International Journal of Ecology 世界生态学, 2018, 7(2), 53-60
Published Online May 2018 in Hans. http://www.hanspub.org/journal/ije
https://doi.org/10.12677/ije.2018.72008

Hans汉斯

Application of BP Neural Network Based on Principal Component Analysis in Algal Bloom Prediction

Jie Xia, Wenqing Wu, Haiyang Xu

School of Science, Southwest University of Science and Technology, Mianyang Sichuan
Email: swust20171001@163.com

Received: Mar. 18th, 2018; accepted: Apr. 2nd, 2018; published: Apr. 9th, 2018

Abstract

With the intensification of water pollution and eutrophication of freshwater ecosystems, large areas of algal bloom have been included, which not only destroy the ecosystems, but also cause huge economic losses. Therefore, it is very important to predict the occurrence of algal bloom according to the physical and chemical factors of water body. Firstly, according to the data of the pond for 1~15 weeks, the main influencing factors of 13 physical and chemical factors affecting the total plankton were analyzed based on principal component analysis (PCA). The main influencing factors of algal blooms were total nitrogen, transparency, dissolved oxygen, ammonium nitrogen, salinity, total phosphorus and dissolved oxygen. Secondly, according to the main seven physical and chemical factors identified as the input layer of BP neural network, the plankton biomass was used as the output layer to predict the occurrence of algal bloom. The results show that the fitting coefficient between the predicted result and the true value of the BP neural network model based on principal component analysis is as high as 0.9912. Therefore, the research method in this paper can effectively predict the occurrence of algal bloom.

Keywords

Algal Bloom Prediction, Physical and Chemical Factors, Principal Component Analysis, BP Neural Network

基于主成分分析的BP神经网络在水华预测中的应用

夏　杰，吴文青，许海洋

西南科技大学理学院，四川 绵阳
Email: swust20171001@163.com

文章引用: 夏杰, 吴文青, 许海洋. 基于主成分分析的 BP 神经网络在水华预测中的应用[J]. 世界生态学, 2018, 7(2): 53-60. DOI: 10.12677/ije.2018.72008

收稿日期：2018年3月18日；录用日期：2018年4月2日；发布日期：2018年4月9日

摘　要

随着淡水生态系统水体污染和富营养化进程的加剧，诱发了大面积水华，其不仅破坏生态系统，而且造成巨大的经济损失。因此，根据水体各个理化因子对水华的发生进行预测就显得尤为重要。首先根据池塘1~15周的数据，利用主成分分析法对影响浮游生物总量的13个理化因子进行主要影响因子分析，得到池塘水华发生的主要影响因子为：总氮、透明度、溶解氧、铵态氮、盐度、总磷，溶氧。其次，根据确定的7个主要理化因子作为BP神经网络的输入层，浮游生物量作为输出层来对水华发生进行预测。结果表明基于主成分分析的BP神经网络模型的预测结果与真实值的拟合系数高达0.9912。为此，本文的研究方法可有效地预测水华的发生。

关键词

水华预测，理化因子，主成分分析，BP神经网络

1. 引言

在我国水产养殖中，池塘养殖产量约占淡水养殖的 70%。随着淡水生态系统水体污染和富营养化进程的加剧，有害蓝藻、轮虫等浮游生物高密度发生，很容易诱发大面积的水华[1]。水华造成严重的环境污染和水体污染，同时区域用水安全也受到威胁，对水体养殖业造成严重的打击[2]。水华的发生是水生态系统中各种理化因子作用的结果，受到总氮、总磷、磷酸盐磷、叶绿素、溶氧、水温等因素的影响。虽然针对不同情况下水生态系统各因子对水华发生已有不少研究成果[3] [4]，但是准确的预测仍然是当前的首要任务。本文利用各个理化因子与水华之间的关系，结合主成分分析和神经网络来判断淡水养殖中水华是否发生，进一步为水华和水体富营养化治理提供理论依据。

预测模型是近几年国内外水华风险研究的重要工具[5] [6]，如回归分析模型、人工神经网络、指数法预测、灰色预测模型。人工神经网络由于具有较强的非线性映射能力，受到广大研究者的青睐。文献[7]利用 RBF 神经网络对叶绿素 a 的浓度做出预测，文献[8]基于深度学习对湖库藻类水华进行预测，文献[9]提出一种过程神经网络的预测模型。文献[10]基于主成分分析的 RBF 神经网络对需水量做出预测。然而，在水华预测中，需要对理化因子进行筛选，选择恰当的因子，以简化神经网络的复杂度，提高预测结果的准确度。

本文通过研究淡水养殖池塘相关主要理化因子，利用主成分分析法确定影响水华发生的主要理化因子。进一步，建立 BP 神经网络模型以理化因子作为输入样本，对浮游生物做出预测，从而更好地控制并预测水华的发生，提高养殖产量，减小环境污染。

2. 研究方法

2.1. 主成分分析法

主成分分析法[11]也称为主分量分析(PCA)，主要为了用较少的变量来解释原始数据的大部分变异，

将彼此相关性很高的变量转化为彼此独立或不相关的变量，是一种将原有的多个变量划分为少数几个综合评价的统计的方法。

1) 基本思想

设有m个样本，每个样本共有n个变量，若X表示指标的总和，x_{ij}表示第i个理化因子对应的第j个属性值，则构成的数据矩阵如下：

$$X=\left(x_{i,j}\right)_{i=1,2,\cdots,m,\ j=1,2,\cdots,m}. \tag{1}$$

设原来的变量指标为$X_1,X_2,X_3,\cdots,X_n$，它们的综合指标为新的变量指标为$Z_1,Z_2,Z_3,\cdots,Z_m$，c_{mp}为相应的系数，则

$$Z_k=c_{k1}X_1+c_{k2}X_2+\cdots+c_{kp}X_p,\ k=1,2,\cdots,m. \tag{2}$$

主成分决定的新变量指标$Z_1,Z_2,Z_3,\cdots,Z_m$分别为原变量$X_1,X_2,X_3,\cdots,X_n$的第一，第二，…，第m个主成分。其中Z_1在变量中所占的贡献率最大，$Z_1,Z_2,Z_3,\cdots,Z_m$对评价系统的贡献率逐渐下降，挑选贡献率较大的作为主成分。

2) 分析过程

① 计算相关系数矩阵$R=\left(r_{ij}\right)_{p\times p}$，其中

$$r_{ij}=\frac{\sum_{k=1}^{n}\left(x_{ki}-\overline{x}_i\right)\left(x_{kj}-\overline{x}_j\right)}{\sqrt{\sum_{k=1}^{n}\left(x_{ki}-\overline{x}_i\right)^2\left(x_{kj}-\overline{x}_j\right)^2}}. \tag{3}$$

② 计算特征值与特征向量

用雅可比行列式对$\left|\lambda I-R\right|=0$进行求解，求出特征值$\lambda_i\left(i=1,2,\cdots,p\right)$，并按照$\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_p\geq 0$排列，然后分别求出对应于特征值$\lambda_i$的特征向量$e_i\left(i=1,2,\cdots,p\right)$。

③ 计算主成份贡献率z_i及累计贡献率Z_i。

主成份z_i的贡献率为：

$$z_i=\lambda_i\Big/\sum_{k=1}^{p}\lambda_k,\ i=1,2,\cdots,p. \tag{4}$$

累计贡献率Z_i：

$$Z_i=\sum_{k=1}^{p}\left(\lambda_k\Big/\sum_{k=1}^{p}\lambda_k\right),\ i=1,2,\cdots,p. \tag{5}$$

④ 计算主成份载荷l_{ij}：

$$l_{ij}=p\left(z_i,x_j\right)=\sqrt{\lambda_j}-e_{ij},\ i,j=1,2,\cdots,p. \tag{6}$$

2.2. BP 神经网络模型

在神经网络中，应用最广泛的是美国加州大学的鲁梅尔哈特和麦克莱兰等人于 1985 年提出 BP 神经网络[12] [13]。该模型是典型的多层网络，具有输入层节点、输出层节点，而且具有一层或多层隐含节点。其核心是通过一边向后传递误差，一边修正误差的方法来不断调节网络参数(权值和阈值)，以实现或逼近所希望的输出、输入映射关系。一个三层的 BP 网络结构如图 1 所示。

BP算法具有梯度性，也称为快速下降法，其迭代基本思想是：从一个初始点w_0出发，计算在点w_0的

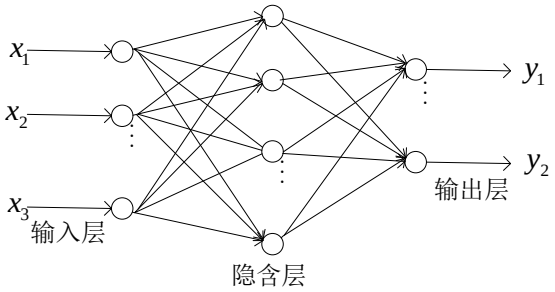

Figure 1. BP neural network topology
图 1. BP 神经网络拓扑结构图

负梯度方向 $-\nabla E(w_0)$，并沿该方向移动一定距离到达新的点 $w_1 = w_0 - \eta \nabla E(w_0)$，只要参数 η 足够小，就能保证 $E(w_1) < E(w_0)$。不断重复这一过程，一定能找到 $E(w)$ 的一个极值。

BP神经网络具体算法如下：

1) 初始化网络及学习参数。为加快网络学习将隐含层和输出层各节点的连接权值、神经元阈值规范化处理，其表达式如下：

$$\begin{cases} x_k^{new} = \dfrac{0.002 + 0.996\left(x_k^{old} - \min x_k^{old}\right)}{\max x_k^{old} - \min x_k^{old}} \\ y_k^{new} = \dfrac{0.002 + 0.996\left(y_k^{old} - \min y_k^{old}\right)}{\max y_k^{old} - \min y_k^{old}} \end{cases}, \tag{7}$$

其中式(7)中，x_k^{old}，y_k^{old}，x_k^{new}，y_k^{new} 分别为处理前后的网络输入与输出。通过数据规范化处理，将输入、输出数据变成 $[-1,1]$ 中的数。

2) 利用数据对网络进行训练，计算网络的输入、输出值。

隐含层各节点的输入、输出值分别为：

$$s_j^k = \sum_{i=1}^{n} a_i^k w_{ij} - \theta_j, \quad a_j^k = \frac{1}{1+\mathrm{e}^{-s_j^k}}, \quad j = 1,2,\cdots,p. \tag{8}$$

输出层各节点的输入、输出值分别为：

$$l_t^k = \sum_{i=1}^{p} b_t^k v_t, \quad b_t^k = \frac{1}{1+\mathrm{e}^{-l_t^k}}, \quad t = 1,2,\cdots,q. \tag{9}$$

3) 误差逆传播，利用梯度下降法对各层连接层及阈值进行调整。

设网络的计算输出为 c_t^k，网络期望输出与计算输出的偏差均方值 E_k 为

$$E_k = \sum_{t=1}^{q} \frac{\left(y_k^k - c_t^k\right)^2}{2}, \tag{10}$$

输出层各节点的误差 $d_t^k = \left(y_t^k - c_t^k\right) c_t^k \left(1 - c_t^k\right)$，隐含层各节点的误差 $h_t^k = \left(\sum_{t=1}^{q} d_t^k v_{jt}\right) b_j^k \left(1 - b_j^k\right)$。

4) 修正权值与阈值。

5) 若网络的全局误差小于指定的值，则算法转入第6步，否则转入第2步。

6) 计算输出层。

7) 计算网络训练误差。

BP神经网络仿真测试结束后，通过计算真实值与输出值的偏差情况，对网络训练的泛化能力进行评价，选取决定系数R^2评价模型的性能，其中

$$R^2=\frac{\left(l\sum_{i=0}^{l}\hat{y}_i y_i-\sum_{i=0}^{l}\hat{y}_i\sum_{i=0}^{l}y_i\right)^2}{\left(l\sum_{i=0}^{l}\hat{y}_i^2-\left(\sum_{i=0}^{l}\hat{y}_i^2\right)^2\right)\left(l\sum_{i=0}^{l}y_i^2-\left(\sum_{i=0}^{l}y_i^2\right)^2\right)},\tag{11}$$

式(11)中，$\hat{y}_i$为预测值，y_i为真实值。决定系数在$[0,1]$内，系数越接近于1，表明训练效果越接近真实值，反之，系数越接近于0，表明训练效果越差。

3. 实例分析

3.1. 池塘水华主要理化因子

为了利用神经网络预测出水华发生，需要收集大量的淡水养殖池塘的历史数据。本文收集 2016 年 MathCup 大学生数学建模竞赛 A 题的数据，以 1 号池为例，对 1 号池 2 个观测点的数据进行平均，最终得到池塘 1 到 15 周各个理化因子与浮游生物量的数据。将浮游生物总量作为主成分分析的因变量，自变量包括总磷、磷酸盐磷、总氮、硝态氮、亚硝态氮、铵态氮、溶解氧、COD、水温、PH、盐度、透明度、总碱度。

根据主成分分析法的具体步骤，利用 MATLAB 对理化因子进行求解。首先对原始观测数据进行标准化处理，其次计算相关系数矩阵，接着计算特征值和特征向量，最后计算各个指标的贡献率。根据各个指标贡献率的大小关系，确定影响水华发生的指标。其中各个理化因子的特征根和贡献率及评价值见表 1，从表 1 可看出，总氮的贡献率为 31.9345%，透明度贡献率为 23.6956%，溶解氧贡献率为 12.5598%，铵态氮贡献率为 10.1669%，盐度贡献率为 8.2344，总磷贡献率为 6.2346，COD 贡献率为 3.6718%，因此贡献率排名前七的累积贡献率高达 96.5177%。为此选取 13 个理化因子中贡献率排名前七的因子代替原变量。

根据主成分分析进行指标筛选，最终筛选出评价值排名前 7 的作为引起水华发生的理化因子，即总氮、透明度、溶解氧(COD)、铵态氮、盐度、总磷，溶氧作为池塘水华的主要因子，并作为 BP 神经网络输入样本。最后，可得到引起水华的 7 个主要指标和浮游生物量数据如表 2。

Table 1. Principal component solution data sheet
表 1. 主成分求解数据表

理化因子	总磷	磷酸盐磷	总氮	硝态氮	亚硝态氮	铵态氮	溶解氧
特征根	0.9392	0.1251	4.7902	0.0186	0.3140	1.5250	1.8840
贡献率	6.2546	0.8342	31.9345	0.1237	0.1251	10.1669	12.5598
评价值	0.0112	−0.2643	0.5756	−0.5047	−0.2229	0.3419	0.4812
排名	6	9	1	11	8	4	3
理化因子	COD	水温	PH	盐度	透明度	总碱度	
特征根	0.5508	0.0027	0.0064	1.2352	3.5543	0.0582	
贡献率	3.6718	0.0428	0.0429	8.2344	23.6956	0.3880	
评价值	−0.0168	−0.8339	−0.5208	0.2568	0.5641	−0.4591	
排名	7	13	12	5	2	10	

Table 2. The main physical and chemical factors of water bloom and biological data sheet
表 2. 水华主要理化因子及生物总量数据表

周数	总磷	总氮	铵态氮	溶氧	COD	盐度	透明度	生物总量
	mg·kg^{-1}	mg·kg^{-1}	mg·kg^{-1}	mg/L	mg/L	mg/L	cm	10^6 个/L
1	10.1891	8.8743	18.7008	5.12	21.90	1.80	28.00	52.73
2	8.3721	8.3905	20.0819	4.16	20.95	2.05	26.00	296.12
3	8.7440	7.8190	23.6535	3.20	20.00	2.30	24.00	274.41
4	9.3626	8.3797	25.2355	4.96	23.40	2.10	25.00	95.94
5	8.4130	8.0672	22.2213	6.72	26.80	1.90	26.00	231.88
6	9.9054	6.4773	30.3525	5.04	27.27	2.00	24.00	172.29
7	5.5932	2.7576	11.2677	3.36	27.73	2.10	22.00	505.65
8	6.9944	4.1013	16.2684	2.88	25.57	2.10	22.00	185.53
9	9.9059	4.8385	21.6749	2.40	23.40	2.10	22.00	36.24
10	10.8805	6.7343	17.6684	3.27	23.08	1.60	21.00	262.41
11	11.8625	8.2514	27.0140	4.14	22.75	1.10	20.00	466.12
12	8.7985	8.1177	12.0263	5.29	24.06	1.30	19.50	290.00
13	12.5384	9.5958	26.7652	6.43	25.36	1.50	19.00	554.71
14	13.6464	8.3491	24.0939	5.56	25.70	1.50	21.00	1064.82
15	14.2669	7.3688	19.7633	4.60	26.03	1.50	23.00	954.71

3.2. 结果与分析

1) 模型的输入、输出参数的选择

本文用MTALAB 2015a进行神经网络训练，其中模型的输入层为各个主要理化因子，输出层为浮游生物总量，具体设计如下：

输入层：总氮、透明度、溶解氧、铵态氮、盐度、总磷，溶氧共7个单元作为输入层。

输出层：浮游生物总量共一个输出层。

2) 模型的结构设计

本文神经网络模型采用三层网络模型，输入参数为池塘1到15周的7个理化因子。对于隐含层节点的选取，通过构造不同隐节点数的网络进行训练，根据各层网络的误差对权值和阈值进行修正，并通过评价系数来判断拟合的效果。经过多次训练后发现，60个隐含节点的网络结构拟合效果最佳，因此网络拓扑结构为$7\times60\times1$。在训练过程中，选取logsig函数为输出层的激活函数，选取trainlm为训练函数。

3) 模型网络参数的选取及参数设定

参数设置为：net.trainparam.epochs = 10000为网络训练的最大次数，net.trainparam.lr=0.2;为学习效率，net.trainparam.show=200为每200轮显示一次。

4) 模型检验与仿真分析

对收集的数据分别对池塘浮游生物总量进行仿真分析，其结果如图2。通过观察浮游生物的预测曲线和实际数据预测曲线，可明显看出两者吻合度较高。此外，通过看评价系数$R=0.9912$可得，评价系数接近于1，表明拟合效果较好，对浮游生物的预测具有一定的使用价值如图3所示。

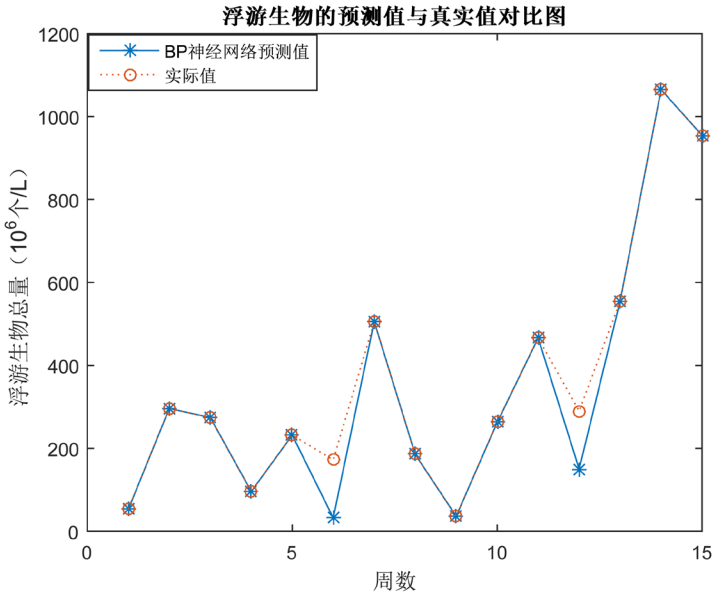

Figure 2. Planktonic predicted values and real value comparison chart
图 2. 浮游生物预测值与真实值对比图

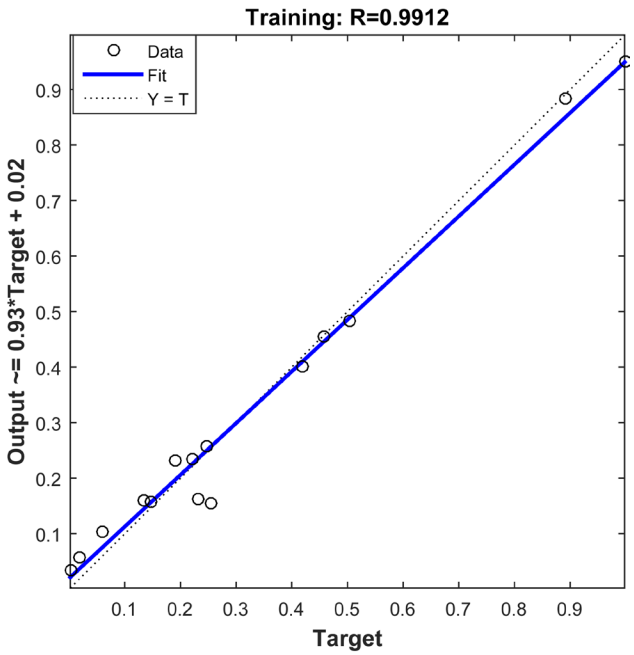

Figure 3. Neural network fitting coefficient map
图 3. 神经网络拟合系数图

4. 结论

本文通过构建影响水华发生的理化因子，并利用主成分分析法筛选出池塘水华主要因子。其次构建水华神经网络预测模型，得出以下结论。

1) 通过主成分分析可得影响水华发生的主要理化因子，包括总氮、透明度、溶解氧(COD)、铵态氮、盐度、总磷。这 7 个理化因子的累积贡献率为 96.5177%。

2) 所构建的神经网络模型，由于神经网络模型具有较强的非线性映射能力，经过训练后的预测值与真实值吻合度较高。

3) 基于主成分分析的 BP 神经网络降低了网络输入的层数，提高了程序运行效率，从而提高了神经网络的性能，对水华预测有较好的效果。

基金项目

西南科技大学理学院创新基金项目“量化投资平台下的交易策略研究”(项目编号：LXCX-05)，主持人：夏杰。

参考文献

[1] 窦明, 谢平, 夏军, 沈晓鲤, 方芳. 汉江水华问题研究[J]. 水科学进展, 2002, 13(5): 557-561

[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, 26(3): 878-883.

[3] 杜桂森, 王建厅, 张为华, 冯伶亲, 刘静. 官厅水库水体营养状况分析[J]. 湖泊科学, 2004, 16(3): 277-281.

[4] 全为民, 严力蛟, 虞左明, 焦荔. 湖泊富营养化模型研究进展[J]. 生物多样性, 2001, 9(2): 168-175.

[5] 郝启文, 王小艺, 许继平, 刘载文, 盛璐, 何多多. 湖库水质监测与水华预警信息系统[J]. 计算机工程, 2013, 39(1): 287-289+293.

[6] 孔繁翔, 马荣华, 高俊峰, 吴晓东. 太湖蓝藻水华的预防、预测和预警的理论与实践[J]. 湖泊科学, 2009, 21(3): 314-328.

[7] 仝玉华, 周洪亮, 黄浙丰, 张宏建. 一种自优化 RBF 神经网络的叶绿素 a 浓度时序预测模型[J]. 生态学报, 2011, 31(22): 6788-6795.

[8] 姚俊杨, 许继平, 王小艺, 黄振芳. 基于深度学习的湖库藻类水华预测研究[J]. 计算机与应用学, 2015, 32(10): 1265-1268.

[9] 李大刚, 王小艺, 刘载文, 许继平, 赵星, 戴军. 过程神经网络水华预测方法研究[J]. 计算机与应用化学, 2011, 28(2): 173-176.

[10] 桑慧茹, 王丽学, 陈韶明, 孙娟, 李司瑾. 基于主成分分析的 RBF 神经网络在需水预测中的应用[J]. 水电能源科学, 2017, 35(7): 58-61.

[11] 苏键, 陈军, 何洁. 主成分分析法及其应用[J]. 轻工科技, 2012, 28(9): 12-13+16.

[12] 万金保, 曾海燕, 朱邦辉. 主成分分析法在乐安河水质评价中的应用[J]. 中国给水排水, 2009, 25(16): 104-108.

[13] 任黎, 董增川, 李少华. 人工神经网络模型在太湖富营养化评价中的应用[J]. 河海大学学报(自然科学版), 2004, 32(2): 147-150.