计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(2):45−54 [doi: 10.15888/j.cnki.csa.008943]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于特征选择和数据增强的电池荷电状态预测①

朱月凡[1], 蒋国平[1], 高　辉[1], 李炜卓[2], 归耀城[2]

[1](南京邮电大学 自动化学院、人工智能学院, 南京 210023)
[2](南京邮电大学 现代邮政学院, 南京 210003)
通信作者: 高　辉, E-mail: 1020051533@njupt.edu.cn

摘　要: 现有基于神经网络的电池荷电状态 (state of charge, SOC) 预测研究大多把重点放在模型结构和相关参数的优化上, 却忽略了训练数据的重要作用. 针对该问题, 文中提出了一种基于特征选择和数据增强的电池 SOC 预测方法. 首先, 方法根据原始电池充放电数据进行特征工程, 并使用排列重要性 (permutation importance, PI) 方法选出对模型预测最有帮助的 7 个特征; 其次, 通过加入高斯噪声来扩大训练数据样本总量, 达到数据增强的目的. 实验使用双向长短时记忆网络 (bidirectional long short-term memory, Bi-LSTM) 作为预测模型, 使用 Panasonic 18650PF 数据集作为训练数据. 使用标准 Bi-LSTM 进行预测时, 平均绝对误差 (mean absolute error, *MAE*) 和最大误差 (max error, MaxE) 分别为 0.65% 和 3.92%, 而在进行特征选择和数据增强后, 模型预测的 *MAE* 和 MaxE 分别为 0.47% 和 2.62%, 表明 PI 特征工程与高斯数据增强方法可以进一步提升电池荷电状态预测模型的精度.

关键词: 电池荷电状态预测; 双向长短时记忆网络; 特征选择; 数据增强; 高斯噪声

引用格式: 朱月凡,蒋国平,高辉,李炜卓,归耀城.基于特征选择和数据增强的电池荷电状态预测.计算机系统应用,2023,32(2):45–54. http://www.c-s-a.org.cn/1003-3254/8943.html

Battery State of Charge Prediction Based on Feature Selection and Data Augmentation

ZHU Yue-Fan[1], JIANG Guo-Ping[1], GAO Hui[1], LI Wei-Zhuo[2], GUI Yao-Cheng[2]

[1](College of Automation & College of Artificial Intelligence, Nanjing University of Posts and Telecommunications, Nanjing 210023, China)
[2](School of Modern Posts, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: The existing research on battery state of charge (SOC) prediction based on neural networks mostly focuses on the optimization of model structure and related parameters, ignoring the important role of training data. A battery SOC prediction method based on feature selection and data augmentation is proposed to overcome this problem. Specifically, feature engineering is carried out according to the original battery charge and discharge data, and seven features that are most helpful to model prediction are selected by the permutation importance (PI) method; then, Gaussian noise is added to expand the total number of training data samples and thereby achieve the purpose of data augmentation. In the experiment, a bidirectional long short-term memory (Bi-LSTM) network is used as the prediction model, and the Panasonic 18650PF dataset is adopted as the training data. When the standard Bi-LSTM model is employed for prediction, the mean absolute error (*MAE*) and the maximum error (MaxE) are 0.65% and 3.92% respectively. After feature selection and data augmentation, the *MAE* and MaxE of model prediction are 0.47% and 2.62% respectively, indicating that the accuracy of the battery SOC prediction model can be further improved by PI feature engineering and the Gaussian data augmentation method.

Key words: state-of-charge prediction; bidirectional long short-term memory (Bi-LSTM) network; feature selection; data augmentation; Gaussian noise

① 基金项目: 国家自然科学基金 (52077107)
收稿时间: 2022-06-28; 修改时间: 2022-07-25; 采用时间: 2022-08-15; csa 在线出版时间: 2022-12-06
CNKI 网络首发时间: 2022-12-07

精确的 SOC 预测作为电池管理系统中极其重要的一环, 可以保障电池免受过充过放带来的损害. 然而, 锂离子电池是一个非常复杂的非线性系统, 它的荷电状态无法直接测量, 只有通过可直接测量的电池端电压、充放电电流和电池表面温度来对其进行估计. 此外, 电池的健康状态、自放电行为都会对 SOC 预测带来不利影响. 因此, 如何准确预测锂离子电池的荷电状态是一个值得研究问题.

SOC 是电池中所存储能量的相对度量, 定义为特定时间点可从电芯提取的电荷量与总容量之比[1]. 目前用来预测 SOC 的方法主要有 4 种: 安时计数法, 开路电压法, 基于模型的方法和基于数据驱动的方法. 安时计数法是使用最广泛的 SOC 预测方法, 它的原理很简单, 用电流的积分代表电量, 这种方法的缺陷也很明显: 如果电流测量不准, 计算误差会随着时间不断增加, 需要不断的校准[2]; 开路电压法的原理是电池 SOC 与电池开路电压有着单调的对应关系[3], 但是开路电压的测量需要先将电池静置 2 h 以上, 因此该方法不适合实时预测 SOC[4]. 基于模型的方法是指建立电池的状态空间模型, 常见的有卡尔曼滤波器及其变体[5,6], 粒子滤波器[7], H 无穷滤波器[8]. 基于模型的方法可以获得较好的预测效果, 但是这些模型需要大量的计算来进行参数辨识, 而且在不同的环境下, 模型也要改变[9,10].

神经网络可以通过智能算法自动学习网络参数并从中获取电池输入数据和 SOC 之间的关系. 近年来进行 SOC 估计的神经网络结构主要有径向基函数神经网络 (radial basis function neural network, RBFNN)[11], 反向传播神经网络 (back-propagation neural network, BPNN)[12] 等. 为了提升神经网络的预测精度, 需要进行大量实验去寻找最优参数, 采用布谷鸟搜索, 粒子群优化等算法可以快速寻找到神经网络相关参数的最优值[13].

为了解决传统神经网络中存在的梯度消失、过拟合等问题, 基于深度学习的 SOC 预测方法引起了学者们的广泛关注. 常用于电池 SOC 估计的深度学习模型有卷积神经网络 (convolutional neural network, CNN) 和循环神经网络 (recurrent neural network, RNN)[14,15], 其中, RNN 的发展, 尤其是对门控循环单元 (gated recurrent units, GRU) 和长短期记忆网络 (long short-term memory, LSTM) 的研究, 为解决时间序列预测问题提供了新的方向[16–19].

Chemali 等人使用 LSTM 训练了一个单一的网络模型, 该模型可以准确预测不同环境温度下的 SOC, 在固定的环境温度下, 模型的 MAE 为 0.573%, 在环境温度从 10℃ 上升到 25℃ 的数据集上, *MAE* 为 1.606%, 证明了该深度学习模型是 SOC 预测的强力工具[20]. 倪水平等人提出了一种基于一维卷积神经网络与长短期记忆网络结合的电池荷电状态预测方法, 该方法首先通过一层一维卷积层从样本数据中提取出高级数据特征, 之后再使用一层 LSTM 来预测 SOC 结果[21]. Bian 等人在 LSTM 的基础上引入了双向循环结构和多层隐藏层, 构建了 SBLSTM 模型, 该模型可以完全利用电池的时间信息来预测 SOC[22].

数据和特征决定了机器学习的上限, 而模型和算法只是逼近了这个上限而已. 然而, 现有的基于深度学习的 SOC 预测研究大多把研究重心放在模型的选择以及模型参数的优化上, 却忽视了数据本身的作用. 对深度学习来说, 海量的适配数据通常较难获得, 因此文中通过其他途径来优化数据, 继而提高预测的性能. 最常见的优化数据方法有特征选择和数据增强. 特征选择的目的是找到和预测目标关联性最强的特征, 这样可以极大的提升模型的预测效果[23]. 数据增强也叫数据扩增, 意思是在不实质性增加数据的情况下, 让有限的数据产生等价于更多数据的价值.

因此, 本文在选定了双向长短时记忆网络 (Bi-LSTM) 作为预测模型的基础上, 又采用了两种数据优化的方法: 使用 PI 特征选择算法来选择特征, 使用加入高斯噪声的方法来增强原始数据. 实验结果表明, 本文提出的 PI 特征工程和高斯数据增强方法可以进一步提升电池荷电状态预测模型的精度. 本文的剩余部分组织如下: 第 1 节介绍本文使用到的模型和方法: Bi-LSTM 模型、PI 算法、高斯数据增强. 第 2 节介绍了电池数据的获取以及实验步骤. 第 3 节是实验结果与分析. 第 4 节作为总结.

1 所用模型与方法介绍

1.1 Bi-LSTM 模型

1.1.1 LSTM 结构

传统的前馈神经网络缺乏使用历史信息的能力, 历史信息在时间序列的数据中是非常重要的, 电池的充放电数据就是很典型的时间序列数据. 因此, 建立一个可以有效利用历史信息的神经网络模型来预测电池的荷电状态是很有必要的. 循环神经网络 RNN 满足这

一要求, 但是普通的 RNN 由于反向传播时的梯度消失问题, 无法捕捉长期的依赖信息. 为了解决该问题, Hochreiter 等通过添加存储单元结构, 提出了 LSTM 循环神经网络[24]. LSTM 循环神经网络的结构如图 1 所示.

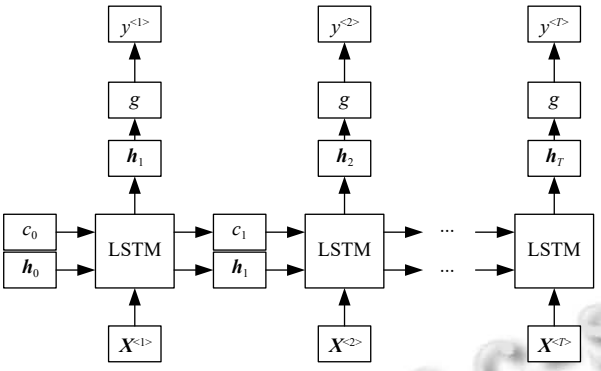


图 1 LSTM 循环神经网络

LSTM 的输入是一条序列 $\boldsymbol{S}=\{\boldsymbol{X}^{<1>}, \boldsymbol{X}^{<2>}, \cdots, \boldsymbol{X}^{<T>}\}$, 长度为 T, 序列中每一个元素是一个样本点 $\boldsymbol{X}=[x_1, x_2, \cdots, x_n]$, n 是样本的维度即特征个数. 在电池荷电状态预测中, 每个样本数据包含电流、电压、温度等 n 个特征, T 个连续的样本构成一条序列, 序列中最后一个元素对应的 SOC 值就是该条序列的标签. $\boldsymbol{h}$ 和 $\boldsymbol{c}$ 是隐藏层状态, 每个 $\boldsymbol{X}$ 对应一个输出 y, y 的计算公式为:

$$y^{<t>}=g(\boldsymbol{W}_{hy}\times\boldsymbol{h}_i+b_y) \tag{1}$$

其中, g 是激活函数, 根据具体情况选择使用哪个激活函数; $\boldsymbol{W}_{hy}$, $\boldsymbol{b}_y$ 是输出层的权重与偏置矩阵. 在电池荷电状态预测中, $y^{<T>}$是这条序列对应的最终预测值.

一个 LSTM 由一个保存历史信息的记忆单元和控制信息流动的门控机制组成. 图 2 显示的是一个 LSTM 单元的结构, 其中包括输入门、遗忘门、更新门和输出门. 输入门决定了当前输入的哪些信息可以被记忆下来, 它的计算公式为:

$$\boldsymbol{i}_t=\sigma(\boldsymbol{W}_{xi}\boldsymbol{X}^{<t>}+\boldsymbol{W}_{hi}\boldsymbol{h}_{t-1}+\boldsymbol{b}_i) \tag{2}$$

遗忘门会根据过去的记忆信息来决定当前输入的哪些信息是可以被忽略的, 其计算公式为:

$$\boldsymbol{f}_t=\sigma(\boldsymbol{W}_{xf}\boldsymbol{X}^{<t>}+\boldsymbol{W}_{hf}\boldsymbol{h}_{t-1}+\boldsymbol{b}_f) \tag{3}$$

更新门用来控制 LSTM 单元的状态信息 $\boldsymbol{c}$ 的更新, 更新方式如下:

$$\boldsymbol{c}'_t=\tanh(\boldsymbol{W}_{xc}\boldsymbol{X}^{<t>}+\boldsymbol{W}_{hc}\boldsymbol{h}_{t-1}+\boldsymbol{b}_c) \tag{4}$$

$$\boldsymbol{c}_t=\boldsymbol{f}_t\odot\boldsymbol{c}_{t-1}+\boldsymbol{i}_t\odot\boldsymbol{c}'_t \tag{5}$$

最后通过输出门得到的输出与 $\boldsymbol{c}_t$ 一起计算得到 $\boldsymbol{h}_t$, 计算公式如下:

$$\boldsymbol{o}_t=\sigma(\boldsymbol{W}_{xo}\boldsymbol{X}^{<t>}+\boldsymbol{W}_{ho}\boldsymbol{h}_{t-1}+\boldsymbol{b}_o) \tag{6}$$

$$\boldsymbol{h}_t=\boldsymbol{o}_t\odot\tanh(\boldsymbol{c}_t) \tag{7}$$

在式 (2)–式 (7) 中, σ是逻辑 S 型函数, 是激活函数; $\boldsymbol{W}_*$和 $\boldsymbol{b}_*$分别代表各个门的权重矩阵和偏置矩阵; $\vec{\boldsymbol{h}}$代表元素积, tanh 是一种可以将输入值归一化到 (−1, 1) 之间的激活函数.

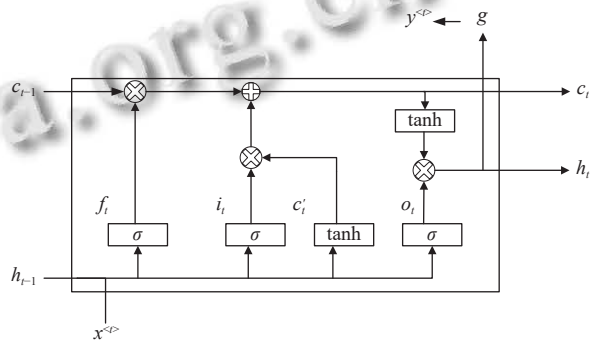


图 2 LSTM 单元结构

1.1.2 双向 LSTM 结构

RNN 和 LSTM 都是单向的, 这意味着他们只能访问正时间方向的输入时间序列. 与单向 RNN 不同, 双向 RNN (BRNN) 由两个独立的循环隐藏层组成, 一个用于正向输入序列, 另一个用于反向输入序列, 如图 3 所示. 其中$\vec{\boldsymbol{h}}$使用正向输入序列计算, $\overleftarrow{\boldsymbol{h}}$使用反向输入序列计算. $y^{<t>}$的计算公式为:

$$y^{<t>}=g\left(\vec{\boldsymbol{h}}_t, \overleftarrow{\boldsymbol{h}}_t\right) \tag{8}$$

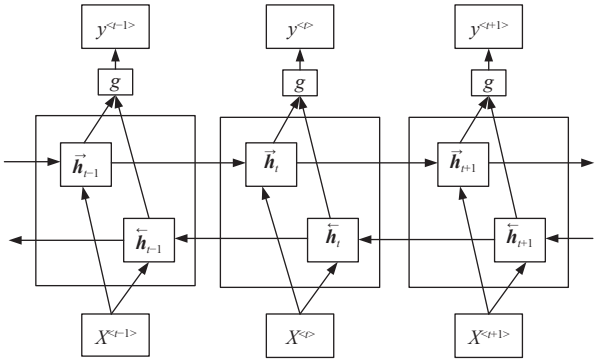


图 3 双向 RNN 结构

使用这种特殊的结构, BRNN 能够以向前和向后的方式捕获序列数据的时间依赖性, 然后将它们反馈

到同一输出层. BRNN 可以使用与单向 RNN 相同的算法进行训练, 因为前向层和后向层之间没有交互作用. 通过将双向结构应用于 LSTM, 可以构造双向 LSTM (Bi-LSTM).

1.2 排列重要性 (permutation importance, PI)

使用神经网络方法来做回归任务时, 特征工程是极其重要的一个环节, 有效的特征可以帮助神经网络模型捕捉到输入与输出之间的关系. 锂离子电池是一个复杂的非线性系统, 然而电池充放电实验中可以获取到的原始数据只有电压、电流和温度, 仅仅有这些是不够的. 此时就需要通过特征工程来根据原始数据构造新的特征, 然而, 并不是所有的特征都能提高模型的预测精度, 这就需要一种方法去找出对 SOC 预测最重要的几个特征. 本文使用 PI 方法来寻找重要特征, 使用该方法寻找重要特征的步骤如下.

Step 1. 用上全部特征训练一个模型.

Step 2. 在验证集上验证模型效果, 计算损失 $loss_{\text{raw}}$.

Step 3. 将验证集的一个特征 f 对应的数据打乱, 再次预测并得到损失 $loss_f$.

Step 4. 将上述得分做差即可得到特征 f 对预测的重要性: $imp_f=|loss_f-loss_{\text{raw}}|$.

Step 5. 依次将每一列特征按上述方法做, 得到每个特征对预测的重要性.

Step 6. 最后根据计算得到的重要性来选取适合的特征.

Step 7. 使用选出的特征再次训练一个模型, 并在测试集上验证其效果.

该方法中的特征重要性计算是在模型训练好后进行的, 所以不需要再训练其他的模型, 这大大节省了时间. 将一个特征的数据随机打乱会降低模型在测试集上的预测效果, 因为打乱后的数据就没有任何实际意义了. 被打乱的特征越是重要, 那么预测的结果就越差, 而那些对预测效果影响不大的特征被打乱后, 模型的预测效果也不会受到太大影响, 因此通过计算打乱数据后的模型预测损失 $loss_f$ 与原损失 $loss_{\text{raw}}$ 的差值所得到的 imp_f 可以表征特征 f 对模型预测的重要性.相对于 PCA 和 EMD 等特征选择方法来说, PI 方法有两个优点, 一是该方法可以获得每个特征对模型重要性的数值度量, 让研究者可以直观地知道哪个特征重要以及为什么重要. 二是在选出重要的几个特征后, 以后在对新数据进行预测时, 只要将需要的特征制作出来就可以直接进行预测了, 而 PCA 方法则每次都需要制作出所有特征再进行降维处理才能进行预测.

1.3 高斯数据增强

在训练神经网络时, 往往会遇到过拟合的问题, 即网络可以学习特定的输入示例及其关联的输出, 而不是学习从输入到输出的一般映射. 这将导致模型在训练数据集上表现良好, 而在新数据上表现不佳, 即泛化能力很差.

目前解决过拟合的方法主要有两个方向. 一是对神经网络做改动, 例如减少神经元个数、减少隐藏层层数和降低学习率等, 这类方法是通过限制网络学习训练数据的能力来增强泛化能力, 虽然有效, 但却无法充分利用神经网络强大的学习能力.

二是对训练数据做改动, 最理想的方式是获取更多更全面的数据, 但往往不容易实现. 一种合理的解决方案是通过给训练数据添加噪声从而增加数据量. 添加噪声意味着网络无法记住训练样本, 因为它们一直在变化, 从而导致网络权重更小且网络更健壮, 泛化误差更低. 噪声意味着从已知样本附近的域中抽取新样本, 从而平滑了输入空间的结构. 这种平滑意味着训练数据更易于网络学习, 从而导致更好, 更快的学习. 高斯噪声常常用来增强训练数据, 高斯噪声是概率密度函数服从高斯分布的一类噪声, 它的平均值为 0, 标准差为 1, 概率密度函数为:

$$f(x)=\frac{1}{\sqrt{2\pi}}\mathrm{e}^{\left(-\frac{x^2}{2}\right)} \tag{9}$$

在对训练数据添加噪声前会将训练数据进行归一化, 本文使用最大最小值方法将训练数据的每一特征列归一化到 [0, 1] 之间. 之后在每一轮迭代训练开始前对训练数据添加高斯噪声, 添加的高斯噪声的幅度要合理的设置, 太小则新数据与原数据几乎相同达不到数据增强的效果, 太大则新数据可能是现实情况中不可能出现的数据, 这种数据是无效的. 在实验过程中需要不断对噪声幅度进行调整来寻找最佳的幅度值.

2 电池数据获取与实验步骤

2.1 电池数据介绍

本文使用 Panasonic 18650PF 数据集[20] 来验证所提方法的有效性. 所使用的电池的详细参数如表 1 所示. 电池采用恒流恒压的方式进行充电, 一开始电流为

1C (2.9 A), 电压为 4.2 V, 当电流降到 50 mA 时停止充电, 此时电池的 SOC 为 1. 然后, 使用驱动循环功率模式 (drive cycle power profiles, 后称工况) 对电池进行放电, 直到电压降至 2.5 V. 在试验中使用了 9 种工况, 包括循环 1、循环 2、循环 3、循环 4、洛杉矶 92 (Los Angeles 92, LA92)、公路燃油经济性驾驶计划 (highway fuel economy test, HWFET)、补充联邦测试程序驾驶计划 (supplemental federal test procedure driving schedule, US06)、城市测功机驾驶计划 (urban dynamometer driving schedule, UDDS)、神经网络 (neural networks, NN). 循环 1–4 包括 US06、HWFET、UDDS 和 LA92 的随机混合. NN 由 US06 和 LA92 驱动循环的部分组合而成, 其设计具有一些额外的动力学特性, 有助于训练神经网络. 每种工况分别在 3 个不同的环境温度下 (0℃, 10℃, 25℃) 进行放电实验并获取相应的数据, 用来验证本文所提方法在不同环境温度下的有效性.

表 1 电池信息

属性	参数
额定容量 (Ah)	2.9
额定开路电压 (V)	3.6
最小/最大电压 (V)	2.5/4.2

各种工况都是模拟现实中电动汽车在不同情况下的行驶特点, 不同工况的区别具体体现在电池放电时端电流的曲线不同, 最终获取的电池数据中, 电流数据是固定的, 只有电压和温度是在放电过程中实时测量获得的. 记录数据的时间间隔为 1 s, 在不同的工况下, 放电所需时间大约在 6 000 s 到 10 000 s 之间. 图 4 显示了 US06、LA92、UDDS 和 HWFET 这 4 种工况在 25℃ 下的电流曲线. 从图 4 可以看出, 各个工况下的电流曲线都表现出了极强的非线性.

Panasonic 18650PF 数据集中的 9 种工况并不能涵盖现实中电动汽车的所有行驶情况, 为了能让本文提出的模型能应用于现实, 本文选取 NN 工况下的数据作为验证集, 包含 11 698 条数据, US06 工况和 HWFET 工况下的数据作为测试集来验证模型效果, 分别包含 4 806 和 7 595 条数据, 其余 6 种工况下的数据作为训练集, 共有 80 924 条数据. 图 5 显示了 US06 工况在 0℃ 和 25℃ 下, 电流 (图 5(a))、电压 (图 5(b))、温度 (图 5(c)) 和电池剩余容量 (图 5(d)) 随时间的变化曲线. 在 25℃ 的情况下, 放电过程中是存在再生制动的, 而 0℃ 的情况下没有出现, 这是因为当温度低于 10℃ 时进行充电会对电池造成较大伤害.

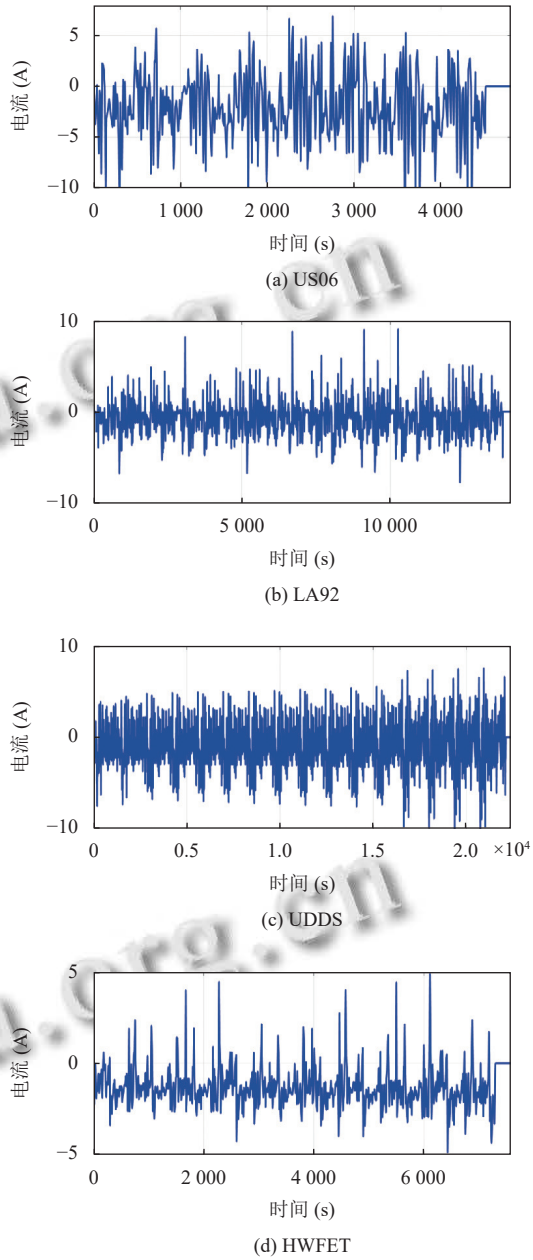


图 4 不同工况下的电流曲线

2.2 实验步骤

2.2.1 数据处理与特征工程

原始数据包括电压、电流和温度, 数据的采集间隔为 1 s, 每个温度下的数据总量大约为 100 000 条. 根据采集的数据进行特征工程, 共制作 n 个特征 [$f1$, $f2$, ⋯, fn]. 此后还要将每个特征分别进行归一化, 本文采用最大最小值归一化方法, 计算方法如下:

$$x_{\text{norm}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{10}$$

其中, x, $x_{\min}$, $x_{\max}$, x_{norm} 分别代表原始值、该特征的最小值、该特征的最大值和归一化后的值. 最后要构造时间序列数据, 本实验设定一条序列的时间跨度为 50 s, 所以序列长度 N_{seq} 为 50.

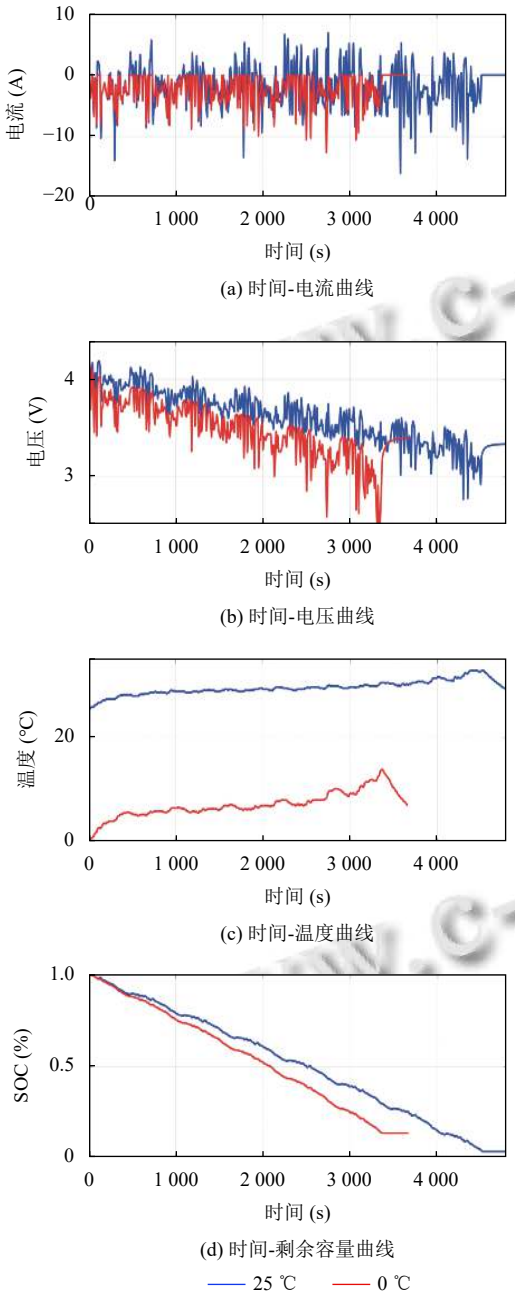


图 5 US06 工况在 25℃ 和 0℃ 下测得的电池放电数据

2.2.2 实验环境和 Bi-LSTM 模型设置

本文使用 Python 的 PyTorch 库搭建 Bi-LSTM 模型, 模型的各个相关参数设置如表 2 所示. 实验硬件设施使用 AMD Ryzen 7 5800U CPU @3.2 GHz, Windows 11 64 位操作系统, 16 GB 运行内存和 500 GB 硬盘. 编程软件为 Jupyter Notebook.

表 2 Bi-LSTM 模型参数设置

参数	值
学习率	0.005
隐藏层神经元个数	128
隐藏层层数	1
批处理大小	128
迭代次数	100
Dropout率	0.8
是否使用双向结构	是

2.2.3 模型训练

本文使用 NN 工况下测得的数据作为验证集, US06 和 HWFET 工况下测得的数据作为测试集, 其余数据作为训练集. 每次训练都要将所有训练数据迭代 100 次, 每次迭代训练包括以下 3 个步骤.

Step 1. 读入所有训练数据并加入噪声, 这样可以保证每次迭代训练的数据都不同.

Step 2. 模型训练.

Step 3. 在测试集上验证此时模型的效果, 计算各个性能指标.

100 次迭代结束后, 取最后 10 次迭代的 *MAE*、*RMSE*、MaxE 的平均值作为该模型最终的效果.

3 实验结果与分析

本文使用平均绝对误差 *MAE*, 均方根误差 *RMSE* 和最大误差 MaxE 来衡量模型预测的效果, 它们的计算公式分别为:

$$MAE = \frac{1}{N}\sum_{k=1}^{N}|y_k - \hat{y}_k| \tag{11}$$

$$RMSE = \frac{1}{N}\sum_{k=1}^{N}\sqrt{(y_k - \hat{y}_k)^2} \tag{12}$$

其中, N 是样本总数, y_k是样本的真实 SOC 值, $\hat{y}_k$是该条样本的预测值.

3.1 使用 PI 寻找重要特征

本文一共根据原始数据制作了 14 个特征, 主要分为 3 类, 原始数据特征 [v, i, $temp$] 和 $charge$、导数类特征 [dv, di, $dtemp$, d^2v, d^2i, d^2temp] 和历史统计特征

[$\int vdt, \int idt, \int tempdt, \int chargedt$]. 特征 *charge* 指前一个采集时间点到当前时间的放电量. 导数类特征可以捕捉电压、电流和温度的短期变化信息, 由当前时间的值与前一个时间的值作差得到. 历史统计特征可以捕捉电压、电流和温度的长期历史行为信息, 历史窗口长度为 10 s.

之后根据 PI 算法计算每个特征的重要性, 如图 6 所示. 从图中可以看出, 最重要的特征是v和$\int vdt$, 这比较容易理解因为从图 5(b) 和图 5(d) 中就可以看出电池电压曲线与 SOC 曲线有着相同的变化趋势. 其次, $\int chargedt$, *charge*, i, $\int idt$, *temp* 有较高的重要性. $\int chargedt$, *charge* 反映了在过去的一段时间 (10 s) 内电池的充放电总量, 由于时间短, 使用安时积分法不得不考虑的累积误差可以忽略不计. i, *temp* 作为可以直接反映电池当前状态的数据, 对 SOC 的预测有着不可忽视的作用. 其余特征的重要性太小, 从 PI 算法的角度来看, 这些特征对 SOC 预测几乎没有作用. 因此本文选择 v, $\int vdt$, $\int chargedt$, *charge*, i, $\int idt$, *temp* 这 7 个特征来训练模型, 称为 PI 特征. 表 3 列出了模型在不同的特征下的 SOC 预测性能. 图 7 显示了在不同特征下模型的预测效果. 从表 3 中可以看出, 当使用所有特征时, 模型的预测效果相较于只用原始特征时已有提升, 提升了 10.7%; 当使用 PI 特征时, 模型的预测效果有更大幅度的提升, 提升了 20%. 这说明 PI 算法的确可以从众多的特征中挑选出对 SOC 预测最重要的特征, 同时也说明了无效的特征会降低模型的性能.

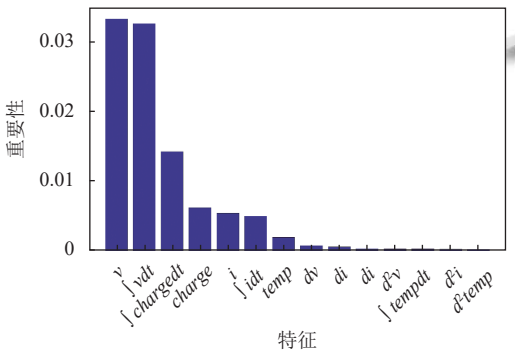


图 6 各个特征的重要性

3.2 高斯数据增强对模型预测的影响

本文中给训练数据加的高斯噪声服从标准正态分布, 平均值为 0, 标准差为 1. 训练时, 在每一个 epoch 开始前都要重新读取数据并加入高斯噪声, 这样训练集数据总量相当于原来的 100 倍. 加入了高斯噪声后模型的预测结果如表 4 和图 8 所示. 从表 4 中可以看出, 加入噪声的幅度不同, 对模型预测效果的影响也不同.噪声幅度为 0.001 时, 模型预测效果并无明显提升, 这是因为新数据和原数据几乎一样, 这并没有达到数据增强的目的, 相当于将一样的数据迭代 100 次, 这与寻常的训练方式并无区别. 噪声幅度为 0.005 时, 模型的预测效果有了较大提升, *MAE* 值和 *RMSE* 值分别为 0.47% 和 0.68%, 相较于没有噪声的情况下提升了 9.6%, 6.8%, 这说明通过添加适当的高斯噪声得到的大量额外数据可以弥补原始训练数据过少的缺点, 从而提升 Bi-LSTM 模型的预测效果. 当噪声幅度增大到 0.01 时, 模型的预测效果反而变差, 这说明加入的噪声幅度不能过大, 这样会导致获取到的新数据失去现实意义, 这种数据相当于异常数据, 会对模型预测产生负面影响.

表 3 不同特征下模型的预测结果 (%)

工况	特征	*MAE*	*RMSE*	MaxE
HWFET	*i*, *v*, *temp*	0.65	0.90	3.26
	PI	**0.52**	**0.73**	**2.62**
	All	0.58	0.82	2.99
US06	*i*, *v*, *temp*	0.61	0.82	4.05
	PI	**0.56**	**0.72**	**2.63**
	All	0.59	0.78	3.00

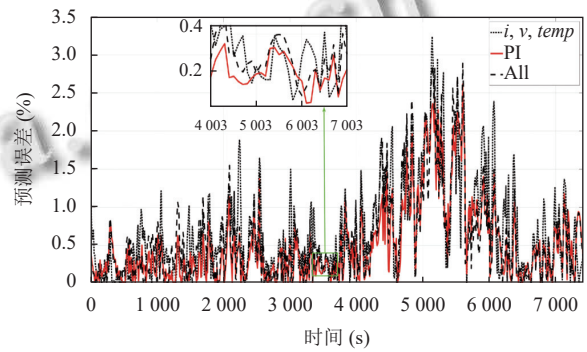


图 7 不同特征下模型的预测误差

表 4 高斯数据增强后的预测效果

工况	噪声幅度	*MAE* (%)	*RMSE* (%)	MaxE (%)
HWFET	无	0.52	0.73	2.62
	0.001	0.51	0.73	2.64
	0.005	**0.47**	**0.68**	**2.61**
	0.01	0.58	0.79	2.87
US06	无	0.56	0.72	2.63
	0.001	0.54	0.72	2.62
	0.005	**0.51**	**0.70**	**2.53**
	0.01	0.61	0.81	3.32

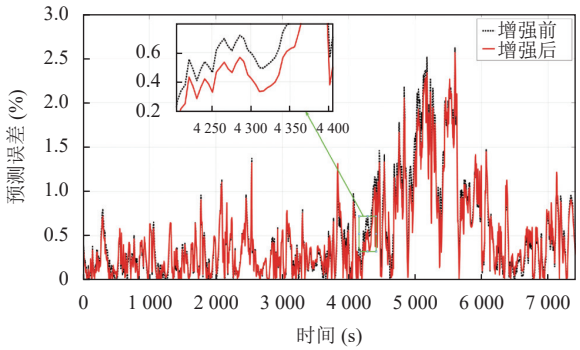


图 8 数据增强后的模型预测误差

此外, 考虑到现实情况中电动汽车上的电流、电压和温度传感器的精度会随着使用时间下降, 这将导致各项数据无法准确测量. 所以进行了额外的实验: 给测试集也加上高斯噪声来验证 Bi-LSTM 模型的鲁棒性. 给测试集加的高斯噪声的幅度较大, 包括 [0.01, 0.05, 0.1], 因为数据集的数据已经归一化到 0 到 1 之间, 所以加入噪声就相当于测量精度为 [1%, 5%, 10%], 测试结果如表 5 所示.

表 5 测试集有噪声时的预测效果

工况	噪声幅度	*MAE* (%)	*RMSE* (%)	MaxE (%)
HWFET	**无**	**0.47**	**0.68**	**2.61**
	0.01	0.62	0.91	4.56
	0.05	0.91	1.20	5.42
	0.1	1.32	1.68	7.92
US06	**无**	**0.51**	**0.70**	**2.53**
	0.01	0.75	0.93	2.75
	0.05	1.05	1.29	3.79
	0.1	1.57	1.96	7.34

从表 5 中可以看出, 测试集加入高斯噪声后, 模型在测试集上的预测效果有所下降, 当高斯噪声的幅度在 0.05 以内时, 模型预测的 *MAE* 维持在 1% 以内, 最大误差在 5% 左右, 这是一个可以接受的值. 但当高斯噪声的幅度达到 0.1 后, 即测量精度只有 10%, 模型在测试集上的预测效果很差, 最大误差达到了 8%. 由此可见, 在实际情况下, 即使锂离子电池的各项数据不能准确测量, 但只要把精度控制在 5% 以内, 使用本模型来预测 SOC 都能达到较好的效果.

3.3 不同环境温度下的 SOC 预测

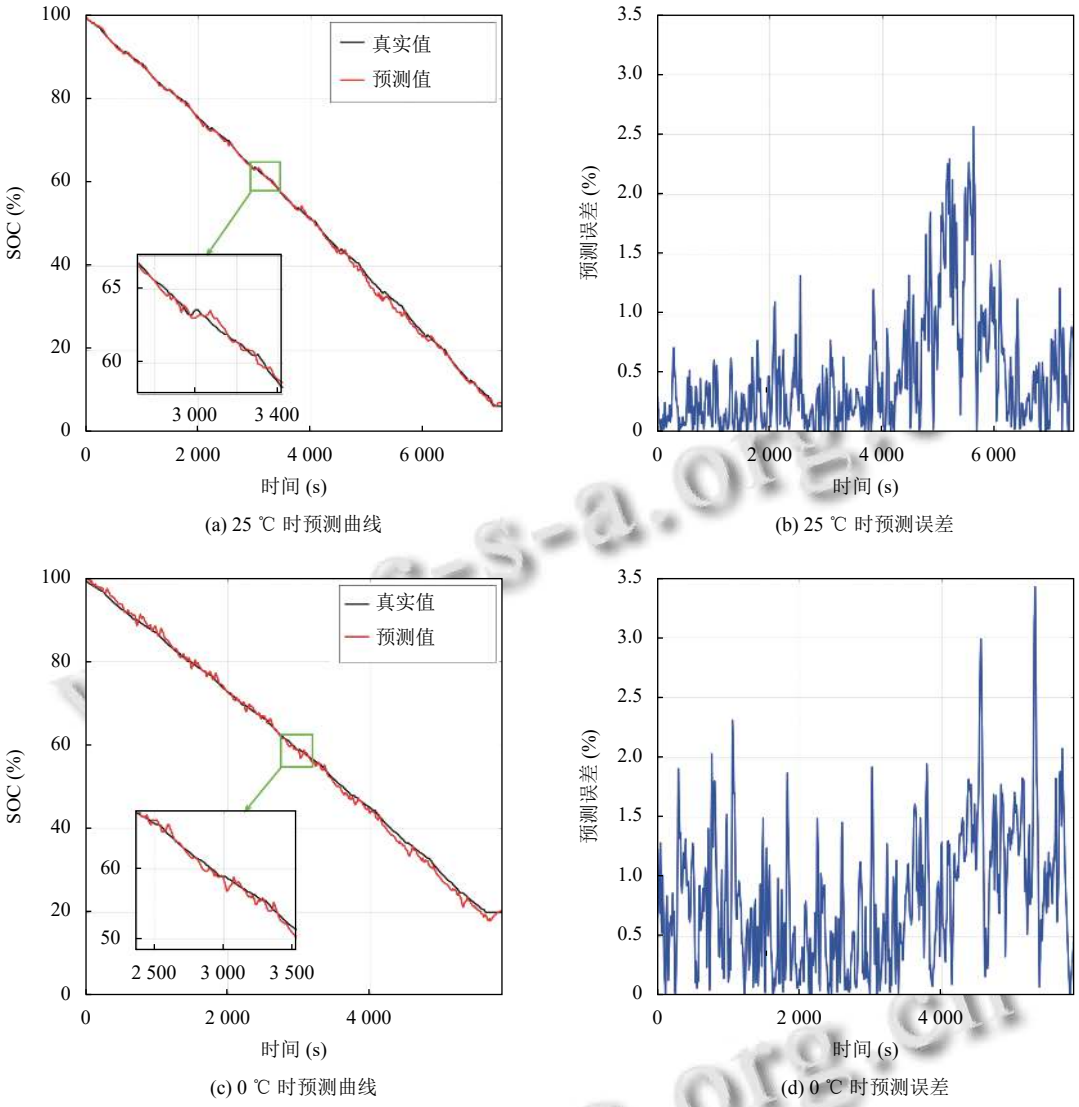
温度对锂离子的电池荷电状态预测有着极大的影响, 当环境温度固定时, 电池在充放电过程中表面温度会随时间变化, 但整体上不会与环境温度相差太大. 电池表面温度一方面可以作为 Bi-LSTM 的一个输入特征来预测 SOC, 此外电池表面温度还影响电池的充放电策略, 比如当电池表面温度过低时给电池充电会对电池造成损害, 因此应当避免在低温下进行充电. 本文研究了模型在 3 个固定的环境温度下的预测效果, 分别是 0℃, 10℃, 25℃. 预测效果如表 6 所示, 预测曲线和误差曲线如图 9 所示. 从表 6 中可以看出, 模型在 3 个不同温度下的预测效果有所不同, 在 HWFET 工况下, *MAE* 分别为 0.47%, 0.84%, 0.80%, MaxE 分别为 2.61%, 4.05%, 3.61%. 由此可见, 温度对模型预测效果有着一定影响, 模型在低温 (0℃, 10℃) 下的预测精度要低于常温 (25℃) 下的预测精度. 作为对比, 文献 [20] 中的 LSTM 神经网络在相同的数据集下的预测结果分别是 0.774%, 0.782%, 2.088% 和 3.692%, 4.047%, 6.687%. 本文提出的方法获得了更好的预测效果.

表 6 不同环境温度下模型的预测结果

工况	温度(℃)	*MAE* (%)	*RMSE* (%)	MaxE (%)
HWFET	25	0.47	0.68	2.61
	10	0.84	1.11	4.05
	0	0.80	0.97	3.61
US06	25	0.51	0.70	2.53
	10	1.09	1.41	5.20
	0	0.85	1.10	4.62

4 结论

本文提出了一种基于特征选择和高斯数据增强的锂离子电池 SOC 预测方法. 该方法的贡献主要有 3 个方面. 1) 根据原始电池数据设计了大量额外的特征, 并使用 PI 特征选择算法选出最有效的 7 个特征, 它们是 v, $\int v dt$, $\int charge dt$, $charge$, i, $temp$, dv. 2) 引入高斯噪声的方法来增强训练数据, 使数据构成的输入空间更加平滑, 提升了 Bi-LSTM 模型的预测效果. 3) 通过预测加入了高斯噪声的测试集, 证明了当测量数据与真实数据的误差在 5% 以内时, 模型依然有较好的预测效果, 即具备较强的鲁棒性. 本文提出的 PI 特征工程与高斯数据增强方法可以为后续研究者在预测 SOC 时提供新的数据改良途径, 从而帮助研究者进一步提升模型的预测效果. 将来的工作重点围绕研究其他类型的噪声对原始数据的增强效果, 并通过其他型号的锂离子电池来验证提出方法的有效性.



(a) 25 ℃ 时预测曲线　(b) 25 ℃ 时预测误差

(c) 0 ℃ 时预测曲线　(d) 0 ℃ 时预测误差

图 9　HWFET 在 25℃ 和 0℃ 下模型的预测曲线和误差曲线

参考文献

1 张照娓, 郭天滋, 高明裕, 等. 电动汽车锂离子电池荷电状态估算方法研究综述. 电子与信息学报, 2021, 43(7): 1803–1815. [doi: 10.11999/JEIT200487]

2 Truchot C, Dubarry M, Liaw BY. State-of-charge estimation and uncertainty for lithium-ion battery strings. Applied Energy, 2014, 119: 218–227. [doi: 10.1016/j.apenergy.2013.12.046]

3 Shen P, Ouyang MG, Lu LG, *et al*. The co-estimation of state of charge, state of health, and state of function for lithium-ion batteries in electric vehicles. IEEE Transactions on Vehicular Technology, 2018, 67(1): 92–103. [doi: 10.1109/TVT.2017.2751613]

4 Barai A, Widanage WD, Marco J, *et al*. A study of the open circuit voltage characterization technique and hysteresis assessment of lithium-ion cells. Journal of Power Sources, 2015, 295: 99–107. [doi: 10.1016/j.jpowsour.2015.06.140]

5 郝世宇, 殷会飞, 杨茹, 等. 基于 AEKF 的锂离子动力电池荷电状态估计. 国外电子测量技术, 2021, 40(9): 49–53. [doi: 10.19652/j.cnki.femt.2102842]

6 朱奕楠, 吕桃林, 赵芝芸, 等. 基于并行卡尔曼滤波器的锂离子电池荷电状态估计. 储能科学与技术, 2021, 10(6): 2352–2362. [doi: 10.19799/j.cnki.2095-4239.2021.0169]

7 Bartlett A, Marcicki J, Onori S, *et al*. Electrochemical model-based state of charge and capacity estimation for a composite electrode lithium-ion battery. IEEE Transactions on Control Systems Technology, 2016, 24(2): 384–399.

8 Aung H, Low KS, Goh ST. State-of-charge estimation of

lithium-ion battery using square root spherical unscented Kalman filter (Sqrt-UKFST) in nanosatellite. IEEE Transactions on Power Electronics, 2015, 30(9): 4774–4783. [doi: 10.1109/TPEL.2014.2361755]

9 骆凡, 黄海宏, 王海欣. 基于电化学阻抗谱的退役动力电池荷电状态和健康状态快速预测. 仪器仪表学报, 2021, 42(9): 172–180. [doi: 10.19650/j.cnki.cjsi.J2107637]

10 武龙星, 庞辉, 晋佳敏, 等. 基于电化学模型的锂离子电池荷电状态估计方法综述. 电工技术学报, 2022, 37(7): 1703–1725.

11 张森. 基于径向基函数网络的 MH/Ni 电池荷电状态预测. 化工学报, 2006, 57(9): 2162–2166. [doi: 10.3321/j.issn:0438-1157.2006.09.029]

12 唐豪, 张振东, 吴兵. 基于 BP 神经网络的 HPPC 低温 SOC 优化估计. 计算机系统应用, 2021, 30(6): 293–299. [doi: 10.15888/j.cnki.csa.007955]

13 Lipu MSM, Hannan MA, Hussain A, *et al*. Optimal BP neural network algorithm for state of charge estimation of lithium-ion battery using PSO with PCA feature selection. Journal of Renewable and Sustainable Energy, 2017, 9(6): 064102. [doi: 10.1063/1.5008491]

14 Shin HC, Roth HR, Gao MC, *et al*. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Transactions on Medical Imaging, 2016, 35(5): 1285–1298. [doi: 10.1109/TMI.2016.2528162]

15 Sherstinsky A. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. Physica D: Nonlinear Phenomena, 2020, 404: 132306. [doi: 10.1016/j.physd.2019.132306]

16 Xiao B, Liu YG, Xiao B. Accurate state-of-charge estimation approach for lithium-ion batteries by gated recurrent unit with ensemble optimizer. IEEE Access, 2019, 7: 54192–54202. [doi: 10.1109/ACCESS.2019.2913078]

17 Yang FF, Li WH, Li C, *et al*. State-of-charge estimation of lithium-ion batteries based on gated recurrent neural network. Energy, 2019, 175: 66–75. [doi: 10.1016/j.energy.2019.03.059]

18 Chinomona B, Chung C, Chang LK, *et al*. Long short-term memory approach to estimate battery remaining useful life using partial data. IEEE Access, 2020, 8: 165419–165431. [doi: 10.1109/ACCESS.2020.3022505]

19 李超然, 肖飞, 樊亚翔. 基于循环神经网络的锂电池 SOC 估算方法. 海军工程大学学报, 2019, 31(6): 107–112. [doi: 10.7495/j.issn.1009-3486.2019.06.020]

20 Chemali E, Kollmeyer PJ, Preindl M, *et al*. Long short-term memory networks for accurate state-of-charge estimation of Li-ion batteries. IEEE Transactions on Industrial Electronics, 2018, 65(8): 6730–6739. [doi: 10.1109/TIE.2017.2787586]

21 倪水平, 李慧芳. 基于一维卷积神经网络与长短期记忆网络结合的电池荷电状态预测方法. 计算机应用, 2021, 41(5): 1514–1521.

22 Bian C, He HL, Yang SK. Stacked bidirectional long short-term memory networks for state-of-charge estimation of lithium-ion batteries. Energy, 2020, 191: 116538. [doi: 10.1016/j.energy.2019.116538]

23 Chandrashekar G, Sahin F. A survey on feature selection methods. Computers & Electrical Engineering, 2014, 40(1): 16–28.

24 Graves A. Long short-term memory. Supervised Sequence Labelling with Recurrent Neural Networks. Berlin: Springer, 2012. 37–45.

(校对责编: 牛欣悦)