

建筑能耗预测的机器学习回归模型研究

李继伟,冯国会,徐 丽

(沈阳建筑大学市政与环境工程学院,辽宁 沈阳 110168)

摘 要 **目的** 使用回归模型对建筑样本进行机器学习训练,筛选出适用于预测建筑能耗的模型。**方法** 以香港地区1 923个商用建筑作为研究对象,通过网络信息收集的方式获取建筑的物理参数、使用参数以及环境参数,采用PCA算法对高维度数据进行降维处理,并对相关数据进行了归一化处理;选取13种常见的机器学习回归模型作为建筑能耗预测模型,采用平均绝对误差MAE、绝对中位差MAD和决定系数 R^2 作为模型性能评价指标,采用StratifiedKfold分层采样法对总样本进行划分,并对划分后样本进行机器学习训练。**结果** Bagging、XGBoost、Random Forest、Extra Trees集成学习回归模型对建筑能耗预测的准确性远优于其他9种模型,其中XGBoost有最小的MAE(6.47)和MAD(2.95),Random Forest有最大的 R^2 (0.97)。**结论** 5种集成学习回归模型中除了分类算法外,Bagging、XGBoost、Random Forest、Extra Trees 4种模型对建筑能耗预测较优。XGBoost对数据较为完整的建筑能耗预测准确度最高,Extra Trees对于数据缺失严重的建筑预测准确度优于XGBoost。

关键词 建筑能耗;机器学习;回归模型;集成学习;性能评价指标

中图分类号 TU111.19⁺5;TP181

文献标志码 A

Study on Machine Learning Regression Model for Prediction of Building Energy Consumption

LI Jiwei, FENG Guohui, XU Li

(School of Municipal and Environmental Engineering, Shenyang Jianzhu University, Shenyang, China, 110168)

Abstract: The purpose of this paper is to choose the appropriate machine learning regressions as building energy consumption prediction model. 1923 commercial buildings in Hong Kong taken as samples, buildings' parameters on their physical features, utilization features and environmental features are collected by Internet. And high dimension data in parameters are adjusted by PCA algorithm. 13 kinds of regressions were chosen as prediction models of building energy consumption, with index MAE, MAD and R^2 to evaluate their performance. The whole samples were divided by StratifiedKfold algorithm, and then trained by machine learning regressions. Prediction results of the building energy consumption by Bagging, XGBoost, Random Forest and

收稿日期:2021-01-25

基金项目:国家自然科学基金项目(51778376)

作者简介:李继伟(1990—),男,博士研究生,主要从事机器学习在建筑能耗预测中的技术方面研究。

Extra Trees model, which are based on ensemble learning, were more accurate than that of other models, XGBoost got minimum value on MAE (6.47) and MAD (2.95) and Random Forest got maximum value on R^2 (0.90801). In addition, the main parameters that affect on building energy consumption are building area, stand alone building, wind parameter in 500 meter height, rental status, point solar radiation, height and number of floors. In addition to the classification algorithm, among the 5 ensemble learning regression models involved in this article, Bagging, XGBoost, Random Forest and Extra Trees perform good at building energy consumption predict. Among them, XGBoost is the best regression to predict the energy consumption of building with complete data, while Extra Trees perform better at samples with lack of data.

Key words: building energy consumption; machine learning; regression model; ensemble learning; performance evaluation index

能源消耗是温室气体排放增加和环境污染日益严重的主要原因之一,其中建筑能耗占能源消耗总量的 35% 以上^[1]。20 世纪 70 年代,欧美发达国家就开始了建筑能耗的研究,并制定了建筑节能标准^[2-5]。我国的建筑能耗研究起步较晚,且数据缺失严重,因此,国内常用的研究手段主要有现场调研和计算机数值模拟。尤其是到了 21 世纪大量经典算法的出现,机器学习逐渐被应用于建筑能耗预测或影响因素的研究,如空间关系、建筑材料、环境因素与建筑能耗之间的关系^[6-7],建筑材料、建筑结构、建筑照明和外部气候条件对建筑能耗的影响^[8-12]。目前用于建筑能耗预测的机器学习回归模型很多,但是针对建筑能耗预测的回归模型研究鲜有报道。而模型的选择直接影响建筑能耗

预测的准确性,因此,笔者以香港地区的商用建筑为研究对象,选取 13 种常用于预测建筑能耗的回归模型,通过对模型预测性能的比较,获取建筑能耗预测的最优回归模型,为建筑能耗预测和能源供应的合理分配提供技术支持。

1 数据来源与处理

1.1 数据来源

建筑能耗通常受其自身特征、使用性能以及外界环境的影响。笔者采用网络信息收集的方式收集了物理参数、使用参数与环境参数。涉及的建筑样本来源于 ESRI 网站,该网站提供了 1 923 个商用建筑样本的能耗数据,具体参数名称与所属类型如表 1 所示。

表 1 参数表
Table 1 Features list

参数序号	参数名称	参数描述
1	sqft	建筑面积
2	number_of_ floors	建筑层数
3	height	建筑高度
4	elevator	电梯数量
5	year	建筑年限
6	EUI	能源使用指数
7	MF1、MF2、Sat1、Sat2、Sun1、Sun2	营业时间
8	CarPark、Exhibition、Famous、FamousBuilding、Hotels、IND、	使用功能
9	Office、OfficeMalls、OfficeOnly、Recreation、StandAloneMalls、wTenant、population、employment、labouring_percent、number_of_family、medium_of_family_house	人口与劳工
10	point_solar_radiation	太阳辐射点
11	CDD	制冷度日数
12	wind	风环境

表中 1-5 为建筑的物理参数;6 为能源使用指数,即目标参数;7-9 为使用参数;10-12 为环境参数,其中风环境参数为预处理后的数据。建筑的物理参数和使用参数中的营业时间数据来源于建筑所属的物业公司网站或维基百科,由于其数据来源较为分散,因此数据存在缺失情况,其他使用参数和 EUI 数据来源于 ESRichina 中的香港建筑物能源使用指数网站以及香港民政总署网站,太阳辐射点、制冷度日数数据通过 Arcgis 和建筑的经纬度坐标计算获得,风环境参数数据来源于香港规划署网站。

1.2 数据预处理

由于参数数值之间差异通常较大,还有个别参数维度较高导致数据冗余,影响拟合过程。因此,笔者在模拟之前对样本数据进行了预处理。数据处理包括风环境参数的降维处理和其他参数的归一化处理两部分。

1.2.1 风环境参数的降维处理

风环境参数数据包含 3 个高度、16 个风向以及 24 个风速区间参数,共计 $3 \times 16 \times 24 = 1\,152$ 个数据。笔者采用主成分分析(Principal Component Analysis, PCA)算法,对所有风环境参数进行降维处理。将同高度 16×24 个风环境参数降维处理为 2 个参数。常用 PCA 算法具体步骤:①将参数数据减去同高度参数的平均值,构建矩阵 X ;②计算矩阵 X 的协方差矩阵及其特征值和特征向量;③选取 2 个特征值最大的特征向量,组成特征向量矩阵 P ;④由步骤①和③构建矩阵 Y ,即 $Y = PX$ 。矩阵 Y 中数据即为降维后的风环境参数。

1.2.2 数据归一化

为了消除各参数间由于数据数量级差异带来的结果误差或不准确,笔者对数据进行了标准化处理。笔者使用的风环境参数数据间由于方差非常小,归一化预处理可以增强数据的稳定性。因此,采用归一化方法对原始数据进行标准化,将各参数数据分别映射

到 $[0,1]$ 之间。

$$x^* = \frac{x - x_{\min}}{x_{\max} - x_{\min}}.$$

(1)

式中: x^* 为处理后的数据,即原始数据标准化得到的结果; x 为原始数据, x_{\min} 、 x_{\max} 为原始数据中最小值和最大值。

2 基于机器学习的能耗预测模型构建

2.1 能耗预测模型

常用于能耗预测的机器学习模型有线性回归、Random Forest、支持向量机、神经网络、梯度提升、k-means、KNN 等^[8,12-15],笔者选取 13 种常见的机器学习回归模型,包括 4 种基于线性回归的线性模型,2 种支持向量机模型,5 种基于决策树的集成学习模型,以及其他 2 种常用回归模型(见表 2)。

表 2 回归模型列表

Table 2 Regressions model list

回归模型		模型所属类型
英文名称	中文名称	
Linear	线性回归	线性
Ridge	岭回归	
Lasso	Lasso 回归	
ElasticNet	弹性网络	支持向量机
SVR	支持向量回归	
Linear SVR	线性支持向量回归	
AdaBoost	自适应提升	集成学习
Bagging	套袋法	
XGBoost	极限梯度提升	
Random Forest	随机森林	神经网络
Extra Trees	极端随机树	
MLP	多层感知机	
KNN	K 近邻	K 近邻学习

2.2 模型性能评价指标

为了评价各机器学习回归模型对建筑能耗的预测准确性,笔者选用平均绝对误差 MAE、绝对中位差 MAD、决定系数 3 种性能评价指标对模型进行评价。

$$MAE(y,y') = \frac{1}{n} \sum_{i=0}^{n-1} |y_i - y_i'|. \quad (2)$$

$$MAD = \text{median}(|y_i - \text{median}(y)|). \quad (3)$$

$$R^2(y,y') = 1 - \frac{\sum_{i=0}^{n-1} (y_i - y_i')^2}{\sum_{i=0}^{n-1} (y_i - \bar{y}')^2}. \quad (4)$$

式中: n 为样本数量; y_i 为实际 EUl 值; y_i' 为预测 EUl 值; \bar{y} 为预测平均 EUl 值。

MAE 和 MAD 值越小,表明回归模型预测值与实测值的误差越小。 R^2 取值范围为 $[0,1]$,越接近 1 代表回归模型预测值的离散程度越低。

2.3 样本数据分组

采用 StratifiedKFold 分层采样法构建机
器学习的训练集与测试集,划分比例为 8:2,
样本分组结果如表 3 所示。

表 3 样本分组结果

Table 3 Result of sample divided

数据集	样本数量/个	参数数量/个
训练集	1 537	42
测试集	385	42

为了保证样本分组后数据结构与总样本
集相似,选取总样本、训练集和测试集中的
EUl、建筑面积、TPU 内人口以及太阳辐射点
4 项参数的分布情况进行统计分析,验证样
本划分的合理性,结果如图 1 ~ 图 4 所示。
由图1 ~ 图4 可以看出,总样本、训练集以及
验证集中 EUl 参数分布相似、数值范围一
致,建筑面积、TPU 内人口以及太阳辐射点 3
项参数数据分布曲线基本相似。因此,样本
划分基本满足机器学习训练的要求。

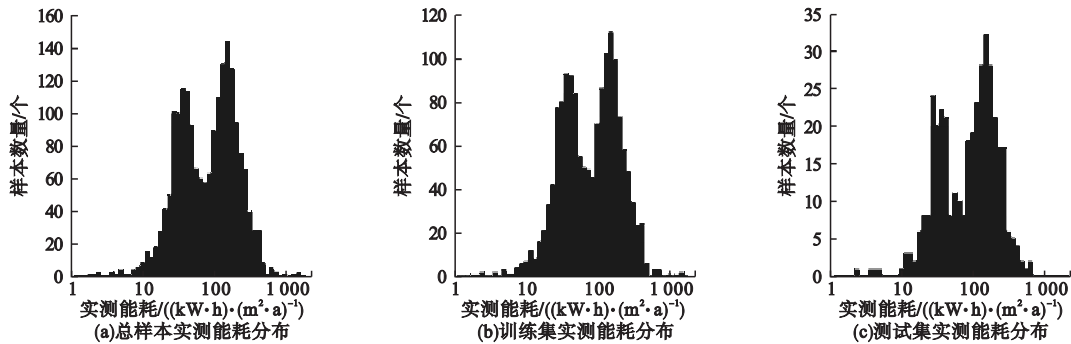


图 1 实测能耗分布

Fig. 1 Distribution of energy consumption

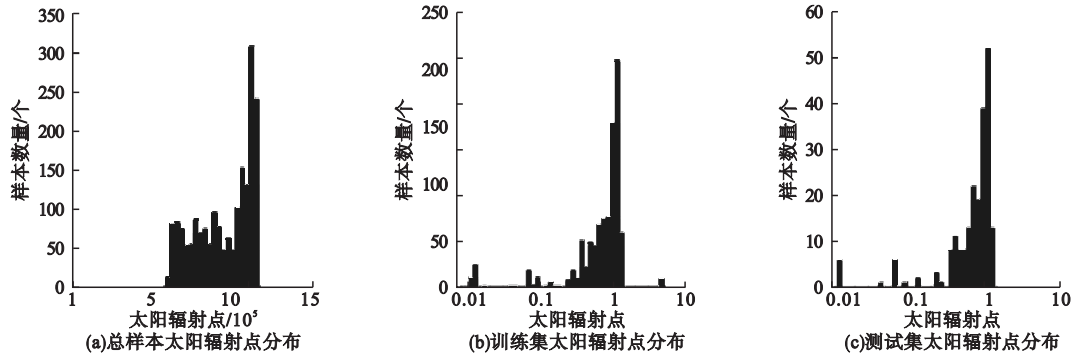


图 2 太阳辐射点分布

Fig. 2 Distribution of point solar radiation

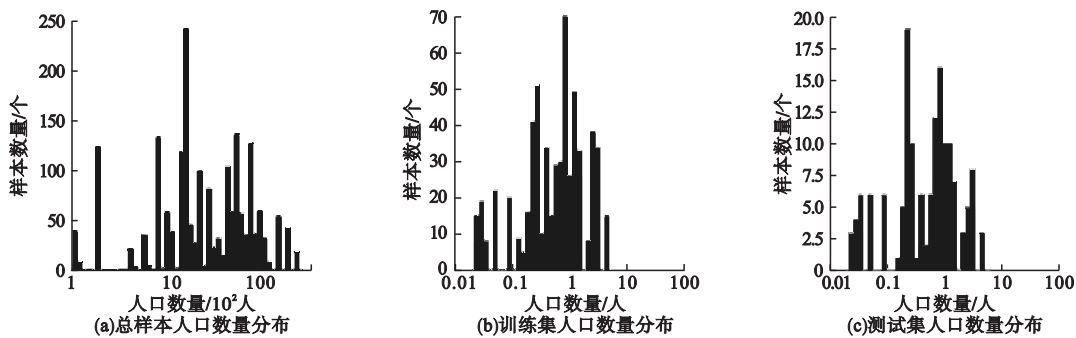


图 3 人口数量分布

Fig. 3 Distribution of population

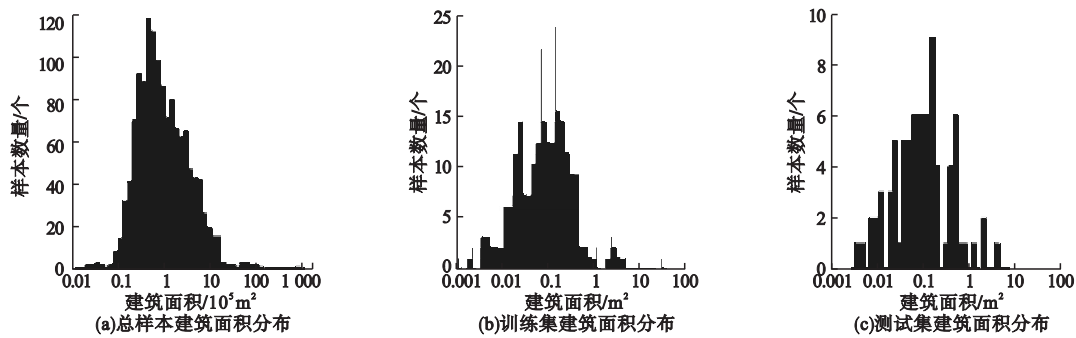


图 4 建筑面积分布

Fig. 4 Distribution of floor area

3 模型验证、分析、优化与应用

3.1 模型验证

根据样本划分的结果,首先采用训练集中样本对 13 种回归模型进行机器学习训练,然后采用测试集样本对训练好的回归模型进行模型验证和校核,采用式(2)~(4)计算模型各项性能评价指标,具体计算结果如表 4 所示。从表 4 可知,基于线性模型的 Linear Regression、Ridge Regressor、Lasso、ElasticNet,基于支持向量机的 SVR、Linear SVR,基于神经网络的 MLP 以及基于 K 近邻学习的 KNN 回归模型对建筑能耗的预测误差较大,对建筑能耗预测能力较强的 Bagging、XGBoost、Random Forest、Extra Trees4 种回归模型都属于集成学习,但是集成学习中的 AdaBoost 回归模型对建筑能耗的预测结果误差很大。其中 XGBoost 有最小的 MAE (6.47)、 MAD (2.95),Random Forest 有最大的 R^2 (0.91)。

表 4 回归模型性能评价指标

Table 4 Performance evaluation index of regression

回归模型	MAE	MAD	R^2
Linear Regression	37.14	24.68	0.62
Ridge Regressor	36.80	24.50	0.62
SVR	56.36	33.96	0.08
Lasso	34.89	23.95	0.64
ElasticNet	48.60	35.64	0.39
Linear SVR	32.80	14.05	0.55
AdaBoost	57.31	56.47	0.42
Bagging	9.51	3.14	0.89
XGBoost	6.47	2.95	0.89
Random Forest	8.60	3.04	0.91
MLP	36.81	21.28	0.59
KNN	45.43	23.61	0.37

3.2 集成学习回归模型的性能分析

从表 4 可知,除了 AdaBoost 回归模型外(见图 5),XGBoost、Bagging、Random Forest、Extra Trees 4 种基于集成学习的回归模型性能较好,对 EUI 的预测较准确(见图 6)。实测能耗与预测能耗越接近 45° 线的点说明预测越准确。

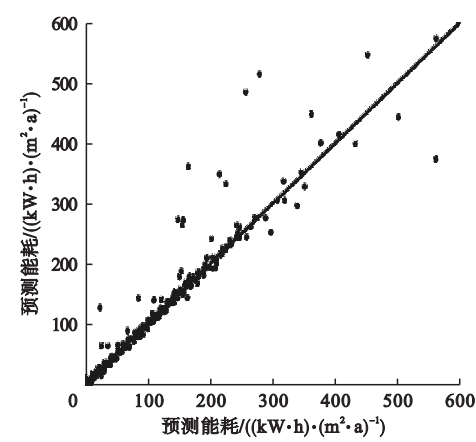


图5 AdaBoost 模型预测的能耗分布

Fig.5 Distribution of energy consumption predicted by AdaBoost

从图5可知,AdaBoost 回归模型预测的能耗与实测能耗误差较大,尤其是实测能耗0~250 (kW·h)/(m²·a)的样本,预测的能耗值主要分布在几条平行于横轴的线附近,实测能耗250~400 (kW·h)/(m²·a)的样本预测较准确,实测能耗大于400 (kW·h)/

(m²·a)的样本预测误差较大。AdaBoost 回归模型采用改变样本权重的基学习器进行样本分类,然后通过改变错误样本权重的基学习器再次进行样本分类,直至样本分类正确为止,将所有基学习器按分类错误率分配权重,形成建筑能耗预测模型,即此时模拟数据与训练集数据吻合最好。因此,该模型模拟结果依赖于数据质量,而在建筑样本收集过程中,存在不可避免的数据缺失或误差,导致样本在处理过程中存在2种错误情形。第1种情形是原本不属于同一类的样本而被错误的按同一类样本进行处理,从而模糊或降低了样本间的差异,导致样本欠拟合;第2种情形是原本属于同一类型的样本被错误的按不同类型样本进行处理,导致样本过拟合。这两种情形都会导致模拟结果不准确,使得模拟结果高度相似或相等,图5中实测 EUI 小于250 (kW·h)/(m²·a)的样本被过拟合或欠拟合。

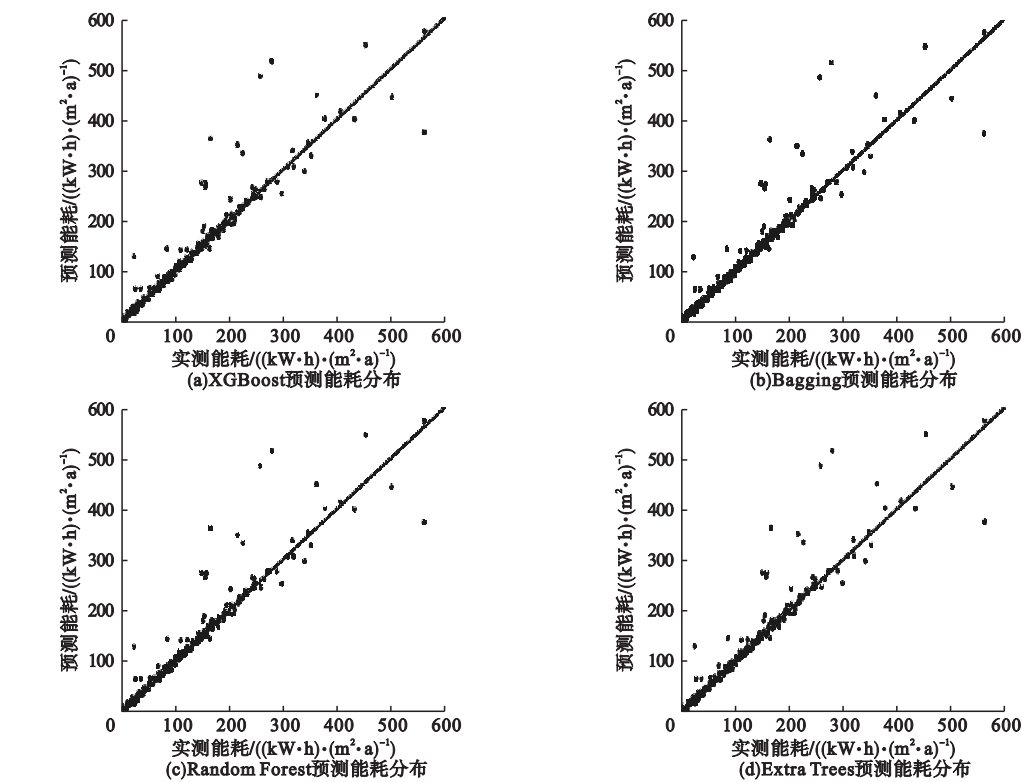


图6 4种回归模型预测能耗分布

Fig.6 Distribution of predicted energy consumption by 4 regressors

从图 6 可知,XGBoost 对实测能耗低于 $400(\text{kW}\cdot\text{h})/(\text{m}^2\cdot\text{a})$ 的建筑样本预测基本准确,但是对于实测能耗大于 $400(\text{kW}\cdot\text{h})/(\text{m}^2\cdot\text{a})$ 的建筑样本预测误差较大。其余 3 种回归模型对实测能耗低于 $300(\text{kW}\cdot\text{h})/(\text{m}^2\cdot\text{a})$ 的建筑样本预测较为准确,但是少量样本预测的能耗与实际能耗存在误差,预测值偏高,但是对于实测能耗高于 $300(\text{kW}\cdot\text{h})/(\text{m}^2\cdot\text{a})$ 的建筑样本,Bagging 回归模型预测较为准确,Random Forest 回归模型的预测结果更离散,误差更大,Extra Trees 回归模型的预测准确性介于两者之间。

这 4 种模型预测结果优于其他 9 种模型,其主要原因在于模型自身通过不同的方式避免过拟合的发生,以 XGBoost 回归模型为例,XGBoost 是一种梯度提升回归模型,每一次的训练都是在残差减小的方向上建立模型,模型中有专门控制精度项和防止过拟合项,同时还考虑了数据的缺失情况。其数学函数为

$$Obj = \sum_i (y_i - \hat{y}_i)^2 + \gamma T + \frac{1}{2} \lambda \| \omega \|^2. \quad (5)$$

式中: y_i 为实测能耗; \hat{y}_i 为使用所有决策树对单一样本的能耗预测值; T 为节点的数量 ω 为节点的数值; γ 为节点增益的阈值; λ 为系数。

Bagging 和 Random Forest 两种回归模型都是对所有建筑样本进行多次有放回的随机抽样构建训练样本集,每次抽样的数量与表 3 中训练集样本数量相同。每个基学习器训练的样本不同,最终的预测能耗取所有训练结果的算术平均值。由于噪声数据占比低,因此大部分基学习器的训练结果基于正确数据,降低了噪声数据对预测结果的影响,相比于 AdaBoost 回归模型可以更有效地避免噪声数据的干扰。Extra Trees 中训练集是总样本,所有基学习器训练的样本相同,都是总样本,因此,高能耗样本比 Random Forest 的训练更充分,样本的预测结果更准确且离散程度更低(见图 6(b)~图 6(d))。

3.3 模型优化

选取 MAE、MAD 两项性能评价指标较小且 R^2 较大的回归模型进行参数调整,进一步提高回归模型对建筑能耗预测的准确性。调参之后 4 种回归模型的性能评价指标数值如表 5 所示。

表 5 调参后性能评价指标

Table 5 Performance evaluation index after features adjustment

回归模型	MAE	MAD	R^2
XGBoost	5.59	2.95	0.97
Bagging	7.22	3.07	0.94
Random Forest	8.60	3.04	0.91
Extra Trees	7.62	3.12	0.94

从表 4 和表 5 可知,除 Random Forest 外,调参后各算法的 3 项性能评价指标都优于调参前,说明调参后模型精度更好,预测能力更强,XGBoost 依然是 4 种回归模型中预测最准确的回归模型。

3.4 模型应用

在 42 项样本参数的基础上构建模型,而实际工作中,收集完整的 42 项参数数据很难实现,而有些参数对能耗的影响可以忽略,仅需收集影响较大的参数,因此,需分析 42 项参数的重要性。由于 Random Forest 回归模型的基学习器中任意节点的分裂取决于样本的参数重要性,即节点增益,计算所有基学习器中相同参数的节点增益,就可以获取各项参数在建筑能耗中的重要性。所以,笔者采用 Random Forest 回归模型进行了参数重要性分析。

通过 Random Forest 回归模型筛选出权重之和接近 0.95 的参数共有 7 项,建筑面积、建筑独立性、风环境参数、出租状态、太阳辐射点、建筑高度以及层数,对应的权重分别为 0.67、0.16、0.05、0.044、0.01、0.01、0.01。基于这 7 项参数,采用 4 种优化后模型预测了所有建筑样本能耗,结果如图 7 所示,各模型的性能如表 6 所示。

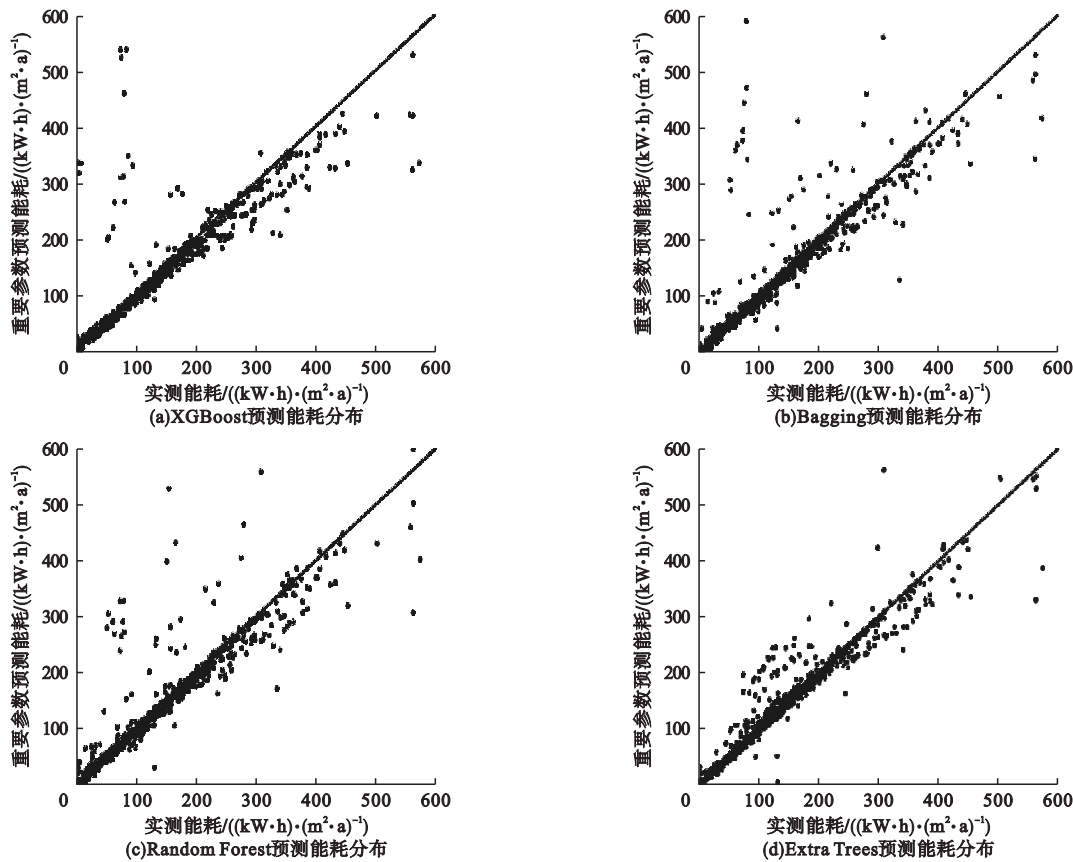


图7 基于重要参数的预测能耗分布

Fig.7 Distribution of predicted energy consumption by important features

表6 基于7项参数的模型性能

4种回归模型均可用于建筑能耗预测。

Table 6 Performance of models by 7 features			
算法类别	MAE	MAD	R ²
XGBoost	10.84	3.37	0.80
Bagging	9.56	1.92	0.85
Random Forest	9.23	1.95	0.87
Extra Trees	7.18	1.30	0.93

4 结 论

由表6可以看出, Bagging、XGBoost、Random Forest、Extra Trees4种回归模型性能评价指标较好,其中 Extra Trees 的3项性能评价指标皆为最优,说明在参数不全或数据缺失的情况下,这4种回归模型仍可以有效预测建筑能耗。图7中预测能耗基本与实测能耗一致,误差在可接受范围内。与图6相比,7项参数能耗预测结果和42项参数预测结果分布情况相似,基本一致,进一步说明这

(1) 基于决策树的集成学习 Bagging、XGBoost、Random Forest、Extra Trees4种回归模型对建筑能耗的预测较为准确,对数据缺失的建筑能耗预测基本准确。

(2) XGBoost 有最小的 MAE (6.47)、MAD (2.95); Random Forest 有最大的 R² (0.91); Extra Trees 有最小的 MAE (7.18)、MAD (1.30) 和最大的 R² (0.93)。

(3) Random Forest 回归模型既可用于建筑能耗的预测,也可用于建筑能耗影响因素重要性的分析。

(4) AdaBoost 分类算法并不适用于建筑能耗的预测。

参考文献

- [1] 中国工程院土木水利与建筑工程学部,清华大学建筑节能研究中心,国家发改委,等.《中国建筑节能年度发展研究报告2015》发布[J]. 发电与空调,2015,36(2):59.
(China Engineering Institute of Civil, Hydraulic and Construction Engineering, Tsinghua University Building Energy Conservation Research Center. State development and reform commission, annual development report of building energy conservation in China 2015 [J]. Power generation and air conditioning, 2015, 36(2): 59.)
- [2] CALEB R, BISTRA D, JEFFREY H, et al. Machine learning approaches for estimating commercial building energy consumption [J]. Applied energy, 2017, 208: 889 - 904.
- [3] 陈超,渡边俊行,谢光亚,等.日本的建筑节能概念与政策[J]. 暖通空调,2002(6):40 - 43.
(CHEN Chao, TOSHIYUKI, XIE Guangya, et al. Japanese concepts and measures of building energy efficiency [J]. Heating ventilating & air conditioning, 2002(6): 40 - 43.)
- [4] 张忠义. 日本关于能源使用合理化法律[J]. 发电设备,1987(9):44 - 45.
(ZHANG Zhongyi. Japanese law on rationalization of energy use [J]. Power equipment, 1987(9): 44 - 45.)
- [5] 王清勤,徐选才,李明玮. 加拿大的建筑节能规范简介[J]. 建筑科学,1997(6):59 - 60.
(WANG Qingqin, XU Xuancai, LI Mingwei. Brief introduction of building energy conservation standards in Canada [J]. Building science, 1997(6): 59 - 60.)
- [6] SAREH N, AFRAM K, SHAHABODDIN S, et al. Estimating building energy consumption using extreme learning machine method [J]. Energy, 2016, 97: 506 - 516.
- [7] YANG Y R, HE M, LI L. A new machine learning based geometry feature extraction approach for energy consumption estimation in mask image projection stereolithography [J]. Procedia CIRP, 2019, 80: 741 - 745.
- [8] 程亚豪,陈焕新,王江宇. 基于机器学习的住宅能耗预测[J]. 制冷与空调,2019,19(5):35 - 40.
(CHENG Yahao, CHEN Huanxin, WANG Jiangyu. Prediction of residential energy consumption based on machine learning [J]. Refrigeration and air-conditioning, 2019, 19(5): 35 - 40.)
- [9] JOHN E S. Using intelligent data analysis to detect abnormal energy consumption in buildings [J]. Energy & buildings, 2006, 39(1): 52 - 58.
- [10] 王艺斐,王宏愿. 建筑能耗预测技术应用[J]. 山东工业技术,2017(12):113 - 114.
(WANG Yifei, WANG Hongyuan. Application of building energy consumption prediction technology [J]. Journal Shandong industrial technology, 2017(12): 113 - 114.)
- [11] 张智超,李振亮,李亚,等. 链牵引式外遮阳百叶窗结构设计及其性能分析[J]. 天津科技大学学报,2013,28(6):52 - 55.
(ZHANG Zhichao, LI Zhenliang, LI Ya, et al. Structure design and performance analysis of the external shading blind with chain traction [J]. Journal of Tianjin university of science & technology, 2013, 28(6): 52 - 55.)
- [12] 冯国会,徐小龙,王悦,等. 以能耗为导向的近零能耗建筑围护结构设计参数敏感性分析[J]. 沈阳建筑大学学报(自然科学版),2018,34(6):1069 - 1077.
(FENG Guohui, XU Xiaolong, WANG Yue, et al. Sensitivity analysis of nearly zero energy buildings envelope design parameters based on energy consumption [J]. Journal of Shenyang jianzhu university (natural science), 2018, 34(6): 1069 - 1077.)
- [13] 田玮,魏来,李占勇,等. 基于机器学习的建筑能耗模型适用性研究[J]. 天津科技大学学报,2016,31(3):54 - 59.
(TIAN Wei, WEI Lai, LI Zhanyong, et al. Building energy models based on machine learning methods [J]. Journal of Tianjin university of science & technology, 2016, 31(3): 54 - 59.)
- [14] 吴蔚沁. 基于机器学习算法的建筑能耗监测数据异常识别及修复方法[J]. 建设科技,2017(9):60 - 62.
(WU Weiqin. Abnormal identification and repair method of building energy consumption monitoring data based on machine learning algorithm [J]. Construction science and technology, 2017(9): 60 - 62.)
- [15] 崔治国,曹勇,武根峰,等. 基于机器学习算法的建筑能耗监测数据预处理技术研究[J]. 建筑科学,2018,34(2):94 - 99.
(CUI Zhiguo, CAO Yong, WU Genfeng, et al. Research on preprocessing technology of building energy consumption monitoring data based on machine learning algorithm [J]. Building science, 2018, 34(2): 94 - 99.)

(责任编辑:徐玉梅 英文审校:唐玉兰)