

# 投影寻踪和BP网络在农村饮用水水质评价中的应用

高颖, 方国华, 林泽昕

(河海大学 水利水电学院, 江苏 南京 210098)

**摘要:** 日渐严重的农村饮用水水源污染问题已经影响到人们的日常生活,对水质变化情况的及时了解并针对其综合状况进行合理评价十分必要。根据农村饮用水水源地特点,选取了粪大肠杆菌、氨氮、总磷、高锰酸盐指数、溶解氧5个主要污染物作为评价指标,建立投影寻踪等级评价模型,计算水质标准等级与水质样本投影值,通过比较投影值得出样本等级。然后用BP神经网络对样本及投影值进行学习并预测,验证计算结果的合理性。以安徽省定远县的11个地表水源地丰水期和枯水期水质进行实际评价。计算结果表明:与常用评价方法相比,模型准确有效,可为水质综合评价提供新方法。

**关键词:** 投影寻踪; 遗传算法; BP神经网络; 水质评价; 农村饮用水

**中图分类号:** TP183; X824

**文献标识码:** A

**文章编号:** 1672-643X(2017)02-0115-05

## Application of projection pursuit and BP neural network in assessment of drinking water quality in rural areas

GAO Ying, FANG Guohua, LIN Zexin

(College of Water Conservancy and Hydropower, Hohai University, Nanjing 210098, China)

**Abstract:** The increasingly drinking water pollution in rural areas have affected people's life. To capture the variation timely and to evaluate the water quality conditions based on monitoring data are of great importance. Based on the rural drinking water source characteristics, 5 major pollutants were selected as the evaluation indexes which included fecal coliform, ammonia nitrogen, total phosphorus, permanganate index and dissolved oxygen, respectively. Projection pursuit evaluation model was established to evaluate water quality standard level and pursuit values of water samples. Then the two projection values would be matched to evaluate the levels of samples. BP neural network was used to study and predict the calculation results and test the rationality 11 surface water quality samples of Dingyuan County in Anhui Province during the wet and the dry seasons were collected for assessment. The results showed that, compared with common assessment methods, this model was of great accuracy and validity, which provided a new way for the comprehensive evaluation of water quality.

**Key words:** projection pursuit; genetic algorithm; BP neural network; water quality assessment; drinking water in rural areas

饮用水水质状况直接关系到国计民生,水资源匮乏和水源地污染引起的社会问题尤其突出,水质综合评价作为表征地表水环境质量的总体状况的重要手段,是分析水体承载力和控制水污染工作的前提<sup>[1]</sup>。目前我国用于水质评价的方法主要分为两大类,一类是传统方法,统称为指数评价法,具体包括单因子污染指数法,综合污染指数法<sup>[2]</sup>;第二类水质评价新方法主要有模糊综合评价法,主观赋权

法,熵权理想点法,人工神经网络等<sup>[3-5]</sup>。上述方法中,第一类指数评价法计算简单,应用最为广泛,但结果往往不能反映水体的综合状况<sup>[6]</sup>;第二类方法中,模糊综合评判<sup>[7]</sup>没有精确、统一的评判标准作为支撑,可能出现偏差;主观赋权法会受限于所积累的知识、经验而产生人为偏差;熵权理想点法<sup>[8]</sup>根据评价指标间的距离测度确定各指标的权重,虽然一定程度上消除了人为赋权的不足,但实用中可能

收稿日期:2017-01-12; 修回日期:2017-02-18

基金项目:江苏省水利科技项目(2015056)

作者简介:高颖(1992-),女,安徽铜陵人,硕士研究生,研究方向:水利水电规划与水利经济。

出现各指标权重的平均化现象。人工神经网络<sup>[9-10]</sup>虽结构简单,但缺乏严密的数学分析,不具备普遍适用性。基于此,本文将神经网络与投影寻踪相结合,先求出水质各级标准与所属等级之间的函数关系,然后将待评价样本的投影函数值与水质等级投影值相比判定水质的所属等级。评价模型用投影寻踪客观确定权重,消除人为因素;用 BP 网络将投影寻踪计算结果进行仿真预测,对评价模型结果进行验证,以实现主观与客观评价法相结合的农村饮用水水质综合评价。

1 BP-PP 水质综合评价模型

投影寻踪可以将多维指标样本降维,综合成单个投影数值,每个独立样本的投影值分散相对均匀,最大程度保持其空间相似性,易于决策。此外,投影寻踪无需另外确定权重,在环境影响因子的污染作用分析方面有很好的适应性。投影寻踪等级评价模型可以通过投影值关联样本水质和标准水质间的对应关系,采用遗传算法求解投影指标函数可避免陷入局部极值点,快速找到最佳投影方向。人工神经网络的种类很多,常见的有感知器网络、BP 网络、RBF 网络、Hopfield 网络等<sup>[11]</sup>。其中,BP 网络因具有反向传播(Back Propagation)的学习算法而不同于其他算法,BP 网络的 3 层结构通过误差的逆传播不断调整网络各层权重及阈值,在解决实际问题中的应用相当广泛,具体结构如图 1 所示。对新建立的 BP 网络进行训练时,任意样本进入网络后,神经元立即进入激活状态,所有输入信息通过传递函数正向传播,经过各中间层到达输出层,最后在输出层获得响应。为了缩小初始值和期望值间的距离,每

次计算产生的误差值  $e$  从输出层反向逐层传播,最后回到输入层。当误差值  $e$  的在逐次优化循环中不断减小时,网络对输入模式响应的报错频率也不断下降<sup>[12]</sup>。含一个神经元的 BP 网络可实现对任意非线性函数的逼近,这种非线性映射能力适用于对水质样本的分类,将样本矩阵与投影值之间的规律解析化,以验证投影寻踪等级评价模型的准确性。

1.1 农村饮用水水质指标选择与分级标准

农村饮用水水源地污染源中占主要贡献的是农业面源污染,例如畜禽养殖粪便、农村生活污水、工农业生产导致的面源污染<sup>[13]</sup>。不经处理的有毒有害物质大面积污染土体和水体,造成细菌、微生物和硝酸盐氮类污染物超标。因而选择粪大肠菌群落数作为典型指标之一用于表征农村饮用水水质。常规指标选取了  $X-MnO_4$ 、 $NH_4-N$ 、DO 和 TP。

各指标对水质的影响程度不同,根据评价指标的实测单位浓度和各级标准浓度计算评价单元的水质状况。表 1 给出了 5 个污染物不同浓度所对应的水质等级。

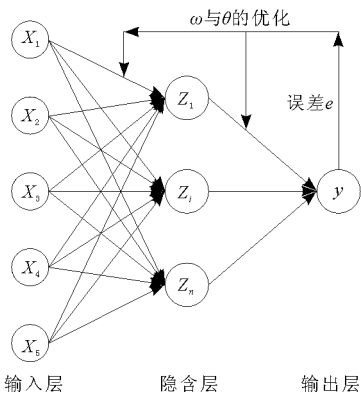


图 1 BP 神经网络优化示意图

表 1 地表水环境质量标准 (GB3838-2002)

| 水质等级 | 物质/(mg · L <sup>-1</sup> ) |            |         |              | 粪大肠菌群<br>个数   |
|------|----------------------------|------------|---------|--------------|---------------|
|      | 高锰酸盐指数                     | 氨氮         | 溶解氧     | 总磷           |               |
| I    | (0,2]                      | (0,0.15]   | [7.5,∞) | (0,0.01]     | (0,200]       |
| II   | (2,4]                      | (0.15,0.5] | [6,7.5) | (0.01,0.025] | (200,2000]    |
| III  | (4,6]                      | (0.5,1.0]  | [5,6)   | (0.025,0.05] | (2000,10000]  |
| IV   | (6,10]                     | (1.0,1.5]  | [3,5)   | (0.05,0.1]   | (10000,20000] |
| V    | (10,15]                    | (1.5,2.0]  | [2,3)   | (0.1,0.2]    | (20000,40000] |

1.2 评价指标集的归一化处理

令样本集表达式为  $\{x(i,j) | i = 1,2,\cdots,n;j = 1,2,\cdots,p\}$ ,其中  $x(i,j)$  为第  $i$  个样本的第  $j$  个指标,  $n,p$  分别为样本数和指标数。对样本数集进行归一化处理,统一量纲和指标数值变化幅度,具体过程如

下:  
对于越小越优的指标:  
$$x(i,j) = \frac{x_{\max}(j) - x(i,j)}{x_{\max}(j) - x_{\min}(j)} \tag{1}$$
  
越大越优的指标:

$$x(i,j) = \frac{x(i,j) - x_{\min}(j)}{x_{\max}(j) - x_{\min}(j)} \tag{2}$$

式中: $x_{\max}(j)$ 、 $x_{\min}(j)$  为第  $j$  个序列中的最大值和最小值; $x(i,j)$  为归一化之后的无量纲实矩阵<sup>[14]</sup>。

1.3 构造投影指标函数

将  $p$  维数据  $\{x(i,j) \mid j = 1,2,\cdots,p\}$  以方向  $a = \{a(1),a(2),a(3),\cdots,a(p)\}$  投影,经向量计算得投影值序列  $z(i)$ :

$$z(i) = \sum_{j=1}^p a(j)x(i,j) \quad i = 1,2,\cdots,n \tag{3}$$

$$\text{s. t.} \quad \sum_{j=1}^p a^2(j) = 1 \tag{4}$$

式中: $a$  为单位投影向量,根据所得一维投影值,画出  $z(i)$  的分布图便于分析。

投影指标函数  $Q(a)$  的数值大小随参数  $a$  而改变,不同的投影方向反映不同的数据结构特征,最佳投影方向就是最大可能暴露多维数据某类特征结构的投影方向,因此可以通过求解投影指标函数最小值问题来估计最佳投影方向<sup>[15]</sup>:

$$\min Q(a) = -S_z \cdot D_z \tag{5}$$

$$S(z) = \sqrt{\sum_{i=1}^n (z(i) - E(i))^2 / (n - 1)} \tag{6}$$

$$D_z = \sum_{i=1}^n \sum_{j=1}^p (R - r(i,j) \cdot u(R - r(i,j))) \tag{7}$$

式中: $S_z$  为投影值序列  $z(i)$  的标准差; $D_z$  为投影值的分布密度; $R$  为给定  $D_z$  的分布范围内窗口半径,通常根据经验确定,本文选定指标数  $p = 5$ ,故取  $R = 4.24$ ;  $u$  为单位布阶函数,用来记录位于窗口范围内的投影值点,当自变量非负时,其值为 1,当自变量为负时,其值为 0<sup>[16]</sup>。

1.4 优化投影指标函数

$Q(a)$  的求解寻优以  $a$  为优化变量,属于复杂非线性求解过程。采用模拟自然界生物优胜劣汰与自然遗传机制的随机搜索算法,利用实数编码技术作用于染色体的交叉变异,全局搜索其最优解<sup>[17]</sup>。适应度函数即为投影指标函数,非线性约束条件为  $a$  是单位向量。

1.5 构建 BP 网络

(1) 为完成图 1 所示的仿真预测,建立 BP 网络,确定网络层数与各层节点数。根据所构造的投影指标函数选定网络为中间层为 1 层的单隐层网络结构,输入层节点数为输入向量的长度  $n$ ,输出层节点数为投影值,即为 1<sup>[18]</sup>。

(2) 以  $x_{ij}$  的输入为开始,进行正向传播计,输出层的期望值为遗传投影寻踪计算结果投影值序列  $z(i)$ 。给定允许误差  $\varepsilon$ ,每次循环计算误差: $e = \frac{1}{2}(y - \hat{y})^2$ ,当满足  $e < \varepsilon$  时输出各层权重及阈值。

(3) 节点传递函数采用正切 S 型函数  $\text{tansig}$ ,训练函数采用动量反传的梯度下降函数  $\text{traingdm}$ 。网络先通过对多个样本的训练学习,在各层权重、阈值稳定之后对样本进行仿真测试。

2 实例应用

2.1 定远县饮用水水源地基本情况

定远县位于安徽省东部,淮河以南,是滁州市土地面积最大、人口最多的下辖县,素以农业大县著称,是全国商品粮油、岩盐资源、石膏资源大县,也是亚洲最大盐化工基地。定远县地表水源较为丰富,有池河、窖河两大水系,除少数河流、湖泊水源污染较为严重外,大多数地表水资源经处理后可以满足饮用水的供给。定远县地下水源含量相对贫乏,大部分地下水源均能满足饮用水水质标准,目前有 3 个正常运行的地下水源地,水质良好。作为地表水水源的水库,除了有供水功能外,均具有灌溉、防洪、养殖功能,总库容之和  $2.83 \times 10^8 \text{ m}^3$ ,汇水面积合计  $599.1 \text{ km}^2$ 。图 2 为定远县主要地表水源地分布。工农业生产和城镇生活造成的长期性污染严重影响了水资源的质量,对供水水质造成威胁,制约着经济的持续稳健发展,影响着人民身体健康。

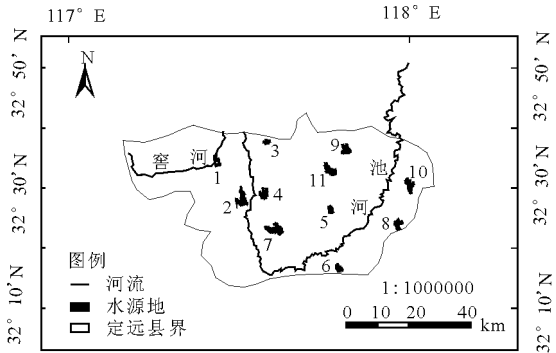


图 2 定远县饮用水水源地示意图

为获取水源地水质监测数据,搜集了 11 个水库型水源地 2012 年 5 月丰水期和 11 月枯水期的水质监测数据,其中 5 月监测数据见表 2。

2.2 投影寻踪等级评价

根据 (GB3838 - 2002)《地表水环境质量标准》<sup>[19]</sup> 中 5 个水质等级,用公式(1)、(2)预处理,得

到由无量纲数据组成的水质标准等级指标矩阵。由公式(3)~(7)计算水质等级样本的最佳投影方向及投影值序列。

用遗传算法优化投影指标函数时,取初始种群规模 $n=100$ ,优秀个体 $n'=10$ ,交叉概率 $P_c=0.8$ ,变异概率 $P_m=0.8$ ,迭代计算17次后得到全局最小值,此时投影方向 $a_1^*=(0.430,0.458,0.359,0.473,0.503)$ ,根据公式(3)可得 $Z_1^*=(2.537,2.188,1.692,1.006,0.096)$ 。

同理对定远县22个水源地水质样本进行上述模型计算,遗传算法的优化过程参数取值相同,可得样本的最佳投影方向 $a_2^*=(0.326,0.441,0.517,0.429,0.499)$ , $Z_2^*=(1.734,1.709,1.697,1.653,1.674,1.766,1.731,1.749,1.731,1.757,1.729,1.741,1.741,1.702,1.741,1.739,1.811,1.731,1.728,1.781,1.778)$ 。

对比 $Z_1^*$ 与 $Z_2^*$ 可得水质样本的水质等级,其中 $Z_2^*(4)$ 和 $Z_2^*(5)$ 均小于 $Z_1^*(3)$ ,因此判定为Ⅳ类水,其余 $Z_2^*$ 均大于 $Z_1^*(3)$ 且小于 $Z_1^*(2)$ ,因此判定为Ⅲ类水。其中,2012年5月的水质样本评价结果见表2。

| 评价单元 | 物质/(mg·L <sup>-1</sup> ) |     |     |      | 粪大肠杆菌个数 | 评价结果    |      |
|------|--------------------------|-----|-----|------|---------|---------|------|
|      | 高锰酸盐                     | 氨氮  | 溶解氧 | 总磷   |         | $Z_2^*$ | 水质等级 |
| 1    | 3.2                      | 0.3 | 5.3 | 0.08 | 5260    | 1.734   | Ⅲ    |
| 2    | 3.2                      | 0.2 | 5.3 | 0.08 | 5300    | 1.709   | Ⅲ    |
| 3    | 3.2                      | 0.3 | 5.3 | 0.07 | 5200    | 1.697   | Ⅲ    |
| 4    | 3.2                      | 0.3 | 5.3 | 0.09 | 5600    | 1.653   | Ⅳ    |
| 5    | 3.4                      | 0.2 | 5.1 | 0.07 | 5200    | 1.674   | Ⅳ    |
| 6    | 3.4                      | 0.3 | 5.2 | 0.07 | 5100    | 1.766   | Ⅲ    |
| 7    | 3.2                      | 0.3 | 5.5 | 0.05 | 4700    | 1.731   | Ⅲ    |
| 8    | 3.2                      | 0.3 | 5.2 | 0.07 | 5300    | 1.749   | Ⅲ    |
| 9    | 3.2                      | 0.3 | 5.2 | 0.07 | 5500    | 1.731   | Ⅲ    |
| 10   | 3.2                      | 0.3 | 5.6 | 0.07 | 5200    | 1.757   | Ⅲ    |
| 11   | 3.4                      | 0.3 | 5.3 | 0.06 | 4600    | 1.729   | Ⅲ    |

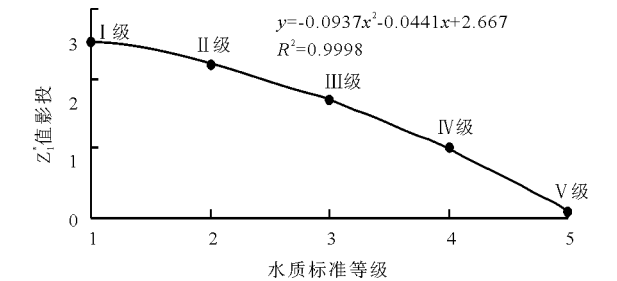


图3 水质标准等级样本投影值 $Z_1^*$ 图

2.3 BP网络仿真计算

将 $Z_2^*$ 作为BP网络输出的期望向量,经反复试验,当隐含层含9个节点时的BP网络预测误差最小,因此构建型式为5-9-1的BP神经网络。

取计算期望误差 $\varepsilon=0.0001$ ,训练速度 $\eta=0.1$ ,迭代次数 $n=2\,000$ 。将水源地水质样本实测值作为网络学习样本,网络在1562次迭代计算后,误差满足精度要求,输出结果。22个样本的计算值与仿真值见图3。网络权值输入层至中间层为 $22\times 5$ 矩阵,中间层至输出层为 $1\times 22$ 行向量,阈值分别为 $9\times 1$ 列向量和 $1\times 1$ 数值。

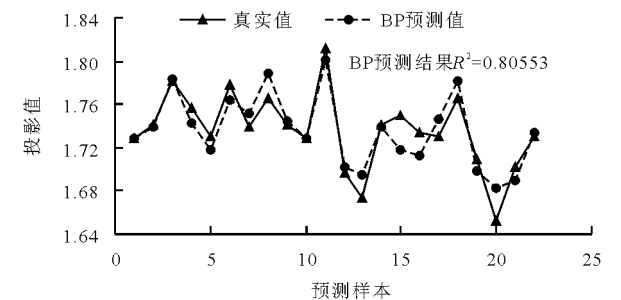


图4 真实值 $Z_2^*$ 与网络预测投影值结果比较

22个仿真计算结果中,BP网络对水质评价投影值的预测相对误差在0~1.63%,精度较高,满足预测需求。模型预测的成功与中间层神经元个数、传递函数、训练函数以及误差值关系密切。

2.4 水质评价结果分析

由于11个水源地的供水条件、能力及范围各不相同,从水质水量相结合的角度综合考察定远县饮用水水质的整体情况。以从11个水源地2012年的总取水量为界定标准,计算枯水期与丰水期水质平均值,则各类水质的量分别占定远县总用水量的比重见表3。由表3可知,定远县Ⅲ类水占比85.7%,Ⅳ类水占比14.3%,水源地内水质无Ⅰ、Ⅱ类水且水质均未超过Ⅳ类。

| 表3 定远县2012年各类水总取水量及占比                  |        |       |        |        |
|----------------------------------------|--------|-------|--------|--------|
| 取水量及占比                                 | Ⅱ类水及以下 | Ⅲ类水   | Ⅳ类水    | V类水及以上 |
| 各类水年取水量/10 <sup>4</sup> m <sup>3</sup> | 0      | 474.5 | 302.95 | 0      |
| 各类水占比/%                                | 0      | 85.7  | 14.3   | 0      |

2.5 不同评价方法结果比较

为了进一步验证评价结果,对评价样本采用多种方法进行计算验证,结果见表4。从4种方法的结果比较来看,单因子指数法只考虑了样本中最劣污染物单项,相对其它3种方法来看,结果偏重,不

能反映水质综合情况。后两种方法综合考虑了多种主要污染物对水体的影响,使得结果更为接水质的实际情况。本文的评价结果与模糊综合评判法、水质指数评价法相比,结果具有一致性,证实该模型正确合理,可用于水质综合评价。此外,对被评价为同一类水质的评价单元,通过比较投影值大小,仍然可以分辨微小程度上水质的好坏。

表 4 不同评价方法结果对比

| 评价单元 | 本法结果 | 单因子评价法 | 模糊综合评判法 | 综合指数评价法 |
|------|------|--------|---------|---------|
| 1    | 3    | 4      | 3       | 3       |
| 2    | 3    | 4      | 3       | 3       |
| 3    | 3    | 4      | 3       | 3       |
| 4    | 4    | 4      | 3       | 3       |
| 5    | 4    | 4      | 4       | 4       |
| 6    | 3    | 4      | 3       | 3       |
| 7    | 3    | 4      | 3       | 2       |
| 8    | 3    | 4      | 3       | 3       |
| 9    | 3    | 4      | 3       | 3       |
| 10   | 3    | 4      | 3       | 3       |
| 11   | 3    | 4      | 3       | 3       |

3 结 论

(1)由定远县 2012 年 5 月的水质样本评价结果可知,评价单元 3、4 未达到饮用水标准,11 个水源地中达到Ⅲ类水质可进入自来水厂最低要求的占 81.8%。近年来定远县大力加强农村饮水安全设施建设,在严格控制对水质水量监测的同时,对取水、制水、供水实施全覆盖管理,并以水质为核心建立完备的质量管理体系,很大程度上确保了饮用水质安全。至 2016 年 4 月,定远县地表水源地达标率上升至 92.8%。

(2)投影寻踪等级评价模型将水质多维指标数据按照空间相似性降至一维,避免主观确定权重,并结合 BP 神经网络对评价结果进行仿真预测验证了结果的合理性。根据训练后 BP 网络的传递函数、权值及阈值,可直接对水质样本进行预测计算投影值,再参照标准水质等级投影值可快速判定水质等级,计算准确并具有可操作性。

参考文献:

[1] 李文攀,朱 擎,李东一,等. 集中式饮用水水源地水质评价方法研究[J]. 中国环境监测, 2015,31(1):24-27.

[2] 刘 琰,郑丙辉,付 青,等. 水污染指数法在河流水质评价中的应用研究[J]. 中国环境监测,2013,29(3):49-55.

[3] 徐 健,吴 玮,黄天寅,等. 改进的模糊综合评价法在同里古镇水质评价中的应用[J]. 河海大学学报(自然科学版),2014,42(2):143-149.

[4] 卢文喜,李 迪,张 蕾,等. 基于层次分析法的模糊综合评价在水质评价中的应用[J]. 节水灌溉,2011(3):43-46.

[5] 向 娜. 基于神经网络和人工蜂群算法的水质评价和预测研究[D]. 广州:华南理工大学, 2012.

[6] 张亚丽,周 扬,程 真,等. 不同水质评价方法在丹江口流域水质评价中应用比较[J]. 中国环境监测,2015,31(3):58-61.

[7] 席文娟,金 婧,钱 会. 改进模糊综合评价法在水质评价中的应用[J]. 水资源与水工程学报,2012,23(3):25-29.

[8] 刘 愚. 基于熵权-理想点法的大坝安全评价模型及应用[J]. 水电能源科学,2016,34(5):73-76.

[9] 韩力群. 人工神经网络理论、设计及应用[M]. 北京:化学工业出版社, 2007.

[10] 高玉琴, 张利昕,吴焕霞. 改进 GA-BP 神经网络评价算法及其应用[J]. 水利经济,2012,30(6):7-10+25.

[11] 毛 健,赵红东,姚婧婧. 人工神经网络的发展及应用[J]. 电子设计工程,2011,19(24):62-65.

[12] 李志峰, 张二艳. BP 神经网络在空气质量评价分级中的探索与应用[J]. 电子技术与软件工程,2014(5):43-44.

[13] 刘红盼,朱 俊,范 东,等. 中国农村饮用水水质现状分析[J]. 环境科技, 2011,24(2):88-91.

[14] 方国华,黄显峰. 多目标决策理论、方法及其应用[M]. 北京:科学出版社, 2011.

[15] 于国荣,叶 辉,夏自强,等. 投影寻踪分类模型的改进及在水质评价中的应用[J]. 四川大学学报(工程科学版),2008,40(6):24-29.

[16] 龙美林,廖 强,罗 畏,等. 改进的投影寻踪模型在水质评价中的应用[J]. 中国农村水利水电,2011(8):16-19.

[17] 周琪慧,方国华,吴学文,等. 基于遗传投影寻踪模型的泵站运行综合评价[J]. 南水北调与水利科技,2015,5(13):985-989.

[18] 闻 新,周 露,李 翔,等. MATLAB 神经网络仿真与应用[M]. 北京:科学出版社, 2003.

[19] 中华人民共和国环境保护总局,国家质量监督检验检疫总局. GB3838-2002 地表水环境质量标准[S]. 北京:中国环境出版社,2002.