DOI:10.3969/j.issn.1003-5060.2024.06.002

基于集成替代模型和遗传算法的地下水污染源信息识别

刘 蒙[1], 骆乾坤[1], 安济民[2], 赵 梦[1], 钱家忠[1]
(1.合肥工业大学 资源与环境工程学院,安徽 合肥 230009; 2.中国人民解放军 93263 部队,辽宁 锦州 121000)

摘 要:准确识别污染源信息是高效治理和修复地下水污染的前提。为解决使用传统模拟优化方法识别污染源信息过程中产生的严重计算负担问题,文章首先建立 BP 神经网络(back propagation neural network,BPNN)模型、支持向量回归(support vector regression,SVR)模型和核极限学习机(kernel extreme learning machine,KELM)模型代替传统模拟优化方法中的地下水水流和溶质运移模拟模型。然后,使用简单平均法和遗传算法(genetic algorithm,GA)计算权重值并建立集成替代模型进一步提高模型精度。最后,将表现最优的基于遗传算法建立的集成替代模型嵌入识别污染源信息的优化模型中。算例结果分析表明,嵌入基于遗传算法建立的集成替代模型的优化模型相较于传统模拟优化模型计算时间由 11 d 大幅度缩短至 39 min,且求解所得的污染源信息参数接近真实值,可用于解决污染源信息识别问题。

关键词:地下水污染;污染源识别;替代模型;遗传算法(GA)
中图分类号:X523 **文献标志码**:A **文章编号**:1003-5060(2024)06-0731-08

Identification of groundwater contaminant source information based on ensemble surrogate model and genetic algorithm

LIU Meng[1], LUO Qiankun[1], AN Jimin[2], ZHAO Meng[1], QIAN Jiazhong[1]
(1. School of Resources and Environmental Engineering, Hefei University of Technology, Hefei 230009, China; 2. Unit 93263 of PLA, Jinzhou 121000, China)

Abstract: Identifying contaminant source information accurately is the premise for efficient remediation of groundwater pollution. In order to solve the problem of computational burden in the process of identifying contaminant source information using traditional simulation-optimization methods, this paper firstly establishes back propagation neural network(BPNN) model, support vector regression(SVR) model and kernel extreme learning machine(KELM) model to replace the groundwater flow and solute transport model in traditional simulation-optimization methods. Then, the simple average method and genetic algorithm(GA) are used to calculate the weight values, and the ensemble surrogate models are established to further improve the accuracy of the model. Finally, the best-performing ensemble surrogate model based on GA is embedded in the optimization model for identifying contaminant source information. The calculation results show that compared with the traditional simulation-optimization model, the calculation time of the optimization model embedded with the ensemble surrogate model based on GA is shortened from 11 d to 39 min and the contaminant source information parameters obtained by the solution are close to the true value, which can be used to solve the contaminant source information identification problem.

收稿日期:2021-06-02;**修回日期**:2021-07-15
基金项目:国家自然科学基金资助项目(41831289);安徽省自然科学基金资助项目(1708085QD82)和中央高校基本科研业务费专项资金资助项目(JZ2018HGTB0251)
作者简介:刘 蒙(1997—),男,安徽阜阳人,合肥工业大学硕士生;
骆乾坤(1984—),女,河北石家庄人,博士,合肥工业大学副研究员,硕士生导师,通信作者,E-mail:QKLuo@hfut.edu.cn;
钱家忠(1968—),男,安徽凤阳人,博士,合肥工业大学教授,博士生导师.

Key words: groundwater contamination; contaminant source identification; surrogate model; genetic algorithm(GA)

地下水污染源信息识别问题是水文地质领域的一个重要研究课题。成功识别地下水污染源的位置、泄露时间和浓度等信息是开展地下水污染治理、地下水环境修复等工作的基础[1]。模拟-优化方法是典型的求解污染源识别问题的数学模拟方法[2]。传统的模拟优化模型包含模拟模型与优化模型 2 个部分。其中,模拟模型模拟含水层中的水流流动和污染物运移过程,优化模型通常以最小化观测值与模型预测值之间的误差等为目标函数来寻找能满足最接近污染物观测数据的解[3]。模拟优化模型多采用进化算法进行求解,这不可避免地需要反复调用模拟模型,由此所造成的严重计算负担已经成为阻碍模拟优化模型求解污染源识别问题的瓶颈。

近年来,替代模型的出现为减轻反复调用模拟模型的计算负担提供了一种有效方法。替代模型经过合理训练能够有效地逼近模拟模型预测结果[4]。在污染源识别领域中,建立替代模型的典型方法有克里金法(Kriging)[5-7]、多项式回归(polynomial regression, PR)[8]、人工神经网络(artificial neural network, ANN)[9-10]等。随着机器学习理论方法的发展完善,以支持向量回归(support vector regression, SVR)[11]、核极限学习机(kernel extreme learning machine, KELM)[12]等为代表的机器学习方法在污染源识别领域中也逐渐兴起。然而,在实际应用中,每一种替代模型都有自身的限制,例如使用 Kriging 模型进行预测估值时,需要反复对模型进行训练,导致训练时间增加,使得减少计算负荷的效果不明显[13]。此外,单个替代模型在进行学习时也存在易陷入局部最优解的缺点[14]。根据集成学习理论,建立由多个单一替代模型组成的集成替代模型可有效避免单一替代模型存在的各项缺点,同时提高模型的预报精度[15]。

建立集成模型需要解决 2 个关键性问题:① 选择性能优良的子模型,子模型的精度越高、多样性越大,则集成模型的准确率越高,BP 神经网络(back propagation neural network, BPNN)模型具有很强的非线性信息处理能力,对于一些线性回归问题有很高的求解精度[16],SVR 模型能够有效捕获模拟模型中复杂的输入与输出关系[17],KELM 模型具有很强的泛化能力和学习速度,对于复杂的线性回归问题能够产生稳定的输出结果[18];② 采取合适的组合策略将子模型组合在一起,对于数值类的回归预测问题,平均法是最常用的分配模型权重的方法,平均法有 2 种,简单平均法根据子模型的个数为模型分配权重值,加权平均法一般从训练样本中学习进而根据各子模型的预测输出赋予子模型权重值[19]。

文献[20]利用简单平均和加权平均 2 种方法建立考虑不同影响因子的 SVR 集成预测模型,相较于 SVR 单一模型有效提高了径流预测的精度,不足之处在于未选定多种差异较大的单一模型开展研究;文献[21]基于后验概率为 Kriging 模型、径向基函数(radial basis function, RBF)模型和最小二乘支持向量机(least squares support vector machine, LSSVM)模型分配权重建立集成替代模型,并成功对地下水污染源释放历史进行识别,但此种方法分配权重易受数据噪声的影响从而导致集成替代模型准确率降低。

综上所述,本文提出采用遗传算法(genetic algorithm, GA)求解权重最优组合,利用子模型的预测输出信息,合理分配权重值,并建立基于 BPNN 模型、SVR 模型和 KELM 模型的集成替代模型。将建立的集成替代模型与采用简单平均法分配权重建立的集成替代模型进行对比,验证新权重分配方法的准确性与适用性。

1 研究模型

1.1 BPNN 模型

BPNN 的核心思想是使用 BP 算法对网络中神经元之间连接权值和神经元阈值进行迭代更新,最终使输出值和真实值的误差达到可接受范围[22]。对训练样本$(x_i, y_i)(i=1,2,\cdots,n)$,给定学习率 η,则训练后神经网络中神经元连接权值和阈值的迭代公式如下:

$$\Delta w_{hj} = -\eta \frac{\partial E_i}{\partial w_{hj}}, \quad \Delta\theta_j = \eta \frac{\partial E_i}{b_h \partial w_{hj}} \tag{1}$$

其中:w_{hj} 为隐层第 h 个神经元与输出层第 j 个神经元的连接权值;θ_j 为输出层第 j 个神经元的阈值;E_i 为样本输出值与真实值的均方误差;b_h 为隐层第 h 个神经元的输出。

本文使用的 BP 神经网络对误差目标函数进行正则化处理,能够有效避免过拟合问题。此外,在本文的实际应用中,通过"试错法"对神经元个数进行不断调整,最终确定神经元个数为 2 000,

此时神经网络预测精度最高。

1.2　SVR 模型

对训练样本$(x_i,y_i)(i=1,2,\cdots,n)$SVR 的回归方程模型可表示为:

$$f(x_i)=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\varphi}(x_i)+b \tag{2}$$

其中:$\boldsymbol{w}$ 为权值向量;$\boldsymbol{\varphi}(x_i)$为 x_i 经核函数映射后的特征向量;b 为偏差。

为使模型预测值 $f(x_i)$与真实值 y_i 之间的误差最小,将回归问题转化为以下优化问题:

$$\begin{aligned}\min\quad & \frac{1}{2}\|\boldsymbol{w}\|^2+C\sum_{i=1}^{m}(\xi_i+\hat{\xi}_i);\\ \text{s.t.}\quad & f(x_i)-y_i\leqslant\varepsilon+\xi_i,\\ & y_i-f(x_i)\leqslant\varepsilon+\hat{\xi}_i,\\ & \xi_i,\hat{\xi}_i\geqslant 0,\quad i=1,2,\cdots,m\end{aligned} \tag{3}$$

其中:C 为正则化系数;ξ_i 和 $\hat{\xi}_i$ 为松弛变量;$f(x_i)$为模型输出值;ε 表示能接受的模型输出值与实际值之间的最大偏差。本文采用拉格朗日乘子法[23]求解上述优化问题。

1.3　KELM 模型

KELM 模型引入一个核函数来代替极限学习机的显示激活函数,其拟合能力和泛化能力显著强于 ELM 模型。本文引入高斯核函数,对训练样本$(x_i,y_i)(i=1,2,\cdots,n)$的高斯核函数表示为:

$$K(x_i,x_j)=\exp\left(-\frac{\|x_i-x_j\|^2}{2\sigma^2}\right) \tag{4}$$

其中,$\sigma>0$,表示核函数的带宽。

KELM 模型的核矩阵可以定义为:

$$\boldsymbol{K}_{\mathrm{KELM}}=\boldsymbol{M}\boldsymbol{M}^{\mathrm{T}} \tag{5}$$

$$\boldsymbol{K}_{\mathrm{KELM}(i,j)}=[\boldsymbol{m}(x_i)]^{\mathrm{T}}\boldsymbol{m}(x_j)=\boldsymbol{K}(x_i,x_j) \tag{6}$$

其中:$\boldsymbol{M}$ 为样本输入在特征空间中的映射矩阵;$\boldsymbol{m}(x_i)$为 x_i 映射后的特征向量。

结合式(6),KELM 回归方程模型表示为:

$$\boldsymbol{F}(x)=[\boldsymbol{m}(x)]^{\mathrm{T}}\boldsymbol{M}^{\mathrm{T}}\left(\boldsymbol{M}\boldsymbol{M}^{\mathrm{T}}+\frac{\boldsymbol{I}}{C}\right)^{-1}\boldsymbol{N}=\begin{bmatrix}\boldsymbol{K}(x,x_1)\\ \vdots\\ \boldsymbol{K}(x,x_n)\end{bmatrix}^{\mathrm{T}}\left(\boldsymbol{K}_{\mathrm{KELM}}+\frac{\boldsymbol{I}}{C}\right)^{-1}\boldsymbol{N} \tag{7}$$

其中:C 为正则化系数;$\boldsymbol{N}=[y_1\quad\cdots\quad y_n]^{\mathrm{T}}$。

1.4　集成替代模型

使用训练样本分别建立 BPNN、SVR 和 KELM 模型,结合 3 种单一替代模型建立集成替代模型,其输出表达式为:

$$\begin{aligned}&C_{\mathrm{s},t}=w_1C_{\mathrm{BPNN},t}+w_2C_{\mathrm{SVR},t}+w_3C_{\mathrm{KELM},t},\\ &\text{s.t.}\quad\sum_{i=1}^{N}w_i=1\end{aligned} \tag{8}$$

其中:$C_{\mathrm{s},t}$为在监测时刻 t 时集成替代模型的输出值;$C_{\mathrm{BPNN},t}$、$C_{\mathrm{SVR},t}$、$C_{\mathrm{KELM},t}$分别为 3 种单一替代模型在监测时刻 t 时的输出值;w_1、w_2、w_3 为权重值,权重之和为 1。

本文使用简单平均法和遗传算法 2 种确定权重值的方法开展对比研究。

1) 简单平均法获取模型权重值。根据单一替代模型个数确定权重值,表达式如下:

$$w_i=\frac{1}{T} \tag{9}$$

其中:w_i 为第 i 个替代模型的权重值;T 为单一替代模型的个数。

2) 遗传算法获取模型权重值。以最小化集成模型预测值与模拟模型输出值的均方根误差(root mean square error,RMSE)为目标函数,建立权重优化模型,求解得到权重值的最优组合。

$$\begin{aligned}&\min\quad E(C_{\mathrm{s},t},C_{\mathrm{actual},t})=\sqrt{\frac{1}{M}\sum_{j=1}^{M}\left(\sum_{i=1}^{N}w_iC_{i,t}-C_{\mathrm{actual},t}\right)^2},\\ &\text{s.t.}\quad\sum_{i=1}^{N}w_i=1\end{aligned} \tag{10}$$

其中:E 为目标函数;$C_{i,t}$为第 i 个单一替代模型在监测时刻 t 时的输出值;$C_{\mathrm{autual},t}$为模拟模型在监测时刻 t 时的输出值;M 为训练样本个数;N 为单一替代模型的个数。

2　算例研究

2.1　算例介绍

假定研究区为长 1 000 m、宽 600 m 的矩形区域,含水层岩性以中粗砂为主,地下水流为潜水稳定流。为了使场地条件更加符合实际情况,采用直接傅里叶变换方法生成 10 000 个均值为 20 m/d、方差为 0.3 的渗透系数随机场,并随机选择其中一个作为场地实际情况。含水层其余参数见表 1 所列。研究区东西为定水头边界,水头值分别为 88.8、100.0 m,南北为隔水边界。

表 1　含水层参数值

参数	数值
纵向弥散度 α_{L}/m	40
横向弥散度 α_{T}/m	9.6
孔隙度 n	0.25
含水层厚度 b/m	30.5

以研究区西南角为坐标原点,将研究区域划分为 30 行、50 列的正方形有限差分网格,基本单元

网格大小为 20 m×20 m。研究区内布置 6 口监测井和 1 口注水井,注水井的注水量为 200 m^3/d。监测井的分布位置和注水井的潜在范围如图 1 所示。假定研究区中存在一个污染源以 200 g/L 的质量浓度从注水井处泄露污染物,污染物为不发生化学转化与生物迁移的保守污染物。模拟时长为 4 a(本文以 360 d 记为 1 a),分为 4 个应力期,每个应力期为 1 a。监测井每隔 90 d 进行 1 次质量浓度监测,每口井监测 16 次。

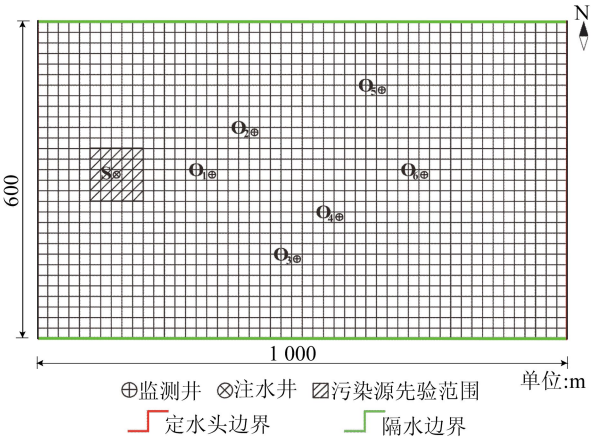


图 1 研究区平面示意图

为获取监测井内污染物观测质量浓度,本文首先使用 MODFLOW 和 MT3DMS 模拟程序在研究区内建立地下水水流及溶质运移模拟模型并求解。真实污染源在第 1 个应力期内从坐标(8,16)处以 200 g/L 的起始质量浓度向场地释放污染物,场地内所有监测井的污染物质量浓度真实穿透曲线如图 2 所示。

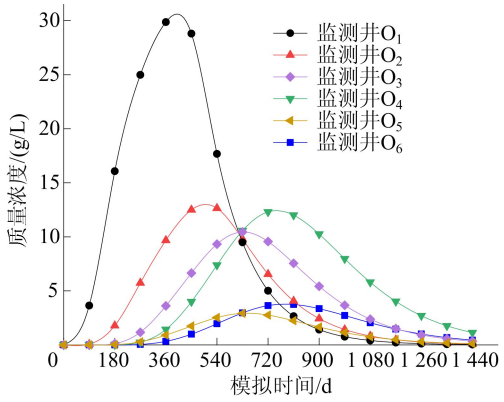


图 2 真实污染源在不同监测点处污染物质量浓度穿透曲线

2.2 替代模型的建立

本文所要识别的污染源未知参数包括污染源的位置(坐标值 X、Y),污染物释放起始和终止时间 T_1、T_2,污染物的释放质量浓度 ρ。使用模拟优化模型识别未知污染源信息时,优化模型中的目标函数一般包含相应的约束条件,约束条件的设置根据具体问题和研究者经验进行确定。本文认为未知污染源存在于场地中央一处 100 m×100 m 的区域内,且认为泄露初始时间和终止时间与真实值的差距不超过 1 d,依据经验估算污染源泄露质量浓度先验范围的上、下限为真实值的 110%、9%。假定上述 5 个参数的先验分布为均匀分布,最终确定参数的先验范围见表 2 所列。

表 2 污染源未知参数先验范围

X	Y	T_1/d	T_2/d	ρ/(g/L)
(6,10)	(14,18)	(0,10)	(360,370)	(180,220)

建立替代模型的过程如下。

1) 抽取训练样本。抽样方法是影响替代模型精度的主要因素之一。拉丁超立方抽样(Latin hypercube sampling,LHS)将参数的先验范围进行分层,能够提高样本对先验范围的覆盖率[24]。研究表明过多的训练样本并不会提高替代模型的精度[25]。本文研究使用 LHS 方法抽取训练样本 80 组,作为替代模型的输入值。

2) 编写模型程序并进行训练。采用 Python 语言编写 3 种替代模型的程序,其中 BPNN 模型、SVR 模型借助 sklearn 工具包编写,KELM 模型为自编程序。分别将 80 组训练样本代入 MODFLOW 和 MT3DMS 模拟程序中,得到 6 口监测井监测到的一系列污染物质量浓度值。利用污染物质量浓度数据对替代模型进行训练。

3) 检验模型。在参数先验范围内重新抽取 20 组检验样本。重复步骤 2)中操作得到对应的污染物监测质量浓度值。将检验样本代入到训练好的替代模型中,得到替代模型输出值,并与监测值进行比较,评估替代模型的精度。

通常采用确定性系数 R^2、平均绝对误差 R_{MAE}和均方根误差 R_{RMSE}这 3 个指标对各替代模型的性能进行评价[26]。R^2 是描述 2 种变量相关性的指标,一般来说,当 $R^2>0.9$ 时,可以认为模型有较好的逼近精度。R_{MAE}和 R_{RMSE}是反映模型预测误差大小的指标,其值越接近于 0,模型越准确。R^2、R_{MAE}和 R_{RMSE}的计算公式如下:

$$R^2=1-\frac{\sum_{i=1}^{n}(y_{i,t}-\hat{y}_{i,t})^2}{\sum_{i=1}^{n}(y_{i,t}-\bar{y}_t)^2} \tag{11}$$

$$R_{\mathrm{MAE}}=\frac{\sum_{i=1}^{n}(y_{i,t}-\hat{y}_{i,t})}{n} \tag{12}$$

$$R_{\mathrm{RMSE}}=\sqrt{\sum_{i=1}^{n}(y_{i,t}-\hat{y}_{i,t})^2/n} \tag{13}$$

其中:$y_{i,t}$ 为监测时刻 t 时污染物质量浓度监测值;$\hat{y}_i$ 为监测时刻 t 时替代模型输出值;$\hat{y}$ 为监测时刻 t 时替代模型输出值均值;n 为检验样本数量。

使用 6 口井监测到的质量浓度数据对 6 口井分别建立 BPNN 替代模型、SVR 替代模型和 KELM 替代模型。对每口井建立的 3 种替代模型的评价指标结果见表 3 所列。

表 3　替代模型评价指标对比结果

井号	替代模型	R^2	R_{MAE}	R_{RMSE}
1	BPNN	0.974 6	0.165 5	0.251 5
	SVR	0.978 6	0.143 1	0.230 7
	KELM	0.982 2	0.134 1	0.210 7
2	BPNN	0.971 2	0.507 6	0.849 3
	SVR	0.988 3	0.321 5	0.540 6
	KELM	0.988 0	0.335 4	0.547 0
3	BPNN	0.969 1	1.001 8	1.831 2
	SVR	0.970 8	0.882 4	1.781 0
	KELM	0.970 2	0.955 2	1.799 6
4	BPNN	0.993 4	0.227 8	0.349 1
	SVR	0.995 9	0.183 0	0.279 0
	KELM	0.996 0	0.188 7	0.275 8
5	BPNN	0.984 7	0.314 8	0.480 1
	SVR	0.990 5	0.230 8	0.377 4
	KELM	0.991 8	0.228 5	0.349 7
6	BPNN	0.944 3	0.201 9	0.331 7
	SVR	0.956 1	0.177 7	0.294 3
	KELM	0.957 7	0.177 3	0.289 0

对表 3 中的数据进行分析可知,第 6 口井的 3 种替代模型的确定性系数在 0.95 左右,其余各口井的替代模型确定性系数均在 0.96 以上,且第 4 口井的替代模型的确定性系数超过了 0.99,说明替代模型输出值十分接近模拟模型输出值。此外,替代模型输出值与模拟模型输出值之间的平均绝对误差和均方根误差都处于较低的水平,第 3 口井中替代模型的平均绝对误差和均方根误差值分别为 1.001 8 和 1.831 2 外,其余井的平均绝对误差和均方根误差值接近于 0。

KELM 和 SVM 模型的性能相差不大,可能是两者都引入了高斯核函数。相对来说,BPNN 模型的性能表现略差,对每口井所建立的 3 种替代模型里,其确定性系数最低,平均绝对误差和均方根误差均最高。但对每口井建立的 BPNN 模型的确定性系数均在 0.94 以上,平均绝对误差和均方根误差值绝大部分在 0.5 以下,因此仍保留 BPNN 模型作为集成模型的子模型之一。

2.3　集成替代模型的建立

基于建立的 3 种单一替代模型,使用 2 种方法确定每口井中单一替代模型的对应权重值。

第 1 种使用简单平均法,利用式(9)计算权重值。使用 3 种单一替代模型作为子模型进行组合建立集成替代模型 1,根据式(9)进行计算,可得每口井中 3 种单一替代模型的权重值均为 0.333 3。

第 2 种利用遗传算法对建立的权重优化模型进行求解,得到对应的权重值后建立集成替代模型 2。权重优化模型目标函数如式(10)所示。计算得到的权重结果见表 4 所列。

表 4　遗传算法求得的权重优化值

井号	BPNN	SVR	KELM
1	0.128 1	0.015 2	0.856 7
2	0.000 9	0.463 0	0.536 1
3	0.104 4	0.675 5	0.220 1
4	0.001 7	0.503 9	0.494 4
5	0.126 7	0.464 7	0.408 6
6	0.008 5	0.193 3	0.798 2

对建立好的集成替代模型 1 和模型 2,使用与单一替代模型相同的检验样本验证集成替代模型的性能。2 种集成替代模型的评价指标结果见表 5所列。

表 5　2 种集成替代模型评价指标结果

井号	模型	R^2	R_{MAE}	R_{RMSE}
1	模型 1	0.981 6	0.141 8	0.206 5
	模型 2	0.986 7	0.101 5	0.161 4
2	模型 1	0.985 6	0.334 2	0.537 8
	模型 2	0.989 4	0.318 4	0.519 8
3	模型 1	0.974 5	0.821 6	1.657 5
	模型 2	0.974 2	0.839 5	1.646 4
4	模型 1	0.996 2	0.181 1	0.269 9
	模型 2	0.995 9	0.181 7	0.272 7
5	模型 1	0.992 0	0.227 5	0.334 7
	模型 2	0.992 7	0.208 7	0.332 5
6	模型 1	0.957 8	0.177 2	0.265 4
	模型 2	0.962 1	0.137 6	0.248 3

从表 5 可以看出,在单一替代模型性能差异较大的 1、2、5、6 号井中,集成替代模型 2 具有比集成替代模型 1 更优的性能表现。在 1、2、5、6 号井中,集成替代模型 2 的 R^2 相较于集成替代模型 1 分别提高了 0.520%、0.386%、0.071%、0.449%,R_{MAE} 分别降低了 28.420%、4.728%、8.264%、22.348%,R_{RMSE} 分别降低了 21.872%、

3.347%、0.657%、6.443%。在单一替代模型性能相近的 3、4 号井中,集成替代模型 1 的性能略好于集成替代模型 2,R^2、R_{MAE} 和 R_{RMSE} 均非常接近。将 2 种集成替代模型的性能和单一替代模型进行进一步对比,所有监测井处集成替代模型 2 的精度均要比单一替代模型的精度高,但在1 号井和 2 号井中,集成替代模型 1 的精度出现了比单一替代模型低的情况。说明简单平均法确定的权重值在子模型性能差异较大的情况下并不合理,基于简单平均法建立的集成替代模型并不总能提高模型精度的目的。此外,从 4 号井中集成替代模型与单一替代模型的比较来看,当单一替代模型的准确度达到足够高时(R^2 均在 0.99 以上,R_{MAE}均低于 0.23,R_{RMSE}均低于 0.35),集成替代模型提高逼近精度的效果并不明显。将 4 号井中性能表现更好的集成替代模型 1 与单一替代模型中表现最好的 KELM 模型进行对比,可以发现集成替代模型 1 的均方根误差和平均绝对误差略有降低,但确定性系数分别为 0.996 2、0.996 0,基本持平。

综上所述,使用遗传算法分配权重值相较于简单平均法更加全面可靠的分配权重值的方法,具有更强的泛化性能。因此,本文选择集成替代模型 2 进行污染源信息的识别。

2.4 识别污染源信息的优化模型

本文采用最小化监测井处质量浓度模拟值与质量浓度真实值的均方根误差作为目标函数,建立以下优化模型,用以求解污染源信息。

$$\min \quad E(X,Y,T_1,T_2,\rho)=\sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N}[\rho_j(t_i)-\rho_j^{\mathrm{act}}(t_i)]^2/D};$$

$$\text{s.t.} \quad \rho_{\min}<\rho<\rho_{\max},\ X_{\min}<X<X_{\max},\ Y_{\min}<Y<Y_{\max},\ T_{\min}<T_q<T_{\max},\ q=1,2 \quad (14)$$

其中:E 为目标函数;ρ 为污染源的质量浓度值;X 为污染源 x 方向的网格位置;Y 为污染源 y 方向的网格位置;T_1、T_2 分别为释放起始时间和终止时间;M 为监测次数;N 为监测井数量;$\rho_j(t_i)$ 为 t_i 时刻第 j 个监测井的计算质量浓度值;$\rho_j^{\mathrm{act}}(t_i)$ 为 t_i 时刻第 j 个监测井的监测质量浓度值;D 为监测次数与监测井数量之积。

使用遗传算法作为污染源信息识别优化模型的求解算法。将训练好的集成替代模型 2 作为子程序嵌入到遗传算法主程序中进行污染源信息识别。遗传算法迭代次数设置为 300,每代种群数量为 200。污染源信息识别结果见表 6 所列。

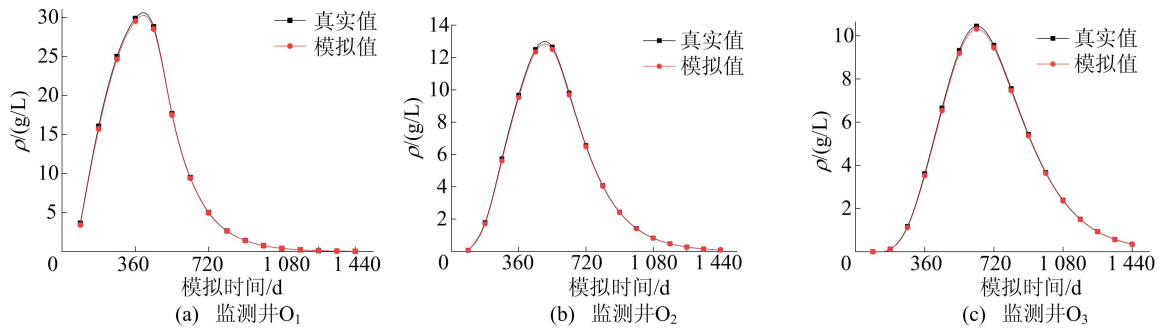
由表 6 可知,嵌入集成替代模型 2 的优化模型可以高效准确识别出污染源的位置(坐标值)、释放起始和终止时间以及泄露质量浓度,其中位置识别结果的相对误差是 0,泄露时长和质量浓度识别结果的相对误差分别为 0.006%和 0.941%。

使用污染源信息识别结果正演计算污染质量浓度的分布,得到不同监测井处污染物质量浓度并与真实值进行对比,结果如图 3 所示。由图 3 可知,不同监测井处污染物质量浓度模拟值均与真实值吻合较好。

计算污染物质量浓度模拟值与真实值之间的均方根误差值仅有0.043,进一步验证了污染源识别结果的准确性。识别用时方面,在配置为 Core i5 处理器、2.9 GHz、8 GiB 内存的 PC 端,使用传统的模拟优化模型求解时长需要 11 d。

表 6 基于集成替代模型 2 的污染源信息识别结果

参数	位置	释放时间/d		ρ/(g/L)	识别用时/min
		T_1	T_2		
真实值	(8,16)	0	360	200.00	
模拟值	(8,16)	2	360	198.12	39



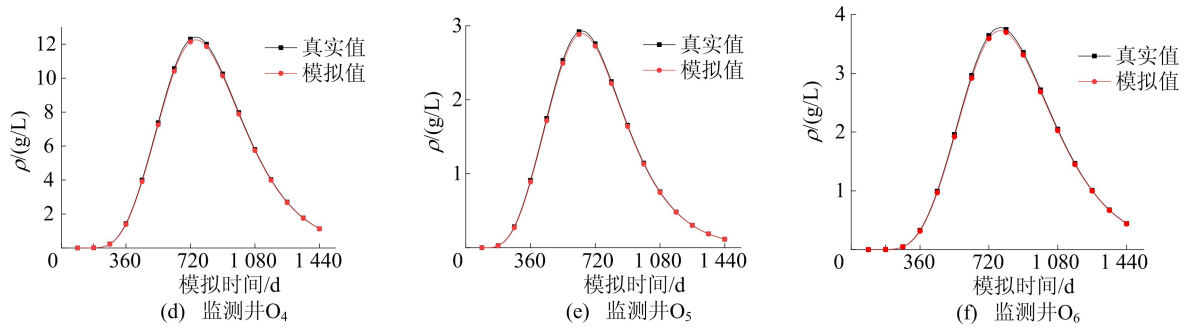


图 3　不同监测点处污染物质量浓度模拟值与真实值比较结果

而在优化过程中将模拟模型替换为集成替代模型 2 对污染源信息识别仅需 39 min。

3　结　　论

本文提出利用遗传算法求解权重最优组合，建立基于 BPNN 模型、SVR 模型和 KELM 模型的集成替代模型，并将其嵌入到优化模型中成功应用于污染源信息识别问题。算例求解结果表明，采用遗传算法分配权重值相较于简单平均法是一种更加可靠有效的分配权重值的方法。基于集成替代模型可以准确高效识别地下水污染源信息，同时，使用集成替代模型代替模拟模型可以将识别时长从 11 d 缩短至 39 min。在本文研究中，建立的求解方法未考虑含水层参数不确定性问题，且未对含水层参数进行同步反演。在下一步研究计划中将会同时反演考虑含水层参数不确定条件下的地下水污染源信息和含水层参数。

[参　考　文　献]

[1] ATMADJA J, BAGTZOGLOU A C. State of the art report on mathematical methods for groundwater pollution source identification[J]. Environmental Forensics, 2001, 2(3): 205-214.

[2] MILNES E, PERROCHET P. Simultaneous identification of a single pollution point-source location and contamination time under known flow field conditions[J]. Advances in Water Resources, 2007, 30(12): 2439-2466.

[3] AYVAZ M T. A linked simulation-optimization model for solving the unknown groundwater pollution source identification problems[J]. Journal of Contaminant Hydrology 2010, 117(1/2/3/4): 4659.

[4] 侯泽宇，卢文喜，王宇. 基于替代模型的地下水 DNAPLs 污染源反演识别[J]. 中国环境科学，2019，39(1)：188-195.

[5] ZHAO Y, LU W X, XIAO C N. A Kriging surrogate model coupled in simulation-optimization approach for identifying release history of groundwater sources[J]. Journal of Contaminant Hydrology, 2016, 185: 51-60.

[6] 张双圣，刘汉湖，强静，等. 地下水污染监测井优化设计及污染源识别[J]. 湖南大学学报学报(自然科学版)，2019，46(6)：120-132.

[7] 苏安玉，濮励杰，付强. 基于 Kriging 模型的三江平原地下水资源评价[J]. 经济地理，2007，27(6)：1018-1020.

[8] HE L, HUANG G H, ZENG G M, et al. An integrated simulation, inference, and optimization method for identifying groundwater remediation strategies at petroleum-contaminated aquifers in western Canada[J]. Water Research, 2008, 42(10/11): 2629-2639.

[9] SRIVASTAVA D, SINGH R M. Breakthrough curves characterization and identification of an unknown pollution source in groundwater system using an artificial neural network (ANN)[J]. Environmental Forensics, 2014, 15(2): 175-189.

[10] SRIVASTAVA D, SINGH R M. Groundwater system modeling for simultaneous identification of pollution sources and parameters with uncertainty characterization [J]. Water Resources Management, 2015, 29: 4607-4627.

[11] ZHANG X S, SRINIVASAN R, LIEW M V. Approximating SWAT model using artifcial neural network and support vector machine[J]. Journal of the American Water Resources Association, 2009, 45(2): 460-474.

[12] SATTAR A M, ERTURUL F, GHARABAGHI B, et al. Extreme learning machine model for water network management[J]. Neural Computing & Applications, 2019, 31(1): 157-169.

[13] 邢贞相，曲睿卓，赵莹，等. 不同替代模型在地下水污染源释放历史反演中适用性研究[J]. 东北农业大学学报，2018，49(12)：59-68.

[14] ZHOU Z H, WU J, TANG W. Ensembling neural networks: many could be better than all[J]. Artificial Intelligence, 2002, 137(1/2): 239-263.

[15] WOLPERT D H, MACREADY W G. An efficient method to estimate Bagging's generalization errror[J]. Machine Leαming, 1999, 35(1): 41-55.

[16] MIRGHANI B Y, ZECHMAN E M, RANJITHAN R S. Enhanced simulation optimization approach using surrogate modeling for solving inverse problems[J]. Environmental Forensic, 2012, 13(4): 348-363.

[17] HOU Z Y, LU W X, CHU H B, et al. Selecting parameter-optimized surrogate models in DNAPL-contaminated aqui-

fer remediation strategies[J]. Environmental Engineering Sciences,2015,32(12):1016-1026.
[18] ZHAO Y,QU R Z,XIN Z X,et al. identifying groundwater contaminant sources based on a KELM surrogate model together with four heuristic optimization algorithms[J]. Advances in Water Resources,2020,138:103540.
[19] KITTLER J,HATEF M,DUIN R,et al. On combining classifiers[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,1998,20(3):226-239.
[20] 代兴兰. 回归支持向量机集成模型在年径流预测中的应用[J]. 长江科学院院报,2015,32(4):12-17.
[21] XING Z X,QU R Z,ZHAO Y,et al. Identifying the release history of a groundwater contaminant source based on an ensemble surrogate model[J]. Journal of Hydrology,2019,572:501-516.
[22] LAXMIDHAR B,SWAGAT K,AWHAN P. On adaptive learning rate that guarantees convergence in feedforward networks[J]. IEEE Transactions on Neural Networks,2006,17(5):1116-1125.
[23] ZHENG Z Q,HUO J Y,LI B B,et al. Model-based Lagrangian multiplier derivation method for depth map coding[J]. Electronics Letters,2018,54(14):880-882.
[24] 张建波,张忠伟,杨洋. 改进拉丁超立方蒙特卡罗模拟[J]. 吉林大学学报(信息科学版),2018,36(4):452-458.
[25] HOU Z Y,LU W X. Comparative study of surrogate models for groundwater contamination source identification at DNAPL-contaminated sites[J]. Hydrogeology Journal,2018,26:923-932.
[26] 李远远,梅红波,任晓杰,等. 基于确定性系数和支持向量机的地质灾害易发性评价[J]. 地球信息科学学报,2018,20(12):1699-1709.

(责任编辑 吴 亮)

(上接第730页)

[35] 王玲,李林,刘小昌,等. 山口岩水库浮游植物季节演替及影响因子分析[J]. 环境科学与技术,2021,44(增刊2):284-291.
[36] WU H,FU S,HU W,et al. Response of different benthic biotic indices to eutrophication and sediment heavy metal pollution,in Fujian coastal water,East China Sea[J]. Chemosphere,2022,307:135653.
[37] GUO C,CHEN Y,XIA W,et al. Eutrophication and heavy metal pollution patterns in the water suppling lakes of China's South-to-North Water Diversion Project[J]. Science of The Total Environment,2020,711:134543.
[38] CHEN Y Y,LIU Q Q. Numerical study of hydrodynamic process in Chaohu Lake[J]. Journal of Hydrodynamics,Ser B,2015,27(5):720-729.
[39] 余秀娟,霍守亮,眷逢宇,等. 巢湖表层沉积物中重金属的分布特征及其污染评价[J]. 环境工程学报,2013,7(2):439-450.
[40] ZHANG T,LI L,XU F,et al. Assessing the remobilization and fraction of cadmium and lead in sediment of the Jialing River by sequential extraction and diffusive gradients in films (DGT) technique[J]. Chemosphere,2020,257:127181.
[41] 王意茹,武晓郯,何静,等. 碳酸盐矿物中稀土元素分馏特征及其获取方法研究进展[J]. 岩矿测试,2022,41(6):935-946.
[42] CHEN M,DING S,LI C,et al. High cadmium pollution from sediments in a eutrophic lake caused by dissolved organic matter complexation and reduction of manganese oxide[J]. Water Research,2021,190:116711.
[43] 王琳杰,余辉,牛勇,等. 抚仙湖夏季热分层时期水温及水质分布特征[J]. 环境科学,2017,38(4):1384-1392.

(责任编辑 胡亚敏)

·信息与动态·

《合肥工业大学学报(自然科学版)》专栏征稿启事

为了贯彻落实党的二十大精神,紧密围绕科教兴国战略、人才强国战略、创新驱动发展战略,《合肥工业大学学报(自然科学版)》设置“机器人与人工智能”“环境污染与防治”两个专栏,面向国内外专家学者征集“机器人与人工智能”“环境污染与防治”领域的原创性学术论文、专题综述;稿件一经录用将优先刊发。

来稿要求政治导向正确、论证充分、具有较强的引领性和创新性。格式要求参见《合肥工业大学学报(自然科学版)》网页投稿指南的征稿简则和投稿模板。

在线投稿网址:http://xbzzs.hfut.edu.cn/xbzk.html。