基于随机森林的混凝土抗压强度预测模型研究

肖丙刚 [1,*]，吴典承 [1]，屈亦挺 [1]，刘小帅 [2]，王承宇 [2]

[1] 中国计量大学信息工程学院浙江省电磁波信息技术与计量学重点实验室，浙江杭州 310018

[2] 中才邦业 (杭州) 智能技术有限公司，浙江杭州 310000

摘要： 针对传统混凝土立方体抗压强度标准检验法受外界影响较大的问题，本研究提出了一种基于随机森林算法的混凝土抗压强度预测模型，通过决定系数等指标进行预测准确度的量化分析，最终优化算法使得决定系数的值达到 0.9 以上。本文首先将水泥、高炉渣掺量、粉煤灰掺量、水含量、粗骨料、细骨料含量以及龄期等数据作为原材料指标进行优化预处理，以优化后的数据集作为输入数据集，并划分数据集用以构建模型，然后通过定义参数网络，执行网格搜索交叉验证来进行模型参数的优化，构建完善各类模型，最后将基于随机森林算法模型的预测结果与基于支持向量回归算法模型、基于决策树算法模型的预测结果进行准确度比较。结果显示，基于随机森林算法的预测模型预测准确度（R^2=0.91547）远高于支持向量回归模型（R^2=0.76802）和决策树模型（R^2=0.87539），并且误差较小（RMSE=5.24087），表明在混凝土抗压强度预测方面，随机森林算法有着极大的优越性，这对于混凝土配料比设计具有重要意义。

关键词： 随机森林；混凝土；抗压强度；预测

DOI: 10.57237/j.cst.2023.04.003

Research on Prediction Model of Concrete Compressive Strength Based on Random Forest

Xiao Binggang[1,*], Wu Diancheng[1], Qu Yiting[1], Liu Xiaoshuai[2], Wang Chengyu[2]

[1]Key Laboratory of Electromagnetic Wave Information Technology and Metrology of Zhejiang Province, College of Information Engineering, China Jiliang University, Hangzhou 310018, China

[2]SINOMA Bonyear (Hangzhou) Intelligent Technology Co. Ltd., Hangzhou 310000, China

Abstract: Concrete compressive strength prediction is a key part of batching ratio design, the traditional concrete cube compressive strength standard test method is vulnerable to external influences, a random forest algorithm is proposed to predict the compressive strength of concrete, by optimizing the data of cement, blast furnace slag mixing, fly ash mixing, water content, high efficiency water reducer dosage, coarse aggregates, fine aggregates content, and age, etc., as raw material indicators. The optimized data set is used as the input data set, and the data set is divided to construct the model, and then the model parameters are optimized by defining the parameter network and performing the grid search cross-validation to construct the perfect models. The results show that the prediction accuracy of the prediction model

基金项目: 2022 年度杭州市重大科技创新项目 (2022AIZD0085, 2022AIZD0016).

*通信作者：肖丙刚, bgxiao@cjlu.edu.cn

收稿日期: 2023-09-08; 接受日期: 2023-10-24; 在线出版日期: 2023-10-28

based on the random forest algorithm (R ²=0.91547) is much higher than that of the support vector regression model (R ²=0.76802) and the decision tree model (R ²=0.87539) and the error is small (RMSE=5.24087), which is of great significance for the research of compressive strength prediction model.

Keywords: Random Forest; Concrete; Compression Strength; Prediction

1 引言

作为建筑中最常用的材料之一，混凝土的抗压强度是其性能的重要评价指标之一。然而，传统的混凝土抗压强度预测方法主要依赖于经验公式，这些方法精度不高且成本高昂，同时缺乏泛化性[1]。

近年来，机器学习算法[2]在建筑工程领域得到广泛应用。国内外众多学者利用机器学习预测混凝土抗压强度取得了一定成果。出现了诸如基于决策树算法的混凝土抗压强度预测模型[3, 4]、基于 BP 神经网络的混凝土抗压强度预测模型[5]、基于 SPSS 回归分析的混凝土抗压强度预测模型、基于支持向量回归的混凝土抗压强度预测模型。徐国强[6]等通过指标权重确定混凝土抗压强度的主要影响因素并建立了 BP 神经网络算法对混凝土抗压强度进行预测；赵明亮[7]等则通过八个影响因素综合考虑建立 BP 神经网络模型对 7 天、28 天混凝土抗压强度进行预测；王继宗[8]等在以上的基础上通过建立多层 BP 神经网络预测 28 天混凝土抗压强度；路佳佳[9]利用交叉验证方法评估 SVR 模型的性能，并调整预测模型的超参数。张浩[10]在 SVM 算法基础上引入随机森林算法。Jui-Sheng Chou [11]通过优化参数改进最小二乘法支持向量回归算法，有效的提高了高性能混凝土抗压强度预测精度[12, 13]。尽管上述学者都在一定程度上提高了混凝土抗压强度的预测准确度，但在预测的精度方面各有不足。为此，我们提出一种基于随机森林的混凝土抗压强度预测模型，随机森林算法因其具有良好的泛化能力和预测精度而备受关注。因此，开展基于随机森林的混凝土抗压强度预测模型研究具有以下意义：①可以充分利用多维特征之间的相互作用关系，提高预测精度和效率，减少试验成本和时间成本；②可以帮助工程师优化混凝土配料比设计，提高混凝土的强度和耐久性；③基于随机森林的混凝土抗压强度预测模型是一种典型的人工智能应用，可以推动混凝土领域的智能化发展，提高建筑工程的质量和效率。

针对此，本研究采用随机森林算法构建一种混凝土抗压强度预测模型，将输入变量拆分数据为特征变量及目标变量，然后划分数据集为训练集和测试集，定义超参数网格，进一步创建随机森林算法模型，执行网格搜索交叉验证优化参数，最后代入模型进行训练，输出得到基于随机森林算法的抗压强度预测值，比对分析支持向量回归模型和决策树模型预测结果，验证随机森林算法模型的准确性及有效性。

2 随机森林算法

随机森林（Random Forest）算法[14]是一种基于决策树的集成学习（Ensemble Learning）算法，它由统计学家 Leo Breiman 等人提出。它是由多棵决策树组成的分类器或回归器，可以用于分类和回归问题。模型图如下图 1 所示，随机森林在决策树的基础上引入了随机性，即每棵决策树都基于随机样本和随机特征进行训练，通过对特征和样本进行随机选择来降低模型的方差，提高模型的泛化能力和减少过拟合的风险。

随机森林算法的核心思想是利用 Bagging（Bootstrap Aggregating）方法来构建多个决策树，并通过随机选择样本和特征来提高模型的多样性。随机森林算法是一种采用 Bagging 方法的机器学习算法，它通过自助法（Bootstrap）放回抽样产生多棵决策树，从中选出随机少数的特征来生成每个节点。这种随机性包括样本的随机抽样和节点特征的随机选择，因此被称为随机森林。通过引入随机性，随机森林能有效地减少过拟合问题，同时提高模型的稳定性和泛化能力。它广泛应用于分类和回归问题，结合了 Bagging 方法和特征随机性，使得模型具备更好的泛化性能，并且减少了人工干预的需求，特别适用于处理高维和大规模数据集。随机森林算法具有许多优点，如高准确率、鲁棒性强、可以处理大量的特征和样本等。

因此，基于随机森林算法的混凝土抗压强度预测

可以用于特征选择，通过统计每个特征在所有决策树中的重要性来进行排序，过度拟合现象在单一决策树模型中得以避免。但在应用随机森林算法时，需要注意调整一些超参数，如决策树数量、特征选择数、节点划分方式等，以获得最佳的预测效果[15, 16]。

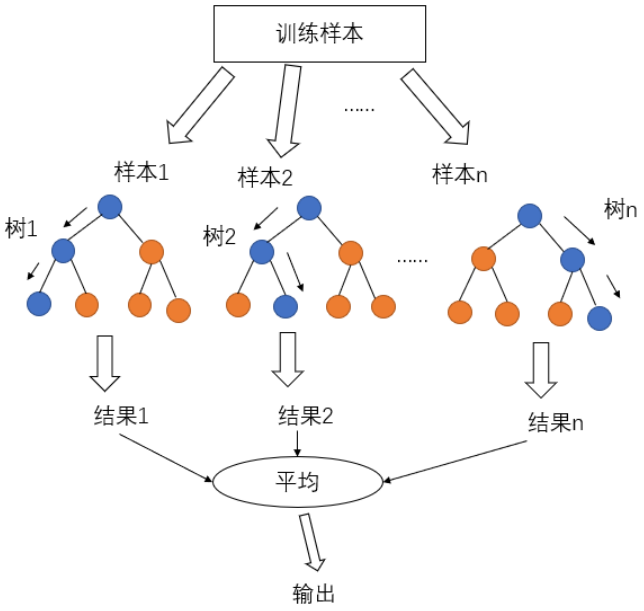

图 1 随机森林模型图

Figure 1 Random Forest Modeling Diagram

3 数据处理和模型构建方法

首先实现对数据集的可视化，方便观察原始数据集的数据情况，并对数据集进行预处理，对一些异常值进行去除，同时将部分数据内容进行格式的转换，以便后续的数据使用。其中数据集来源于论文[17]。最后对本文中所构建的三种算法模型进行了具体的构建过程描述，其中，支持向量回归算法与随机森林算法均通过超参数网格搜索法优化参数，并最终使用最优的参数结果进行模型的训练，得出预测结果。

3.1 数据可视化与预处理

3.1.1 数据可视化

通过数据可视化，可以将抽象的数据转化为具有形状、颜色、大小等视觉属性的图形，从而使数据更易于理解和解释，可以提供更直观的数据表达方式，帮助发现数据之间的模式、趋势、关联性和异常值，以及支持数据驱动的决策。

通过选择合适的可视化方法，以及调整图形的属性和布局，可以根据具体的数据和目标制作出有意义和有效的数据可视化。在研究中，主要对数据进行了可视化展示，分析了各个主要组成成分的直方图分布图如下图 2 所示：

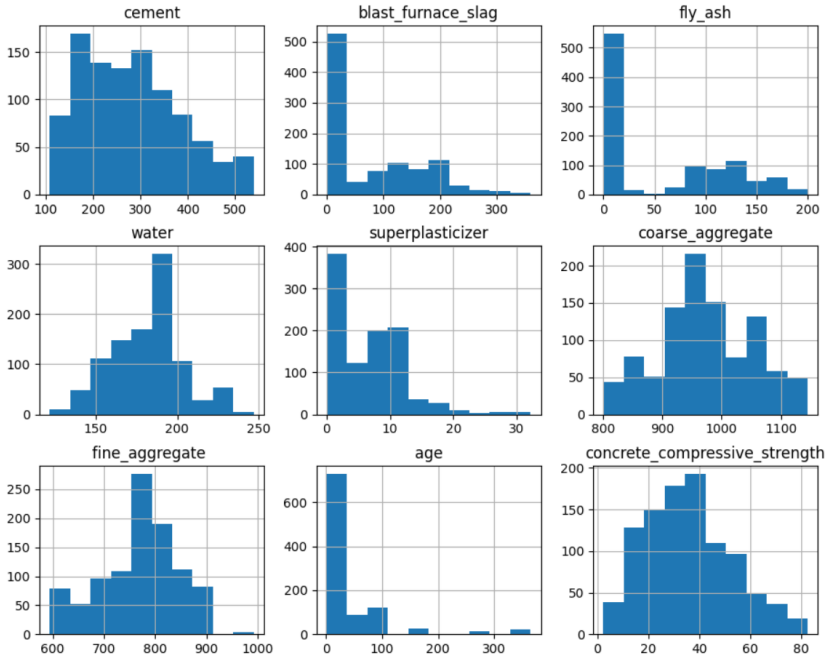

图 2 主要组成成分直方图

Figure 2 Histogram of major components

据图从左到右、从上到下依次数据集中是水泥、高炉矿渣、粉煤灰、水、减水剂、粗骨料、细骨料、期龄、水泥混凝土抗压强度的数据分布情况，得出水泥、水、粗骨料、细骨料、水泥混凝土抗压强度的数据分布较为分散；高炉矿渣、粉煤灰、减水剂、期龄的数据分布较为集中的初步结论。

分析了各主要组成成分之间关系图如下图 3 所示以及各主要组成成分之间相关性系数图如下图 4 所示：

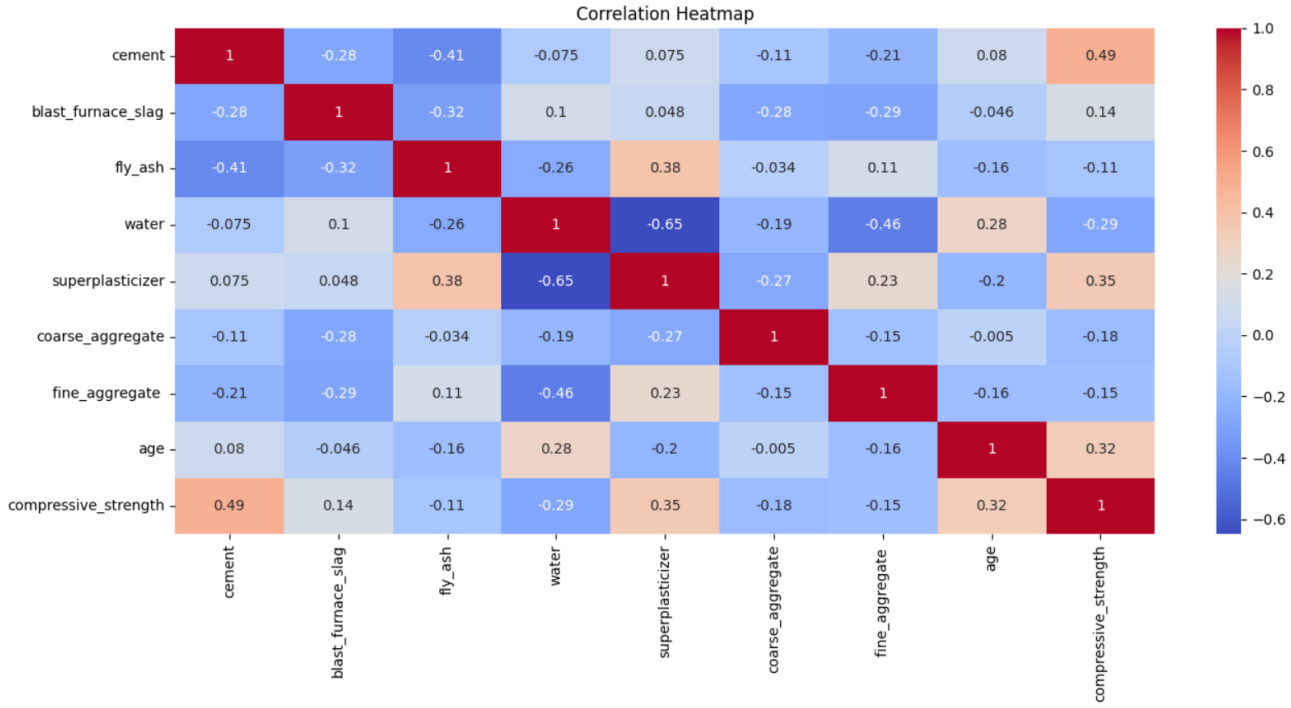

图 3 主要组成成分关系图

Figure 3 Relationship diagram of the main components

图 4 主要组成成分之间相关性系数

Figure 4 Correlation coefficients between major components

得出了水泥、水、减水剂、期龄与水泥混凝土抗压强度之间关系较大；高炉矿渣、粉煤灰、粗骨料、细骨料与水泥混凝土抗压强度之间关系较小；与水泥混凝土抗压强度关系最大的为水泥的初步结论。

3.1.2 数据预处理

数据预处理是指在应用机器学习算法之前对数据进行处理和清洗，以消除数据中的噪声、填补缺失值、规范化数据格式等操作，以提高机器学习模型的性能和准确性。下面是数据预处理的具体操作：①数据清洗：删除不完整的数据、重复的数据、异常值等，以保证数据集的质量和准确性；②缺失值填补：对于缺失的数据，可以通过插值法、均值填充、中位数填充、众数填充等方法进行填补；③特征选择：选择与目标变量相关性较高的 K 个特征变量，以提高模型的预测准确率。可以通过相关性分析、信息增益等方法进行特征选择。在本研究中用 data.drop 和 data 将数据集划分为特征变量和目标变量；④数据转换：将数据集转换为算法可以处理的格式，对于类别型变量，可以使用独热编码或者标签编码进行转换，对于数值型变量，可以进行归一化或者标准化处理；⑤数据集划分：将数据集划分为训练集和测试集，以便对机器学习模型进行训练和测试。本研究中用 train_test_split 将数据集划分为训练集和测试集。

3.2 随机森林模型构建

本研究在决策树的基础上引入随机森林算法模型，并对其参数进行优化，先后通过读取 CSV 文件数据集，拆分数据集为特征变量和目标变量，并按照 8:2 的比例划分训练集和测试集。最终定义随机森林回归算法模型并进行训练，设置参数如下：每棵决策树最大深度限制为 15，每棵决策树在节点分裂时考虑的特征数量限制为 5，设置构建决策树数量为 300，限制叶子节点的最小样本数量为 1 以及随机数种子的数量为 42.限制决策树的最大深度可以有效防止过拟合，限制决策树在节点分裂时考虑的特征数量可以提高模型的多样性，限制叶子节点的最小样本数量可以避免生成过小的叶子节点，而确定的随机数种子数量可以保证每次代码运行得到的训练集和测试集的划分方式都是相同的，保证了结果复现的可能性。

通过不断调整优化上述这些参数，可以达到对随机森林模型的复杂度、泛化性和随机性的控制，最终通过观察结果值达到提高模型预测准确性的目的。经实践表明，随机森林算法在混凝土预测中有可以有效处理高纬特征和大量样本数据，具有较强的泛化能力，能有效处理噪声和异常值等优点。

4 模型评估指标与结果分析

通过使用均方根误差、平均绝对误差、决定系数以及平均绝对百分比误差对随机森林算法模型、决策树算法模型、支持向量回归算法模型进行准确度分析，比较各类指标数据，最终得出结论：随机森林算法模型准确性及泛化性最好。

4.1 模型评价指标

在使用随机森林等算法预测混凝土抗压强度时，需要选择合适的指标参数来评价模型的性能和预测效果。参照以往经验论文[18]，引入以下指标参数进行模型评估：

(1) 均方根误差（Root Mean Squared Error，RMSE）：用于评价预测值与真实值之间的差异，公式如下(1)：

$$\mathrm{RMSE}=\sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i-y_i)^2}{n}} \tag{1}$$

其中 n 为样本容量，$\hat{y}_i$为实际值，y_i为预测值，下同。

(2) 平均绝对误差（Mean Absolute Error，MAE）：也是用于评价预测值与真实值之间的差异，公式如下(2)：

$$\mathrm{MAE}=\frac{\sum_{i=1}^{n}|\hat{y}_i-y_i|}{n} \tag{2}$$

(3) 决定系数（Coefficient of Determination，R^2）：用于评价模型的拟合程度，取值范围在 0 到 1 之间，越接近 1 说明模型的拟合程度越好，公式如下(3)：

$$\mathrm{R}^2=1-\frac{\sum_{i=1}^{n}(\hat{y}_i-y_i)}{\sum_{i=1}^{n}(\hat{y}_i-\bar{y}_i)} \tag{3}$$

(4) 平均绝对百分比误差（Mean Absolute Percentage Error，MAPE）：实际上是一种相对误差度量值，不会因为目标变量的全局缩放而改变，该指标能通过绝对值来避免正误差与负误差相对抵消。

4.2 实验结果比对分析

为验证基于随机森林算法模型的预测准确性，建立了支持向量回归算法模型，采用网格搜索法寻找并输出最优超参数组合、建立了 CART 算法中的决策树

回归模型，使用 Scikit-learn 中的 DecisionTreeRegressor 类实现强度预测，对以上模型进行对比分析，选取 RMSE、R^2、MAE、MAPE 评估模型的性能。最终的各个模型误差结果如下表 1 所示。

表 1 模型误差对比

Table 1 Model Error Comparison

选用模型	RMSE	R^2	MAE	MAPE
支持向量回归	8.68178	0.76802	6.37847	21.31176%
决策树	6.36302	0.87539	3.83305	12.73162%
随机森林	5.24087	0.91547	3.34062	11.15702%

从上述表格结果对比易得出结论：随机森林算法模型均方根误差最小，决定系数最接近 1，平均绝对误差最小，平均绝对百分比误差最小。结合下图 5-7 各模型预测值与真实值散点图可得知随机森林算法模型的预测精度最高，即随机森林算法在工程上具有良好的前景。

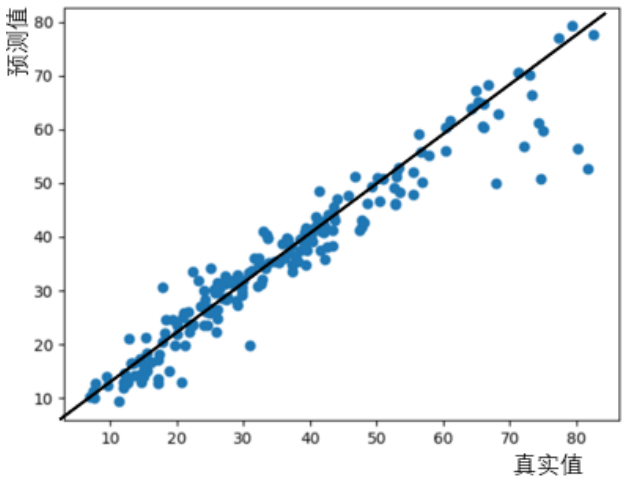

图 5 随机森林模型散点图

Figure 5 Random Forest Model Scatterplot

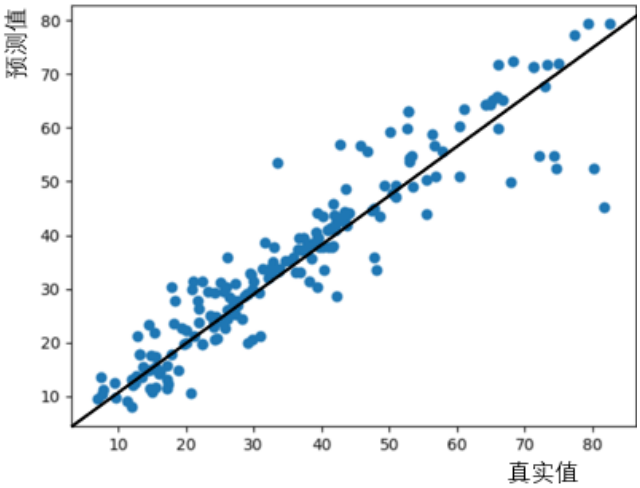

图 6 决策树模型散点图

Figure 6 Decision Tree Model Scatterplot

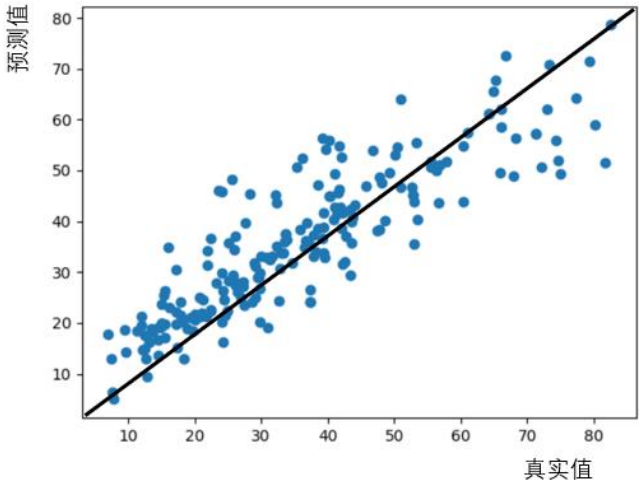

图 7 支持向量回归模型散点图

Figure 7 Support vector regression model scatterplot

5 结论

本研究构建了基于随机森林算法的混凝土抗压强度预测模型，并描述了对应的构建流程和步骤，利用部分数据集作为测试集验证了模型的准确性和可靠性，通过执行网格搜索交叉验证来优化参数，最终确定每棵决策树最大深度（max_depth）为 15，每棵决策树在节点分裂时考虑的特征数量（max_features）为 5，设置构建决策树数量（n_estimators）为 300，限制叶子节点的最小样本数量（min_samples_leaf）为 1 以及随机数种子的数量（random_state）为 42，使得决定系数提高到了 0.91547，均方根误差降低到了 5.24087，通过对比支持向量回归算法模型以及决策树算法模型，不难得出在混凝土抗压强度预测方面，基于随机森林的模型算法拥有更高的准确性。又因其能够通过集成多个决策树来提高预测性能，同时对于数据中的噪声和异常值也具有一定的鲁棒性，使得基于随机森林算法的预测模型优势更为明显。

参考文献

[1] 罗广彬, 洪成雨, 程志良等. 基于BP和GA-BP神经网络的混凝土抗压强度预测研究 [J]. 混凝土, 2023, No. 401(03): 37-41.

[2] 周宜松, 赵传萍, 黄耀明等. 基于机器学习技术的混凝土抗压强度预测研究 [J]. 安阳工学院学报, 2022, 21(06): 91-5.

[3] 邓初晴, 郑夷洲, 刘学等. 基于决策树算法的深圳地区混凝土回弹测强曲线的研究探索 [J]. 建筑监督检测与造价, 2022, 15(03): 36-41.

[4] 田欣. 决策树算法的研究综述 [J]. 现代营销 (下旬刊), 2017, (01): 36.

[5] 陈洪根, 龙蔚莹, 李昕等. 基于 BP 神经网络的粉煤灰混凝土抗压强度预测研究 [J]. 建筑结构, 2021, 51(S2): 1041-5.

[6] 徐国强, 苏幼坡, 韩佃利等. 基于 BP 神经网络的绿色混凝土抗压强度预测模型 [J]. 混凝土, 2013, No. 280(02): 33-5+49.

[7] 赵明亮, 水中和, 周华新等. 中低强度等级混凝土抗压强度的 BP 神经网络模型预测研究 [J]. 混凝土, 2021, No. 377(03): 35-8.

[8] 王继宗, 倪鸿光, 何锦云等. 混凝土强度预测和模拟的智能化方法 [J]. 土木工程学报, 2003, 55(10): 24-9.

[9] 路佳佳. 基于交叉验证的集成学习误差分析 [J]. 计算机系统应用, 2023, 32(01): 302-9.

[10] 张浩, 朱吉鹏, 卓德才等. 基于随机森林和支持向量机的混凝土抗压强度预测模型研究 [J]. 工程与建设, 2022, 36(06): 1784-8+815.

[11] CHOU J S, ANH-DYC P. Smart artificial firefly colony algorithm-based support vector regression for enhanced forecasting in civil engineering [J]. Computer-Aided Civil and Infrastructure Engineering, 2015, 30(9): 715-729.

[12] MOZUMDER R A, ROY B, LASKAR A I. Support Vector Regression Approach to Predict the Strength of FRP Confined Concrete [J]. Arabian Journal for Science and Engineering, 2017, 42(3): 26-27.

[13] 曹斐, 周彧, 王春晓等. 一种改进的支持向量回归的混凝土强度预测方法 [J]. 硅酸盐通报, 2021, 40(01): 90-7.

[14] 王奕森, 夏树涛. 集成学习之随机森林算法综述 [J]. 信息通信技术, 2018, 12(01): 49-55.

[15] 崔晓宁, 王起才, 张戎令等. 基于随机森林的高性能混凝土抗压强度预测 [J]. 兰州交通大学学报, 2021, 40(06): 1-6+14.

[16] 吴贤国, 刘鹏程, 陈虹宇等. 基于随机森林的高性能混凝土抗压强度预测 [J]. 混凝土, 2022, (01): 17-20+4.

[17] YEH I-C. Modeling slump of concrete with fly ash and superplasticizer % J Computers and Concrete [J]. 2008, 5(6): 59-62.

[18] DEROUSSEAU M A, LAFTCHIEV E, KASPRZYK J R, et al. A comparison of machine learning methods for predicting the compressive strength of field-placed concrete [J]. Construction and Building Materials, 2019, 228(C): 10-13.