第39卷，第5期
2019年5月

光 谱 学 与 光 谱 分 析
Spectroscopy and Spectral Analysis

Vol. 39, No. 5, pp1503-1508
May, 2019

# 基于 LLE-BPNN 的小麦岛海水硝酸盐含量分析

王雪霁[1,2]，胡炳樑[1*]，于 涛[1]，刘青松[1,2]，李洪波[1,2]，范 尧[1]

1. 中国科学院西安光学精密机械研究所，陕西 西安 710119
2. 中国科学院大学，北京 100049

**摘 要** 水中过量的硝酸盐会造成部分水生生物难以存活、引发人类尤其是婴儿患病等危害，因此硝酸盐浓度成为水质检测中的一项重要指标。传统的硝酸盐浓度测量方法操作复杂、反应缓慢，近年许多研究人员开始通过紫外可见(UV-Vis)光谱技术结合人工神经网络(ANN)的方法对水中硝酸盐的含量进行测量。提出了一种将流形学习(manifold learning)方法中的局部线性嵌入(LLE)与反向传播神经网络(BPNN)相结合的建模方法，用以得到硝酸盐光谱曲线与浓度间的关系，实现对青岛市崂山区小麦岛海水中硝酸盐浓度快速准确的定量分析。实验选取了过滤后的小麦岛海水配置59组不同浓度的加标溶液，采用实验室自主研制的光谱分析仪采集这些样本的光谱测量值，通过标准正态变换(SNV)方法对测得硝酸盐溶液的光谱数据进行校正处理，有效降低了由仪器本身或环境带来的噪声影响；选取预处理后的光谱数据的前1 500维处理后进行对比实验，以解决使用BPNN对全部2 048维数据建模时内存不足的问题，再通过网格搜索结合十折交叉验证的方法优化LLE中的邻近点数 $k$ 和嵌入维数 $d$，得到最优参数值 $k=15$，$d=3$，实现对实验数据的降维处理；通过BPNN将降维后的训练集光谱信息与其对应的浓度信息进行建模，实现对预测集硝酸盐浓度定量分析，引入决定系数($R^2$)和预测均方根误差(RMSEP)评价建模效果，与直接使用BPNN建模预测的结果比较，改进方法的 $R^2$ 由0.926 3提升至0.992 8，RMSEP由0.442 5下降到0.280 4，建模预测程序的运行时间由327 s缩短至0.5 s。采用这59组数据的全部2 048维进行LLE-BPNN建模时，得到 $R^2=0.995\ 7$，RMSEP $=0.136\ 5$，在用时相近的前提下，相比仅使用前1 500维时的建模精度更好。分析结果表明，LLE-BPNN的方法可实现对海水中硝酸盐浓度的快速预测，使预测精度得到显著提升，同时能大幅降低预测时间。

**关键词** 硝酸盐浓度；紫外可见光谱技术；局部线性嵌入；反向传播神经网络

**中图分类号**：X55 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2019)05-1503-06

## 引 言

水中过量的硝酸盐会使水生昆虫或鱼类难以存活。原因是藻类和其他植物使用硝酸盐作为肥料来源，水中过量的硝酸盐会使藻类的生长不受限制，而大量的藻类会引起水中溶解氧的剧烈波动[1]。当藻类死亡和分解时，高含量的有机物质和分解的生物体会耗尽水中可用的溶解氧，从而导致其他生物如鱼类和贝类的死亡。就像溶解氧、温度和pH一样，水中硝酸盐的含量是由自然过程和人为干预决定的。由于人类不得当的举动，如农业活动、人类废弃物和工业污染排放，都可能会导致水体中硝酸盐含量的升高。

近年，紫外可见(ultraviolet/ visible，UV-Vis)光谱技术依赖回归模型及化学法标准，被广泛的用于海水中硝酸盐含量及其他参数的测量，因为测量程序简单，且可以在不产生二次污染的基础上做到快速检测[2]。故本工作采用紫外可见光谱技术对从青岛市崂山区小麦岛采集到的加标海水进行测量分析，计算并预测相关样品中硝酸盐的浓度。

研究中使用反向传播神经网络(back propagation neural network，BPNN)对样本进行训练学习分析。BPNN是人工神经网络(artificial neural network，ANN)中一种被广泛应用于解决多分类及预测问题的多层前馈神经网络。ANN是一种数据驱动、自适应强、计算灵活的工具，可以高精度的捕捉到任何物理过程中的非线性的、复杂的潜在特征[3]。近些年，许多研究人员将其使用在水质评价上[4-6]。但基于ANN的方法收敛速度慢，处理数据量大时需要占用许多内存且十

**收稿日期**：2018-03-26，**修订日期**：2018-08-02

**基金项目**：国家重点研发计划项目(2017YFC1403700)资助

**作者简介**：王雪霁，1992年生，中国科学院西安光学精密机械研究所博士研究生 e-mail：wangxueji@opt.cn

*通讯联系人 e-mail：hbl@opt.ac.cn

分耗时，甚至不能在普通的计算机上完成计算。因此，本工作使用流形学习(manifold learning)方法中的局部线性嵌入(locally linear embedding，LLE)对将要进行学习的数据进行了降维处理，不但减少了程序计算运行所需的时间和内存，并使得预测精度产生了大幅度的提高。

## 1 原 理

### 1.1 局部线性嵌入算法

局部线性嵌入(locally linear embedding，LLE)是一种非线性降维的方法，它运用局部线性关系的组合表示全局非线性结构，从而达到对数据降维的目的[7]。该方法可以在低维空间内重构数据点并保持数据之间的邻近关系，通过将高维冗余的非线性数据转化为保持原始结构的低维线性数据，有效地解决了由于数据结构发生改变而导致的模型精度下降问题。

通常，LLE 可以总结为三步[8]：

首先，为高维数据集 $X=[x_1, x_2, \cdots, x_N]\in R^{D\times N}$ 中的每个向量 $x_i$ 根据两两之间欧氏距离找到 $k$ 个近邻点；

然后，利用每个数据点的近邻点计算其线性系数(即局部流形结构)，可表示为 $w_i=[w_i^1, \cdots, w_i^j, \cdots, w_i^k]\in R^{1\times k}$。其映射函数为式(1)

$$w_i = \arg\min_{w_i} \| x_i - \sum_{j=1}^{k} w_i^j x_j \|^2, \quad \text{s.t.} \quad \sum_{j=1}^{k} w_i^j = 1 \tag{1}$$

最后，重建线性系数($w$)以获得在低维空间的嵌入，见式(2)。

$$\min\varepsilon(Y) = \sum_{i=1}^{N} \| y_i - \sum_{j=1}^{k} w_i^j y_j \|^2 \quad \text{s.t.} \quad \sum_{i=1}^{N} y_j = 0,\ \frac{1}{N}\sum_{i=1}^{N} y_i y_i^{\mathrm{T}} = I \tag{2}$$

最终得到低维数据集 $Y=\{y_1, y_2, \cdots, y_n\}$，$y_i\in R^m$。

LLE 算法的参数优化处理分为两部分：邻近点数 $k$ 和嵌入维数 $d$，通常要求 $d$ 小于 $k$。其中 $k$ 是 *LLE* 算法中最重要的参数，$k$ 过小或者过大，都将破坏输出向量的原始结构。

### 1.2 反向传播神经网络

反向传播神经网络 BPNN 是一种受生物启发的计算方法，用于对大规模非线性系统的建模[9]。该方法基于梯度下降(gradient descent)策略，以目标的负梯度方向对参数进行调整[10]。BPNN 具有自学习的特点，于是该方法有较强的适应性和鲁棒性[11]，被广泛使用在许多领域[12-13]，如函数逼近、模式识别、图像处理、预测及其他领域。因实验用的光谱仪可以实现对水质的多参数测量，因此选用了基于神经网络的多组分方法，以验证在使用多组分方法时对单一成分浓度预测的精度。

通常模型可分为三层：输入层、隐藏层和输出层，有一个隐藏层的 BPNN 是最常用的结构，可以表示为图 1。

隐藏层输出 $H$

$$H_j = f(\sum_{i=1}^{n} w_{ij} x_j - a_j) \quad j = 1, 2, \cdots, l \tag{3}$$

式(3)中 $f$ 是隐藏层的激励函数。

整个网络的预测输出[见式(4)]

$$O_k = \sum_{j=1}^{l} H_j w_{jk} - b_k \quad k = 1, 2, \cdots, m \tag{4}$$

则网络的预测误差为式(5)

$$e_k = Y_k - O_k \quad k = 1, 2, \cdots, m \tag{5}$$

具体运算时，BPNN 算法先将输入的训练样本提供给输入层神经元，逐层将信号向前传递，直至计算出输出层的结果；然后计算输出层的误差，将该误差反向传递给隐藏层的神经元；最后由隐藏层神经元的误差再对权值和阈值进行相应调节。循环该迭代过程，直到达到终止条件为止。

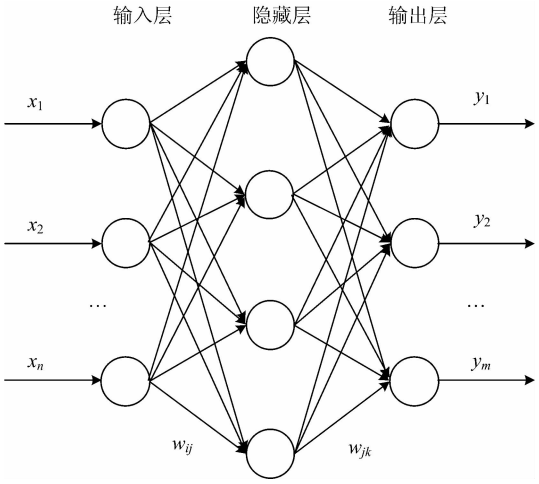


**图 1 BPNN 结构示意图**

**Fig. 1 Structure chart of BPNN**

### 1.3 模型评价

选择决定系数(coefficient of determination，$R^2$)和预测均方根误差(root mean square error of prediction，RMSEP)作为模型的评价指标，二者呈正相关。决定系数 $R^2$ 表现了模型的稳定性，而 RMSEP 值可用来表征所建预测模型的预测性能，其值越小表示该模型的预测精度越高，其计算公式为式(6)。其中 $\bar{c}_i$ 表示包括训练集和预测集所有样本 $c_i$ 的均值，$c_i$ 是实际测量值，$\hat{c}_i$ 是通过模型预测的结果，$m$ 是用于检验模型的预测集样本数。

$$\mathrm{RMSEP} = \sqrt{\frac{\sum_{i=1}^{m}(\hat{c}_i - c_i)^2}{m}} \tag{6}$$

## 2 实验与讨论

### 2.1 实验流程

首先配制不同浓度的溶液、采集其光谱数据，对数据进行简单的预处理，包括去除暗背景、扣除海水基底等，以降低或者消除外界干扰对光谱产生的影响；再采用 LLE 算法对光谱数据进行降维处理，然后通过 BPNN 建立训练集样本浓度与光谱值关系的模型，采用测试集样本进行浓度预测，评价预测结果，并与不降维而直接使用 BPNN 建模预测的结果进行比较。

### 2.2　实验仪器及样本采集

实验中所使用的光谱仪为课题组自主研发的光谱分析仪样机，采用了双光路主动校正和连续谱精细获取技术，同时，采用特征点邻域多波长位置实现大量程适应性调节，避免了单波长无法实现多源误差综合校正补偿的缺陷，并改善了传统单光路(无校准光路)在连续长时间业务应用时，由光源的不稳定性引起的测量误差。其探测器范围为：180～1 100 nm，光谱采样间隔约为 0.45 nm，因此测量将会产生 2 048个波长点，原理图如图 2。另外，实验中采用了 Ocean Optics 公司生产的微型氘-钨卤 UV-Vis-NIR 光源（型号：DT-MINI-2-GS)。

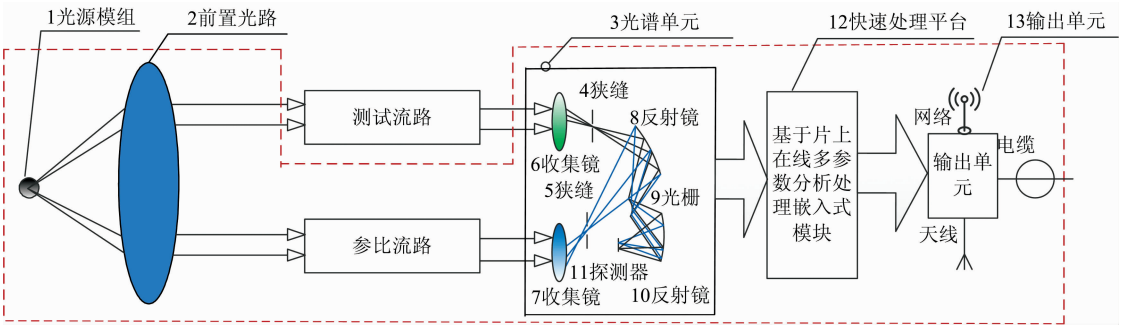


**图 2　双光路主动校正连续谱精细获取光谱仪原理图**

**Fig. 2　Schematic diagram of the dual optical path active correction continuous spectral fine acquisition spectrometer**

采用硝酸钾(Sigma-Aldrich，Co.)、过滤后的小麦岛附近较纯净的海水(即假设其不含硝酸盐)配制 59 个不同浓度的样本，浓度范围为 0.01～5 mg·$L^{-1}$。在室温(23±1) ℃条件下对配制水样进行透射光谱测量，实验装置由光谱仪、光源、笔记本电脑组成，采集时将配置好的溶液倒入比色皿中，将比色皿放入两端分别连接光谱仪和光源的专用固定底座内。对实验数据扣除暗背景影响并采用与过滤后的海水做基线校正，将每个样本扫描 10 次，取其平均值作为该浓度样本最终的测量值。经处理后的光谱图如图 3。

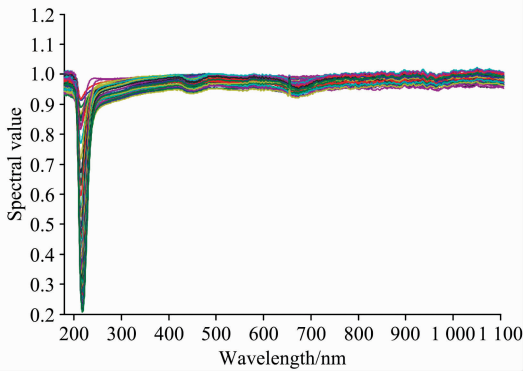


**图 3　59 个样本的光谱图**

**Fig. 3　Spectra of 59 samples**

从 59 个样本中等间隔选取 10 个作为测试集样本，其余设置为训练集样本，测试集的选取及编号如表 1 所示。

### 2.3　数据预处理

由于仪器本身和环境的影响，可能会使获得的光谱产生基线漂移，合适的光谱预处理可以有效降低或者消除这些外界干扰对光谱产生的影响，从而提高模型的预测精度和稳健性。这里采用标准正态变换(standard normal variate，SNV)算法对光谱测量值进行处理，其变换公式为式(7)

$$x_{ij}^{*}=\frac{x_{ij}-\bar{x}_i}{s_i} \tag{7}$$

其中，$i=1, 2, \cdots, m$，$m$ 为变量数(即波长点数)；$j=1, 2, \cdots, n$，$n$ 为样本数；$\bar{x}_i$ 为样本在第 $i$ 个变量处光谱值的平均值，$s_i$ 表示样本在第 $i$ 个变量处光谱值的标准差。

**表 1　测试集样本浓度表**

**Table 1　Test set sample concentration table**

| serial number | Concentration /(mg·$L^{-1}$) | serial number | Concentration /(mg·$L^{-1}$) |
|---|---|---|---|
| 1 | 0.03 | 6 | 2.4 |
| 2 | 0.09 | 7 | 3.0 |
| 3 | 0.6 | 8 | 3.6 |
| 4 | 1.2 | 9 | 4.2 |
| 5 | 1.8 | 10 | 4.8 |

### 2.4　光谱数据降维

采用 LLE 的方法对预处理后的硝酸盐样本数据降维。由于电脑内存限制，在直接使用 BP 网络建模时，选取了样本数据的前 1 500 个波长点(共 2 048 个)，其对应波长范围约为 180～870 nm。为了与所提出的 LLE-BPNN 模型进行对比，在使用 LLE 降维时将只选用这 1500 个波长点，而后将 LLE 算法降维后的训练集数据作为 BPNN 的输入。因此，在进行 LLE 算法的参数优化时，所选取的是 59 个样本数据的前 1 500 个波长点进行降维。

首先用网格搜索法结合十折交叉验证法优化 LLE 的两个参数 $k$ 和 $d$。网格搜索法的优点是可以指定模型空间的范围，并且平等的考虑所有可能的解决方案[14]。为了获得更精确的参数值，参数优化的范围需要足够大同时步长需要足够小，选定邻近点数 $k$ 和嵌入维数 $d$ 的搜索范围分别为[5，20]，[1，20]，步长均为 1。将不同参数下 LLE 降维得到的 59 个样本数据当作 BPNN 的输入，并进行十折交叉验证，其

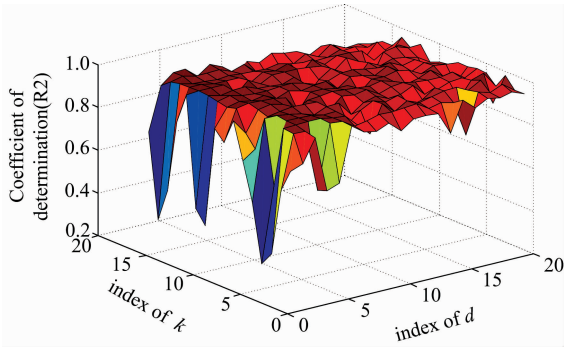


**图 4 不同 $k$ 值和 $d$ 值时的 LLE-BPNN 预测的决定系数**

**Fig. 4 $R^2$ of LLE-BPNN prediction with different $k$ and $d$**

预测结果的决定系数如图 4。

其中，当 $k=15$，$d=3$ 时有最大的决定系数 $R^2=0.996\ 7$，且满足 $k>d$ 的条件。故将这两个参数邻近点数 $k=15$、嵌入维数 $d=3$ 代入 LLE 算法中，对硝酸盐样本数据进行降维。

### 2.5 建模预测

选取具有 12 个隐藏层节点的 BP 神经网络，并设定最大训练次数 1 000、误差容限 $1\times10^{-5}$。最终得到直接使用 BPNN 对这 10 个样本的预测效果如图 5(a)，相对误差(relative error，RE)如图 5(b)；而通过 LLE-BPNN 计算的 10 个预测集样本的预测效果如图 5(c)，RE 如图 5(d)。由图 5 可看出使用改进方法后，对硝酸盐浓度的预测精度显著提高。

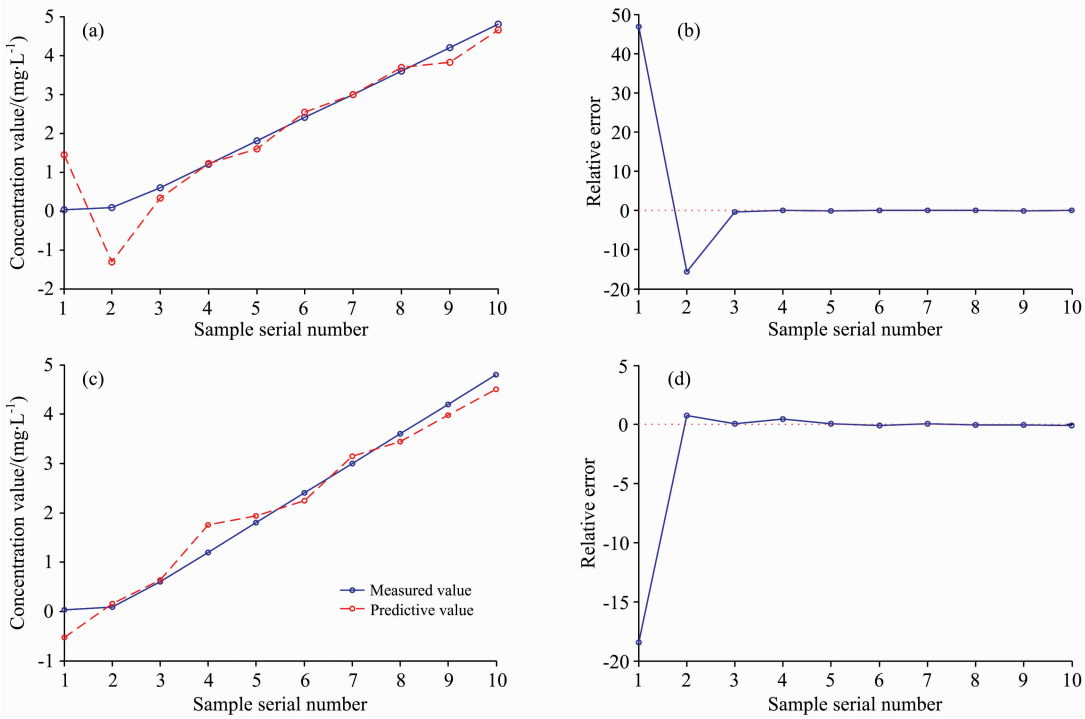


**图 5 (a) BPNN 预测结果对比；(b)BPNN 预测结果相对误差；**
**(c) LLE-BPNN 预测结果对比；(d) LLE-BPNN 预测结果相对误差**

**Fig. 5 (a) BPNN prediction results comparison chart; (b) RE of BPNN prediction results;**
**(c) LLE-BPNN prediction results comparison chart; (d) RE of LLE-BPNN prediction results**

用这两种方法建模预测这十个测试样本后各评价指标如表 2 所示，由表可以看出，经 LLE 降维后的模型对 10 个测试样本预测效果较好。在使用 LLE 降维后，$R^2$ 增大且 RMSEP 减小，说明建模的精度和稳定性相比仅使用 BP 网络建模得到了显著的提升。比较两种方法计算的用时，使用改进算法后用时仅为原来的 1/650，提高了算法的速度，从而能更好地满足实时监测的业务需求。以上分析表明采用 LLE 降维能够很好地保持数据的特征结构，在去除大量冗余信息和噪声信息的同时，优化了数据间的内在特征，从而使得预测精度提高且预测时间减少。

**表 2 两种方法预测硝酸盐浓度的评价**

**Table 2 Evaluation of nitrate concentration prediction by the two methods**

| | $R^2$ | RMSEP | Elapsed time/s |
|---|---|---|---|
| BP | 0.926 3 | 0.442 5 | 327 |
| LLE-BP | 0.992 8 | 0.280 4 | <0.5 |

实验验证了使用全部 2 048 个波长点进行 LLE-BP 建模并对比 10 个预测集样本预测的效果，仍使用 2.4 中参数优化后的得到的最优参数，即 $k=15$，$d=3$，其预测结果如图 6(a)，相对误差如图 6(b)。

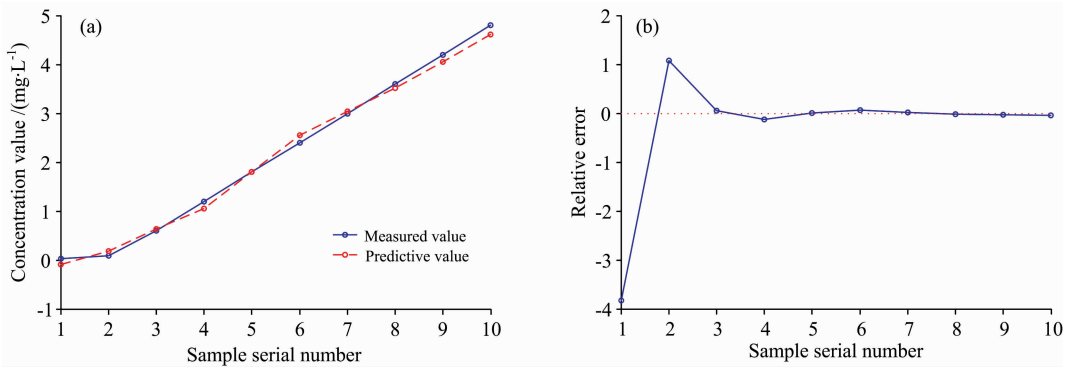


**图 6　(a) LLE-BPNN 预测结果对比(2 048 个波段)；(b) LLE-BPNN 预测结果相对误差(2 048 个波段)**

**Fig. 6　(a) LLE-BPNN prediction results comparison chart (2 048 bands);**
**(b) relative error chart of LLE-BPNN prediction results (2 048 bands)**

计算得到此时的 $R^2=0.995\ 7$，RMSEP=0.136 5，相比仅使用其中 1 500 个波长点建模，其预测精度又有所提升。证明了本研究所提出的 LLE-BPNN 模型既能节约计算运行的时间，同时使预测精度得到大幅度提升，且对计算运行所需的内存要求较低，适用于对硝酸盐浓度实时精确计算。

## 3　结　论

通过所提出的局部线性嵌入结合反向传播神经网络(LLE-BPNN)的方法，实现了对小麦岛海水中的硝酸盐浓度进行快速定量分析。通过 UV/Vis 光谱技术得到硝酸盐溶液的光谱数据，使用 SNV 对其进行预处理后，LLE 对光谱数据进行降维，再建立基于 BPNN 的硝酸盐浓度预测模型。该方法的预测结果为 $R^2=0.992\ 8$，RMSEP=0.280 4，与直接使用 BPNN 建模预测的结果相比，$R^2$ 增大且 RMSEP 减小，且建模预测程序的运行时间由 327 s 缩短到 0.5 s 左右。以上分析表明：采用 LLE-BPNN 的方法，可以实现对海水中硝酸盐快速精确的定量分析。同时也从对这 10 个预测样本的预测中发现对于小浓度样本，用该优化的神经网络建模预测的效果较差，其相对误差较大，在之后的工作中，将会重点对这一问题进行分析研究。

## References

[1] ZHANG Yi-wen, LUO Jian-zhong, CHEN Yu-yang(张懿文，罗建中，陈宇阳). Guangdong Chemical Industry(广东化工), 2015, 42(14): 99.
[2] Hu Yingtian, Wen Yizhang, Wang Xiaoping. Sensors and Actuators, B: Chemical, 2016, 227: 393.
[3] Hao Z, Huang X, Zhu C. Proc. Int. Conf. Nat. Comput., 2008, (3): 394.
[4] Siddiquee M, Hossain M. Neural Computing and Applications, 2015, 26(8): 1979.
[5] Meilin W, Youshao W, Jidong G. Ecotoxicology, 2015, 24(8): 1632.
[6] Ostadaliakari K, Shayannejed M, Chorbanizadehkharazi H. Ksee Journal of Civil Engineering, 2016, 21(1): 1.
[7] BAO Cui-mei, HAN Xiao-chun, YI Hui, et al(薄翠梅，韩晓春，易　辉，等). CIESC Journal(化工学报), 2016, 67(3): 925.
[8] Chen Y, An S, Dong J. et al. Optical and Quantum Electronics, 2016, 48(11): 488.
[9] Yang T M, Fan S K, Fan C, et al. Environmental Monitoring and Assessment, 2014, 186(8): 4925.
[10] ZHOU Zhi-hua(周志华). Machine Learning(机器学习), 2016.
[11] Fan Dayong, Yang Jiachen, Zhang Junbao, et al. IEEE Journal of Translational Engineering in Health & Medicine, 2017, PP(99): 1.
[12] HOU Ya-li, LI Tie(侯亚丽，李　铁). Journal of Detection & Control(探测与控制学报), 2008, 30(1): 53.
[13] Yen C T, Huang Y J. Multimedia Tools and Applications, 2015, 75(16): 1.
[14] Jiang M, Spikes K T. Geophysical Journal International, 2013, 195(1): 315.

# Analysis of Nitrate in Seawater of Wheat Island Based on LLE-BPNN

WANG Xue-ji[1, 2], HU Bing-liang[1*], YU Tao[1], LIU Qing-song[1, 2], LI Hong-bo[1, 2], FAN Yao[1]

1. Laboratory of Spectral Imaging Technique, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, China
2. University of Chinese Academy of Sciences, Beijing 100049, China

**Abstract** Excessive nitrate in water may influence some aquatic organisms' survival and cause harm to humans, especially infants. Therefore, nitrate concentration becomes an important indicator in water quality monitoring. Due to the complexity of operation and slow response of conventional methods for measuring nitrate concentration, many researchers have begun to use ultraviolet/visible (UV-Vis) spectroscopy combined with artificial neural network (ANN) methods to measure nitrate content in water. This paper proposes a modeling method combining locally linear embedding (LLE) in manifold learning with back propagation neural network (BPNN). The relationship between the spectral curve of nitrate and the concentration was obtained, so that a rapid and accurate quantitative analysis of the nitrate concentration in the wheat island of Laoshan District, Qingdao was achieved. In the experiment, we selected 59 groups of spiked solutions with different concentrations of filtered wheat island seawater, and collected spectral measurements of these samples using a laboratory-developed spectrum analyzer, with standard normal variate (SNV) method calibrating spectral data of measured nitrate solution to reduce the noise caused by the instrument itself or the environment. First 1 500-dimensional of the pre-processed spectral data was used to avoid insufficient memory when using the entire 2 048-dimensional data to build BPNN model, and a control experiment was performed. Then the number of neighboring points $k$ and the embedding dimension $d$ in the LLE were optimized by the grid search combined with the ten-fold cross validation method, obtaining the optimal $k=15$, $d=3$. Then the dimension of the experimental data was reduced. The spectral information of the reduced-dimensional training set and its corresponding concentration information were modeled by the BPNN to achieve a quantitative analysis of the nitrate concentration in the prediction set. Coefficient of determination ($R^2$) and root mean square error of prediction (RMSEP) were introduced to evaluate modeling effects. And compared with the predicted results obtained by only using BPNN modeling, $R^2$ of our improved method increased from 0.926 3 to 0.992 8, and RMSEP decreased from 0.442 5 to 0.280 4, and prediction modeling program run time decreased from 327 s to about 0.5 s. In addition, we used all 2 048 dimensions of the 59 data sets for LLE-BPNN modeling, with $R^2=0.995\ 7$ and RMSEP=0.136 5, which was improved compared to the modeling accuracy when only using the first 1 500 dimensions, while elapsed time was similar. The analysis results above showed that using the LLE-BPNN method can achieve a rapid prediction of nitrate concentration in seawater, while significantly improving prediction accuracy and reducing prediction time.

**Keywords** Nitrate concentration; Ultraviolet/visible spectral technology; Locally linear embedding; Back propagation neural network

(Received Mar. 26, 2018; accepted Aug. 2, 2018)

* Corresponding author