

基于 LightGBM- 带注意力机制的 LSTM 组合模型的光伏瞬时发电量预测研究

金丹^{1,2} 王婧瑜^{1,2}

1. 华电云南发电有限公司以礼河发电厂 云南曲靖 654200

2. 云南电网有限责任公司曲靖会泽供电局鲁纳供电所 云南曲靖 654200

摘要: 为提高光伏电站瞬时发电量预测的精度与可靠性,本文基于机理知识对光伏瞬时发电数据展开特征工程,同时基于四分位距准则实现对输入特征数据中离群点的清洗,然后以 8:2 比例划分训练数据与测试集,再以 8:2 比例将训练数据划分训练集与验证集;分别使用随机森林、SVR、LightGBM、LSTM、带注意力机制的 LSTM(A-LSTM)以及本文提出的基于线性加权法对 LightGBM 与 LSTM(LGB-L)、LightGBM 与 A-LSTM(LGB-AL)组合模型对光伏瞬时发电量进行回归预测,使用 RMSE、MAPE 两个指标进行评估,并以 RMSE 作为主要指标。评估结果表明,本文提出的 LGB-AL 线性加权组合模型总体表现最佳,从测试集的评估表现来看,LGB-AL 模型在 MAPE 上相较其它模型总体降低 5.62%~35.41%,在 RMSE 上相较其它模型总体降低 3.16%~17.23%。

关键词: 光伏瞬时发电量; LightGBM; LSTM; 注意力机制; 线性加权法

引言

作为世界第一大清洁能源的太阳能相对煤炭石油等能源来说是可再生、无污染的,只要有太阳就有太阳能,所以太阳能的利用被很多国家列为重点开发项目。近十年来,随着科学技术的进步,光伏发电发展迅速,我国对于光伏发电也越发重视,光伏基地建设和屋顶分布式光伏开发建设提速,各界对光伏发电项目投资明显加快^[1]。尽管光伏发电有许多优点,但光伏发电系统也存在一些问题,比如如何最大化发电。然而最严重的问题是,光伏发电的功率输出不稳定,并且受阳光强度的影响很大^[2]。光伏发电受多种天气因素影响导致其功率的波动性较大,在接入电网时会给电力系统带来较大的调度问题,同时对电网的安全和运营造成不小的冲击。当夜晚没有光线时,光伏板不会发电。然而,一旦白天天气发生变化,对电力输出的影响也很大。当晴天有云时,光伏发电的输出功率将大幅波动。因此,为了更好地利用太阳能,大力发展光伏发电,必须掌握光伏发电规律,准确预测光伏发电的输出功率^[3]。

时间序列预测方法^[4]广泛应用于光伏发电预测的早期阶段,但其预测精度较差。灰色预测^[5]方法也被用于预测,但结果并不稳定。新出现的机器学习算法^[6],如支持向量机^[7](SVM)和随机森林^[8],具有更好的预测精度。神经网络^[9]

算法具有更好的预测效果,由于其强大的学习能力,可以反映复杂天气条件下光伏发电的变化,但只使用神经网络通过时间序列预测来预测数据往往在面对复杂数据时考虑不到数据之间的依赖性,从而导致分割精度低,而机器学习在这方面就有着相对较好的表现。

由于单一模型在拟合问题和泛化问题上表现较差,近些年混合预测模型在光伏发电量的预测中应用逐渐成为主流,特别是加入 LSTM 后对解决梯度爆炸或消失的问题有较好的作用。

文献^[10]将 XGBoost 模型、LightGBM 模型、ConvLSTM 模型组合出 XG-LG-CL 短期光伏发电量混合预测模型,同时使用高阶特征融合方法增加特征维数,对比单一模型预测准确率得到了一定的提升;文献^[11]提出一种基于注意力机制的分布式光伏电站电量短期预测模型的 A-HNN (attention-hybrid neural network),在分布式电站预测中取得了比单一模型较高的预测精度;文献^[12]基于 FCM-RF 算法(模糊均值-随机森林)和 LSTM 实现对分布式光伏电站的中长期发电量预测,预测性能 MAPE 稳定在 2.2%,RMSE 稳定在 2,具有较好的预测效果。LSTM 神经网络相比其他模型,MAPE 减少到 4%,R-square 提高到 0.93;文献^[13]基于 FTS (fuzzy time series)提出了两种改进算法

WEIGHTED-FTS 和 FIG-FTS, 在短期光伏预测方面, 两种模型预测误差小于 20%。且 FIG-FTS 在超短期光伏预测方面性能更好。

1 方法理论

1.1 LightGBM 原理

LightGBM^[14] 算法是 Boosting 算法的新成员, LightGBM 算法设计理念主要是采用了分布式的 GBDT, 使用基于直方图的决策树算法。在训练速度和内存消耗上远远优于 XGBoost^[15], 同时保留了较高的预测精度。LightGBM 在处理分类特征时需要 OneHot 编码预处理直接支持类别特征, 其优势主要在于降低每次拆分增益的成本, 使用减法直方图进一步加速, 进一步降并行学习的通信成本。直方图算法如图 1-1 所示。

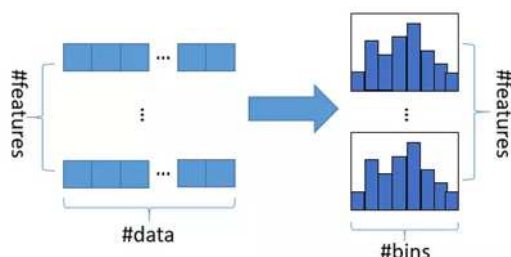


图 1-1 直方图算法

LightGBM 的数学原理主要是在 XGBoost 的基础上融入了 Histogram 算法、GOSS 算法、EFB 算法以及 leaf-wise 策略。采用 Histogram 算法来减少候选分类点的数量, 再引入 GOSS 算法和 EFB 算法来减少样本数量和特征数量, 通过这三个算法的引入 LightGBM 生成一片叶子需要的复杂度大大降低, 从而节省了计算时间, 如图 1-2 所示。

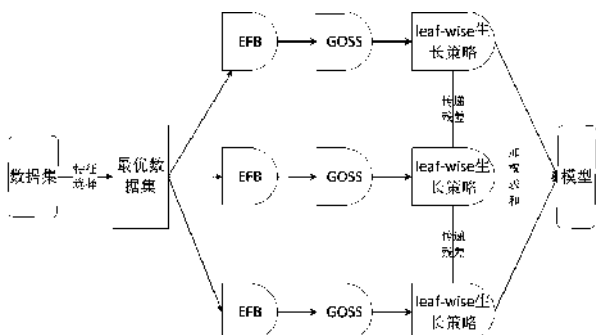


图 1-2 LightGBM 算法原理图

LightGBM 使用 leaf-wise 策略, 每次在当前的叶子节点时, 找到分裂增益最大的叶子节点来分裂, 而不是所有节点都分裂, 以此来提高精度, 如图 1-3 所示。另外在过拟合问题上 LightGBM 也有所改进, 在 Leaf-wise 分裂叶子节点

处增加了最大深度限制, 提高计算效率的同时防止过拟合。

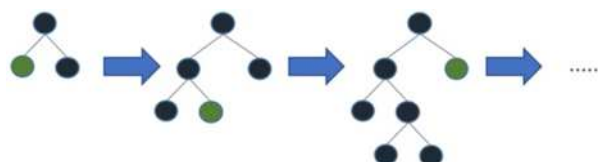


图 1-3 leaf-wise 生长策略

1.2 LSTM 原理

本文使用经典时序回归算法 LSTM^[16](long short-term memory), 其是 RNN 的一种变体, 属于一种时间递归神经网络, 适合于处理和预测有关时间序列的数据, 特别是时序依赖比较强的数据。对于长期记忆中梯度消失的问题, LSTM 以细胞状态做基础将加法运算代入网络中, 从而在时间序列预测中替代了 RNN。即使是较早的时间步信息 LSTM 也能够将其传递到较晚的时间步, 从而综合较早时间步的信息来对当前做预测, 从而克服了短期记忆的影响。

LSTM 算法以细胞状态和细胞内的“门结构”作为算法基础。细胞状态作为传递信息的道路, 通过细胞内“门”结构控制 LSTM 的记忆。“门”结构在训练时学会选择保留或遗忘 LSTM 的信息, 最后得到训练好的模型, 如图 1-4 所示。

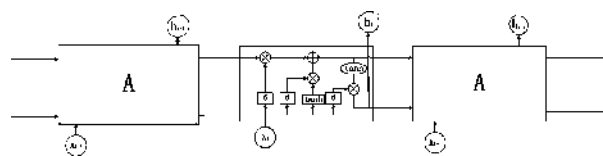


图 1-4 LSTM 算法原理图

在 LSTM 单元中, 遗忘、记忆和输出信息是通过门控函数 Sigmoid 和 tanh 激活函数运算实现, 如下所述:

(1) 选择性遗忘: 决定之前的获得的信息, 有多少信息要流向下一时间步, 具体公式如式 (1) 所示:

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

(2) 选择性记忆: 选择性记忆, 决定哪些信息需要存储下来。而这个过程就包含两个过程, 具体公式如式 (2)(3) 所示:

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

$$\tilde{C}_t = \tanh(w_c \cdot [h_{t-1}, x_t] + b_c) \quad (3)$$

(3) 新单元状态: 首先对信息进行选择性遗忘, 然后对输入进行选择记忆, 进而得到新单元状态, 具体公式如式 (4) 所示:

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \quad (4)$$

(4) 输出: 输出是根据当前信息, 上一个输出, 和当前输入决定的。具体公式如式 (5)(6) 所示:

$$o_t = \sigma \left(w_o \cdot [h_{t-1}, x_t] + b_o \right) \quad (5)$$

$$h_t = o_t \cdot \tanh(C_t) \quad (6)$$

1.3 注意力机制

注意力机制的原理为: 在对给定的目标, 通过生成一个权重系数对输入进行加权求和, 来识别输入中哪些特征对于目标是重要的, 哪些特征是不重要的; 为实现注意力机制, 需要把输入的原始数据看作 Key-Value 键值对的形式, 根据给定的任务目标中的查询值 Query 计算 Key 与 Query 之间的相似系数, 可以得到 Value 值对应的权重系数, 之后再用权重系数对 Value 值进行加权求和, 即可得到输出。分别将 Q、K、V 表示为 Query、Key、Value, 注意力权重系数 W 的公式如式 (7) 所示:

$$W = \text{softmax}(QK^T) \quad (7)$$

将注意力权重系数 W 与 Value 做点积操作 (加权求和) 得到融合了注意力的输出, 公式如式 (8) 所示:

$$\text{Attention}(Q, K, V) = W \cdot V = \text{softmax}(QK^T) \cdot V \quad (8)$$

注意力模型的具体结构如图 1-5 所示。

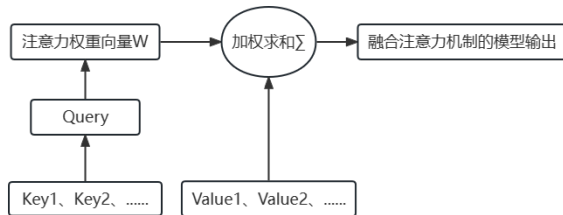


图 1-5 注意力模型结构图

1.4 线性加权法

线性加权法是一种评价函数法, 是指通过对各独立结果的重要性赋予对应的权重系数, 进行线性组合进行寻优的求解多目标规划的一种方法, 即归结为求 $\min_x \in X \sum_{k=1}^m \omega_k f_k(x)$ 的最优解, 其中 $\omega_k \geq 0 (k = 0, 1, 2, \dots, m)$ 为权系数, $\sum_{k=1}^m \omega_k = 1$ 。

1.5 数据归一化

训练 LSTM 网络模型, 为避免每批次的输入数据分布不一致, 缓解过拟合, 同时加快网络的训练速度, 需要对数据归一化处理, 本文采取 MinMaxScaler 归一化, 具体公式如式 (9)(10) 所示:

$$X_{std} = \frac{X - X_{\min}(axis=0)}{X_{\max}(axis=0) - X_{\min}(axis=0)} \quad (9)$$

$$X_{scaled} = X_{std} \times (max - min) + min \quad (10)$$

式 (7) 中, $X_{\min}(axis=0)$ 为此维度中最小值组成的行向量; $X_{\max}(axis=0)$ 为此维度中的最大值组成的行向量; max 为区间最大值, 默认是 1; min 为区间最小值, 默认是 0; X_{std} 为标准化结果; X_{scaled} 为归一化结果。

1.6 评估方法

为了更充分的分析光伏发电预测的, 本文采取均方根误差 (RMSE) 与平均绝对百分比误差 (MAPE) 作为评估模型性能的指标, 具体评价公式如式 (11)(12) 所示:

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n}} \quad (11)$$

$$MAPE = 100\% \times \frac{1}{n} \sum_{i=1}^n \left| \frac{\hat{y}_i - y_i}{y_i} \right| \quad (12)$$

上式中, RMSE 为均方根误差; MAPE 为平均绝对百分比误差; $\hat{y}_i$ 为预测值; y_i 为实际值; n 为样本数。

本文以 RMSE 作为主要评估指标, MAPE 作为辅助评估指标。

2 实证分析

2.1 数据描述及划分

本文使用的实验数据来源于 2018 “创客中国” 国家电投大数据及智能应用创新创业大赛, 数据是从某光伏设备采集的 9000 条连续数据, 共有 21 个特征变量, 其中 “发电量” 为需要预测的目标变量。

从数据集中随机划分 20% 的数据 (1800 条) 为测试集, 剩下 80% 的数据 (7200 条) 作为训练集, 再从训练集中划分 20% 的数据 (1440 条) 作为验证集, 训练集中剩下 80% 的数据 (5760 条) 作为最终训练集。

2.2 数据处理

通过对各特征数据的可视化分析, 可以查看到存在明显的离群点。例如, 特征 “电流 B” 数据绝大部分小于 100, 但存在极少数大于 600 的离群点, 如图 2-1 所示; 特征 “平均功率” 数据绝大部分小于 20000, 但存在极少数大于 140000 的离群点, 如图 2-2 所示。对于这些离群点, 本文采取四分位距准则进行识别, 同时对于离群值的处理, 考虑到该数据集具有时序性, 可采取用前一个正常值替换的策略进行修正, 而非直接剔除。

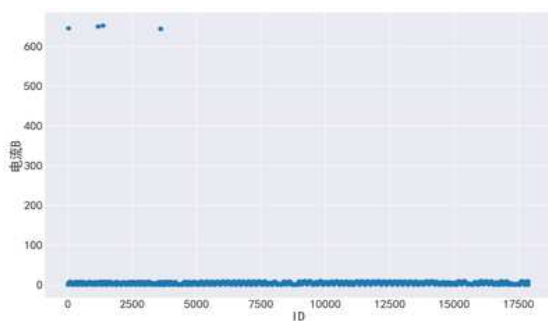


图 2-1 “电流 B” 特征散点图

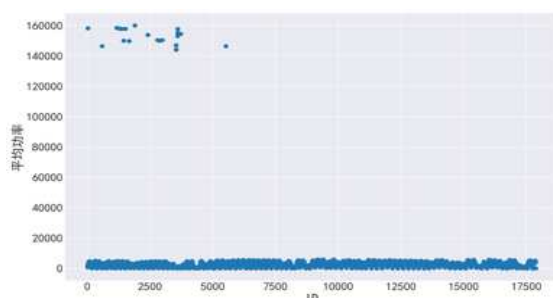


图 2-2 “平均功率” 特征散点图

2.3 特征工程

光伏发电即运用电池板的光生伏特效应，将光能转化为电能。通过对光伏发电系统的工作原理的了解以及对已有特征数据的分析，展开特征工程生成新的特征，找到对光伏发电具有重要影响的因素。

光照强度随时间变化过程中较为缓慢，为此功率的变化也是较为缓慢的，将前一时刻的功率作为现一时刻发电量的输入特征具有一定的积极影响。为此计算“前一时刻功率”特征数据。

温度对光伏发电存在一定影响，板温减去现场温度可以理解为电池板在发电过程中自身产生的热量，如图 2-3 所示，从局部可以看出温差与发电量存在较强的相关性。为此，计算“温差”特征数据，即板温 - 现场温度。

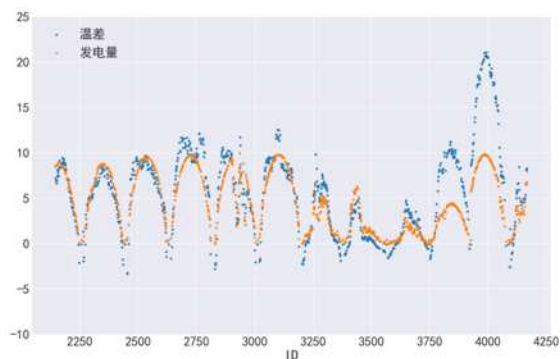


图 2-3 温差与发电量散点图

由于一天中日照角度的变化，已经提供的光照强度不能直接反应板面所受光照。转换效率计算方法：光伏板输出功率 / 日照功率，其中光伏板输出功率可用电压 A 代表，已知转换效率的情况下，电压 A / 转换效率可以表示板面实际光光照。为此计算特征：实际板面所受光照，即电压 A / 转换效率。

从光伏电站的瞬时发电量数据可看出具有明显的周期性，周期大致在 180 左右，一个周期代表一天，由于每天的天气会有差别，导致光照强度的形状改变，从而发电量形状也会发生一定改变，如图 2-4 所示。

根据发电量与时间的周期性变化规律，定义一个距离周期内峰值距离 (dis2peak) 的特征^[17]，如图 2-5 所示。同时计算相同距离的特征“平均功率”的均值及标准差，共计算 3 个新特征。

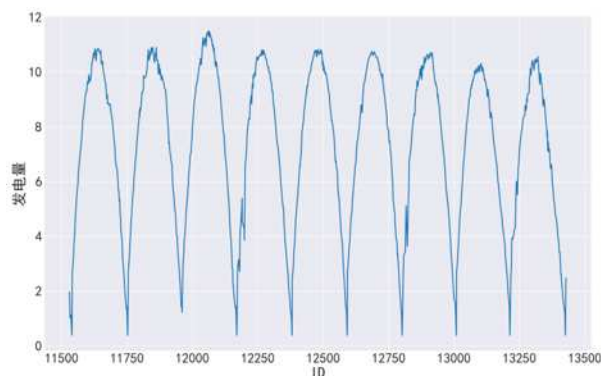


图 2-4 发电量时序曲线图

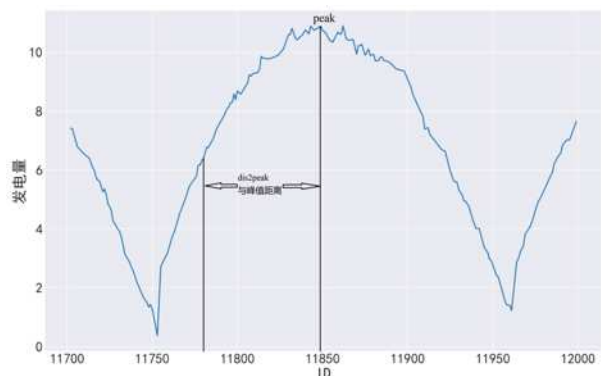


图 2-5 单个周期内峰值距离

2.4 模型训练

基于 LightGBM 算法训练发电量回归模型，由于数据存在不均匀，同时减少因划分数据造成的过拟合风险，采用 5 折交叉验证对训练集进行交叉训练，该过程为：将训练集拆分为 5 个连续的折叠 (folds) (不打乱数据顺序)，然后将每

1 个折叠作为验证数据，剩余的 4 个形成训练数据，以交叉验证的方式对训练数据进行 5 次回归模型训练，使用每次训练的回归模型对测试集进行预测采用“macro-average”方法计算 5 次对测试集预测的均值。

本文训练 LSTM 模型的隐含层使用了 260 个单元的 LSTM 单层网络，激活函数设置为 tanh；输出层激活函数为 relu；损失函数选择 mse，优化器选择 adam。

在上述 LSTM 参数设置的基础上，添加注意力机制，训练 Attention-LSTM（即带注意力机制的 LSTM，下文简称 A-LSTM）模型，根据经验设置注意力层数为 2000。

组合模型预测在大多数机器学习任务中对提高预测结果具有重要作用，本文基于线性加权法分别对 LightGBM 与 LSTM（下文简称 LGB-L）组合模型、LightGBM 与 A-LSTM（下文简称 LGB-AL）组合模型的预测结果进行线性加权组合，考虑到神经网络模型在预测过程中存在一定的随机误差，为此基于 LightGBM 模型对验证集预测的权重变化，自动寻找最优线性权值，如图 2-6 所示，当 LightGBM 的线性权值为 0.4 时，加权组合模型的 RMSE 最小，即以该线性权值进行线性加权组合。

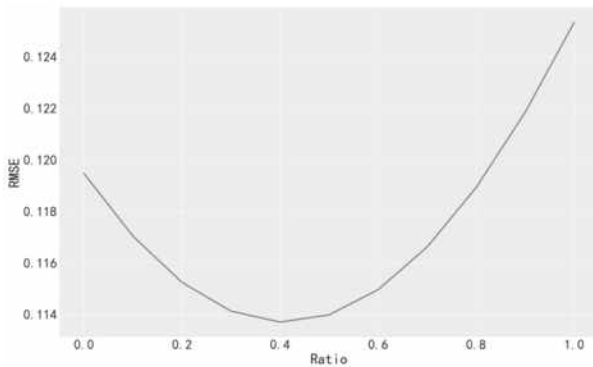


图 2-6 LightGBM 对验证集的预测结果在不同比例下的组合模型 RMSE 曲线图

此外，使用随机森林、支持向量机（SVR）算法分别训练回归预测模型，从多方面比较不同类型的模型预测效果。

2.5 评估结果对比

为验证本文模型的可靠性，设置了验证集与测试集预测结果的对比实验，分析 7 个模型在验证集与测试集的 RMSE 和 MAPE。结果如表 2-1 所示，7 个模型在验证集与测试集的预测结果表现基本一致，没有出现明显过拟合，说明本文对 7 个模型的训练具有可靠性。

表 2-1 多个模型在验证集与测试集的 MAPE、RMSE

模型	MAPE		RMSE	
	验证集	测试集	验证集	测试集
随机森林	0.032488	0.040394	0.143569	0.137843
SVR	0.081691	0.065961	0.148190	0.179241
LightGBM	0.054868	0.031143	0.126823	0.119601
LSTM	0.022015	0.025137	0.122144	0.124896
A-LSTM	0.021586	0.023702	0.119010	0.112911
LGB-L	0.021017	0.021983	0.115041	0.110076
LGB-AL	0.020903	0.021363	0.113906	0.109772

为验证本文特征工程的有效性，比较分析 7 个模型在特征工程前、后的 RMSE 和 MAPE。结果如表 2-2 所示，本文使用的 7 个模型中，随机森林、LightGBM、LSTM、A-LSTM 以及提出的组合模型 LGB-L 和 LGB-AL 在特征工程后的测试集的 MAPE 与 RMSE 均减小，说明本文所开展的特征工程对模型预测效果的提升总体具有有效性。

表 2-2 多个模型在特征工程前、后的测试集预测结果的 MAPE、RMSE

模型	测试集 MAPE		测试集 RMSE	
	特征工程前	特征工程后	特征工程前	特征工程后
随机森林	0.040394	0.036975	0.137843	0.126275
SVR	0.065961	0.073611	0.179241	0.189365
LightGBM	0.031143	0.031049	0.119601	0.112502
LSTM	0.025137	0.024764	0.124896	0.115174
A-LSTM	0.023702	0.022160	0.112911	0.109562
LGB-L	0.021983	0.021249	0.110076	0.107930
LGB-AL	0.021363	0.020054	0.109772	0.104524

为验证本文所提出的线性加权组合模型的有效性，比较分析 7 个模型在特征工程后的验证集与测试集的 RMSE 和 MAPE。结果如表 2-3 所示，LGB-L、LGB-AL 两个组合模型的 MAPE、RMSE 均小于其它单模型，LGB-AL 相较于 LGB-L 的 MAPE、RMSE 均减小；从最终测试集的评估表现来看，LGB-AL 组合模型在 MAPE 上总体降低 5.62%~35.41%，在 RMSE 上总体降低 3.16%~17.23%。

表 2-3 多个模型在特征工程后的验证集与测试集的 MAPE、RMSE

模型	MAPE		RMSE	
	验证集	测试集	验证集	测试集
随机森林	0.029734	0.036975	0.133495	0.126275
SVR	0.097640	0.073611	0.155478	0.189365
LightGBM	0.040256	0.031049	0.125331	0.112502
LSTM	0.021641	0.024764	0.118689	0.115174
A-LSTM	0.020665	0.022160	0.115225	0.109562
LGB-L	0.020627	0.021249	0.113397	0.107930
LGB-AL	0.019630	0.020054	0.111006	0.104524

3 结论

本文通过对 LightGBM 集成模型及 LSTM 神经网络模型的的研究分析,提出一种基于 LightGBM-带注意力机制 LSTM(LGB-AL) 的线性加权组合模型。

在对光伏发电数据分析探索中,对数据离群值展开清洗,同时根据光伏发电的原理以及机理知识,展开特征工程,挖掘有效信息,提升了数据的总体质量;在对多个单模型以及组合模型的训练预测评估中,研究结果表明,本文提出的 LGB-AL 线性加权组合模型的预测效果与单一模型相比得到了显著提高,为此对提高光伏电站瞬时发电量预测的精度与可靠性具有现实性意义。

参考文献:

- [1] 肖佳,梅琦,黄晓琪,蒋龙,张建平. “双碳”目标下我国光伏发电技术现状与发展趋势[J]. 天然气技术与经济,2022,16(05):64–69.
- [2] Wang H, Sun J, Wang W. Photovoltaic Power Forecasting Based on EEMD and a Variable-Weight Combination Forecasting Model[J]. Sustainability, 2018, 10(8):2627.
- [3] 陈美珍,柳扬,徐胜彬,郭俊锋,张永强,林金阳. 基于天气类型及改进的小波神经网络的光伏发电短期预测[J]. 贵州大学学报(自然科学版),2022,39(01):70–77+88.
- [4] Contreras J, Espinola R, Nogales F J, et al. ARIMA Models to Predict Next-Day Electricity Prices[J]. IEEE Transactions on Power Systems, 2003, 18(3):1014–1020.
- [5] 蔡方中,张磊磊. 基于改进灰色预测模型的供应链韧性评价与预警研究[J]. 工业技术经济,2022,41(12):100–107.
- [6] Xiong Z, Li J, Zhu H, et al. Ultimate Bending Strength Evaluation of MVFT Composite Girder by using Finite Element Method and Machine Learning Regressors[J]. 2022.
- [7] Joachims T. Making Large-Scale SVM Learning Practical[J]. Technical Reports, 1998, 8(3):499–526.
- [8] R D í az-Uriarte, SAD Andr é s. Gene selection and classification of microarray data using random forest[J]. Bmc Bioinformatics, 2006, 7(1):3.
- [9] 张皓,涂雅培,高瑜翔,唐军,黄天赐. 基于多重模糊神经网络的 PID 温度控制算法[J/OL]. 西华大学学报(自然科学版):1–8[2022–12–26].
- [10] 王俊杰,毕利,张凯,孙鹏翔,马训德. 基于多特征融合和 XGBoost-LightGBM-ConvLSTM 的短期光伏发电量预测[J/OL]. 太阳能学报:1–7[2022–12–22].
- [11] 崔佳豪,毕利. 基于混合神经网络的光伏电量预测模型的研究[J]. 电力系统保护与控制,2021,49(13):142–149.
- [12] 方鹏,高亚栋,潘国兵,马登昌,孙鸿飞. 基于 LSTM 神经网络的中长期光伏电站发电量预测方法研究[J]. 可再生能源,2022,40(01):48–54.
- [13] Serrano Ardila, V.M.; Maciel, J.N.; Ledesma, J.J.; Ando Junior, O.H. Fuzzy Time Series Methods Applied to (In)Direct Short-Term Photovoltaic Power Forecasting. Energies 2022, 15, 845. [Google Scholar] [CrossRef]
- [14] Qi M. LightGBM: A Highly Efficient Gradient Boosting Decision Tree[C]// Neural Information Processing Systems. Curran Associates Inc. 2017.
- [15] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System[J]. ACM, 2016.
- [16] Gers, Felix A, Schmidhuber, et al. Learning to Forget: Continual Prediction with LSTM[J]. Neural Computation, 2000.
- [17] 彭曙蓉,郑国栋,黄士峻等. 基于 XGBoost 算法融合多特征短期光伏发电量预测[J]. 电测与仪表,2020,57(24):76–83.