潘瑜. 基于天气数据和历史工况的城市供热系统 LightGBM 预测模型[J]. 智能计算机与应用,2025,15(12):191-197. DOI:10.20169/j.issn.2095-2163.24052602

基于天气数据和历史工况的城市供热系统 LightGBM 预测模型

潘 瑜
(南京审计大学 商学院, 南京 211800)

摘 要: 本研究旨在通过融合实时天气数据和换热站历史工况数据,运用先进的 LightGBM 模型对城市供热系统的热负荷进行高精度预测。通过研究山东省济南市一座热源厂及 8 个换热站组成的供热系统,收集其历史运行数据和天气信息并进行预处理和特征工程,利用皮尔逊相关系数法,成功筛选出与目标变量高度相关的关键特征。本研究主要使用的是 LightGBM 模型,通过贝叶斯调参方法优化模型的超参数,并设立 3 个基准模型:决策树(Decision Tree,DT)、随机森林(Random Forest,RF)和梯度提升决策树(Gradient Boosting Decision Tree,GBDT)作为比较模型进行对比分析。实验结果表明,LightGBM 模型在预测换热站瞬时热量方面表现优异,其平均绝对误差(MAE)、均方根误差(RMSE)和平均绝对百分比误差(MAPE)这 3 项指标均低于其他模型,这一结果验证了该模型的高效性和准确性。本研究为城市供热系统的智能调控提供了新的解决方案,有助于减少能源浪费,提高供热舒适度,推动智慧城市的发展。

关键词: 城市供热系统; LightGBM 模型; 热负荷预测; 贝叶斯调参; 机器学习
中图分类号: TP181 **文献标志码**: A **文章编号**: 2095-2163(2025)12-0191-07

Accurate thermal load prediction for urban heating systems using LightGBM and integrated data

PAN Yu
(School of Business, Nanjing Audit University, Nanjing 211800, China)

Abstract: The aim of this study is to use the advanced LightGBM model to predict the heat load of urban heating system with high precision by integrating real-time weather data and historical working condition data of heat exchange station. In this paper, by studying the heating system composed of a heat source plant and eight heat exchange stations in Jinan, Shandong Province, collecting its historical operation data and weather information, and conducting pre-processing and feature engineering, we successfully screened out the key features highly correlated with the target variables by using Pearson correlation coefficient method. In this study, the LightGBM model is mainly used, and the hyperparameters of the model are optimized by Bayesian tuning method, and three benchmark models: Decision Tree (DT), Random Forest (RF) and (Gradient Boosting Decision Tree,GBDT) are set up as comparison models for comparative analysis. The experimental results show that LightGBM model has excellent performance in predicting instantaneous heat of heat exchange station, and its Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) are all lower than other models, which verifies the efficiency and accuracy of the model. This study provides a new solution for the intelligent regulation of urban heating system, which helps to reduce energy waste, improve heating comfort, and promote the development of smart cities.

Key words: urban heating system; LightGBM model; heat load forecasting; Bayesian tuning; machine learning

0 引 言

随着持续提高的城镇化率,对于中国的北方地区,冬季供暖是确保民生和城市正常运转的关键。然而传统的供热调控方式往往存在着反应滞后,精度较差的问题,导致难以及时有效地应对气温波动,这不仅会造成热能浪费,同时还会降低供热的舒适度。随着人工智能技术的快速发展,将其应用于供热系统的优化调控具有重大的理论与现实意义。城市供热系统通常呈现的是“源-网-站-户”的分层结构,热源厂通过一次管网将热量输送至各个换热站,再由换热站将热量传递给小区的居民用户。在气温

基金项目: 国家自然科学基金(72072092)。
作者简介: 潘 瑜(2001—),男,硕士研究生,主要研究方向:大数据驱动的商务分析与决策优化。Email:ulysess062@163.com。
收稿日期: 2024-05-26

降低时,热源厂会采取提高供水温度和增大流量等措施,以确保热量输送足以维持用户室温。然而,单纯依赖人工经验很难及时准确地调节供热参数。本文探讨将机器学习算法与实时工况数据相结合,动态预测同一供热系统中多个换热站的热负荷需求,并基于预测结果智能调节供热量,实现供热系统的精准控制。通过 8 个换热站的历史运行数据和气象数据,构建瞬时热负荷的预测模型,结合未来天气预报数据,计算出未来 24 h 内每个换热站的热量需求。图 1 为该换热系统的拓扑图。

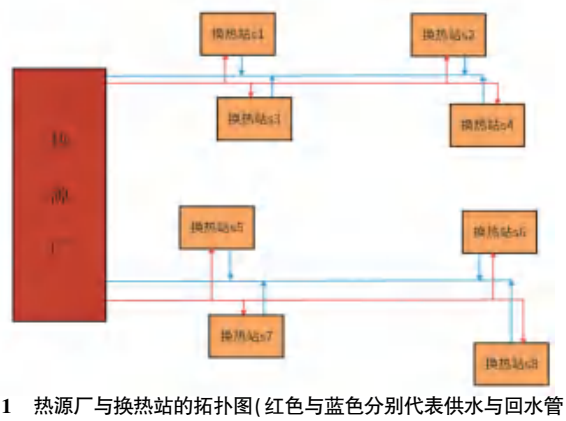

图 1 热源厂与换热站的拓扑图(红色与蓝色分别代表供水与回水管)

Fig. 1 Topology diagram of heat source plant and heat exchange station (red and blue represent water supply and return pipes respectively)

目前,国内外学者在供热负荷预测方面已经取得了一定的成果。随着能源管理技术和预测技术的快速发展,多种先进的算法被应用于供热负荷领域,预测精度也在不断提高。Eiman 等[1]采用支持向量机和 FFA (Firefly Algorithm)算法。建立了 7 个不同时间范围的 SVM-FFA 预测模型。将 SVM-FFA 模型与遗传规划、人工神经网络和 SVM 模型进行比较。结果表示 SVM-FFA 效果最优。张腾达等[2]则针对室外气温和风力等关键因素,利用长短期记忆神经网络(LSTM)对区域供热系统进行了深入研究,展示了时间序列数据在负荷预测中的重要作用。谢文举等[3]提出了一种基于粒子群(Particle Swarm Optimization,PSO)、经验模态分解(Empirical Mode Decomposition,EMD)、卷积神经网络和长短时记忆神经网络的混合预测模型。使用 EMD 分解供热负荷序列,利用卷积神经网络来挖掘输入数据与不同模态之间的耦合关系。利用长短期记忆网络对不同模态分量进行预测,并使用 PSO 算法优化参数。Sajjadi 等[4]提出了一种基于极限学习机(Extreme Learning Machine, ELM)的短期、多步预测模型,用于预测连接至集中供暖系统的消费者的热负荷。该模型能够捕捉数据中的非线性关系,在训练速度、预测精度和泛化能力方面表现出色。马文菁等[5]则通过引入 PSO 算法对 ELM 的输入权值、隐含层阈值进行优化, PSO-ELM 预测结果比 ELM 预测结果更接近实际值。果泽泉等[6]以湖北省十堰市一个集中供热热力站为对象,分别采用随机森林、XGBoost、BP 神经网络、SVR、LSTM 神经网络 5 种方法进行预测模型训练及测试,使用 PSO 算法优化各模型参数,在此基础上针对不同模型在不同短期负荷预测情景下的表现进行对比研究。张玲等[7]提出了一种基于随机森林和卷积神经网络的热负荷预测模型,通过降低数据维度和简化网络结构,有效提高了预测精度和泛化能力。孙玉芝等[8]提出了一个双层 LSTM 基础模型和平滑预处理方案相结合的方法。顾伟等[9]提出了基于改进果蝇算法优化的广义回归神经网络结合天气因素的短期热负荷预测方法。Yu 等[10]介绍了实际 BDES 的两个案例研究,利用 XGBoost、LSTM、GRU 和添加注意力的 LSTM/GRU 5 种数据驱动模型,研究了多输入多输出预测策略的性能。Habib 等[11]提出了一种结合无监督和监督机器学习的数据挖掘与数据驱动建模方法,使用一种多特征人工神经网络(ANN)回归模型用于 DH 负荷预测,得出结论在室外温度作为预测的参数时,开发准确的负荷预测模型需要更多指标,Davy 等[12]使用机器学习和专家建议(Expert advice)进行区域供热网络运营热负荷预测的方法,探讨了不同机器学习算法的应用效果,为提高预测精度提出了一种基于专家建议的集成方法。Wang 等[13]提出了一种组合预测模型来克服单一模型的局限性。在该模型中,选取 4 个经典的单一模型作为个体模型进行预测,利用组合预测误差算法的最小平方和来计算各个模型的权重系数。

从以上综述可知,供热负荷预测领域具有多样化的研究方法和策略。从支持向量机、萤火虫算法到深度学习模型如 LSTM、CNN,以及混合预测模型和集成方法,不同算法和技术在提升预测精度、鲁棒性和泛化能力方面均取得了显著成果。特别是随着人工智能和大数据技术的发展,供热负荷预测将更加依赖于先进的算法和模型,以实现更高效、准确的能源管理和优化。

1 模型框架

如图 2 所示,本文的核心研究流程如下:本文汇集了热力站、各换热站的历史运行数据以及每日的

天气信息,对这 3 部分数据进行综合整合,并开展预处理工作,主要目的是填补数据中的缺失值。基于已有数据,本文设计了新的数据特征,并对热负荷与各参数间的关联性进行了深入分析,从中筛选出与目标变量(即热负荷)高度相关的特征。对选定的特征进行归一化处理后,进一步将数据划分为训练集和测试集。构建预测模型时,选择了 LightGBM 作为主要工具,并利用筛选出的特征进行模型训练。为了优化模型的性能,使用贝叶斯调参方法对 LightGBM 模型进行参数调整。贝叶斯调参可以有效地探索模型参数的空间,找到使模型性能最优的参数组合。为了进行性能对比,额外建立了决策树(DT)、随机森林(RF)[14] 和梯度提升决策树(GBDT)[15]3 个模型作为基准模型。这些模型也将经过相同的预处理、特征选择和归一化步骤,并使用相同的训练集进行训练。为确保预测精度的严格评估,选用了 3 个关键的绩效指标:平均绝对误差(MAE)、均方根误差(RMSE)和平均绝对百分比误差(MAPE),以全面衡量预测结果的准确性。将这些指标应用于所有模型,包括 LightGBM、DT、RF 和 GBDT,并进行比较,以评估不同模型的性能。

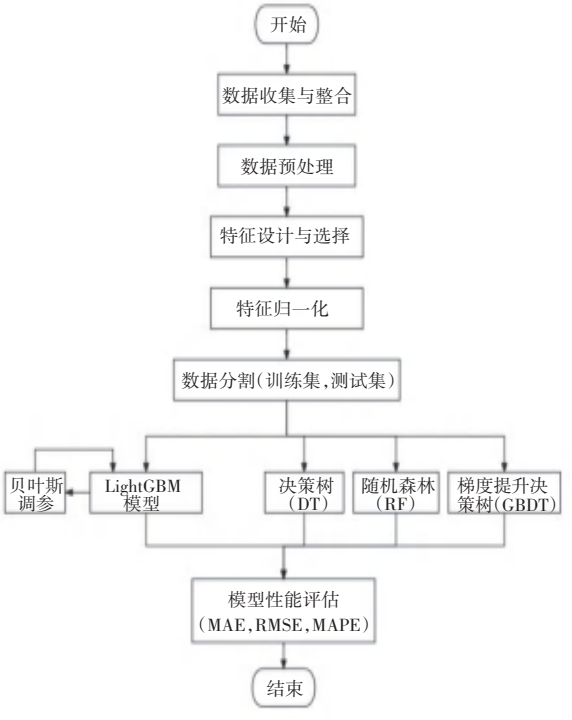

图 2　研究框架

Fig. 2　Research framework

2　数据收集与预处理

2.1　Data sample 数据样本

本文的实验数据来源于山东省济南市的一座热源厂和 8 个换热站从 2023 年 12 月 12 日到 2024 年 1 月 1 日 20 天内记录的天气信息与历史工况,该数据集包括 3 部分,分别为各换热站数据,热源厂数据以及每天的天气数据。其中换热站数据包括各个换热站供水与回水的温度和压力,以及每个时间段瞬时流量和瞬时温度。热源厂数据包括供水和回水的温度,压力以及瞬时流量。天气数据包括当天的气温、风力等级、日出时间、日落时间和具体天气情况。

2.2 数据处理与特征工程

本次实验收集到的数据主要分为 3 部分,其中各数据集的时间都为分钟格式,与热源厂和天气数据不同的是,换热站的数据包含 8 个换热站的数据,因此按分钟将热源厂数据和天气数据合并到换热站数据中,后续预测的结果也是每分钟各个换热站的瞬时热量。

在应对数据集中出现的缺失值时,本文采取的是滚动平均法来进行有效的填充。该方法的核心在于计算一个固定大小为 10 的窗口内的数据平均值,并将这些计算得出的平均值填补到原始数据中的缺失值位置。这种方法特别适用于具有时间序列特性的数据,因为其能够充分利用每个换热站的历史数据趋势,为缺失值提供合理的估计。滚动平均法不仅能够有效地捕捉数据中的季节性和趋势特征,还能够将这些特征融合到缺失值的填充过程中,从而确保填充值的准确性和合理性。作为一种常见且高效的数据预处理方法,滚动平均法能够充分利用数据的内在特性,为后续的数据分析和建模工作提供更为完整和可靠的数据基础。

由于供热系统本身的复杂性,气象数据和系统的运行状态可能都会对预测产生一定的影响,本文采用皮尔逊相关系数法对每个特征进行相关性分析,来决定哪些特征作为预测模型的特征进行训练。该方法的计算表达如下式:

$$r=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}} \tag{1}$$

其中 $\bar{X}$ 和 $\bar{Y}$ 分别表示样本均值。

在探讨两个变量之间的相关性时,若其绝对值在(0,0.3)的范围内,通常认为这两个变量之间的相关性较弱;若位于(0.3,0.7)区间,则表示这两个变量存在中等程度的相关性;而当其绝对值接近或达到(0.7,1)区间时,则表明这两个变量之间存在强烈的相关性。如图 3 所示,其中 LAG 为创建的滞

后特征。而日出时间、一网回水压力、供水瞬时流量、回水瞬时流量、回水压力等特征对应的相关系数都较小,属于弱相关,因此不作为模型的训练特征。

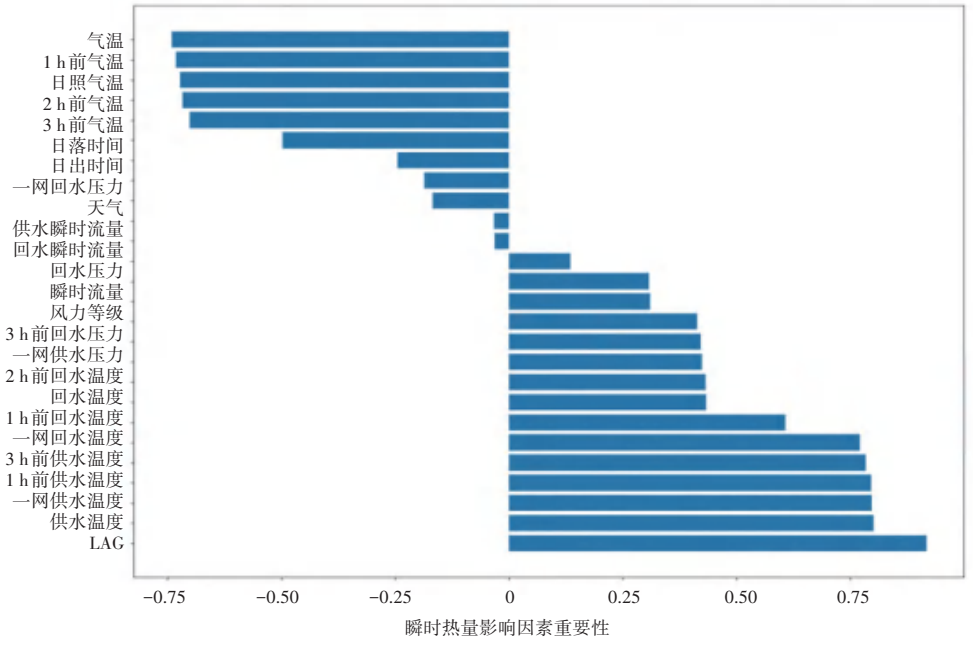

图 3 不同变量与瞬时热量的相关系数

Fig. 3 Correlation coefficients between different variables and instantaneous heat

在完成数据预处理和特征方程后,按时间范围将数据集分割为训练集和测试集,其中测试集为 2023 年 12 月 12 日到 2023 年 12 月 27 日的数据,验证集为 2023 年 12 月 28 日到 2024 年 1 月 1 日的数据,具体见表 1。

表 1 训练集和测试集的数据分布

Table 1 Data distribution of training set and test set

Dataset	Time Range	Features	Label	Samples
Training	2023/12/12-2023/12/27	筛选后的特征	瞬时热量	184 320
Validation	2023/12/28-2024/01/01	筛选后的特征	瞬时热量	57 600

3 预测模型

3.1 模型介绍

本文构建的预测模型为 LightGBM 模型,利用筛选出的特征进行模型训练,并使用贝叶斯调参方法对该模型进行参数调整。同时,额外建立了决策树(DT)、随机森林(RF)和梯度提升决策树(GBDT) 3 个模型作为基准。这些模型也将经过相同的预处理、特征选择和归一化步骤,并使用相同的训练集进行训练。

其中,决策树(DT)是一种通过构建二叉树结构来对数据进行分类的概率分析模型[16]。这种方法利用树状图的形式,将数据的特征作为树的节点,并根据这些特征将数据集划分为不同的子集,最终达到分类或回归的目的。随机森林(RF)则是一种通过集成学习中的 bagging 方法[17],显著提升单个决策树预测性能的算法[18]。在随机森林中,多个决策树被并行地训练,并且每个树都基于原始数据的一个随机子集和一个特征子集进行构建。梯度提升决策树(GBDT)则是基于 Boosting 思想发展而来的一种集成学习算法。GBDT 通过迭代地训练多个决策树,这些树被加性地组合在一起,形成一个强大的预测模型,采用前向分步算法,通过梯度下降法来优化损失函数,并找到最优的决策树组合方式。由于 GBDT 算法每一次输入的数据都依赖上一次训练的输出结果,因此训练过程需要多次对整个数据集进行遍历,会导致算法运算的时间成本增加。为了优化这一问题,2017 年微软公司提出了 LightGBM(Light Gradient Boosting Machine)算法。

LightGBM 算法是对 GBDT 算法的显著改进,尤

其在处理海量数据方面展现出了卓越的性能。在实际应用中，LightGBM 算法因其高效性和准确性而备受推崇。同时，丰富的参数调节选项，使用户可以根据具体需求进行模型性能优化。在模型训练的过程中，LightGBM 模型采用 GOSS 方法来对训练过程进行加速，该技术通过在计算样本梯度大小时只保留梯度较大的样本，来减少决策树的训练时间。LightGBM 的算法可以表示为：

$$\text{obj}^{(t)} = \sum_{i=1}^{n} l(y_i, H_{t-1}(x_i)) + \sum_{i=1}^{t-1} \Omega(h_i) \qquad (2)$$

式中：l 表示损失函数，H_{t-1} 表示前 $t-1$ 轮迭代的结果，Ω 表示正则化项，h_i 表示第 i 棵树，y_i 为预测目标值。模型迭代过程如图 4 所示。

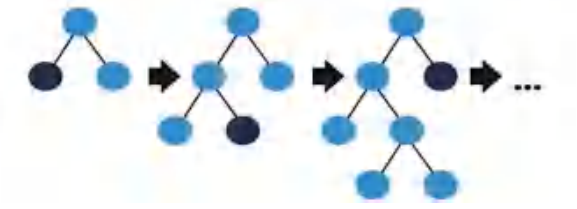

图 4　LightGBM 模型迭代过程

Fig. 4　LightGBM model iteration process

3.2　参数设置

LightGBM 算法的超参数大致可以划分为几个核心类别，包括核心参数、学习控制参数、IO 参数、目标参数和度量参数等。在这些参数中，有几个特别关键的超参数对于模型的表现具有重要影响。num_leaves，其决定了每棵树上的最大叶子数目，初始值设为 31。min_data_in_leaf，这个参数定义了每个叶子节点所需的最少样本数，初始值设为 20。learning_rate 即学习率，其控制着模型学习过程中的步长，初始值通常为 0.01。此外，feature_fraction 和 bagging_fraction 也是影响模型性能的关键参数。feature_fraction 代表了在每次迭代中随机选取的特征比例，其初始值设为 1。bagging_fraction 决定了在迭代过程中使用的数据比例，其初始值同样设为 1。

为了进一步提升 LightGBM 模型的性能，本文采用了贝叶斯调参方法[19]来进行超参数的优化。这种方法能够有效地探索模型参数的空间，帮助发现能使模型性能达到最优的参数组合。通过贝叶斯调参，本文期望能够提升模型的复杂度和拟合能力，进而增强模型的预测效果。该方法具体是采用 Python 中 bayes_opt 库的 BayesianOptimization 函数来寻找最优参数。最终得到最佳的超参数设置，见表 2。

表 2　贝叶斯调参后的最佳超参数

Table 2　Optimal hyperparameters after Bayesian tuning

超参数	最佳取值
num_leaves	32.00
min_data_in_leaf	18.00
learning_rate	0.01
feature_fraction	0.42
bagging_fraction	0.61

4　实验结果与分析

4.1　评估标准

为了验证 LightGBM 模型的有效性，本文采用平均绝对误差（MAE）、均方根误差（RMSE）和平均绝对百分比误差（MAPE）这 3 个指标作为评估标准，这些指标在机器学习和数据分析领域中被广泛认可，其能够全面、准确地衡量模型预测值与真实值之间的差异。对比 LightGBM 模型与另外建立的 3 个基准模型的指标，观察 LightGBM 模型在各项指标上的表现是否优于基准模型。平均绝对误差（MAE）是预测误差绝对值的平均，MAE 越小，说明模型预测结果越准确。均方根误差（RMSE）是预测误差平方的平均值的平方根。其对于较大的误差更为敏感，RMSE 越小，说明模型在整体上表现越好。平均绝对百分比误差（MAPE）是预测误差百分比的平均值。由于 MAPE 是基于百分比的，因此能够消除不同数据集之间量纲差异的影响，使得不同数据集之间的模型性能更加公平。MAPE 越小，说明模型在百分比尺度上的预测精度越高。

4.2　结果分析

本文成功构建了一个 LightGBM 模型，旨在预测某工人系统中 8 个换热站分钟级的瞬时热量，在使用了贝叶斯调参方法对超参数进行调整后，得到效果最优的预测模型。图 5 为该预测模型预测结果与实际值的时间序列对比图，通过对比预测值与实际值的时间序列，可以观察到除了换热站 s5（其预测值与实际值的最大差值不超过 1 ℃），大部分模型拟合效果都较好，预测值与实际值的偏差不超过 0.5 ℃，因此可以确定该模型的拟合效果较好。

在验证了该模型的拟合效果后，将模型的 3 个评价指标：MAPE、RMSE 和 MAE 与建立的基准模型进行对比，从图 6 的柱状图可以看出，LightGBM 模型的 3 项评估指标均呈现出最低值，这一结果凸显了 LightGBM 模型卓越的性能，证明其在实际应用中的高效性和准确性。

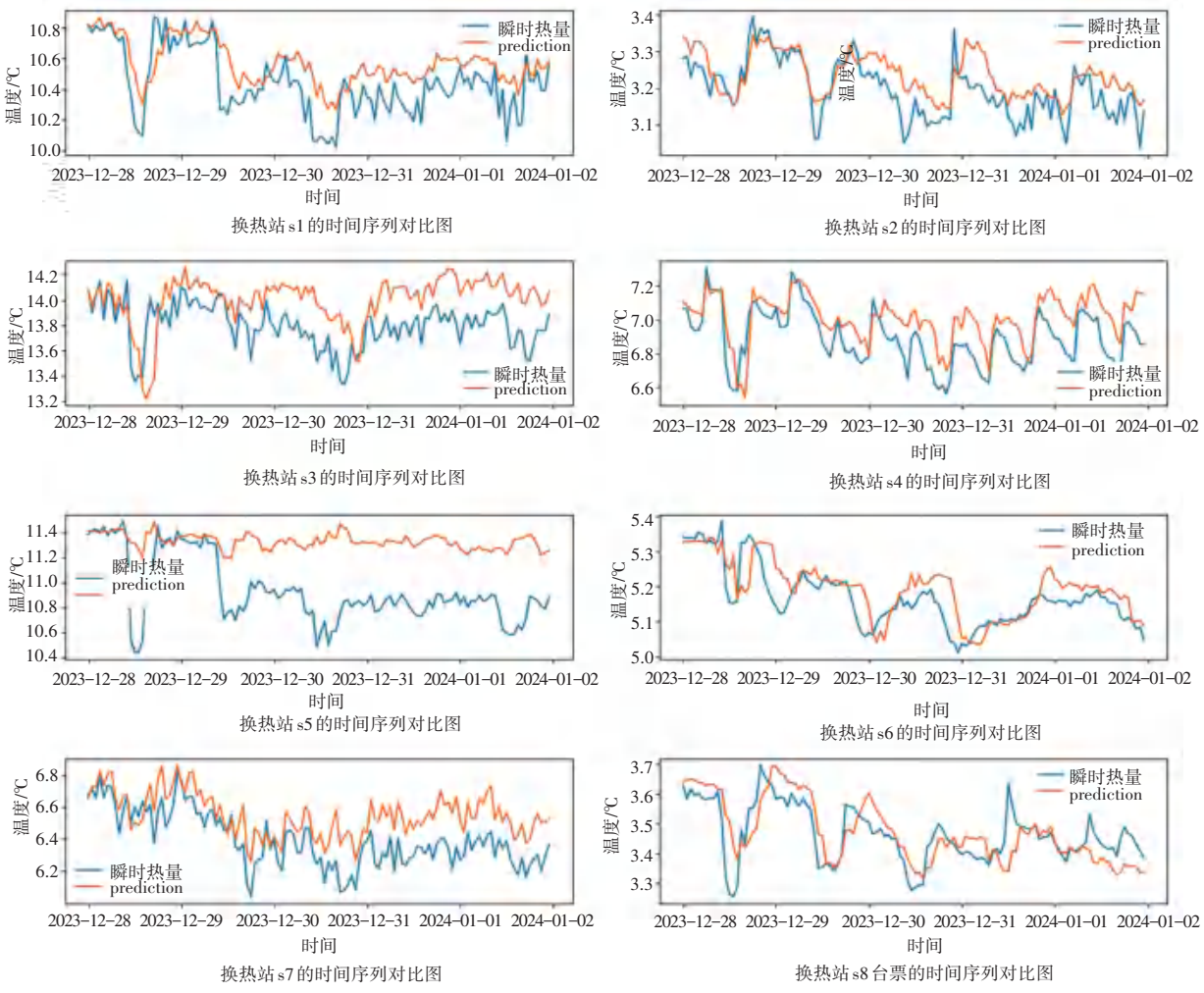

图 5 预测值与实测数据时间序列对比图

Fig. 5 Time series comparison between predicted value and measured data

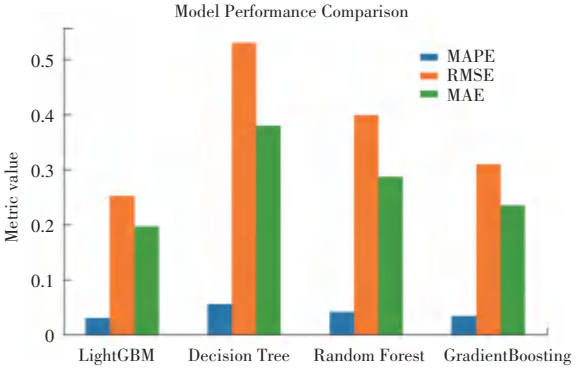

图 6 不同模型的评估指标对比图

Fig. 6 Comparison of evaluation indicators of different models

5 结束语

精准预测换热站瞬时热量对减少能源损耗,提高用户舒适度具有重大意义,同时还有助于智慧城市的建设。通过不断优化预测模型,能够更准确地掌握热源厂的热力输出,从而实现能源的高效利用和合理调配。为了更精准地预测换热站瞬时热量,本文引入了 LightGBM 模型,并针对 8 个换热站未来 24 小时的瞬时热量进行了深入研究。研究结果表明,LightGBM 模型在预测换热站瞬时热量方面展现出了卓越的性能。其预测值与实际值之间的误差极小,充分证明了该模型的准确性和可靠性。与其他 3 种基准模型相比,LightGBM 模型在 3 项关键评价指数 MAPE、RMSE 和 MAE 上均取得了最低值,进一步巩固了其在预测换热站瞬时热量方面的领先地位。

通过构建模型相关特征,优化模型的超参数,本文实现了对瞬时热量的高精度预测,结果表明这一方法不仅可以对热源厂实时调节供热策略起到指导作用,还可以为供热系统自动化控制奠定基础,有望为供热系统的智能调控带来新的契机。

参考文献

[1] EIMAN T, KEIVANI A, SHAMSHIRBAND S, et al. Prediction of heat load in district heating systems by support vector machine

with firefly searching algorithm[J]. Energy, 2016, 95: 266-273. DOI:10.1016/j.energy.2015.11.079
[2] 张腾达, 李琦, 陈波. 基于 LSTM 的热力站短期热负荷预测研究[J]. 计算机仿真, 2022, 39(9): 507-512.
[3] 谢文举, 薛贵军, 白宇. 基于 PSO-CNN-LSTM 的短期热负荷预测模型[J]. 计算机仿真, 2024, 41(4): 102-107.
[4] SAJJADI S, SHAMSHIRBAND S, ALIZAMIR M, et al. Extreme learning machine for prediction of heat load in district heating systems[J]. Energy and Buildings, 2016, 122: 222-227. DOI: 10.1016/j.enbuild.2016.04.021
[5] 马文菁, 郭晓杰, 曹姗姗, 等. 基于 PSO-ELM 组合算法的热力站负荷预测研究[J]. 暖通空调, 2024, 54(3): 157-162. DOI: 10.19991/j.hvac1971.2024.03.24
[6] 果泽泉, 何波, 何强, 等. 集中供热热力站短期热负荷预测模型对比研究[J]. 区域供热, 2024(1): 146-158. DOI:10.16641/j.cnki.cn11-3241/tk.2024.01.023
[7] 张玲, 王美萍, 田琦, 等. 基于卷积神经网络的换热站热负荷预测研究[J]. 建筑节能(中英文), 2023, 51(5): 67-72.
[8] 孙玉芝, 杜向宁. 基于深度学习的热负荷预测研究[J]. 自动化仪表, 2022, 43(12): 92-96. DOI:10.16086/j.cnki.issn1000-0380.2022050136
[9] 顾伟, 薛贵军, 李水清. 结合天气因素的 HTS-FOA-GRNN 短期热负荷预测研究[J]. 机械设计与制造, 2022(1): 81-83. DOI:10.19356/j.cnki.1001-3997.20211116.073
[10] YU H, ZHONG F, DU Y, et al. Short-term cooling and heating loads forecasting of building district energy system based on data-driven models[J]. Energy and Buildings, 2023, 298: 113513. DOI:10.1016/j.enbuild.2023.113513
[11] HABIB M, TIMOUDAS T O, DING Y, et al. A hybrid machine learning approach for the load prediction in the sustainable transition of district heating networks[J]. Sustainable Cities and Society, 2023, 99: 104892. DOI:10.1016/j.scs.2023.104892
[12] DAVY G, DE SOMER O, JOHANSSON C, et al. Operational thermal load forecasting in district heating networks using machine learning and expert advice[J]. Energy and Buildings, 2018, 162: 144-153. DOI:10.1016/j.enbuild.2017.12.042
[13] WANG Y, LI Z, LIU J, et al. A novel combined model for heat load prediction in district heating systems[J]. Applied Thermal Engineering, 2023, 227: 120372. DOI: 10.1016/j.applthermaleng.2023.120372
[14] BREIMAN L. Random Forests[J]. Machine Learning, 2001, 45(1): 5-32. DOI:10.1023/A:1010933404324
[15] FRIEDMAN H. Greedy function approximation: A gradient boosting machine[J]. The Annals of Statistics, 2001, 29(5): 1189-1232. DOI:10.1214/aos/1013203451
[16] 栾丽华, 吉根林. 决策树分类技术研究[J]. 计算机工程, 2004,30(9): 94-96.
[17] WITTEN I H, FRANK E. Data mining: Practical machine learning tools and techniques[J]. Acm Sigmod Record, 2017, 46(2): 64-65.
[18] CUTLER A, CUTLER D, STEVENS J. Random Forests[J]. Machine Learning, 2004, 45(1):157-176.
[19] SNOEK J, LAROCHELLE H, ADAMS R P. Practical Bayesian optimization of machine learning algorithms[C]//Proceedings of Advances in Neural Information Processing Systems. NeurIPS, 2012:2951-2959. DOI:10.48550/arXiv.1206.2944