

基于 Stacking 算法的锂电池荷电状态估计

王志军¹
WANG Zhijun

摘要

为提高锂电池荷电状态 (SOC) 的估计精度, 文章提出了一种基于 Stacking 融合算法的 SOC 预测方法。以新能源汽车锂电池数据为基础, 首先对原始数据进行充放电状态筛选、异常值剔除及缺失值填充等预处理, 并通过偏相关性分析构建了 9 个与 SOC 高度相关的特征参数。随后, 分别构建随机森林 (RF)、XGBoost、LightGBM 三种单一模型进行 SOC 预测, 并引入 Stacking 融合框架以集成多模型优势, 提升模型的鲁棒性与泛化能力。实验结果表明, Stacking 模型与单个基础模型相比表现优异, 具备良好的实时性能。研究表明, 融合算法在动力电池 SOC 估计中具有较高的应用潜力, 为锂电池管理系统的智能化提供了技术支撑。

关键词

锂电池; 电池 SOC 估计; 集成学习; Stacking 算法

doi: 10.3969/j.issn.1672-9528.2026.01.036

0 引言

锂离子动力电池的荷电状态 (SOC) 是衡量电池剩余可用容量与运行安全性的关键指标^[1], 其估计精度直接决定电池剩余使用时间、充放电控制与热失控预警等核心功能。在实际运行过程中, SOC 无法通过任何传感器直接测量, 只能依赖电压、电流与温度等外部可观测信号进行推断; 然而, 该映射关系随循环老化、环境温度及负载工况的变化呈现强非线性和时变特征, 使得高精度估计长期成为电池管理系统 (BMS) 的技术瓶颈。若 SOC 估计偏差过大, 将导致续航预测失准、过充或过放风险增加, 甚至诱发热失控事故; 而准确的 SOC 则是后续健康状态 (SOH) 与功率状态 (SOP) 评估的基础, 对于提升电池寿命、保障行车安全、优化整车能量管理具有不可替代的重要意义。特别是在国家大力扶持新能源^[2]、电池梯次利用和车网协同发展的背景下, SOC 估计的精度与稳定性已成为影响产业链上下游协同效率与运营安全的关键因素。

传统方法如安时积分、开路电压和卡尔曼滤波等, 在实际道路条件下暴露出静置要求严苛、误差累积或模型参数敏感等局限, 难以适配复杂多变的工况^[3-5]。数据驱动的随机森林、XGBoost、LSTM、BP 神经网络等机器学习模型凭借非线性拟合与抗噪能力在 SOC 预测领域已经取得进展^[6], Lee 等^[7]用时序卷积网络和 Xgboost 的混合模型对锂电池荷电状态进行估计; 国内研究中, 孔祥创等^[8]利用 BP 神经网络弥

补 EKF 算法的估计误差实现对新能源汽车电池 SOC 估计的改进。商云龙等^[9]利用模糊神经网络 (fuzzy neural network, FN) 建立误差预测模型, 并且修正扩展卡尔曼滤波来测量噪声协方差, 结果表明该方法可以应对实际运行中各种复杂工况下的电池 SOC 预测。马向平等^[10]将极限学习机算法与安时积分法相结合, 解决了安时积分法估计 SOC 时的不精确度随时间误差逐渐变大的缺点, 提高了 SOC 估计精度。王超等^[11]创造性地将递推最小二乘法与无迹卡尔曼滤波相结合, 研究结果表明该模型的平均误差约为 1.41%。但任何单一模型都难以兼顾偏差与方差, 其泛化稳定性仍受限制。Stacking 通过分层融合异质模型、用元学习器二次学习基模型输出, 可有效汲取多模型优势并抑制过拟合^[12], 已在医学诊断、农业产量预测等领域验证性能^[13-14], 然而在动力电池 SOC 估计场景中尚缺乏系统研究。

本文基于大规模实车运行数据, 提出一种“随机森林-XGBoost-LightGBM+动态权重元学习器”的 Stacking-SOC 估计框架: 首先分别训练三种基模型, 以交叉验证生成一级预测并构造元特征; 随后设计工况自适应的线性回归元学习器, 融合不同工况下的模型输出; 最终通过独立测试集评估模型在常温、高温及高速等典型场景的泛化性能。实验结果显示, 所提方法在均方根误差和平均绝对误差指标上较基础单一模型表现优异, 可满足 BMS 实时应用需求。本研究具体贡献如下:

(1) 系统验证 Stacking 在 SOC 估计中的可行性, 为集成学习在电池管理领域开辟新路径;

(2) 构建温区、速度和循环深度多样的数据集, 为后

1. 大唐三门峡发电有限责任公司 河南三门峡 472000

续研究提供可复现基准;

(3) 从理论与实验视角证明异质模型协同可提升鲁棒性, 兼顾精度与实时性。预期成果将直接服务于续航管理、寿命预估与热安全保护, 助力新能源汽车的可持续发展。

1 模型结构设计

1.1 集成学习概述

集成学习通过组合多个学习器来共同决策, 以降低方差或偏差、提升泛化能力。根据基模型间的依赖关系, 可分为并行 Bagging 与串行 Boosting。Bagging 在样本或特征维度引入随机扰动, 通过多数投票或平均抑制过拟合, 训练过程易并行; Boosting 迭代地调整样本权重, 使后续模型聚焦前一轮误差, 逐步减少整体偏差, 精度高但更易过拟合。两类方法互补, 为后续模型融合奠定基础。

1.2 基学习器设计

1.2.1 随机森林 (RF)

RF 是典型 Bagging 实现: 对原始数据做自助采样, 在每个树节点随机选取特征划分, 训练多棵决策树, 分类用投票、回归取均值。双重随机化提升了模型多样性与抗噪声能力, 并行性好, 但可解释性及对时空相关数据的处理仍有限。

1.2.2 XGBoost

XGBoost 在 GBDT 框架上引入二阶梯度与 L1/L2 正则化, 显式约束模型复杂度; 结合缓存优化、列块化存储等工程手段显著加速训练。目标函数由损失项和正则项组成, 叶节点数与权重共同衡量复杂度。其精度得益于残差拟合, 但对噪声敏感, 需要合理剪枝与正则化强度。

1.2.3 LightGBM

LightGBM 采用三项关键改进提升大数据场景效率:

- (1) 直方图离散化将连续特征桶化, 减少遍历开销;
- (2) Leaf-wise 且深度受限的生长策略, 只扩展增益最大的叶节点, 兼顾精度与规模;
- (3) 梯度单边采样 (GOSS) 保留大梯度样本、随机降采小梯度样本, 加速训练并节省内存。

1.3 Stacking 融合策略

为兼顾模型多样性与残差纠正能力, 本文构建两层 Stacking 架构, 其模型结构如图 1 所示。在第一层, 置入 RF、XGBoost 与 LightGBM 三种异质基模型, 通过 K 折交叉验证生成 Out-of-Fold 预测, 构成新的特征空间。在第二层:

将上一级输出输入次级学习器, 由其自动学习各基模型最优组合权重, 输出最终 SOC 估值。

该结构既保留各模型对不同数据模式的捕获能力, 又通过元学习器动态调整权重抑制局限。实验证明, 在常温、高

低温等典型工况下, 该融合方法的均方根误差与平均绝对误差较基础单一模型表现优异, 可满足 BMS 实时需求。

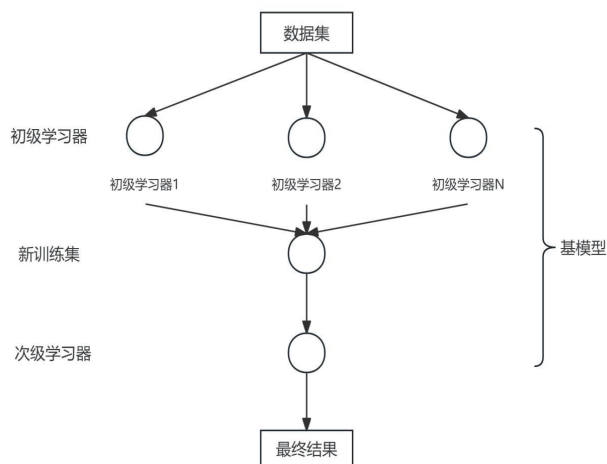


图 1 Stacking 融合模型示意图

2 数据预处理及特征参数提取

2.1 数据预处理流程

本文使用的数据来自百度飞桨电动汽车的公开数据, 该数据集包含充放电状态、电压电流、车辆运行状态、速度、温度等多种数据, 为了方便建模进行 SOC 的预测, 本章需要对数据进行预处理并选择与预测目标 SOC 相关的数据字段, 并建立相关的特征参数。

基于以上数据基本信息, 本研究采用四阶段数据处理流程为:

(1) 充放电分离

根据数据定义, 提取出车辆状态 =1 (车辆处于启动状态) 且充电状态 =3 (车辆处于未充电状态) 的数据作为研究的基础数据, 经过提取处理后得到的汽车行驶部分数据记录。

(2) 缺失与异常处理

本部分的处理具体包括: 将数据按时间顺序进行排列, 对重复无用数据进行剔除; 对于缺失值采用该指标的平均值进行填充; 用箱型图法对数据集中存在的异常值进行判定并剔除。

(3) 数据片段划分

在连续 3 min 内如果 SOC 出现状态的突变则触发片段划分, 即以状态突变点作为不同数据片段的分界点。通过网络搜索合理的 SOC 变化值, 本文将一定时间内 SOC 的突变阈值定为 3%, 在此规则下划分数据, 最终得到了车辆放电的 11 562 个数据片段。

2.2 特征工程建立

本研究基于运行数据中的基础数据, 按照 SOC 变化 1% 对数据进行分组, 对同一 SOC 数值下的所有数据进行统计指标计算并以此建立特征工程, 所有特征如表 1 所示。

表 1 特征名称及定义汇总

特征名称	定义
总电压平均值 /V	连续且同一 SOC 数据下的总电压的平均值
总电压最大值 /V	连续且同一 SOC 数据下的总电压的最大值
总电压方差 /V ²	连续且同一 SOC 数据下的总电压的离散程度
总电压之和 /V	连续且同一 SOC 数据下的总电压的累加和
总电流平均值 /A	连续且同一 SOC 数据下的总电流的平均值
总电流最大值 /A	连续且同一 SOC 数据下的总电流的最大值
总电流之和 /A	连续且同一 SOC 数据下的总电流的累加和
平均速度 /(km·h ⁻¹)	连续且同一 SOC 数据下的速度的平均值
最大速度 /(km·h ⁻¹)	连续且同一 SOC 数据下的速度的最大值
速度方差 /(km ² ·h ⁻²)	连续且同一 SOC 数据下的速度的离散程度
单体电池最高电压 /V	电池组内所有子电池的电压最高值
单体电池最低电压 /V	电池组内所有子电池的电压最低值
单体电池最高温度 /℃	电池组内所有子电池的温度最高值
单体电池最低温度 /℃	电池组内所有子电池的温度最低值
行程时间 /s	在 1% 的 SOC 范围内所行驶的时间长度
行驶里程 /km	在 1% 的 SOC 范围内，车辆所行驶的总里程数
循环次数 /次	数据所在放电片段的片段序号

2.3 特征相关性分析

在多变量关系复杂交织的研究背景下，准确识别各变量与目标变量之间的独立作用尤为关键。偏相关性分析正是一种能够剔除其他变量干扰、揭示变量之间真实联系的有效工具。通过在控制其他变量不变的条件下考察某一变量与目标变量之间的“净”相关性，该方法能够更客观地反映变量间的本质关系，为后续建模、因果推断或策略制定提供更加可靠的依据。因此，开展偏相关性分析有助于明确关键影响因素，避免因共线性或伪相关而造成的误判。

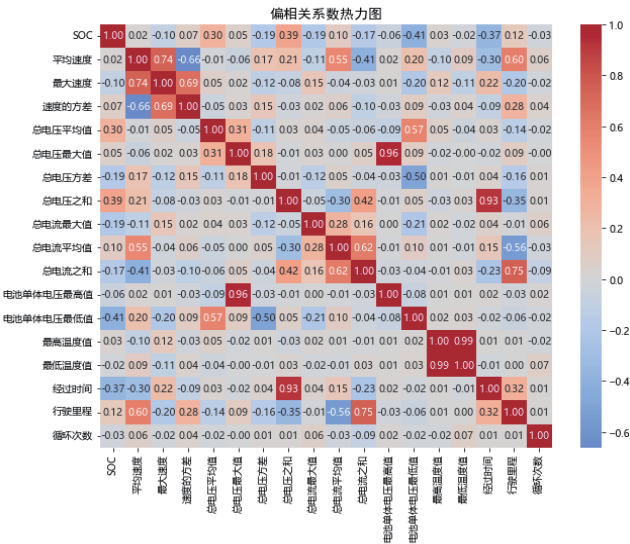


图 2 偏相关性分析热力图

本研究采用单辆车的放电数据，共计 37 687 条记录，基于偏相关分析方法，对 SOC 及其相关影响因素进行了系统性分析，结果如图 2 所示。结果表明，部分变量与 SOC 存在较强的偏相关关系。其中，总电压之和与 SOC 呈现出最强的正相关性，偏相关系数为 0.39；总电压平均值亦表现出较强的正相关性，系数为 0.30。电池单体电压最低值与 SOC 之间则呈现出较强的负相关关系，系数为 -0.41；而经过时间与 SOC 的偏相关系数为 -0.37，显示出中等程度的负相关性。此外，总电压方差和总电流最大值与 SOC 之间均存在一定程度的负相关性，偏相关系数分别为 -0.19。相比之下，其他变量如平均速度、最高 / 最低温度值及循环次数等，与 SOC 之间的相关性接近于零，说明其对 SOC 的预测作用可能较弱或不具显著性。基于上述结果，本文最终选取行驶里程、总电流最大值、车辆状态、充电状态、电池单体电压最高值与最低值、总电压平均值、总电压之和、车速以及经过时间作为后续分析与建模的特征参数。

3 SOC 与模型效果评估

3.1 评价指标

为对比本研究使用的 Stacking 模型与单一的随机森林、XGBoost、LightGBM 的效果优劣，在模型对比阶段选取平均绝对误差（MAE）、均方误差（MSE）、决定系数（R²）指标作为评价指标，计算公式分别为：

$$MAE = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i|$$
(1)

$$MSE = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2$$
(2)

$$R^2 = (1 - \frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2})$$
(3)

式中： $\hat{y}_i$ 代表第 i 个样本的预测值； y_i 代表第 i 个样本的真实值； $\bar{y}$ 代表真实值的平均值。

3.2 预测结果及模型对比

交叉验证是一种常用的模型评估方法，旨在降低模型对数据划分的依赖性，提升其对未知数据的泛化能力。其核心在于通过多次划分训练集与验证集，缓解过拟合并减少因单次划分带来的随机误差。本研究采用 5 折交叉验证，并按照 9:1 的比例划分训练集和测试集，以平衡模型评估的准确性与计算效率。

在此基础上，将堆叠模型与 3 个基础模型的预测结果进行对比，如表 2 所示。通过对比三种单一模型（随机森林、

XGBoost、LightGBM)与Stacking融合模型的预测性能,可得出如下结论:在所有单模型中,随机森林表现最优,其决定系数 R^2 为0.982 2, MSE为0.564 9,均优于XGBoost(0.981 9, 0.572 6)和LightGBM(0.981 4, 0.590 0)。Stacking模型在此基础上进一步提升了预测精度,其 R^2 达到0.982 9, MSE降至0.541 4,相较最优单模型提升了4.16%。由此可见,Stacking模型在综合多个基模型优势方面具有明显优势。未来可通过引入注意力机制或残差加权策略,进一步增强融合效果,提升模型精度并降低误差。

表2 模型对比结果

	MAE	MSE	R^2
随机森林	0.582 6	0.564 9	0.982 2
XGBoost	0.616 4	0.572 6	0.982 2
LightGBM	0.620 0	0.590 0	0.981 4
Stacking	0.606 8	0.541 4	0.982 9

4 结论

本文以新能源汽车的实际运行数据为基础,围绕电池SOC的预测问题开展研究。通过对原始数据进行清洗与预处理,构建了9个与SOC相关的特征参数。在模型构建方面,分别建立了RF、XGBoost、LightGBM三种单一模型,并引入Stacking融合算法以提升预测性能。实验结果表明,随机森林在单模型中表现最佳,Stacking融合模型在整体预测精度上进一步提升,验证了模型融合方法在SOC预测中的有效性。最后,结合模型评价指标对预测结果进行了对比分析,明确了本研究的优势与不足,并为后续引入更高效的融合策略提出了改进方向。

参考文献:

- [1] 郑大宇,高煜琨,董静,等.基于AFFRLS-AUKF的多工况下锂离子电池SOC估计[J].哈尔滨商业大学学报(自然科学版),2025,41(3):336-345.
- [2] 刘文超,孟锦豪,张涛,等.基于动态电化学阻抗谱的锂电池荷电状态估计[J/OL].高电压技术,1-13[2026-01-09].<https://doi.org/10.13336/j.1003-6520.hve.20250221>.
- [3] 孟繁雨,吴道明,夏洋.提高基于安时积分法估算动力电池SOC精度的方法研究[C]//中国汽车工程学会年会论文集,2023:316-321.
- [4] 孙艳艳,周雪松,游祥龙,等.基于开路电压法的电池荷电状态估算修正[J].内燃机与配件,2019(19):225-226.
- [5] 陈志宣,王浩,陆玲霞,等.基于GRU软测量与卡尔曼滤波的电池SOC快速估计[J].电源技术,2025,49(4):740-749.

- [6] 孙腾,冯丹,胡利明.国外新能源汽车发展现状及对我国发展的启示[J].化工时刊,2018,32(9):30-35.
- [7] LEE J H, LEE I S. Hybrid estimation method for the state of charge of lithium batteries using a temporal convolutional network and XGBoost[J]. Batteries, 2023, 544.
- [8] 孔祥创,赵万忠,王春燕.基于BP-EKF算法的锂电池SOC联合估计[J].汽车工程,2017,39(6):648-652.
- [9] 商云龙,张承慧,崔纳新,等.基于模糊神经网络优化扩展卡尔曼滤波的锂离子电池荷电状态估计[J].控制理论与应用,2016,33(2):212-220.
- [10] 马向平,靳皓晴,朱奇先,等.锂离子电池荷电状态的在线融合估计方法[J].兰州理工大学学报,2020,46(5):78-84.
- [11] 王超,陈德海,王显朝,等.基于无迹Kalman滤波算法的动力电池荷电状态估计[J].汽车安全与节能学报,2020,11(3):379-387.
- [12] 滕曦.基于Stacking融合模型的新能源汽车动力电池SOC预测研究[D].重庆:重庆工商大学,2022.
- [13] 段春艳,刘千拓,王佳洁,等.基于Stacking集成学习的乳腺癌智能诊断预测模型[J].同济大学学报(自然科学版),2025,53(6):976-984.
- [14] 任彩玉,郭小燕,刘立群,等.应用Stacking集成学习模型短期预测临夏州马铃薯产量[J].东北农业大学学报,2025,56(2):156-167.

【作者简介】

王志军(1974—),男,河南三门峡人,专科,工程师,研究方向:储能磷酸铁锂电池寿命因素分析。

(收稿日期:2025-07-21 修回日期:2026-01-15)