DOI:10.19651/j.cnki.emt.2211703

基于混合策略改进的 LightGBM 土壤污染预测模型*

李文杰 王占刚
(北京信息科技大学信息与通信工程学院 北京 100101)

摘 要: 社会经济快速发展,然而土壤污染中重金属污染所占的比例越来越大,对生态环境和人们的生命健康造成了巨大的威胁。针对以上问题,提出一种基于混合策略改进的土壤重金属污染预测模型,即先采用随机森林选出最优特征子集,再利用随机搜索对 LightGBM 参数进行优化,最后通过训练得到的 LightGBM 模型预测土壤的内梅罗综合污染指数,从而得出土壤重金属污染状况。以我国华北平原的某区域为研究区,并与 RS-LightGBM、LightGBM、SVR 模型的预测结果进行对比。结果表明,所提模型的均方误差、平均绝对误差相比于 LightGBM 模型分别降低了 69.09%、39.09%,决定系数相比于 LightGBM 模型提高了 6.11%。上述结果表明本论文提出的模型可以有效应用于土壤重金属污染预测研究中。

关键词: 污染预测;LightGBM;随机森林;随机搜索
中图分类号: TP391;TN **文献标识码:** A **国家标准学科分类代码:** 520.20

Improved LightGBM soil pollution prediction model based on mixed strategy

Li Wenjie Wang Zhangang
(School of Information and Communication Engineering, Beijing Information Science & Technology University, Beijing 100101, China)

Abstract: With the rapid development of social economy, the proportion of heavy metal pollution in soil pollution is increasing, posing a huge threat to the ecological environment and people's life and health. Aiming at the above problems, this paper proposes a soil heavy metal pollution prediction model based on the improved mixing strategy, that is, the optimal subset of characteristics is selected by random forest, and then the LightGBM parameters are optimized by random search, and finally the Nemero comprehensive pollution index of the soil is predicted by the trained LightGBM model, so as to obtain the soil heavy metal pollution status. A certain area of the North China Plain in China is used as the research area, and the prediction results of RS-LightGBM, LightGBM and SVR models are compared. The results show that the mean squared error and mean absolute error of the proposed model are reduced by 69.09% and 39.09% respectively compared with the LightGBM model. The coefficient of determination is 6.11% higher than the LightGBM model. The above results show that the proposed model can be effectively applied to the prediction of soil heavy metal pollution.

Keywords: pollution prediction;LightGBM;random forest;random search

0 引 言

土壤作为环境的重要组成部分以及动植物生存生长的载体,其污染程度对人类生命健康起主导性作用[1]。近几十年,经济快速发展的同时也产生了大量重金属污染,重金属通过大气扩散、交通运输等方式传播从而造成严重的土壤污染,对生态、经济的健康发展造成巨大威胁。因此,重金属污染防治逐渐受到社会的广泛关注并成为近几年的研究热点。

由于土地管理的迫切需求,研究者们进行了大量关于污染预测的相关研究。何云山[2]使用基于随机森林(random forest, RF)和遗传算法改进的支持向量回归(support vector regression,SVR)算法对土壤重金属污染状况进行预测,但是 SVR 算法的性能受到核函数和惩罚因

收稿日期:2022-10-14
* 基金项目:国家重点研发计划课题(2018YFC1800203)、北京市科技创新服务能力建设-基本科研业务费(市级)(科研类)(PXM2019_014224_000026)项目资助

子的影响;任加国等[3]利用反向传播(back propagation,BP)神经网络模型,对样本中重金属和多环芳烃含量的部分缺失值进行预测,但是 BP 神经网络有有权值和阈值初值过于随机化、稳定性和准确性差等缺点;段嘉欣[4]利用多元统计分析方法中主成分分析法来确认土壤重金属污染的主要来源,但是该方法面对数据量大、数据复杂多变等问题,通过线性变换对原始数据进行降维,提取数据的主要特征,未考虑数据不符合高斯分布的特点,导致准确率较低,结果往往不尽如人意。罗一茗[5]采用地统计学的方法对土壤中的重金属进行评价和预测,然而传统的地统计学方法需要进行复杂的参数调整,难以构建合适的半变异函数和拟合研究对象的空间结构,从而导致插值精度低,插值结果不确定性大等问题;李杨[6]采用神经网络对土壤重金属含量预测及污染风险研究,但神经网络存在易陷入局部最优值、输出值太依赖于输入样本等缺陷。通过上述分析可以看出,现在形成了以神经网络和统计学为主的两类污染预测方法。但是神经网络训练需要消耗很长的时间,多个超参数会影响预测的精度,而且其输出结果难以解释;统计学方法需要对数据进行平稳性假设,与实际数据情况偏离较大,所以该方法不适合广泛使用。

为解决现有预测方法存在的问题,本文将高效率的轻量级梯度提升树(light gradient boosting machine,LightGBM)应用到土壤重金属污染预测领域[7],并且结合 RF 和随机搜索(randomized search,RS)构建出基于混合策略改进的 LightGBM 土壤重金属污染预测模型。该模型利用多种土壤重金属含量来预测该区域的内梅罗综合污染指数,最后与多种不同预测方法进行比较,从而验证模型的有效性[8]。实验结果表明所提模型的预测结果具备更高精度,并且极大缩短模型参数寻优时间,为土壤污染预测领域提供一种高效率的新方法。

1　研究方法

1.1　基本原理

RF 是通过有放回地从数据集中抽取 m 个样本,共抽取 N 次从而分别构建 N 个决策树[9]。每次抽取中未抽到的数据称为"袋外数据"(out-of-bag,OOB)。RF 特征选择就是利用袋外数据误差来度量特征变量的相对重要性,然后通过一定的方法对特征进行提取[10]。

LightGBM 是 2017 年微软提出的一个基于梯度提升决策树(gradient boosting decison tree,GBDT)的改进算法[11],具有训练速度快、占用内存少、准确率高等优点。LightGBM 在使用直方图算法(Histogram)寻找最佳分裂点的基础上,通过单边梯度采样算法(gradient-based one-side sampling,GOSS)和互斥特征绑定算法(exclusive feature bundling,EFB)来降低训练学习过程中样本数量和特征数量[12]。其中 EFB 算法将互斥的特征捆绑在一起,从而降低特征维度[13]。

RS 由 Bergstra 等[14]在 2012 年提出。该算法不需逐个计算所有可能参数组合,只需随机选择每个超参数的可能值并将其随机组合,计算随机抽取的超参数组合即可[15],计算量和寻优时间大大减少。

1.2　基于混合策略改进的 LightGBM 预测模型构建

综合前文对数据处理方法和预测模型的介绍,本文所提出的污染预测模型,如图 1 所示。模型的具体实现步骤如下:

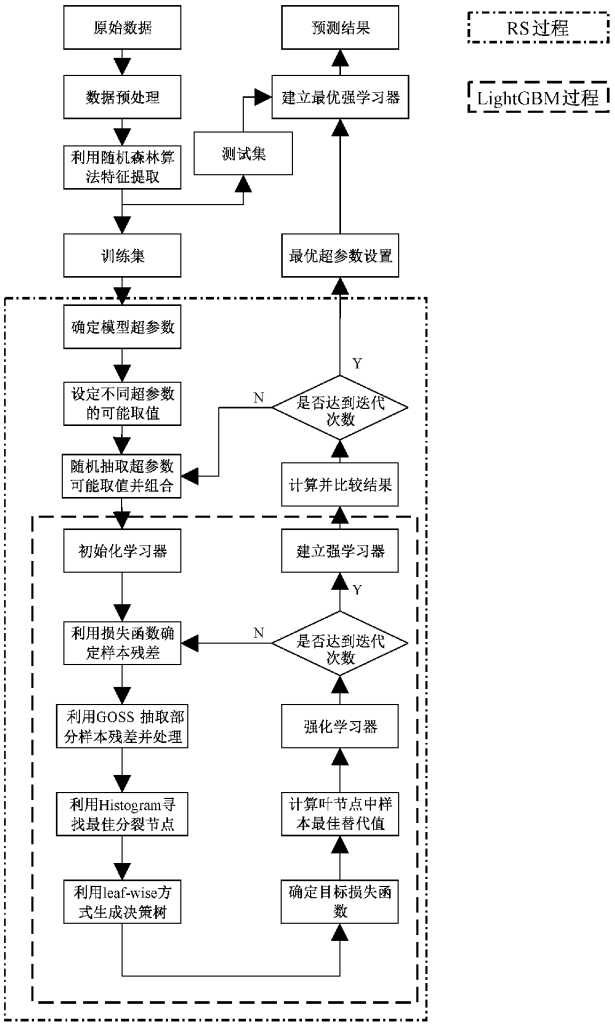

图 1　基于混合策略改进的 LightGBM 模型流程

1)将原始数据进行预处理,包括对缺失值、异常值的处理。

2)采用 RF 提取出新的特征集合,形成新的数据集。首先计算每个特征的重要性:

$$imp = \frac{1}{N}\sum_{i-1}^{N}(err2_{ji} - err1_i) \tag{1}$$

其中,$err1_i(i = 1,2,3,\cdots,N)$ 表示使用 N 组 OOB,分别计算每棵决策树的误差;$err2_{ji}(i = 1,2,3,\cdots,N)$ 表示在其他特征不变的情况下,对第 j 个特征添加噪声,并计算每棵决策树的误差。然后将特征根据重要性降序排列,

并从居于首位的特征开始逐次加入下一个特征，计算每一个特征向量组合的袋外误差，其计算公式如下：

$$err = \frac{1}{N}\sum_{i=1}^{N} err_i \tag{2}$$

其中，err_i 表示利用每一次特征组合后的 OOB 计算得到的每棵决策树的误差。重复上一步，直到所有的特征都参与完毕。最后将使用 err 最小的特征组合作为最终选择特征[16]。

3)将新的数据集分为训练集和测试集。

4)使用训练集对 LightGBM 进行训练，并通过 RS 确定模型的最佳超参数组合，得到最终的模型。

假设 X^s 为输入空间，其中 s 表示特征维度，包括 Pb、Zn、Cu 等特征；$\{x_1, x_2, \cdots, x_n\}$ 为 n 个独立同分布实例，其中 $\boldsymbol{x}_i$ 是 s 维向量，且属于 X^s 空间。在每次梯度提升迭代中，模型输出损失函数负梯度表示为 $\{g_1, g_2, \cdots, g_n\}$。决策树使用信息量最大(信息增益最大)的特征值将数据分割到左右子节点。对于 GBDT，信息增益通常用分裂后方差来测量[17]。假设 O 为决策树某一固定节点上的训练数据集，则在该节点上特征 j 且特征值为 d 处的方差增益表示为：

$$V_{j|O} = \frac{1}{n_o}\left(\frac{\left(\sum\limits_{x_i \in O: x_{ij} \leqslant d} g_i\right)^2}{n^j_{l|O}(d)} + \frac{\left(\sum\limits_{x_i \in O: x_{ij} > d} g_i\right)^2}{n^j_{r|O}(d)}\right) \tag{3}$$

其中，$n_o = \sum I(x_i \in O)$；$n^j_{l|O} = \sum I[x_i \in O: x_{ij} \leqslant d]$；$n^j_{r|O} = \sum I[x_i \in O: x_{ij} > d]$。

遍历每一个特征值，特征 j 决策树算法会选择特征值 $d_j^* = \arg\max_d V_j(d)$，并且计算最大的增益 $V_j(d_j^*)$，然后将数据根据特征 j 的分割点 d_j^* 分割到左右子节点。

此时，使用 GOSS 算法降低样本数量，首先将训练实例根据梯度进行降序排列。保留训练实例前 $a\%$ 大梯度实例，组成一个新实例集 A。其次对于剩余 $(1-a\%)$ 实例 A^c，使用随机抽取方法，从中抽取 $b^* \mid A^c \mid$ 个实例形成新实例集 B，从而降低样本数量。最后根据方差增益 $\tilde{V}_j(d)$ 对新组成的实例集 $A \cup B$ 进行拆分。$\tilde{V}_j(d)$ 计算公式如下：

$$\tilde{V}_j(d) = \frac{1}{n}\left(\frac{\left(\sum\limits_{x_i \in A_l} g_i + \frac{1-a}{b}\sum\limits_{x_i \in B_l} g_i\right)^2}{n^j_l(d)} + \frac{\left(\sum\limits_{x_i \in A_r} g_i + \frac{1-a}{b}\sum\limits_{x_i \in B_r}^{x_i \in B_r} g_i\right)^2}{n^j_r(d)}\right) \tag{4}$$

其中，$A_l = x_i \in A: x_{ij} \leqslant d$；$A_r = x_i \in A: x_{ij} > d$；$B_l = x_i \in A: x_{ij} \leqslant d$；$B_r = x_i \in A: x_{ij} > d$；系数 $\frac{1-a}{b}$ 是修正因子，用来放大小梯度实例带来的信息增益。

其中，a、b 就是 LightGBM 模型超参数的一部分，同时还要确定模型的其他超参数，如 n_estimators、min_samples_split、min_samples_leaf 等。此时使用 RS 得到 LightGBM 的最优超参数组合，给每个超参数设定各自可能取值；然后从每个超参数可能取值中随机抽取并组合，组合结果作为本次 LightGBM 的参数；最后使用五折交叉验证获得的均方误差，并将其作为此次 RS 结果的评价标准，在达到训练次数时，将均方误差最小的参数组合作为最优参数组合。

5)使用测试集进行测试，验证模型的预测效果。

1.3　模型性能评价指标

为了对土壤重金属污染预测模型性进行全面评价，本实验选取决定系数(R^2)、均方误差(mean squared error，MSE)和平均绝对误差(mean absolute error，MAE)3 个指标[18]，其计算公式分别为：

$$R^2 = 1 - \frac{\sum\limits_{i=1}^{n}(f_i - y_i)^2}{\sum\limits_{i=1}^{n}(\bar{y}_i - y_i)^2} \tag{5}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(f_i - y_i)^2 \tag{6}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\mid f_i - y_i \mid \tag{7}$$

其中，y_i 为实测值；f_i 为预测值；$\bar{y}_i$ 为均值；n 为实测值个数。

2　实验与结果分析

2.1　数据预处理

研究数据集源于 2020 年华北平原某区域采集的 322 个样本点，每个样本点包含 8 类土壤重金属含量，这 8 类土壤重金属分别是 Cd、Hg、As、Pb、Cr、Cu、Zn、Ni。

使用 Python 中 pandas 库的 info 函数对数据统计分析，发现数据集不存在缺失值。使用箱形图检查数据中是否有异常值，其判断标准为：把小于 $Q_1 - 1.5IQR$ 或大于 $Q_3 + 1.5IQR$ 的值判定为异常值[19]。其中，Q_1 和 Q_3 分别为 25%分位数和 75%分位数；$IQR = Q_3 - Q_1$。

图 2 为数据处理前后的箱形图。未处理数据箱形图如图 2(a)所示，从图中可以看出数据集中存在较多异常值，有许多数据在大于 $Q_3 + 1.5IQR$ 的位置。实验中把异常值当作缺失值，并使用均值插补法处理。如图 2(b)所示，为异常值处理之后的数据箱形图，可以看出经处理后的数据大部分处于上下边缘之内，仅存在少量异常值聚集在上下边缘之外近距离范围内。

为减少数量级较大和量纲较多带来的影响，对数据集做标准化处理：

$$X_{ij} = \frac{x_{ij} - \overline{x}_i}{\overline{\sigma}_i}, i = 1,2,3,\cdots,n; j = 1,2,3,\cdots,p \tag{8}$$

其中，x_{ij} 为第 i 个样本的第 j 个金属的含量值；$\overline{x}_i$ 和 $\overline{\sigma}_i$ 分别为第 i 个样本的所有指标的样本均值和标准差。处理所得的部分数据，如表 1 所示。

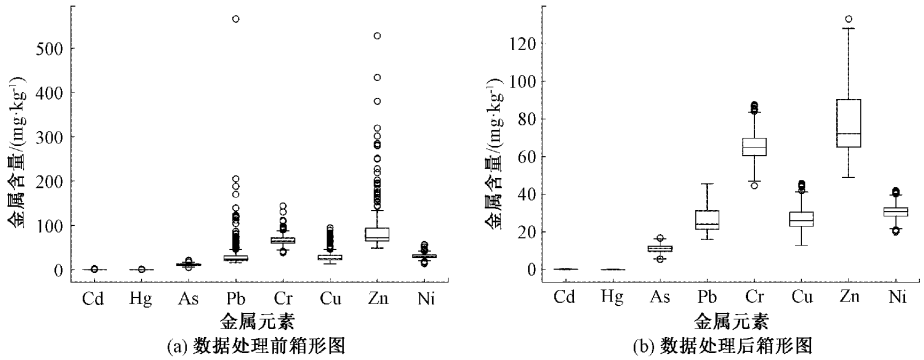

图 2 数据处理前后箱形图

表 1 标准化处理之后的数据

序号	Cd	Hg	As	Pb	Cr	Cu	Zn	Ni
0	0.541 617	3.164 893	−1.127 840	1.168 181	2.456 910	0.824 415	0.476 438	1.149 031
1	−1.092 112	−0.840 986	0.572 049	−1.334 248	−0.176 877	−0.409 655	−0.576 726	0.558 350
2	−0.683 680	−0.306 869	−1.263 271	−0.217 092	−1.287 199	−1.264 952	−0.594 989	−0.942 965
3	−0.071 032	−1.108 044	−0.861 649	−0.568 198	−0.228 520	1.015 838	1.176 518	0.706 020
4	0.133 184	−0.573 927	0.618 749	−0.823 548	−0.086 502	−0.022 736	−0.461 060	0.509 126

2.2 特征提取

经过 RF 特征选择,首先对 8 种土壤重金属元素进行评分,得到重要性评分结果,如图 3 所示,重要性由高到低依次是 Pb、Zn、Cu、As、Cd、Ni、Cr、Hg。通过了解当地的实际情况,与本文研究结果对比可以看出,模型得到的重要性特征与实际结果契合。

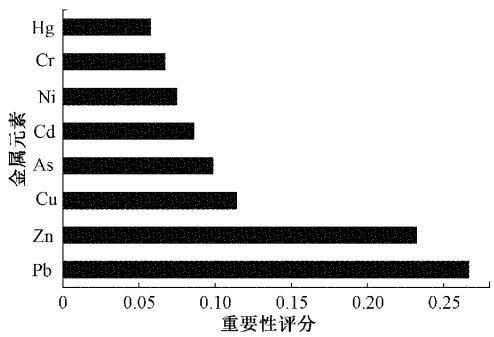

图 3 不同重金属的重要性系数

其次对特征进行选择。当特征向量组合为:Pb、Zn、Cu、As 时,其 *err* 最小。经查阅资料和结合当地情况得知 Pb、Zn、Cu、As 对该地区土壤重金属污染贡献率较高,所以选择特征 Pb、Zn、Cu、As。

2.3 模型参数调优

经过特征提取后,将 Pb、Zn、Cu、As 组成最优特征子集。将数据集按 17∶3 的比例随机划分为 274 组训练集和 48 组测试集。

基于混合策略改进的 LightGBM 预测模型中的 RS,使用 Scikit learn 中 RandomizedSearch。同时使用网格搜索(GridSearch,GS)作为对比实验,验证 RS 的高效性[20]。分别给参数 n_estimators、min_samples_split、min_samples_leaf、max_features、max_depth 设置可能的取值范围,经过 RS 和 GS 优化,分别得到模型最优参数组合,其他参数采用算法默认值,如表 2 所示。

使用不同寻优方法所得最优参数组合分别构建不同 LightGBM 预测模型,在测试集上预测,得到如表 3 所示结

表 2 LightGBM 参数及优化值

参数	参数范围	优化值	
		RS	GS
max_depth	[5, 10, 15, 20, 25, 30]	15	10
max_features	['auto','sqrt']	'auto'	'auto'
min_samples_leaf	[1,2,5,10]	5	1
min_samples_split	[2,5,10,15,100]	100	2
n_estimators	[100,200,300,400,500,600,700,800,900,1 000, 1 100, 1 200]	300	100

果。可以看出 RS-LightGBM 相比于 GS-LightGBM：MSE 和 MAE 两个指标分别降低 15.0%、1.66%，R^2 指标提高 0.52%，且 GS 使用时间是 RS 使用时间的 246.99 倍。分析原因，GS 在所有可能的参数组合里寻找使得预测结果最好的参数组合，所以使用的时间比较长，准确度比较低。综合比较运行时间和准确率，认为 RS 为最优的参数优化方法。

表 3　不同优化方法得到模型预测结果比较

模型	MSE	MAE	R^2	所用时间
RS-LightGBM	0.001 7	0.029 6	0.973 9	5.49 s
GS-LightGBM	0.002 0	0.030 1	0.968 9	22 min 36 s

2.4　模型结果分析

构建 RS-LightGBM、LightGBM、SVR 模型，并将其与所提模型作对比。为提高模型对比公平性和可靠性，4 个模型使用相同训练集和测试集，且训练集和测试集之比为 17∶3。将它们的超参数设置方法概括如下：

基于混合策略改进的 LightGBM 和 RS-LightGBM 使用 RS 获得超参数。通过实验发现，它们 RS 获得的超参数是一样的，其中 n_estimators 为 300，min_samples_split 为 100，min_samples_leaf 为 5，max_features 为'auto'，max_depth 为 15，其他参数采用算法默认值。

LightGBM 模型和 SVR 模型根据算法特性和调参经验进行人工调参，每一个参数调整到其值增大或减小都会降低模型预测精度时停止。对于 LightGBM 模型，num_leaves 为 29，subsample 为 0.6，feature_fraction 为 0.7，learning_rate 为 0.01，其他参数采用算法默认值。对于 SVR 模型，kernel 为'rbf'，C 为 0.06，gamma 为'auto'，degree 为 60，其他参数采用算法默认值。基于混合策略改进的 LightGBM、RS-LightGBM、LightGBM、SVR 模型的内梅罗综合污染指数预测结果对比如图 4 所示。

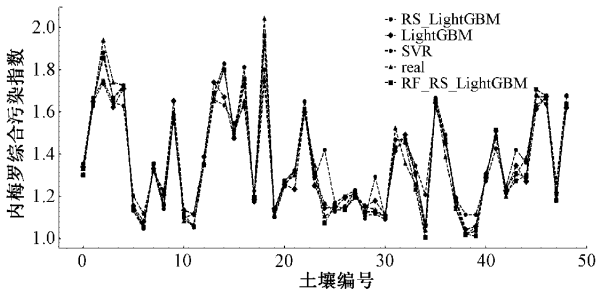

图 4　内梅罗综合污染指数预测结果对比

从图 4 的对比结果图可以清晰的看出，经过超参数优化以及对数据集特征提取之后获得的基于混合策略改进的 LightGBM 模型（RF-RS-LightGBM），其预测效果相比 LightGBM 模型以及传统的 SVR 模型提高很多，其预测结果更接近于标准值 real，具有更好的拟合效果。证明 RS 得到 LightGBM 的最优参数组合，解决了人工寻找 LightGBM 多个超参数的主观性，缩短了寻参时间，提高了预测的准确率。同时，RF 特征提取减少了数据之间的冗余，降低了不同数据之间的相关性，防止预测模型过拟合，并且通过降维加快了整个模型的运行速度。

为了评估基于混合策略改进的 LightGBM 模型与其他预测模型的性能，本文分别从 R^2、MSE 和 MAE 进行模型评估，如表 4 所示。由表 4 可知，在 R^2、MAE 及 MSE 指标中，基于混合策略改进的 LightGBM 预测模型（RF-RS-LightGBM）均优于 RS-LightGBM、LightGBM、SVR 模型。首先 LightGBM 模型相比 SVR 模型：MSE 和 MAE 两个指标分别降低 33.73%、26.25%，R^2 指标提高 4.77%，证明 LightGBM 模型相比于传统的土壤重金属预测模型性能有所提高。经过 RS 得到 RS-LightGBM 模型相比 LightGBM 模型：MSE 和 MAE 两个指标分别降低 60.0%、31.89%；R^2 指标提高 5.27%，证明 RS 相比于人工调参准确率有大幅提升。基于混合策略改进的 LightGBM 模型相比 RS-LightGBM 模型：MSE 和 MAE 两个指标分别降低 22.73%、10.57%；R^2 指标提高 0.80%，证明经过特征提取解决了多特征之间高相关性产生的高维问题，提高了预测精度。上述分析表明基于混合策略改进的 LightGBM 模型预测精度最高，更具有可靠性。

表 4　不同预测模型结果比较

模型	MSE	MAE	R^2
RF-RS-LightGBM	0.001 7	0.029 6	0.973 9
RS-LightGBM	0.002 2	0.033 1	0.966 2
LightGBM	0.005 5	0.048 6	0.917 8
SVR	0.008 3	0.065 9	0.876 0

为进一步直观说明各个模型在土壤重金属污染预测方面表现，将基于混合策略改进的 LightGBM 模型等预测结果的绝对误差使用箱形图表示，如图 5 所示。从图 5 中可以看出，基于混合策略改进的 LightGBM 模型预测绝对误差整体小于其他模型，且所有绝对误差集中分布在 0.10 内。而 RS-LightGBM、LightGBM 和 SVR 的 Q_3 和最大值

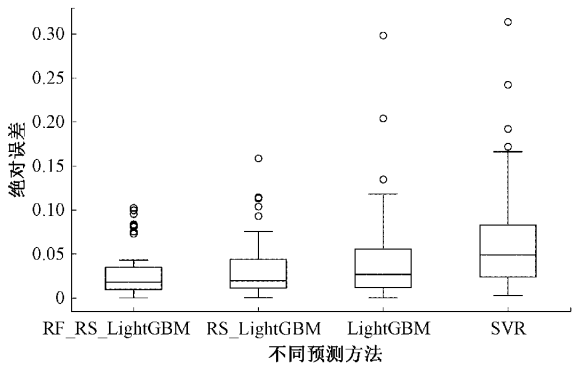

图 5　不同模型绝对误差箱形图

都大于基于混合策略改进的 LightGBM,且呈递增关系。进一步证明基于混合策略改进的 LightGBM 模型预测效果优于其他模型。

3 结 论

本文提出基于混合策略改进的 LightGBM 的土壤重金属污染预测模型,该模型改善了现有模型准确率低,易过拟合,调参耗时等问题,并通过实验得到如下结论:

与 GS 相比,RS 所得效果优于 GS,并且时间是它的 1/246.99,从而减少了不必要目标函数评估,提高了参数搜索效率。

RF 可对输入参数重要性进行评估,对影响土壤重金属污染特征进行筛选,从而确定对土壤重金属污染影响更显著的因素。经过 RF 特征提取之后的 RS-LightGBM 模型相比 RS-LightGBM 模型:MSE 和 MAE 两个指标分别降低 22.73%、10.57%;R^2 指标提高 0.80%,从而为后续污染评估、防治工作提供参考。

提出将基于混合策略改进的 LightGBM 模型应用于土壤重金属污染预测,与传统 LightGBM 模型相比:MSE 和 MAE 两个指标分别降低 69.09%、39.09%;R^2 指标提高 6.11%,所提模型预测精度最高,在 3 种定量评估指标中均表现最佳。

参考文献

[1] 费徐峰,任周桥,楼昭涵,等.基于贝叶斯最大熵和辅助信息的土壤重金属含量空间预测[J].浙江大学学报(农业与生命科学版),2019,45(4):452-459.

[2] 何云山.区域土壤重金属污染预测模型研究与应用[D].北京:北京信息科技大学,2021.

[3] 任加国,龚克,马福俊,等.基于 BP 神经网络的污染场地土壤重金属和 PAHs 含量预测[J].环境科学研究,2021,34(9):2237-2247.

[4] 段嘉欣.川南某地富硒土壤重金属污染特征及预测预警[D].绵阳:西南科技大学,2021.

[5] 罗一著.基于地统计学的鄱阳湖经济区城市土壤中重金属污染评价与预测[D].南昌:南昌大学,2019.

[6] 李杨.基于神经网络的土壤重金属含量预测及污染风险研究[D].昆明:昆明理工大学,2017.

[7] 张天一,苏华,杨欣,等.基于 LightGBM 的全球海洋次表层温盐遥感预测[J].遥感学报,2020,24(10):1255-1269.

[8] 余东昌,赵文芳,聂凯,等.基于 LightGBM 算法的能见度预测模型[J].计算机应用,2021,41(4):1035-1041.

[9] 姚锐,惠萌,李俊,等.基于随机森林的局部放电特征提取和优选研究[J].华北电力大学学报(自然科学版),2021,48(4):63-72.

[10] 陈维刚,张会林.基于 RF-LightGBM 算法在风机叶片开裂故障预测中的应用[J].电子测量技术,2020,43(1):162-168.

[11] KE G L, MENG Q, FINLEY T, et al. LightGBM: A highly efficient gradient boosting decision tree[C]. Proceedings of the 31st Annual Conference on Neural Information Processing Systems(NIPS), Long Beach, California: Curran Associates Inc,2017:3149-3157.

[12] 高治鑫,包腾飞,李扬涛,等.基于贝叶斯优化 LightGBM 的大坝变形预测模型[J].长江科学院院报,2021,38(7):46-50,57.

[13] 刁宁昆,马怀祥,刘锋.一种改进 LeNet5 结合 LightGBM 的滚动轴承故障诊断方法[J].国外电子测量技术,2022,41(1):140-145.

[14] BERGSTRA J, BENGIO Y. Random search for hyper-parameter optimization [J]. Journal of Machine Learning Research,2012,13(2):281-305.

[15] 孙永壮,黄鋆.多任务深度学习技术在储层横波速度预测中的应用[J].地球物理学进展,2021,36(2):799-809.

[16] 王斌,何丙辉,林娜,等.基于随机森林特征选择的茶园遥感提取[J].吉林大学学报(工学版),2022,52(7):1719-1732.

[17] 石欣,田文彬,冷正立,等.基于 CFD 和 LightGBM 算法的建筑室内温度全局预测模型[J].仪器仪表学报,2021,42(1):237-247.

[18] 谷宇峰,张道勇,阮金凤,等.一种用于油气储量评估中渗透率预测新模型[J].地球物理学进展,2022,37(2):588-599.

[19] 马良玉,於世磊,赵尚羽,等.基于随机搜索算法优化 XGBoost 的过热汽温预测模型[J].华北电力大学学报(自然科学版),2021,48(4):99-105.

[20] 孙斌,储芳芳,陈小惠.基于贝叶斯优化 XGBoost 的无创血压预测方法[J].电子测量技术,2022,45(7):68-74.

作者简介

李文杰(通信作者),硕士研究生,主要研究方向为深度学习、机器学习、数据挖掘分析等。

E-mail:3484999798@qq.com

王占刚,博士,教授,主要研究方向为时空模型与可视化、数据挖掘分析等。