doi:10.3969/j.issn.1672-1292.2017.02.011

基于多元回归和神经网络途径的太湖富营养化指标分析与预测

王凯翔

(南京师范大学计算机科学与技术学院,江苏 南京 210023)

[**摘要**] 水体富营养化是目前太湖面临的一个重大环境问题,有效地预测湖泊的水质变化对防治富营养化至关重要. 现有的预测方案很多依赖于一些常见的与目标变量相关的因素,针对太湖这一具体对象,难以全面准确地分析相关因素之间的联系以及预见目标变量的变化趋势. 本文首先寻找出太湖富营养化的所有可能的影响因素,提供较为全面的备选变量以期更好地查找出目标变量的相关因素. 选择总氮这一富营养化重要评价指标作为分析和预测的对象,根据各影响因子与总氮之间的相关系数筛选出影响程度较高的因素. 分别运用多元线性回归分析方法和 BP 神经网路预测方法对总氮的变化进行预测研究,并将两种方法的预测性能进行比较. 结果表明,从较为全面的变量中选出的影响因子可较好地预测总氮的变化情况,多元线性回归和 BP 神经网络方法的实验结果都较好,从拟合优度和均方误差的角度看,BP 神经网络的预测效果更好.

[**关键词**] 富营养化,预测,相关系数,多元线性回归分析,BP 神经网络
[**中图分类号**]TP18　[**文献标志码**]A　[**文章编号**]1672-1292(2017)02-0070-05

Taihu Lake Eutrophication Index Analysis and Prediction Based on Multiple Regression and Neural Network

Wang Kaixiang

(School of Computer Science and Technology, Nanjing Normal University, Nanjing 210023, China)

Abstract: Eutrophication of body of water is a major environmental problem of the Taihu lake, and the eutrophication prediction is an effective early warning method to know the change of lake water quality. Now there are a lot of prediction schemes depending on the existing known factors related to the target, and it is difficult to fully accurate analysis of the relationship between the related factors and predict the change trend of target variable. In this paper, firstly, we find out all of the possible affecting factors of the eutrophication of Taihu lake, then we can obtain a better target variable of related factors with a more comprehensive variables provided. We select the eutrophication of total nitrogen as the object of analysis and forecasting. The factors are selected by their correlation with total nitrogen. We conduct the prediction study for the change of total nitrogen with multiple linear regression analysis method and BP neural network prediction method, and compare the performances of the two methods. The results show that the variables selected can well predict the changes of total nitrogen. The experimental results obtained from multiple linear regression and BP neural network method are both accurate, from the perspectives of the goodness of fit and the mean square error, and the BP neural network is better.

Key words: eutrophication, forecast, correlation coefficient, the multivariate linear regression analysis, the BP neural network

富营养化是当前湖泊水库面临的一种严峻的水环境问题[1]. 水体中营养过剩是藻类过度繁殖的主要原因,湖泊水体的生态平衡也因此被破坏. 城市工业生产以及周边居民生活所产生的污染物由城市地表径流携带流入湖中[2],随着温室气体排放加剧,全球变暖以及区域气候灾害的不断出现,水体的富营养化

收稿日期:2016-11-09.
基金项目:2014 年国家级大学生创新训练项目(201410319038Z).
通讯联系人:王凯翔,硕士研究生,研究方向:数据库与数据挖掘. E-mail:15651805876@163.com

程度日益加剧. 太湖流域是我国长江三角洲地区重要的经济中心[3],工业化程度和城镇化程度很高,向水体中排放的污染物的量也很大,由于缺乏较为有力的预防手段和治理措施,太湖的富营养化形势日趋严峻. 如何缓解和遏制太湖富营养化是迫在眉睫的关键问题.

为了应对湖泊富营养化越来越严重的趋势,许多研究者开始探索使用数学模型来预测湖泊富营养化. 当前湖泊的富营养化模型主要分为 3 个大类,分别为统计学模型、单一营养物质负荷模型、复杂的生态动力学模型. 其中,统计学模型主要是建立在水体各水质监测变量数据统计分析的基础之上[4]. 统计学模型可以有效地提供水体水质的大致变化趋势,能够快速地对富营养化的影响因子作出评价.

目前,对富营养化的具体成因和内部机理还不是非常清楚,但是有相当多的研究表明水体富营养化是由多种环境因素共同作用的结果:不仅包括了氮、磷等元素[5],还涉及到温度、pH 值等其他环境因素的共同作用. 由于环境中的影响因素很多,各因素之间存在着复杂的关系[6],通过其中的一些影响因素对另一些具有决定意义的核心指标进行建模并预测,一直是学者们关注的重点.

现有的大量预测方案多是按照某些文献中提出的一些常见的影响因素去预测目标变量,但不同的湖泊有着自己的特点,对一个具体的目标变量而言,不同的湖泊在影响因素的选择上会有一定的差异. 如果不能很好地确定出具体的影响因素,就很难准确作出预测.

本文首先寻找出太湖富营养化所有可能的影响因素,采集太湖 2009-2010 年间在自动监测点的部分污染数据. 论文以总氮这一富营养化重要评价指标作为分析和预测对象. 对太湖富营养化的影响因子之间的相关关系进行二元相关分析,再根据各影响因子与总氮之间的相关系数筛选出影响程度较高的因素. 论文的实验以 2009 年及 2010 上半年的数据作为训练样本,以相关系数筛选的因素构建多元线性回归分析模型以及在实际运用中最为广泛的 BP 神经网络训练模型,然后根据 2010 下半年中影响因素的值对这一段时间研究对象的值进行预测,并与 2010 年下半年研究对象的实际值进行比较,评估对目标变量预测的准确性. 最后分析和比较两种预测方法的性能.

1 多元线性回归和 BP 神经网络

1.1 多元线性回归

回归分析研究的主要对象是客观事物变量之间的统计关系,它是建立在对客观事物进行大量实验和观察的基础上,用来寻找隐藏在那些看上去不确定的现象中的统计规律性的统计方法. 回归分析方法是通过建立统计模型研究变量间相互关系的密切程度、结构状态及进行模型预测的一种有效工具[7].

设随机变量 y 与一般变量 $x_1,x_2,\cdots,x_p$ 的线性回归模型为:

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p+\varepsilon, \tag{1}$$

式中,$\beta_0,\beta_1,\beta_2,\cdots,\beta_p$ 是 $p+1$ 个未知参数,β_0 称为回归常数,$\beta_1,\beta_2,\cdots,\beta_p$ 称为回归系数; y 称为被解释变量(因变量);$x_1,x_2,\cdots,x_p$ 是 p 个可以精确测量并控制的一般变量,称为解释变量[8].

多元线性回归模型建立后,必须对其进行各项检验,主要包括 F 检验、t 检验、回归系数的置信区间以及拟合优度检验. 通过一系列检验之后,才能将所得到的回归模型用于预测.

1.2 BP 神经网络模型

人工神经网络是在人类对其大脑神经网络认识理解的基础上人工构造的能够实现某种功能的神经网络. 它是理论化的人脑神经网络的数学模型,是基于模仿大脑神经网络结构的功能而建立的一种信息处理系统. 它实际上是一个大量简单原件相互连接而成的复杂网络,具有高度的非线性,能够进行复杂的逻辑操作和非线性关系实现的系统[9].

反向传播网络(简称 BP 网络)是将 W-H 学习规则一般化,对非线性可微分函数进行权值训练的多层网络. 如图 1 所示,BP 神经网络算法由两部分构成:信息的正向传递与误差的反向传播. 在正向传播过程中,输入信息从输入经隐含层逐层计算传向输出层,每一层神

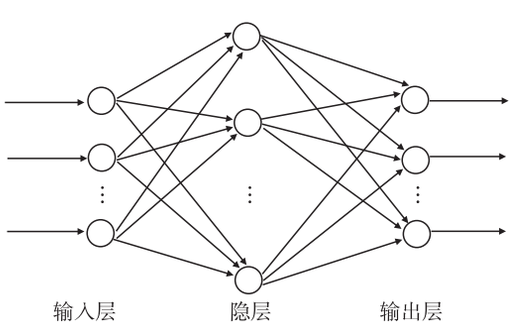

图 1 BP 神经网络

Fig. 1 BP neural network

经元的状态只影响下一层神经元的状态. 如果在输出层没有得到期望的输出,则计算输出层的误差变化值,然后转向反向传播,通过网络将误差信号沿原来的连接通路反传回来修改各层神经元的权值直至达到期望目标.

2　总氮预测的具体方法

2.1　目标变量的因子选择

本文针对太湖的具体情况选取了 pH 值(pH)、悬浮物(SS)、透明度(SD)、溶解氧(DO)、高锰酸盐指数(COD_{Mn})、生化需氧量(BOD_5)、氨氮(W_{NH_4N})、挥发酚(V_{phen})、总磷(TP)、叶绿素(W_{Chla})、硫化物(W_S)、石油类(W_{oils})、水温(W_{temp})、电导率(W_{cond})、化学需氧量(COD_{Cr})、氟(W_F)、砷(W_{As})、铅(W_{Pb})、铜(W_{Cu})、锌(W_{Zn})、总氮(TN)这 21 个与总氮可能存在相关关系的变量. 这些变量几乎涵盖了太湖水体所有与富营养化存在较强相关关系的影响因子,能有效保证挑选出来的变量可对总氮的变化趋势进行较为准确地刻画.

本文采用江苏省环境监测中心在太湖监测点 2009-2010 两年的监测数据对这 21 个变量进行相关性分析,得出总氮和其余 20 个变量的相关系数 r 和 p 值如表 1 所示.

表 1　各变量之间的相关关系

Table 1　The relationship between the variables

		pH	SS	SD	DO	COD_{Mn}	BOD_5	W_{NH_4N}	V_{phen}	TP	W_{Chla}
TN	r	−0.333	0.043	−0.062	−0.048	0.325	0.385	0.775	0.117	0.439	−0.005
	p	0.000	0.352	0.183	0.297	0.000	0.000	0.000	0.011	0.000	0.914
		W_S	W_{oils}	W_{temp}	COD_{Cr}	W_{cond}	W_F	W_{As}	W_{Pb}	W_{Cu}	W_{Zn}
TN	r	0.097	0.219	−0.201	0.397	0.360	−0.128	0.087	−0.213	0.006	0.095
	p	0.038	0.000	0.000	0.000	0.000	0.006	0.061	0.000	0.900	0.041

本文选取与总氮相关系数大于 0.3 的变量视为与其有关联,由表 1 可知,与总氮具有相关性的监测变量有 pH 值、高锰酸盐指数、生化需氧量、氨氮、总磷、电导率、化学需氧量,相关系数分别为-0.333、0.325、0.385、0.775、0.439、0.397、0.360.

2.2　基于多元线性回归的总氮分析与预测

本文将总氮作为因变量即分析与预测的目标变量,将各影响因素作为自变量,具体标记情况如表 2 所示. 本文采用多元线性回归分析方法对因变量和自变量间的关系进行回归分析[10].

表 2　各影响因素的具体标记情况

Table 2　All the factors of specific markers

pH	COD_{Mn}	BOD_5	W_{NH_4N}	TP	COD_{Cr}	W_{cond}
x_1	x_2	x_3	x_4	x_5	x_6	x_7

将 2009 年和 2010 年上半年的数据进行多元线性回归分析,得到如下回归分析预测方程:

$$y=-6.070\ 4+0.342\ 7x_1+0.421\ 3x_2-0.044\ 7x_3+2.026\ 9x_4-0.034\ 7x_5-1.126\ 5x_6-0.030\ 6x_7. \quad (2)$$

由回归结果可得与 F 统计量对应的 p 值为 $1.740\ 8\times10^{-23}$,由此判定该回归方程极显著. 回归方程相关系数为 0.866 3,其决定系数大于 0.7,该回归模型是可接受的.

2.3　基于 BP 神经网络的总氮分析与预测

根据影响因素和研究对象的个数,本文将预测模型的输入层设置为 7 个神经元(即影响因素的个数),输出层设置为 1 个神经元(即目标对象的个数). 对于隐含层神经元个数的确定,目前尚无理论依据和有效方法,本文将网络层数设置为 2 层,隐含层节点数设置为 14.

学习方法是神经网络的核心部分,本文采用 BP 神经网络模型,采取的训练函数主要是“traingd”,对应的学习方法是“标准梯度下降法”[11]. 本次数据处理采用 MATLAB2014a 的神经网络工具箱. 核心代码如下:

```
net=newff(p,t,14,{'logsig','purelin'},'traingd');%设置训练参数
net.trainParam.epochs=100000;%设置最大训练次数为 10 万次
net.trainParam.goal=0.0001;%设置最小均方差
net=train(net,p,t);%网络训练
```

3 实验

复相关系数 R 可用来反映模型对样本观测值的拟合程度,均方误差可有效地反映模型的预测准确度,故本文采用相关系数和均方误差来反映模型建立的优劣.

3.1 基于多元线性回归的预测结果

图 2 反映的是多元线性回归模型的预测情况,预测值的相关系数和均方误差分别为 0.866 3 和 1.121 8,这说明利用 pH 值、高锰酸盐指数、生化需氧量、氨氮、总磷、电导率、化学需氧量能较好地预测出总氮的值. 当然,若已知除总磷之外的 6 个影响因素和总氮的值,也可预测出总磷.

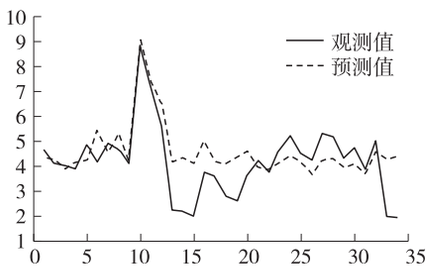

图 2 多元回归预测效果

Fig. 2 Multiple regression prediction effect

3.2 基于 BP 神经网络的预测结果

图 3 和图 4 分别反映的是 BP 人工神经网络模型的训练情况和预测效果,经过 8 629 次训练,训练的相关系数和均方误差分别为 0.950 1 和 0.582 6. 测试的相关系数和均方误差分别为 0.931 4 和 0.702 4. 结合图像可发现,本文训练出来的 BP 神经网络预测模型的性能较好,可用于总氮变化趋势的预测. 相比于多元线性回归模型,神经网络模型的训练和预测效果更好.

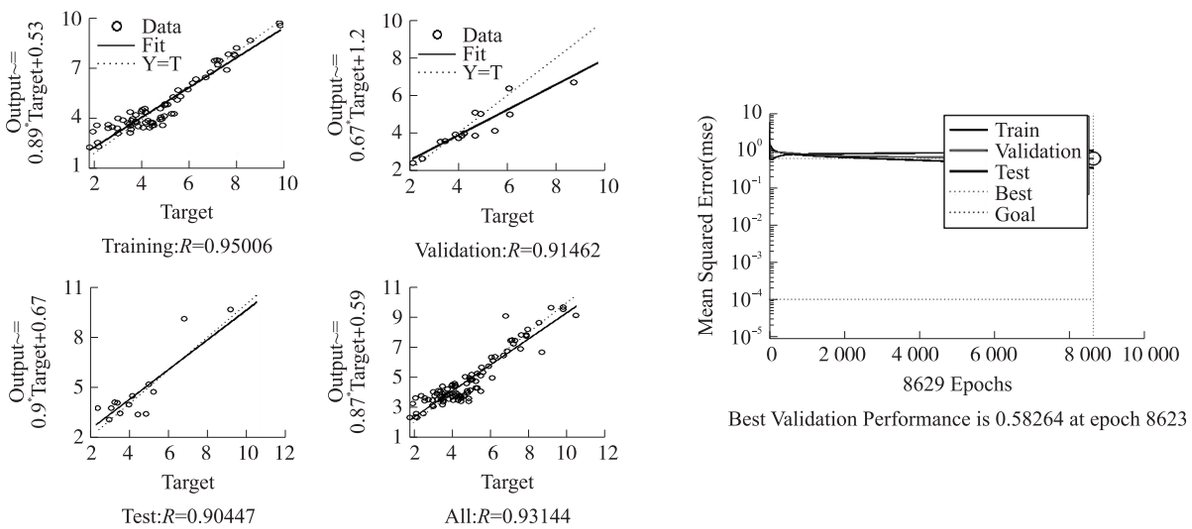

图 3 神经网络训练效果

Fig. 3 Neural network training effect

4 结语

论文在全面考虑影响总氮的相关因素的基础上,通过二元相关性方法来选取与目标对象存在较强相关性的影响因素,运用多元线性回归和 BP 人工神经网络两种不同的方法对富营养化核心指标总氮进行了分析与预测. 研究结果表明:(1)在较为充分的变量上选择出来的影响因素能够较好地预测总氮变化趋势;(2)多元线性回归方法和 BP 神经网络方法在对总氮变化趋势进行分析和预测时,从拟合优度及均方误差角度考虑,BP 神经网络预测模型的预测性能要优于多元线性回归模型.

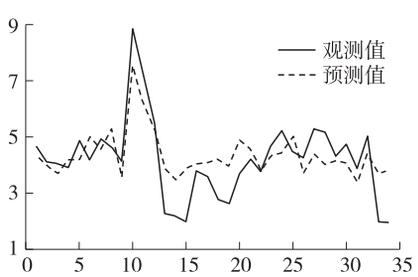

图 4 神经网络预测效果

Fig. 4 Neural network prediction effect

本文存在的问题以及未来的研究方向在于:(1)本文在预测总氮的变化趋势时仅仅考虑了水体中其他相关的影响因子,并未考虑到底泥以及气候变化等其他因素,所呈现出来的模型仍不够精确;(2)多元线性回归模型和两层神经网络模型都是最基础的模型,接下去将会尝试使用非线性多元回归模型和深度学习方法来进行更好的预测.

[参考文献](References)

[1] 陈小锋,揣小明,杨柳燕. 中国典型湖区湖泊富营养化现状、历史演变趋势及成因分析[J]. 生态与农村环境学报,2014,30(4):438-443.
CHEN X F,CHUAI X M,YANG L Y.Status quo,historical evolution and causes of eutrophication in lakes in typical lake regions of China[J]. Journal of ecology and rural environment,2014,30(4):438-443.(in Chinese)

[2] 钱奎梅,陈宇炜,宋晓兰. 太湖浮游植物优势种长期演化与富营养化进程的关系[J]. 生态科学,2008,27(2):65-70.
QIAN K M,CHEN Y W,SONG X L. Long-term development of phytoplankton dominant species related to eutrophicarion in Lake Taihu[J]. Ecological science,2008,27(2):65-70.(in Chinese)

[3] 崔嘉宇,郁建桥,吕学研,等. 太湖富营养化指标 BP 人工神经网络预测模型的建立[J]. 环境研究与监测,2014,27(3):50-54.
CUI J Y,YU J Q,LÜ X Y. Establishment of BP artificial neural network prediction model for eutrophication index of Lake Taihu[J]. Environmental study and monitoring,2014,27(3):50-54.(in Chinese)

[4] 韩菲,陈永灿,刘昭伟. 湖泊及水库富营养化模型研究综述[J]. 水科学进展,2003,14(6):786-790.
HAN F,CHEN Y C,LIU Z W. Advance in the eutrophication models for lakes and reservoirs[J]. Advances in water science,2003,14(6):786-790.(in Chinese)

[5] 胡小贞,耿荣妹,许秋瑾,等. 太湖流域流经不同类型缓冲带入湖河流秋、冬季氮污染特征[J]. 湖泊科学,2016,28(6):1 194-1 203.
HU X Z,GENG R M,XU Q J,et al. Charactistics of nitrogen pollution of rivers flowing through different lake buffer strips in autumn/winter,Taihu Basin[J]. Journal of lake sciences,2016,28(6):1 194-1 203.(in Chinese)

[6] 吴雅丽,许海,杨桂军,等. 太湖水体氮素污染状况研究进展[J]. 湖泊科学,2014,26(1):19-28.
WU Y L,XU H,YANG G J.Progress in nitrogen pollution research in Lake Taihu[J]. Journal of lake sciences,2014,26(1):19-28.(in Chinese)

[7] 何晓群. 多元统计分析[M]. 北京:中国人民大学出版社,2011.
HE X Q. Multivariate statistical analysis[M]. Beijing:China Renmin University Press,2011.(in Chinese)

[8] 何晓群,刘文卿. 应用回归分析[M]. 北京:中国人民大学出版社,2015:57-90.
HE X Q,LIU W Q. Application regression analysis[M]. Beijing:China Renmin University Press,2015:57-90.(in Chinese)

[9] 丛爽. 面向 MATLAB 工具箱的神经网络理论与应用[M]. 合肥:中国科学技术大学出版社,2009.
CONG S. Neural network theory and application with MATLAB toolboxes[M]. Hefei:Press of University of Science and Technology of China,2009.(in Chinese)

[10] 李祚泳,汪嘉杨,金相灿,等. 基于进化算法的湖泊富营养化投影寻踪回归预测模型[J]. 四川大学学报(工程技术版),2007,39(2):1-8.
LI Z Y,WANG J Y,JIN X C,et al. Evolution algorithm based forecasting model for lake eutrophication using PPR[J]. Journal of Sichuan university(engineering and technology edition),2007,39(2):1-8.(in Chinese)

[11] 崔东文. BP 神经网络模型在湖泊富营养化程度评价中的应用[J]. 云南水利水电,2011(1):61-67.
CUI D W.Appliance of BP neural network model in evaluation of lake eutrophication degree[J]. Journal of water conservancy and hydropower in Yunnan,2011(1):61-67.(in Chinese)

[责任编辑:严海琳]