DOI: 10.19797/j.cnki.1000-0852.20210513

基于随机森林法的弥河-潍河流域地下水质量评价研究

林艳竹[1]，韩 忠[2]，黄林显[3,4]，邢立亭[3,4]，梁 浩[5]，侯金霄[3]

（1.中国地质环境监测院(地质灾害技术指导中心)，北京 100081；
2.山东省第六地质矿产勘查院，山东 威海 264209； 3.济南大学 水利与环境学院，山东 济南 250022；
4.山东省地下水数值模拟与污染控制工程技术研究中心，山东 济南 250022；
5. 山东省国土空间生态修复中心，山东 济南 250014）

摘 要：准确掌握地下水的环境质量状况是合理确定地下水资源开发策略和有效进行地下水资源保护的重要前提。通过随机森林(random forest)法构建弥河-潍河流域地下水质量评价模型，结果表明：(1)随机森林法在进行地下水水质分类时具有分类精度高、泛化能力强等特点，且在进行超参数优化后，其分类精度会进一步提高，证明将随机森林法应用于地下水质量评价是可行的，并且其综合性能要优于逻辑回归模型；(2)研究区地下水水样均为Ⅳ类和Ⅴ类水，说明水质状况整体较差；(3)通过分类指标重要性评价可以看出，研究区地下水水质的主要影响指标为硝酸盐、总硬度和溶解性总固体，而此类指标的主要来源是蔬菜种植化肥的不合理使用及河流污染入渗，因此要进一步加强对蔬菜种植污染排放及河流水质的监测和控制。

关键词：机器学习；随机森林法；弥河-潍河流域；地下水质量评价

中图分类号：P641.2；P33；TV11 **文献标识码**：A **文章编号**：1000-0852(2023)03-0060-05

0 引言

地下水质量评价作为水体污染评估的有效手段，能够定量、定性评估地下水水体的污染状况，同时也是地下水环境风险分析、污染源确定和水资源保护的一项重要基础性工作[1-2]。针对地下水质量评价，国内外学者探讨了各种方法：梁乃森等[3]通过将GIS技术与改进的模糊综合评价模型相结合，建立了基于GIS技术的地下水水质模糊综合评价模型；李松青等[4]利用遗传算法-BP神经网络法对开封市深埋地下水水质进行了评价；EGBUERI等[5]利用地下水污染指数(PIG)、生态风险指数(ERI)和层次聚类算法对研究区地下水水质进行评价；SU等[6]采用了一种集对分析(SPA)-马尔科夫链的模型对西安市地下水质量进行评价。

上述研究虽然从不同角度对地下水质量评价方法进行了探讨，并取得了很多有益成果，但也均存在一定局限性：模糊数学法的评价结果较符合实际情况，但其权重和隶属度等相关参数的确定主要取决于人的主观经验[7]；神经网络等评价方法的定性评价结果往往不够直观，且其模型构建及计算过程相对复杂[8]。近些年，随着机器学习模型的发展，随机森林法由于具有操作简单、预测精度高且能够对评价指标重要性进行识别等特点而得到了广泛应用[9-10]。在水文地质领域，杨光等人利用随机森林法对黑河中游地下水埋深变化及其成因进行了分析[11]；BAUDRON等通过随机森林法对地下水水样的所属含水层层位进行了识别[12]。

收稿日期：2021-12-08 网络首发日期：2023-06-13
网络首发地址：https://kns.cnki.net/kcms2/detail/11.1814.P.20230612.1349.014.html
基金项目：国家自然科学基金资助项目(41772257)；山东省自然科学基金资助项目(ZR2019MD029)；山东省高校院所创新团队项目(2021GXRC070)；院科研基金项目(801KY202004)
作者简介：林艳竹(1989—)，女，山东烟台人，硕士，工程师，主要研究方向为地下水科学与工程。E-mail：linyanzhu_u@163.com
通信作者：黄林显(1982—)，男，山东青岛人，博士，副教授，主要研究方向为地下水科学与工程。E-mail：stu_huanglx@ujn.edu.cn

但是,已有的研究对随机森林法超参数设置及误差影响因素的讨论尚不够深入,并且缺少在地下水质量评价中的应用。基于此,本文将随机森林法应用于弥河-潍河流域地下水质量评价中,以期获得一种操作简单、评价精准的评估模型,能够为未来地下水质量评价的研究提供一定的参考,同时为弥河-潍河流域地下水资源的开发利用提供科学依据。

1　研究区概况及数据来源

1.1　研究区概况

研究区属弥河-潍河流域,行政区划上属潍坊市,是我国重要的蔬菜生产基地。区内的供水水源绝大部分依靠地下水,地下水在研究区的经济发展中发挥着不可替代的作用。因此准确查明研究区内的地下水污染现状和污染风险,对当地地下水资源的合理开发与保护具有重要意义。区内主要地下水供水水源地有8处(见图1),主要分布在弥河、白浪河、潍河和汶河四大流域冲洪积区,属于鲁西北平原松散岩类水文地质区和鲁中南低山丘陵碳酸盐岩类水文地质区。

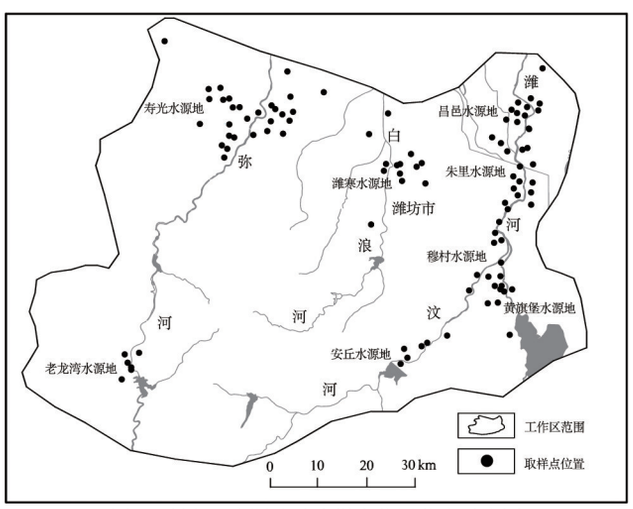

图1　研究区水源地及取样点位置分布示意图

Fig.1　The distribution of water sources and sampling points in study area

1.2　数据来源

本次研究利用2017年5月(枯水期)和2017年9月(丰水期)两个时期的地下水水源地水质数据进行研究区地下水质量评价研究(取样点位置分布如图1所示)。其中,枯水期采集水样109组,丰水期采集水样108组,共217组水样;水质检测与分析方法严格依据《地下水质量标准》(GB/T 14848—2017)的相关规定进行[13],并利用碳酸平衡、电荷平衡进行检验。根据研究区地下水污染现状,综合考虑不同指标对水质的影响,选取锌、氯化物、硫酸盐、氟化物、硝酸盐、COD、总硬度(以$CaCO_3$计)、溶解性总固体和PH值9个主要影响当地水质的污染指标作为评价因子。

2　研究方法

随着机器学习方法的兴起,人们开始广泛利用逻辑回归(LR)、支持向量机(SVR)、人工神经网络(ANN)和随机森林(RF)等方法进行分类模型的建立和训练[14-16]。其中,由加州大学伯克利分校的BREIMAN等在2001年提出的随机森林算法由于训练速度快、泛化能力强且不容易陷入过拟合而得到了广泛的应用[17]。随机森林算法通过集成学习的思想将多棵决策树合并到一起,它的基本单元就是决策树,核心思想是集成学习[18]。

2.1　决策树和集成学习

(1)决策树。决策树属于单学习器,其能够对大量无规则的样本进行递归分析并推导出以树状结构为表示形式的分类规则,最终实现对未知样本数据的分类和预测。决策树包括根节点、内部节点和叶节点,其中根节点表示样本数据总集,内部节点表示对象的属性,叶节点代表决策的结果[19]。分类时,在树的内部节点处对某一属性值进行判断,并根据判断结果决定进入哪个分支节点,直到到达叶节点处,得到分类结果。决策树示意图如图2所示。

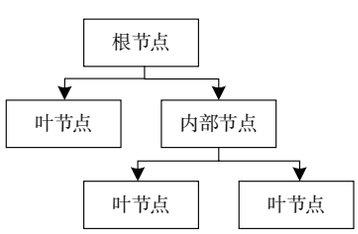

图2　决策树示意图

Fig.2　Schematic diagram of decision tree

(2)集成学习。集成学习算法的基本思想是将多个单学习器组合,形成一个预测效果更好的集成学习器。集成学习器能提高单学习器对离群点和异常值的处理能力,从而达到比单学习器更强的泛化能力。

2.2　随机森林算法

随机森林算法由一系列决策树单学习器组成,通过多个单学习器对输入样本进行简单投票并判断其分类,再集合各单学习器的结果得出随机森林的最终结果[20-21]。随机森林算法流程如图3所示。

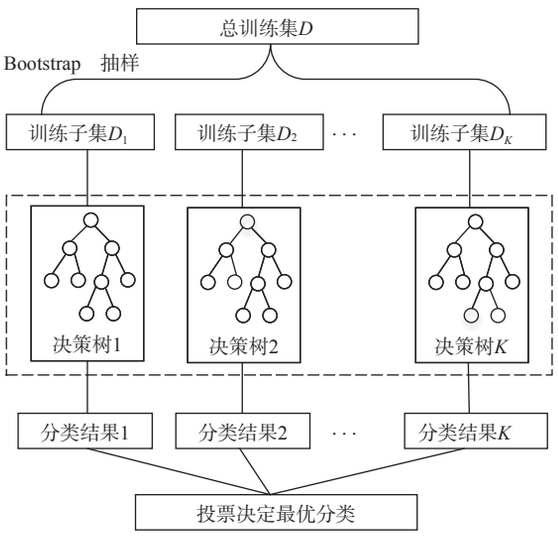

图3 随机森林算法示意图

Fig.3 Schematic diagram of random forest algorithm

(1)通过Bootstrap自助法从总训练集D中有放回的随机抽出K个训练子集($D_1, D_2, \cdots, D_k$),作为K棵决策树。

(2)在决策树的每个节点上从M个指标中随机选取m个指标($M>m$),再从m个指标中选取Gini指数最小的特征作为切分点进行分裂,让决策树充分生长直到每个叶子节点的不纯度(即Gini指数)达到最小,同时在分枝生长的过程中不进行剪枝的操作。Gini指数的计算过程如下所示:

$$Gini(K) = 1 - \sum_{i=1}^{k} p_i^2 \tag{1}$$

式中:k表示将训练子集分为k个不同的类别;p_i表示每个类别在第K个训练子集中出现的概率。需要说明的是在进行节点划分时,Gini指数越小,分类结果越纯粹,说明划分的结果越优;Gini指数越大,则代表分类结果越混乱;如Gini指数为0时,说明只有一个分类。而在衡量每个指标的重要性时,主要依据Gini指数的变化程度来反映每个指标的重要程度,Gini指数越大说明其重要性越大[25]。

(3)重复步骤(2)遍历预建的K棵决策树,并由此K棵决策树组成随机森林。

(4)采用投票的方式确定待测样本的分类结果,K棵决策树所得票最多的类为最终的类别。其分类公式为:

$$H(X) = \arg\max \sum_{i=1}^{K} I\left(h_i(x) = Y\right) \tag{2}$$

式中:$H(X)$是最终的分类结果;argmax代表求取函数最大值自变量点集;Y为目标分类;$h_i(x)$代表单棵决策树的分类结果;$I()$为示性函数。

3 模型构建及评价结果

3.1 训练数据集及测试数据集

训练样本的好坏直接决定着随机森林算法进行地下水质量评价的准确度。本次研究利用《地下水质量标准》(GB/T 14848—2017)中推荐使用的F值评价法生成训练样本。利用F值评价法对研究区的217组水质样品进行地下水质量等级划分,结果如图4所示。通过图4可以看出研究区内没有Ⅰ类、Ⅱ类和Ⅲ类地下水,其中Ⅳ类水质的数量为69组,Ⅴ类水质的数量为148组,说明研究区地下水污染整体较为严重。

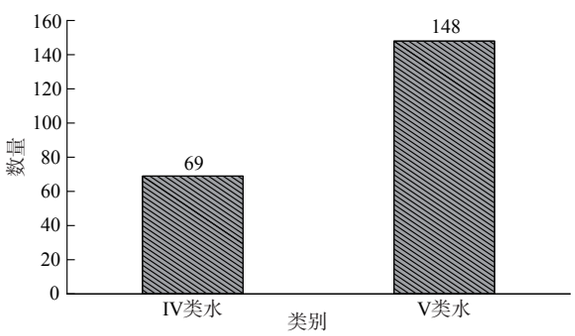

图4 地下水质量等级划分

Fig.4 Groundwater quality classification

从217组水质样品中随机抽取70%(152组)作为训练数据集,将没有被抽中的剩余30%(65组)作为测试数据集,以此考察随机森林水质分类模型的准确性和稳定性。模型训练过程中,将训练数据集中每个水样的9个水质评价因子及F值评价法水质分类结果输入到随机森林模型,通过不断调节优化模型参数,使模型的水质分类模拟值与F值评价法水质分类结果尽可能地相符合,完成对随机森林分类模型的训练优化。利用训练好的模型,输入测试数据集中每个水样的9个水质评价因子,即可实现对地下水水质的分类模拟。

3.2 随机森林模型超参数优化

为了防止随机森林模型在决策树构建过程中出现超参数随机选择及过拟合等问题,本研究首先利用sklearn机器学习库中的网格搜索算法对模型中的相关超参数(决策树数量、决策树最大深度和最大特征数等)进行参数优化,以此提高随机森林模型的分类精度和效率。网格搜索算法是一种模

型超参数优化技术,通过遍历给定的参数组合来优化模型表现。

随机森林算法中决策树的数量通常是越多效果越好,但如果决策树的数量过多,会带来较大的计算负担,计算时间也会随之相应的增加;同时,当决策树的数量达到一个临界值后,再增加树的数量模型的分类效果并不会很显著地提升。网格搜索算法对决策树数量的优化结果如图5所示。通过图5可以看出,随着决策树数量的增加误差不断减小;当决策树数量达到60时,误差变化基本处于稳定状态,并且当数量为70时误差最小。

利用网格搜索算法进一步对其它超参数进行优化,优化结果如表1所示。

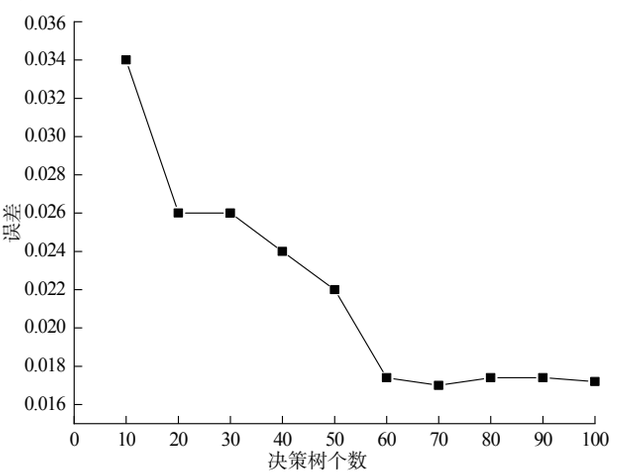

图5　决策树个数与误差的关系

Fig.5　The relationship between the number of decision trees and error

表1　随机森林超参数优化结果

Table1　Random forest hyperparameter optimization results

超参数	数值	超参数	数值
决策树数量	70	决策树最大深度	3
最大特征数	3	内部节点再划分所需最小样本数	50
叶子节点最少样本数	20		

3.3　评价结果及分析

为了验证超参数优化后随机森林模型的分类效果,分别利用未进行超参数优化的随机森林模型和逻辑回归模型对测试数据集进行地下水水质分类,并对分类结果进行比较(见图6)。逻辑回归(Logistic Regression)是一种有监督的统计学习方法,主要用于对样本进行分类,是机器学习中做分类任务常用的方法。逻辑回归由于模型简单、可解释性和可扩展性强而被广泛应用于数据挖掘,疾病诊断,经济预测等领域[23]。

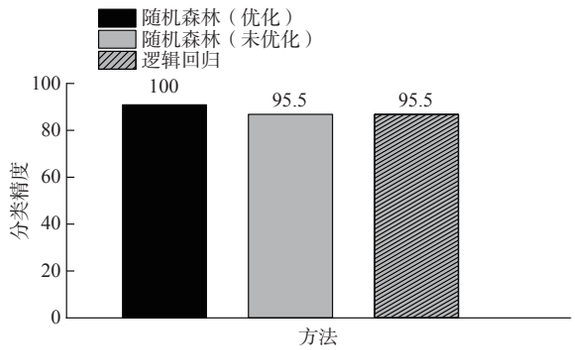

图6　不同方法的分类精度

Fig.6　Classification accuracy of different methods

通过图6可以看出,未进行超参数优化的随机森林模型和逻辑回归模型的分类精度均为95.5%,而超参数优化后随机森林模型的分类精度最高,能够达到100%的分类正确率。分析其原因主要是随机森林模型通过超参数优化后可以选择最佳的决策树数量,防止冗余决策树所引起的过拟合现象的发生,因此能够有效提高模型的分类精度;此外,虽然未优化随机森林模型与逻辑回归模型的分类精度均为95.5%,但逻辑回归模型需要更多的计算量才能获得较好的分类效果,计算效率较低。通过上述分析可以看出,进行超参数优化后的随机森林模型能够很好地处理地下水质量评价中的多特征分类问题,并且具有较高的效率和准确度。

3.4　分类指标重要性评价

通过地下水质量评价能够识别出研究区的地下水污染状况,但无法揭示出不同评价指标之间的相对重要性。随机森林模型的一个很明显的优势就是可以通过Gini指数来评估每个指标对于水质分类结果的重要性。Gini指数越大,表示评价指标的相对重要程度越高。研究区地下水水质评价指标的相对重要性如图7所示。

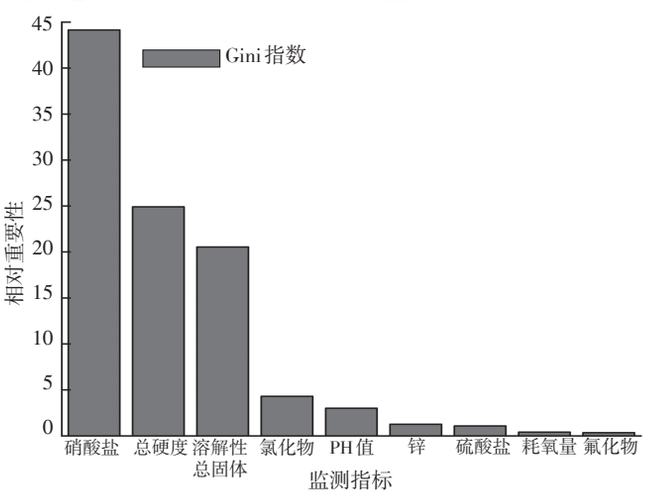

图7　地下水水质指标的相对重要性排序

Fig.7　Ranking of relative importance of groundwater quality indicators

从图7可以看出，研究区地下水水质评价指标的重要性排序为硝酸盐>总硬度>溶解性总固体>氯化物>PH值>锌>硫酸盐>耗氧量>氟化物。其中，硝酸盐、总硬度和溶解性总固体的重要性最为突出，其Gini指数均超过了20%，是研究区地下水质量的主控因素；其它几种评价指标的Gini指数均小于5%，说明重要性较低。硝酸盐的Gini指数为44.14%，说明其重要性最为突出；研究区是我国重要的蔬菜生产基地，蔬菜种植过程中化肥的过度使用造成大量氮肥进入地下水造成严重的硝酸盐污染。总硬度和溶解性总固体的Gini指数分别为24.9%和20.54%；研究区中溶解性总固体和总硬度呈现出随河流流向不断增加的趋势，分析其原因是区内地下水普遍接受河水补给，地表污染物通过河流进入地下水体逐渐累积的结果。因此要进一步加强对蔬菜种植污染排放及河流水质的监控，防止研究区地下水水质的恶化。

4 结论

本文构建了一个进行地下水质量评价的随机森林分类模型，并从弥河-潍河流域的217组地下水水质样品中随机抽取70%（152组）作为训练数据集，将剩余的30%（65组）数据作为测试数据集；此外，为了防止过拟合的出现，提高模型的分类精度和计算效率，采用网格搜索算法对随机森林模型的超参数进行了优化，得出如下结论：

（1）随机森林模型在进行地下水水质分类时具有分类精度高及泛化能力强等特点，且在进行超参数优化后，其分类精度会进一步提高，证明将随机森林模型应用于地下水质量评价是可行的，并且其综合性能要优于逻辑回归模型。

（2）此外，利用训练好的随机森林水质分类模型，工作人员在不需要具备丰富地下水水质评价知识的前提下，可以比较方便地进行地下水质量评价，更加具有实用性。

（3）研究区地下水水质状况整体较差，均为Ⅳ类和Ⅴ类水。通过分类指标重要性评价可以看出，研究区地下水水质的主要影响指标为硝酸盐、总硬度和溶解性总固体；其中，硝酸盐的主要来源是蔬菜种植过程中化肥的过度使用造成大量氮肥进入地下水造成污染；溶解性总固体和总硬度则主要是由于地表污染物通过河流进入地下水体逐渐累积的结果。因此要进一步加强对蔬菜种植污染排放及河流水质的监控，防止研究区地下水水质的恶化。

需要指出的是本研究的水质样本仅有Ⅳ类和Ⅴ类两种类别，无法充分验证随机森林模型进行水质评价时的分类效果，因此在以后的研究中要采用更加全面的训练数据，此外如何克服随机森林算法特征选择的随机性及进一步提升算法的通用性是未来工作所要重点解决的问题。

参考文献：

[1] 方运海, 郑西来, 彭辉, 等. 基于模糊综合优化模型的地下水质量评价[J]. 地学前缘, 2019, 26(4):301-306.

[2] 姬亚琴, 杨鹏年. 孔雀河流域包头湖农场地下水质量评价[J]. 人民黄河, 2015(10):86-88.

[3] 梁乃森, 钱程, 穆文平,等. 大牛地气田区地下水水质模糊综合评价[J]. 水文地质工程地质, 2020, 47(3): 52-59.

[4] 李松青, 王心义, 姬红英, 等. 基于遗传算法-BP神经网络法的深埋地下水水质评价[J]. 水电能源科学, 2019, 37(1):49-52,16.

[5] EGBUERI J C. Groundwater quality assessment using pollution index of groundwater (PIG), ecological risk index (ERI) and hierarchical cluster analysis (HCA): A case study[J]. Groundwater for Sustainable Development, 2020, 10: 100292.

[6] SU F, WU J, HE S. Set pair analysis-Markov chain model for groundwater quality assessment and prediction: A case study of Xi'an city, China[J]. Human and Ecological Risk Assessment: An International Journal, 2019, 25(1-2): 158-175.

[7] 康小兵, 李科, 朱志强, 等. 基于融合权重的云模型在西昌某地区地下水水质评价中的应用[J]. 节水灌溉, 2019(7):62-67.

[8] 闫滨, 姜秀慧, 钟占华,等. 基于改进权重的综合水质标识指数法的大伙房水库上游水质评价研究[J]. 沈阳农业大学学报, 2019(3): 314-323.

[9] CHENG L, CHEN X, DE Vos J, et al. Applying a random forest method approach to model travel mode choice behavior[J]. Travel behaviour and society, 2019(14): 1-10.

[10] PAUL A, MUKHERJEE D P, DAS P, et al. Improved random forest for classification[J]. IEEE Transactions on Image Processing, 2018, 27(8): 4012-4024.

[11] 杨光, 粟晓玲. 基于随机森林的黑河中游地下水埋深变化及成因[J]. 水土保持研究, 2017, 24(1):109-114.

[12] BAUDRON P, ALONSO S F, GAECIA J L et al. Identifying the origin of groundwater samples in a multi-layer aquifer system with Random Forest classification[J]. Journal of Hydrology, 2013(499): 303-315.

[13] GB/T 14848—2017, 地下水质量标准[S]. 北京: 中国标准出版社, 2018.

[14] MUELLER N, LEWIS A, ROBERTS D, et al. Water observations from space: Mapping surface water from 25 years of Landsat imagery across Australia[J]. Remote Sensing of Environment, 2016, 174: 341-352.

[15] SINGH K P, BASANT N, GUPTA S. Support vector machines in water quality management[J]. Analytica chimica acta, 2011, 703(2): 152-162.

（下转第70页）