发因素的影响,因此可以给预测模型增加人工修正量. 该人工修正量建立在湖泊突变环境因素的实时监测和工作人员丰富经验的基础上,估算环境突变带来的影响程度,并将其作为人工修正量叠加到模型输出上,对预测结果进行修正.

② 模型预测较长周期藻类叶绿素 a 浓度效果

差. 作为一种数据驱动模型,BP神经网络本质是从大量数据中提取其中蕴含的规律. 然而,水华预测建模的数据来源是2009年9月～2009年11月的监测数据,实为短期监测数据,并不能体现或完全体现藻类中长期的生长规律,所以模型也不可能准确预测较长时期的藻类叶绿素a浓度. 因此,实现水华的中长期预测,需要更全面的环境参数信息及更长时段的监测数据,这也是作者今后的研究方向.

4 结 论

水华的生消伴随着各种物理、化学和生物过程,各种无机物质和有机物质之间相互作用并互为因果,不同湖库诱发水华的主导环境因子种类和阈值存在明显差异. 因此,孤立地分析单个环境因子,或简单地研究个别环境因子与水华之间的对应关系都是片面的,不能系统地反映水华发生过程的全貌,因此,水华预测的生态学方法研究陷入了困境. 作者将水体环境因子的高频实测数据与BP神经网的数据驱动方法相结合,构建了巢湖水华的短期动态预测模型,并取得了较好的预测效果:在分析BP神经网络自身局限性的基础上,深入研究了建模过程中数据预处理、网络拓扑结构设计、训练模式选择等问题,提出了水华预测模型的输入输出变量、网络层数、隐含层神经元数、学习速率、激励函数、训练次数等模型参数的确定方法,该建模方法适用于所有湖库,对建立统一的水华数据驱动模型具有一定的借鉴作用.

参考文献:

[1] Quan Weimin, Yan Lijiao. Advance in study of lake eutrophication models [J]. Biodiversity Science, 2001, 9(2):168 - 175.

[2] 陈云峰,殷福才,陆根法. 水华爆发的突变模型:以巢湖为例[J]. 生态学报,2006,14(3):878 - 883.
Chen Yunfeng, Yin Fucai, Lu Genfa. The catastrophic model of water bloom: a case study on Lake Chaohu [J]. Acta Ecologica Sinica, 2006,14(3):878 - 883. (in Chinese)

[3] Lui Gibert C S, Li W K, Leung Kenneth M Y, et al. Modelling algal blooms using vector autoregressive model with exogenous variables and long memory filter [J]. Ecological Modelling, 2007,200(22):130 - 138.

[4] 雷蕾. 人工神经网络在大气污染预报中的应用研究[D]. 北京:北京工业大学,2007.
Lei Lei. Research on air pollution forecasting based on artificial neural networks [D]. Beijing: Beijing University of Technology, 2007. (in Chinese)

[5] 屈忠义. 基于人工神经网络理论的区域水:土(盐)环境预测研究[D]. 呼和浩特:内蒙古农业大学,2003.
Qu Zhongyi. Research on the forecast of regional soil-water (salt) environment based on artificial neural networks theory [D]. Huhhot: Inner Mongolia Agricultural University, 2003. (in Chinese)

[6] Uykan Z, Guzelis C, Celebi M E, et al. Analysis of input output clustering for determining centers of RBFNN[J]. IEEE Tras on NN, 2000,11(9):851 - 858.

[7] Nielsen H. Komogorovs mapping neural network existence theorem[C]// Proceedings of the International Conferenceon Neural Networks. New York: IEEE Press,1987,1(3):11 - 13.

[8] Michele S, Lawrence W, Harding J R. Developing an empirical model of phytoplankton primary production: a neural network case study [J]. Ecological Modeling, 1999,120(59):213 - 223.

[9] 刘国东,丁晶. BP网络用于水文预测的几个问题探讨[J]. 水利学报,1999,23(1):5 - 7.
Liu Guodong, Ding Jing. Discussion on problems of BP neural networks applied to hydrological prediction[J]. Journal of Hydraulic Engineering, 1999,23(1):5 - 7. (in Chinese)

[10] 郑高峰. 带动量项的BP神经网络收敛性分析[D]. 大连:大连理工大学,2005.
Zheng Gaofeng. BP neural network convergence analysis with momentum term [D]. Dalian: Dalian University of Technology, 2005. (in Chinese)

(责任编辑:匡梅)