基于 XGBoost 与 Lasso 模型进行电负荷数据预测

尤高琳
北京宇信科技集团股份有限公司，北京 100015

[摘要]本篇文章旨在探讨工厂电力负荷预测的算法和模型，以保证工厂生产系统的效率和稳定性。首先，我们分析了工厂电力负荷的特点和影响因素，主要包括工厂电力负荷的历史数据，以及基于历史数据衍生出特征。然后，我们提出了一种基于机器学习的电力负荷预测模型，该模型能够根据历史数据预测未来一段时间内的电力需求。通过对比不同的机器学习算法，我们发现 Lasso 回归模型在预测精度和稳定性方面表现最好。最后，我们通过实验验证了该模型的有效性和实用性，为工厂电力负荷管理提供了重要的参考依据。
[关键词]电力负荷预测；Lasso 回归模型；机器学习模型
DOI：10.33142/sca.v7i3.11524　　中图分类号：F83　　文献标识码：A

Electricity Load Data Prediction Based on XGBoost and Lasso Models

YOU Gaolin
Beijing Yusys Technologies Co., Ltd., Beijing, 100015, China

Abstract: This article aims to explore the algorithms and models for predicting factory power load, in order to ensure the efficiency and stability of the factory production system. Firstly, we analyzed the characteristics and influencing factors of factory power load, mainly including historical data of factory power load, and derived features based on historical data. We propose a machine learning based power load forecasting model that can predict future electricity demand for a period of time based on historical data. By comparing different machine learning algorithms, we found that the Lasso regression model performs the best in terms of prediction accuracy and stability. Finally, we validated the effectiveness and practicality of the model through experiments, providing important reference for factory power load management.
Keywords: electricity load forecasting; Lasso regression model; machine learning models

引言

目前，我国电力市场针对大工业企业的用电收费是按照电量分档收费，且在一天内按照峰谷平不同时段执行不同的价格，这样的话，预测一家企业的用电负荷，基于预测的情况合理安排企业的生产活动，从而能够达到降低企业耗电成本，从这点而言准确地预测企业的电力负荷对企业还是有很强的经济意义。

对一家企业的电负荷的预测实际上是时间序列数据的预测。传统的时间序列预测模型主要有经典的时间序列模型包括移动平均模型、自回归模型、自回归移动均模型等模型，传统的时间序列预测方法非常依赖参数模型的正确选择，正确选择参数模型在很大程度上决定了预测结果的准确率[1]。

随机机器学习和深度学习的兴起，有不少模型如 XGBoost、LSTM 等模型在时间序列预测中取得了不错的效果。在李钢等人的文中是根据钢铁企业的不同工序分量的特征来进行负荷预测，并采用组合预测的方法将分类工序的负荷预测值累计加权汇总[2]。在亓晓燕等人的文中是采用了基于 LSTM 和 SVM 的模型融合后来对钢铁企业电力进行了电力负荷的端基预测[3]。在常乐等人的文中是先对数据集进行聚类，再基于不同分类数据使用 XGBoost 模型对电力负荷进行预测，从而达到提高预测精度[4]。在陈振宇等人的文中分别使用了 LSTM 和 XGBoost 模型来对超短期的电力负荷进行预测，再根据误差倒数法进行加权组合，修正单一模型误差较大的时序数据，从而达到以降低单一预测模型误差[5]。

时间序列模型预测也是和其他预测模型一样，需要使用自变量对因变量进行预测，有些原始数据集中会给去自变量，有些原始数据只有时间和数值两列，这里如果没有自变量的话，就需要我们自己构造出合适的自变量。在钢铁厂电负荷预测中是使用了铁矿石期货价格、螺纹钢期货价格、日期类型、电价、气温、湿 度、风速、风向等自变量来对预测值的进行预测[3]；另一部分电负荷时间序列就只有时间列和负荷值两列，这里需要构造出自变量来对负荷量进行预测[2-4, 5]，本文所使用的某电子厂的电负荷数据就是只有时间列和负荷列两列，这里需要构造出自变量。

本文通过的构造出自变量，通过 XGBoost 模型来进行变量筛选，使用 Lasso 模型来进行数据预测。

1 模型、评价指标及运行环境

1.1 模型介绍

XGBoost 是一个优化的分布式梯度增强库，旨在实现

高效，灵活和便携。它在 Gradient Boosting 框架下实现机器学习算法。XGBoost 提供并行树提升（也称为 GBDT, GBM），可以快速准确地解决许多数据科学问题。XGBoost 是构建了多个弱学习器进行集成从而变成一个强学习器的树提升模型。一棵树代表一个函数，树的增加相当于在不改变原有模型的基础上学习新函数，使用新的函数去对前一棵树的预测值与真实值误差进行拟合，不断进行迭代从而降低误差。训练结束得到树的数量为 t，预测某样本分数时则根据样本特征找到每棵树上对应的叶子节点，将这些叶子节点对应的分数相加即可得到样本预测值。

Lasso 全名 least absolute shrinkage and seletion operator。最小收缩算子法。是一种可以建立广义线性模型并且可以筛选变量的方法，功能强大，效果好。最开始这个统计模型是应用在地理学（geophysics）领域。后来被斯坦福统计 Robert Tibshirani 在 1996 年提出应用到医学领域模型构建中。在生物信息领域主要用 Lasso 进行筛选与预后相关基因，并构建预后模型。该模型最大的特点就是引入了惩罚项 λ，这个参数可以对模型变量进一步筛选，使模型不至于过于复杂，从而提高其泛化能力。

1.2 评价指标

模型评价指标是使用了平均绝对百分比误差（Mean Absolute Percentage Error），其公式如下所示：

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

其中 A_t 是实际值，F_t 是预测值，t 是时间序列个数。

1.3 运行环境

运行主机硬件环境是个人电脑 Lenovo K4e-IML CPU 是 Intel(R) Core(TM) i7-10510U，内存 16G，操作系统环境是 windows 10，软件环境是 Python 3.9 开发工具使用 Spyder，相关的软件包及版本信息如下：scikit-learn1.0.2 、pandas 1.4.4、numpy1.21.5、XGBoost1.7.4。开发程序在 Spyder 中进行调试测试。

2 数据情况

本次接收到两个数据文件，其中 8.20-10.23 进线总有功功率（间隔：10 分钟）V1.xlsx 是某电子厂的用电负荷数据，这部分数据是下面用于负荷预测的源数据；202110-202209 分时电量&电价 V1.xlsx 是电价数据，在本次负荷预测中没有用到。

某电子厂负荷数据的基本情况是 144 行 66 列，其中行是 0:00:00 到 23:50:00 每十分钟一行共 144 行，列是从 8.20 到 10.23 共 66 列。

某电子厂负荷数据的基本统计信息，从描述性统计来看一下这些数据情况，主要是统计了每天的记录数 count、平均值 mean、标准差 std、最小值 min、最大值 max 以及 25%、50%、75%分位值的情况。

通过计算可以看出负荷的平均值 mean、最小值 min、最大值 max 以及 25%、50%、75%分位值的折线图情况，而且数据从 8 月份到 9 月初负荷呈下降趋势，9 月初到 10 月份负荷趋势基本稳定。通过计算可以看出负荷的日内标准差基本在 3 左右，表明数据离散程度较低，每天的负荷数据稳定。

3 数据处理及特征选择

3.1 数据预处理

第一步为了后续数据处理，先将原始数据处理成了两列，日期时间+负荷，结果是 9360 行*2 列的数据。

3.2 特征构造

某电子厂的数据处理后只有时间和负荷值这两列，这是一个时间序列预测的问题。时间序列的当前发生的值受到历史数据的影响，即在特征构造的时候可以选择历史数据作为预测当前值的特征。在通过讨论滑动窗口在时间序列相似性降维算法中的实际应用情况，在李峰等人的文中他们提出了一种自适应确定滑动窗口宽度的新方法，通过对时序特征值分布函数挖掘，发现时间序列中的有效特征点，进而确定一组合适的滑动窗口宽度；最后根据序列的变化情况来决定最优的滑动窗口宽度，对原数据维度进行简约[6]。在李旭芳等人的文中，提出遥测时序滑动窗口的动态分割流程，规避了由于时间窗口宽度固定时，数据的局部信息不能被充分提取出来，可以根据数据的特点，来划分更为准确合理的非等长的时间窗口集合[7]。

那么我们需要选择多长的时间窗口的值来更好地预测当前值呢，一个是根据经验进行选择合适历史数据，这个对算法开发人员的业务素养要求较高，一般不容易选择初合适的历史期数。另一个方法是使用程序循环运行，寻找出 MAPE（平均绝对百分比误差为 0%表示完美模型，大于 100%则表示劣质模型。）最小值对应的历史期数。本次循环取了 6 个到 144 个历史数据来预测当前值，评价指标选择的是 MAPE，模型使用的 XGBoost 回归模型进行预测，得到 33 个历史数据的时候 MAPE 的值最低约为 1.063%。

33 个特征的重要性经过计算后，其中 ptt1_1、ptt1_24、ptt1_7、ptt1_3 的分值比较大即这几个特征在模型预测中的贡献比较大。

3.3 特征选择

经过轮询计算，选用 33 个历史数据时 MAPE 值最低，我本次特征构选择了 33 个历史数据来预测当前值，构造特征其中前 33 列是历史数据作为自变量，最后 1 列值是因变量，其中 ptt1_1 是距离 ptt1 相邻的历史数据，ptt1_33 是距离 ptt1 间隔 32 个时间周期的历史数据，其他字段的含义依次类推。

4 模型预测与优化

4.1 模型预测

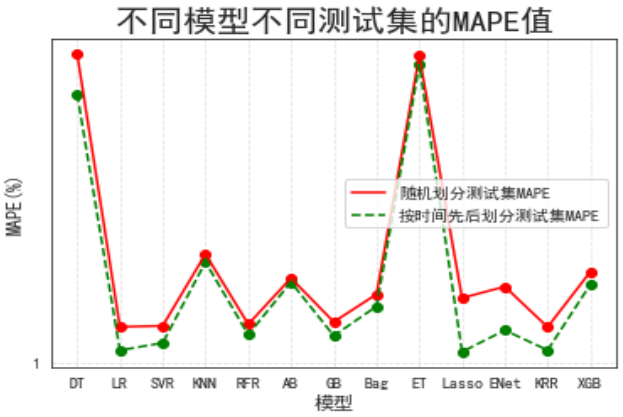

图 1　不同模型不同测试集的 MAPE 值

训练集和测试集的划分按照 8∶2 的原则来划分，划分方式分为随机方式和时间先后方式，评价指标选择 MAPE，重点关注测试集的 MAPE 值。如图 1 所示，红色折线是按照随机方式划分的测试集的 MAPE 值，蓝色折线是按照时间先后方式划分的测试集的 MAPE 值，按时间先后顺序划分训练测试集的 MAPE 的值要比按随机划分训练测试集小，所以我们训练测试集的划分选择按时间先后顺序来划分。X 轴是不同的模型，其中模型名称使用了缩写，部分模型缩写对应的模型全称是 DT-Decision Tree、LR-Linear Regression、AB-Ada Boost、GB-Gradient Boost、Bag-Bagging、ET-Extra Tree、XGB-XGBoost，其中 Lasso 回归预测的 MAPE 最小，最终选择使用按照时间先后顺序划分训练测试集，使用 Sklearn 中 Lasso 模型来进行数据预测。

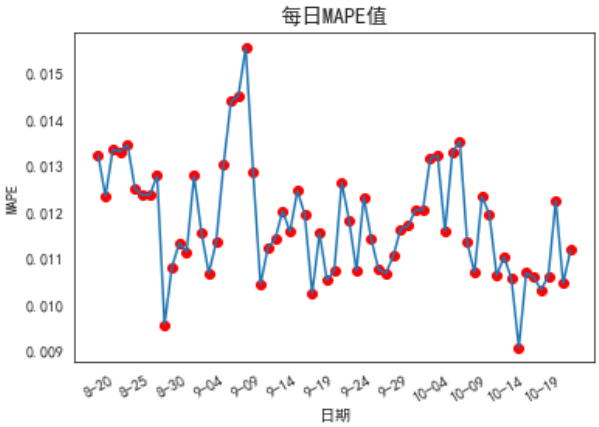

图 2　每日 MAPE 值

经过模型预测，可以看出 Lasso 模型的预测值和实际值的差异很小，逐日计算的 MAPE 的基本稳定在 1.2%左右，如图 2 每日 MAPE 值即按照日期来评估，模型在不同日期上的预测效果很稳定。

4.2 模型优化

scikit-learn 通过交叉验证基于最小角回归算法来公开设置 Lasso 模型中的 alpha 参数的，通过 model_selection 中的 GridSearchCV 包自动调节得到合适的 alpha。如图 3 所示，优化前后逐日计算的 MAPE，优化前多日 MAPE 的平均值在 1.17%左右，优化后多日 MAPE 的平均值在 1.05%，即按照日期来评估，模型在优化后的 MAPE 值平均降低了 0.12%，说明经过参数优化，模型的预测效果得到了一定的提升，如图 3 逐日优化前后 MAPE 趋势图所示。

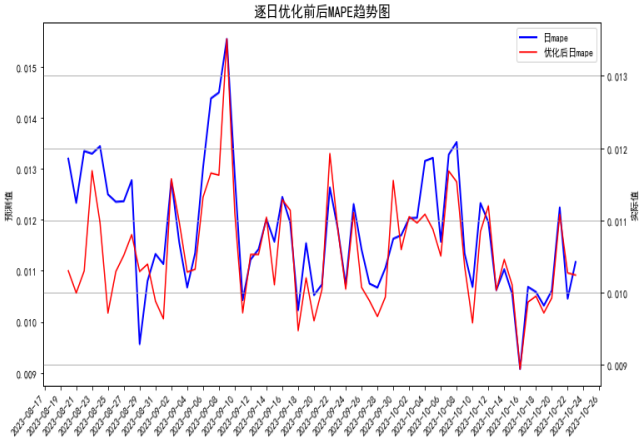

图 3　逐日优化前后 MAPE 趋势图

5 问题和不足

一是数据集的规模有限，只有 3 个月的数据，在更大的数据集上进行模型的预测效果还需要进一步验证，实际上我们可以搜集更长时间范围的数据，这样的话，可以用更多的数据集来进行训练预测模型。第二，对于某电子厂的负荷的预测，影响负荷的因素应该还有别的变量，后续做优化模型的时候还可以再加入新的变量来，其中胡欣在进行电力负荷预测的时候就考虑到了日前竞标负荷、真实负荷、日前市场节点电价、实际市场节点电价、日前发电电价、实际发电电价、日前阻塞电价、实际阻塞电价、日前边际损失电价、实际边际损失电价、温度、湿度等特征[8]。第三，还要考虑到在实际工业生产中还需要校验预测负荷数据和实际负荷之间的差异，在实际生产要合理地使用电力负荷的预测数据。

[参考文献]

[1]杨海民,潘志松,白玮.时间序列预测方法综述[J].计算机科学,2019,1(46):22-27.
[2]李钢,杜欣慧,裴玥瑶,等.基于改进密度峰值聚类的超短期工业负荷预测[J].电测与仪表,2021,5(58):159-163.
[3]亓晓燕,刘恒杰,侯秋华,等.融合LSTM和SVM的钢铁企业电力负荷短期预测[J].山东大学学报(工学

版),2021,51(4):91-98.
[4]常乐,汪庆年.基于优化聚类分解与XGBOOST的超短期电力负荷预测[J].理论与方法,2022,5(41):46-51.
[5]陈振宇,刘金波,李晨,等.基于LSTM与XGBOOST组合模型的超短期电力负荷预测[J].电网技术,2020,2(44):614-619.
[6]李峰,肖建华.时间序列相似性分析中滑动窗口宽度的确定[J].计算机科学与探索,2009,3(1):105-112.
[7]李旭芳,段春林,张冬波,等.遥测数据时间序列滑动窗口动态分割技术[J].飞行器测控学报,2015,34(4):345-349.
[8]胡欣,冯杰,徐先峰,等.基于特征选择实现多因素电力负荷预测[J].自动化仪表,2022,3(43):159-163.
作者简介：尤高琳（1987,9—），女，毕业院校：中国人民大学，所学专业：概率论与数理统计，当前就职单位：北京宇信科技集团股份有限公司，职务：业务专家。