

考虑数据分类的建筑电能耗集成预测方法

唐倩倩,李康吉,魏伯睿,王莹

(江苏大学 电气信息工程学院,江苏 镇江 212013)

摘要:建筑侧各类可再生能源的应用日益普及,建筑电能耗预测在用能供需平衡、电网稳定运行、尖峰需求响应等方面发挥越来越重要作用。尽管诸多数据驱动模型在能耗预测方面获得广泛应用,当前仍缺乏预测精度高、泛化能力强的短期预测模型。针对该问题,提出一种基于建筑物能耗特点并结合数据挖掘技术的分类集成式能耗预测方法。首先,采用递归特征消除法对数据进行特征筛选,并用模糊C均值聚类算法对训练集数据进行聚类,使用K最邻近法对验证集和测试集数据进行归类;选择5种结合智能优化算法的混合数据驱动模型作为子学习器,分别对每类数据做预测,最后使用多元线性回归法进行结果集成。经3个建筑电力用能案例验证,此集成预测模型精度均优于单个子模型,具有适用不同建筑类型和用能尺度的预测潜力。

关键词:建筑;电能耗预测;数据分类;递归特征消除法;模糊C均值聚类算法

An ensemble forecasting method for building electricity consumption considering data classification

TANG Qianqian, LI Kangji, WEI Borui, WANG Ying

(School of Electrical and Information Engineering, Jiangsu University, Zhenjiang 212013, China)

Abstract: The application of various types of renewable energy on the building side is becoming more and more popular. Forecasting of building electricity consumption plays an increasingly important role in the balance of energy supply and demand, stable grid operation, and peak demand response. Although many data-driven models have been widely used in energy consumption prediction, there is still a lack of short-term prediction models with high prediction accuracy and strong generalization ability. In order to solve this problem, a classification and integration energy consumption prediction method based on the characteristics of building energy consumption and combined with data mining technology is proposed. Firstly, the recursive feature elimination method is used to screen the features of the data, and the fuzzy C-means clustering algorithm is used to cluster the training set data, meanwhile, K-nearest neighbor method is used to classify the validation set and test set data. Then, five hybrid data-driven models combined with intelligent optimization algorithms are selected as sub-learners, and each type of data is predicted respectively. Finally, multiple linear regression method is used to integrate the results. The accuracy of the ensemble prediction model is better than that of single sub-model, and it has potential to predict different building types and energy use scales.

Key words: building; forecast of electricity energy consumption; data classification; recursive feature elimination method; fuzzy C-means clustering algorithm

0 引言

据国家统计局数据,2021年国内一次能源生产总量为40.8亿tce,同比增长2.8%。以建筑用电脱碳^[1]为目标,各类可再生能源在建筑行业的应用日益普及。建筑电能耗的精确预测在建筑用能管理、电网稳定运行、尖峰需求响应等方面发挥越来越重要作用。

近年来,诸多数据驱动模型在能耗预测方面获得广泛应用。神经网络、支持向量机(support vector machine, SVM)以及各类混合模型均在具体应用中

取得不错的表现。文献[2]提出一种混沌时间序列和SVM相结合的混合预测模型,验证了该模型性能优于SVM。文献[3]提出了一种基于教与学(teaching-learning-based optimization, TLBO)混合模型的短期预测方法,提高建筑能源短期预测精度。实践表明单一预测模型对预测场景敏感,针对此问题,集成预测方法是更好的选择。文献[4]构建了一种集成预测模型,引入自适应增强集成策略分别提升反向传播(back propagation, BP)神经网络、支持向量回归、遗传编程和径向基神经网络的性能,并通过数据分组方法来提高预测精度。文献[5]提出了一种以极端梯度提升和前馈神经网络为基础模型的集成算法,与其他单个模型相比,预测误差明显降

低。文献[6]提出了利用变分模态分解的方法对数据进行序列分解,同时用差分进化算法来优化极限学习机(extreme learning machine, ELM)的预测集成模型。文献[7]提出一种采用最优加权组合法结合长短期记忆网络(long short-term memory, LSTM)的预测模型。分析集成模型发现:建筑用能对自身规律和数据属性缺乏充分考虑,集成式能耗预测模型有进一步提高精度的潜力。

综上,本文提出一种考虑用能时间规律并结合数据挖掘技术的建筑电能集成预测方法,通过两种种群智能算法,即 TLBO 和粒子群优化算法^[8](particle swarm optimization, PSO),优化子学习器参数。所提模型通过3个不同案例验证其有效性。

1 模型框架

1.1 数据预处理

数据预处理包括特征选择和数据分类。本文采用递归特征消除法(recursive feature elimination, RFE)对原始数据进行特征选择,考虑建筑物能耗特点结合模糊 C 均值聚类算法(fuzzy c-means algorithm, FCM)和 K 最邻近法(k-nearest neighbor, KNN)进行数据分类。

1.1.1 特征选择

本文的特征选择方法 RFE 是通过反复构建模型,对数据进行多次训练,根据模型误差结果选出最好(或最差)的特征。这个过程中特征被消除的次序就是特征的排序,以此寻找最优特征组合。

1.1.2 数据分类

本文的数据分类含两步:首先按建筑物自身电力能源使用模式,根据专家经验将数据集划分为工作日和非工作日;然后将 FCM 和 KNN 两种方法相结合,对数据集进行特性挖掘和分类。模型训练所用的数据为已知历史数据,使用 FCM 寻找聚类中心并分类;模型测试所用的数据在实际应用中为实时采集数据。本文运用 KNN 方法判断与哪类的训练数据最接近,根据已知 FCM 分类标签进行归类。FCM 是一种无监督聚类方法,基本思想是使被划分到簇的对象间相似度最大,不同簇之间相似度最小,具体公式和算法细节见文献[9];KNN 是一种监督分类方法。假设样本有 K 个邻近样本,如果大部分都属于某个已知类别,那么这个样本也属于该类别,具体公式和算法细节见文献[10]。

引入聚类有效性指标来评估初始聚类数 n , 有效性指标公式如下

$$X_B = \frac{\frac{1}{n} \sum_{i=1}^k \sum_{j=1}^n u_{ij}^m \|v_i - v_j\|^2}{\min_{i \neq j} \|v_i - v_j\|^2} \quad (1)$$

式中: k 为聚类数; n 为样本总数; μ_{ij} 为样本 x_j 对聚类中心 v_i 的隶属度; m 为模糊因子,当 X_B 指标越小,聚类效果越好。

1.2 子模型预测

单个模型自身存在某些局限性,因此数据预测部分采用多种属性结构差异大的模型作为子学习器,提升模型整体泛化性能。

LSTM 是一种特殊的循环神经网络,非常适合时间序列预测任务^[11-12]。LSTM 由多个单元组成,每个单元主要有3个门结构。其中输入门、输出门分别用来读取、输出和修正参数,遗忘门用来选择性遗忘上一时刻的单元状态并修正参数。计算公式如下

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \quad (2)$$

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \quad (3)$$

$$C_t = f_t C_{t-1} + i_t (\tanh(W_c[h_{t-1}, x_t] + b_c)) \quad (4)$$

$$O_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \quad (5)$$

$$h_t = O_t \tanh(C_t) \quad (6)$$

式中: σ 为 sigmoid 函数; $\tanh$ 为双曲正切激活函数; f 、 i 、 C 、 O 分别为遗忘门、输入门、更新后的细胞状态和输出门; h_{t-1} 为上一个 LSTM 单元的输出; h_t 为最终输出; x_t 为当前时刻的输入; b_i 、 b_f 、 b_c 、 b_o 为偏置向量; W_i 、 W_f 、 W_c 、 W_o 为权重矩阵; C_{t-1} 为前一单元的状态; C_t 为当前细胞状态。

BP 神经网络是一种按误差反向传播训练的多层前馈网络,求解方法是梯度下降法,利用梯度搜索技术,以期使网络的实际输出值和期望输出值的误差均方差为最小;SVM 建立一个超平面作为决策曲面,使正例和反例之间的隔离边缘被最大化;ELM 是一类基于前馈神经网络构建的机器学习方法,适用于监督学习和非监督学习问题;RF 本质上是使用随机方法生成多个决策树的分类器。为了提高预测精度,采用 TLBO 算法来优化 BP 和 SVM 的参数,采用 PSO 算法优化 ELM 的参数。以上5个模型的部分参数设置如下:LSTM 初始学习率为 0.005、学习率衰减速度为 0.2、迭代次数为 250、隐含层节点数为 200、BP 迭代次数为 100、学习率为 0.001,网络结构为 5-18-1, SVM 的核函数使用径向基核函数, ELM 网络参数设置为 5-50-1, RF 决策树个数设置为 500。

1.3 集成策略及流程

经上述5个子学习器训练学习,使用 MLR 方法进行结果集成,其数学表达式为

$$Y_o = \sum_{i=1}^5 \partial_i y_i \quad (7)$$

式中： y_i 为初级预测每一类的预测输出； ∂_i 为加权因子； Y_o 为每一类的最终输出。加权因子由多元线性回归方法确定，将所得的最终输出按时间序列串行整合，即可得出能耗预测结果。

集成策略流程如图1所示。

2 案例分析

本文采用不同时间尺度、能耗规模的预测场景对模型的预测精度和泛化能力进行分析。具体包括案例A、B、C。案例A为能源预测竞赛数据，共计4 208组，时间步长为1 h，数据特征多，完整性好，用于模型精度的基本验证；案例B为美国怀俄明大学信息技术中心大楼2017年9—12月的实测数据，共计2 500组，时间步长为1 h，代表单栋教育类建筑物的短期用能场景；案例C为仪征市2013年9月至2016年9月每日电能数据，共计1 090组，时间步长为1 d，数据特征较少，代表能量规模较大的长期电力用能场景。这3个案例的原始特征如表1所示。

表1 3个案例的原始特征

Table 1 Original features of three cases

案例	特征
A	前1 h历史电能耗 y_1 ，前2 h历史电能耗 y_2 ，室内干球温度 t ，太阳能辐射 s ，时间的正余弦值 s_h 、 c_h 和节假日标志 f
B	前1 h历史电能耗 y_1 ，前2 h历史电能耗 y_2 ，温度 t ，湿度 h ，压力强度 p ，时间的正余弦值 s_h 、 c_h 和节假日标志 f
C	前1 d历史电能耗 d_1 ，前2 d历史电能耗 d_2 ，前7 d历史电能耗 d_3 ，温度 t 和节假日标志 f

2.1 数据特征选择

为获得最优数据集，用RFE对3个案例进行特

征选择，最终的特征选择结果如表2。

表2 3个案例特征选择结果

Table 2 Results of three cases features selection

模型	A	B	C
LSTM	y_1, s, f, c_h, t	y_1, t, p	d_1, d_2, t, f, d_7
TLBO-SVM	y_1, y_2, s_h, f, s, c_h	$y_1, y_2, s_h, f, t, p, c_h$	d_1, d_2, f, t, d_7
PSO-ELM	y_1, y_2, s_h, c_h	y_1, c_h, y_2, h, p	d_1, d_2, t, f, d_7
RF	$f, t, y_1, s, y_2, c_h, s_h$	$f, p, c_h, s_h, y_1, t, y_2$	f, d_1, d_2, d_7, t
TLBO-BP	$y_1, f, t, s, c_h, s_h, y_2$	y_1, y_2, s_h, c_h, f, t	d_2, d_1, d_7, t, f

数据集按照5:3:2划分为训练集、验证集和测试集，并按照节假日标志 f 将案例各数据划分为工作日和非工作日。在此基础上，通过聚类有效性指标确定聚类个数，分别对训练集进行FCM聚类分析，并使用KNN参照聚类结果对验证集和测试集数据进行分类。最终案例A数据一共分为6类，案例B数据分为8类，案例C因数据样本过少，故仅按照节假日标志 f 将案例C各数据划分为工作日和非工作日2类。

2.2 结果与分析

为验证此模型性能，本研究对3个案例分别做几组对比试验：①采用相同预处理方法，对单个模型进行预测比较；②与现有预测模型预测结果相比较。重复试验操作20次，取平均值作为最终结果。

2.2.1 与单个模型比较

表3为3个案例单个模型和本文所提出的集成模型的预测结果的误差分析。

从表3可以看出，本文的集成模型预测值均为最佳精度，采用相同预处理方法，3个案例的预测结果与实际电能耗的误差： M_{APE} 分别为0.82%、0.74%、1.77%； M_{AE} 分别为5.16、15.27、16.85； R_{MSE} 分别为7.47、26.72、24.44。此外，模型预测值和真实值也进行了相关性分析。比较而言，本文集成模型的预测相关系数值最大，3个案例分别为0.998 5、0.983 7、0.930 5。图3为3个案例集成模型的能耗数据预测。

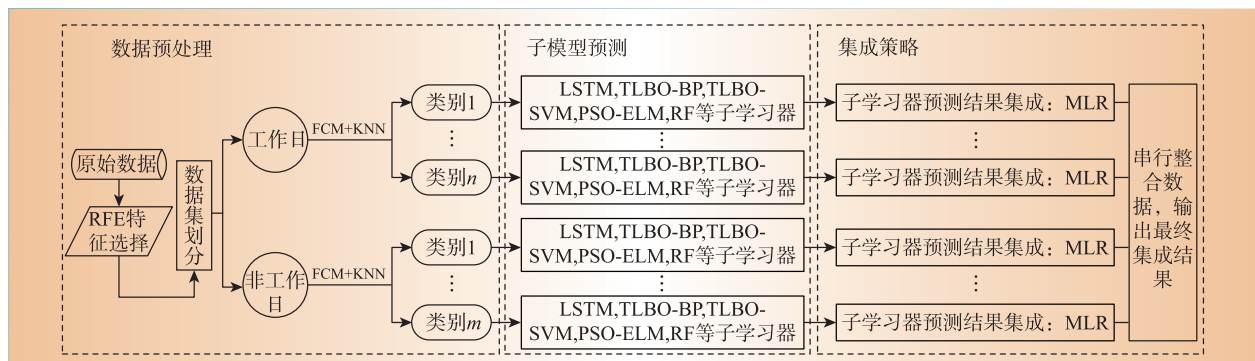


图1 提出的集成策略框架

Fig. 1 Framework of the proposed ensemble strategy

表3 单个模型及集成结果性能比较

Table 3 Performance comparison results between single model and ensemble

案例	模型	$M_{APE}/\%$	M_{AE}	R_{MSE}
A	LSTM	1.73	11.19	17.02
	TLBO-SVM	1.52	9.84	16.13
	PSO-ELM	0.92	5.97	8.64
	RF	2.05	13.15	19.35
	TLBO-BP	1.55	10.11	15.99
	本文集成策略	0.82	5.16	7.47
B	LSTM	1.31	27.52	40.78
	TLBO-SVM	1.79	37.37	58.62
	PSO-ELM	1.07	22.20	31.80
	RF	1.45	30.21	47.81
	TLBO-BP	1.50	31.36	46.15
	本文集成策略	0.74	15.27	26.72
C	LSTM	4.23	42.20	57.88
	TLBO-SVM	9.04	98.54	159.14
	PSO-ELM	1.78	16.86	24.46
	RF	5.49	56.86	79.57
	TLBO-BP	4.40	44.71	60.60
	本文集成策略	1.77	16.85	24.44

注: R_{MSE} 和 M_{AE} 的单位: 案例A为kWh, 案例B为kWh, 案例C为104 kWh, 表4同。

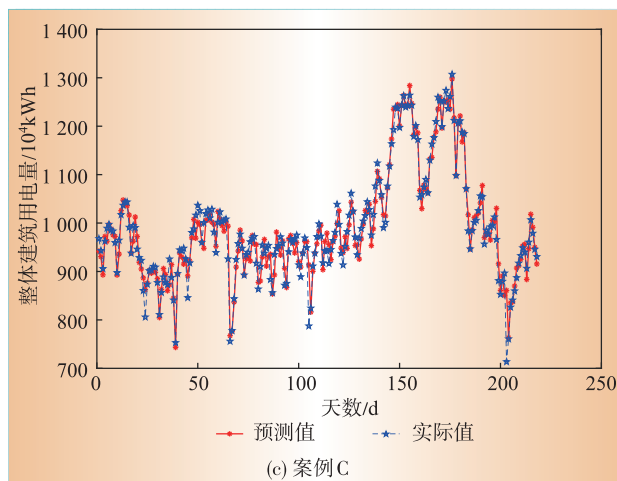
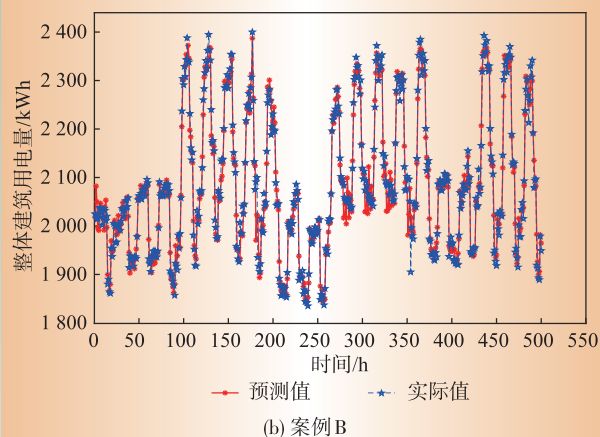
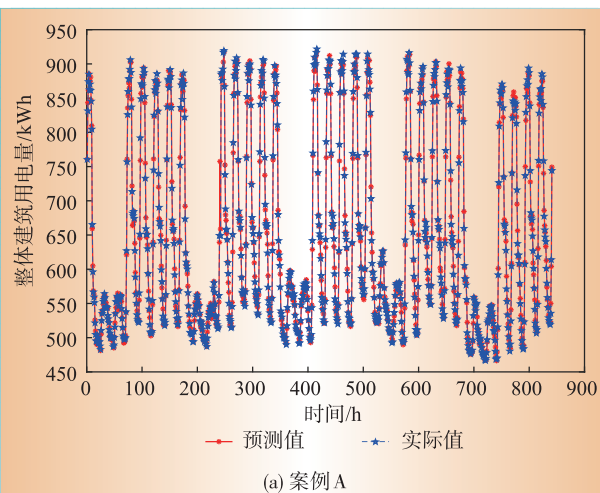


图3 集成模型的能耗数据预测

Fig. 3 Prediction of energy consumption data from ensemble model

2.2.2 与现有预测模型对比

为进一步验证此集成模型性能,选择与本文使用相同数据集的预测模型^[13-16]进行对比,如表4所示。从表4看出,本文所提出的集成策略在不同尺度上均取得了更好的预测精度。

表4 与现有预测模型的误差比较

Table 4 Comparison of error with existing prediction models

案例	模型	$M_{APE}/\%$	M_{AE}	R_{MSE}
A	模型1 ^[13]	1.48	9.29	14.96
	模型2 ^[14]	1.42	9.10	14.40
	模型3 ^[15]	1.03	9.69	6.31
	本文集成策略	0.82	5.16	7.47
B	模型1 ^[13]	1.04	21.72	32.64
	模型2 ^[14]	1.31	31.99	23.48
	模型3 ^[15]	1.04	21.54	29.36
	本文集成策略	0.74	15.27	26.72
C	模型2 ^[14]	3.96	38.84	52.27
	模型4 ^[16]	3.72		
	本文集成策略	1.77	16.85	24.44

3 结束语

根据建筑用能规律,本文提出一种基于建筑物能耗特点的分类集成式能耗预测方法。通过多个不同时间尺度建筑电力用能案例进行验证,得到以下结论。

(1) 数据预处理方法充分考虑用能规律,数据聚类方式能适用于不同建筑用能类型的预测场景。

(2) 集成预测方法采用多样化的子模型,解决了单个模型对预测场景敏感的问题,提高了预测模型的鲁棒性和泛化能力。

今后将通过深入研究建筑物能源需求模式,引入迁移学习等工具,进一步提高在历史数据稀缺情况下的建筑电能耗预测准确性。D

参考文献:

- [1] 叶水泉. 建筑能源低碳化的思考与实践[J]. 电力需求侧管理, 2015, 17(2): 5-8.
YE Shuiquan. Thinking and practice of low carbonization building energy sources [J]. Power Demand Side Management, 2015, 17(2): 5-8.
- [2] 王力, 张超. 机场能耗的时间序列混合预测方法[J]. 中国民航大学学报, 2017, 35(6): 31-35.
WANG Li, ZHANG Chao. Hybrid model of time series for airport energy prediction [J]. Journal of Civil Aviation University of China, 2017, 35(6): 31-35.
- [3] LI K, XIE X, XUE W, et al. A hybrid teaching-learning artificial neural network for building electrical energy consumption prediction [J]. Energy and Buildings, 2018, 174: 323-334.
- [4] XIAO J, LI Y, LING X, et al. A hybrid model based on selective ensemble for energy consumption forecasting in China [J]. Energy, 2018, 159: 534-546.
- [5] CHEN K, JIANG J, ZHENG F, et al. A novel data-driven approach for residential electricity consumption prediction based on ensemble learning [J]. Energy, 2018, 150: 49-60.
- [6] LIN Y, LUO H, WANG D, et al. An ensemble model based on machine learning methods and data preprocessing for short-term electric load forecasting [J]. Energies, 2017, 10(8): 1 186.
- [7] 张鹏飞, 胡博, 何金松, 等. 基于时空图卷积网络的短期空间负荷预测方法[J]. 电力系统自动化, 2023, 47(13): 78-85.
ZHANG Pengfei, HU Bo, HE Jinsong, et al. Short-term spatial load forecasting method based on spatio-temporal graph convolutional network [J]. Automation of Electric Power Systems, 2023, 47(13): 78-85.
- [8] 林勇, 邹晶晶, 左郑敏, 等. 基于改进 PSO 算法的 Logistic 模型在饱和负荷预测中的应用[J]. 电力需求侧管理, 2015, 17(5): 6.
LIN Yong, ZOU Pinjing, ZUO Zhengmin, et al. Application of the logistic model based on improved particle swarm optimization algorithm in saturated load forecasting [J]. Power Demand Side Management, 2015, 17(5): 6.
- [9] 张凯, 冯剑, 刘建华, 等. 基于模糊 C 均值聚类算法的用电行为模式分类[J]. 电力需求侧管理, 2022, 24(3): 6.
ZHANG Kai, FENG Jian, LIU Jianhua, et al. Power consumption behavior pattern classification based on fuzzy C-mean clustering algorithm [J]. Power Demand Side Management, 2022, 24(3): 6.
- [10] 曹瑜, 王楠, 徐志超. Spark 框架结合分布式 KNN 分类器的网络大数据分类处理方法[J]. 计算机应用研究, 2019, 36(11): 5.
CAO Yu, WANG Nan, XU Zhichao. Network big data classification processing method based on Spark and distributed KNN classifier [J]. Application Research of Computers, 2019, 36(11): 5.
- [11] 庞传军, 张波, 余建明. 基于 LSTM 循环神经网络的短期电力负荷预测[J]. 电力工程技术, 2021, 40(1): 175-180, 194.
PANG Chuanjun, ZHANG Bo, YU Jianming. Short-term power load forecasting based on LSTM recurrent neural network [J]. Electric Power Engineering Technology, 2021, 40(1): 175-180, 194.
- [12] 张磊, 王小明, 吴红斌, 等. 基于 FCM 和 LSTM 的光伏功率超短期预测[J]. 供用电, 2023, 40(1): 10-17.
ZHANG Lei, WANG Xiaoming, WU Hongbin, et al. Ultra-short-term prediction of PV power based on FCM and LSTM [J]. Distribution & Utilization, 2023, 40(1): 10-17.
- [13] LI K, TIAN J, XUE W, et al. Short-term electricity consumption prediction for buildings using data-driven swarm intelligence based ensemble model [J]. Energy and Buildings, 2020, 231: 110 558.
- [14] TIAN J, LI K, XUE W. An adaptive ensemble predictive strategy for multiple scale electrical energy usages forecasting [J]. Sustainable Cities and Society, 2021, 66: 102 654.
- [15] LI K, ZHANG J, CHEN X, et al. Building's hourly electrical load prediction based on data clustering and ensemble learning strategy [J]. Energy and Buildings, 2022, 261: 111 943.
- [16] LI K, XIE X, XUE W, et al. Hybrid teaching-learning artificial neural network for city-level electrical load prediction [J]. Science China (Information Sciences), 2020, 63(5): 212-214.

作者简介:

唐倩倩(1995), 女, 安徽合肥人, 硕士研究生, 研究方向为建筑电能耗预测;

李康吉(1979), 男, 江苏无锡人, 教授, 博士研究生导师, 研究方向为数据驱动的多尺度能量系统建模与控制;

魏伯睿(2000), 男, 河南商丘人, 硕士研究生, 研究方向为基于迁移学习的建筑电能耗预测。

(责任编辑 水 鹤)