Data Enhancing for Accurate State-of-Charge Estimation with Ensemble Trees

Okemakinde Femi, Kim Jonghoon
Energy Storage Conversion Lab., Chungnam National University

ABSTRACT

Researchers commonly train models using normalized raw data, yet the data quality significantly impacts the estimation model's performance. This study presents a data feature engineering technique aimed at augmenting the battery variables by manipulating existing data, thus generating additional data automatically accessible to the estimation model. An XGBoost model was developed to verify the integration of this technique, followed by performance analysis. The state-of-charge estimation outcomes derived from the XGBoost models, validated by simulations, underscored the efficacy and reliability of the feature engineering technique. This research showcases the importance of data quality and the potential of data enhancement in improving state-of-charge estimation model accuracy and reliability.

1. Introduction

Machine learning has demonstrated significant industrial impacts across various sectors, such as healthcare, construction, finance, and agriculture, benefiting from its problem-solving capabilities. Battery data can be acquired through laboratory simulations of different driving profiles, while large volumes of complex data can be obtained from onboard batteries using an online BMS. This underscores the potential of machine learning techniques in the field of battery technology for accurate SOC estimation. Numerous studies have shown that the data-driven approach offers more precise SOC estimation than conventional and model-based methods. Researchers have continuously explored various machine learning models for SOC estimation, including support vector machines (SVM), neural networks (NN), and Gaussian process regression (GPR)[1]. Among these models, neural networks have been widely employed due to their ability to extract intricate details from input data.

This further proves the significance of supplying the best dataset for improving proposed estimation models. The study aims to optimize the weights and biases of machine learning models under diverse circumstances. These parameters are crucial for deriving SOC values from computations on the input dataset. To achieve this, a feature engineering technique was integrated into the estimation process, enabling the creation of entirely new data from the available dataset, thereby enriching the dataset for improved learning.

2. Datasets and the Enhancing Technique

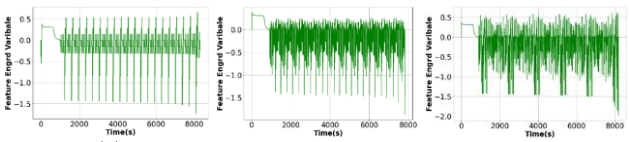

Fig.1 FEV of training data for the driving profiles at room temperature: (a) DST; (b) US06; (c) FUDS.

2.1 Training and Testing Dataset

The model was trained using a dataset from the Battery Research Group of the Center for Advanced Life Cycle Engineering (CALCE)[2] of tests performed on an A123 LiFePO4 battery. Experiments were conducted simulating three drive cycles: the dynamic stress test (DST), US06 highway driving schedule, and federal urban driving schedule (FUDS), across six temperatures from -10℃ to 40℃.

The testing data was obtained by using LG M50L cylindrical cells. The cells were fully charged at the initial stage and left stationary for two hours. Subsequently, a discharge process lasting 12 minutes at a current of 0.5 C rate ensued, resulting in a 5% SOC decrease. Following each discharge, a two-hour rest period was observed. This cycle was repeated until the cells reached 0% SOC and covered temperatures ranging from -10℃ to 45℃.

2.2 Dataset Processing and Enhancing

The preprocessing phase focused on selecting the three primary battery features: current, voltage, and temperature. The SOC reference values were obtained through the Coulomb counting method. Input data for the models were structured into a data frame represented as X = [I, V, T], and the reference SOC value was arranged into an array format denoted as y = [SOC], as depicted in Eq. (1).

$$\mathbf{X} = \begin{bmatrix} I_1 & V_1 & T_1 \\ I_2 & V_2 & T_2 \\ I_3 & V_3 & T_3 \\ \vdots & \vdots & \vdots \\ I_n & V_n & T_n \end{bmatrix}; \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} \quad (1)$$

Feature engineering was employed in this study to generate new battery variable samples for the estimation model. A novel feature termed the feature-engineered variable (FEV) was created by dividing the current (I) by the voltage (V). The FEV variables obtained for the training data at each drive cycle at room temperature are shown in Fig. 1.

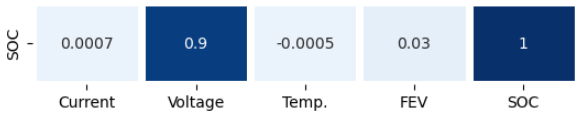

Fig.2 Correlation analysis of each variable to the SOC.

This technique transformed the input into four features in the form X = [I, V, T, FEV], as shown in Eq. (2).

$$\left\{ \mathbf{FEV} = \begin{bmatrix} I_1 \\ I_2 \\ I_3 \\ \vdots \\ I_n \end{bmatrix} \div \begin{bmatrix} V_1 \\ V_2 \\ V_3 \\ \vdots \\ V_n \end{bmatrix} = \begin{bmatrix} FEV_1 \\ FEV_2 \\ FEV_3 \\ \vdots \\ FEV_n \end{bmatrix} \right\}; \mathbf{X} = \begin{bmatrix} I_1 & V_1 & T_1 & \mathbf{FEV_1} \\ I_2 & V_2 & T_2 & \mathbf{FEV_2} \\ I_3 & V_3 & T_3 & \mathbf{FEV_3} \\ \vdots & \vdots & \vdots & \vdots \\ I_n & V_n & T_n & \mathbf{FEV_n} \end{bmatrix} \quad (2)$$

A correlation analysis was conducted to highlight the importance of the feature-engineered variable relative to other variables such as current, voltage, and temperature. The results, visualized using a heatmap in Fig. 2, indicate that only the voltage variable demonstrates a stronger correlation with SOC than the feature-engineered variable. In contrast, the temperature and current variables exhibit lower correlation values with SOC than the feature-engineered variable.

3. XGBoost Models' Implementation and Results

The optimal values for the XGBoost hyperparameters were determined through a GridSearch approach. Setting the maximum tree depth to infinity and the number of estimators to 3000 allowed maximum flexibility in the model's complexity. These values were chosen for the training process due to the ability to strike the best balance between model performance and computational efficiency. This study evaluated SOC estimation from two perspectives: employing a conventional XGBoost model and an XGBoost model trained and tested with the data-enhancing technique. This comparative analysis was undertaken to determine the effectiveness of feature engineering in contrast to the regular datasets. The estimation performances were assessed using two metrics: the RMSE and the MAE.

In the conventional XGBoost model, the inputs consisted of current, voltage, and temperature. Conversely, the inputs for the data-enhanced XGBoost model included voltage, current, temperature, and the feature-engineered variable obtained through feature engineering. Both models showcased robust performance in estimating the SOC for the testing dataset. However, the data-enhanced model demonstrated superior performance compared to the conventional model. As depicted in Fig. 3, illustrating the estimation results, it is clear that the data-enhanced model surpassed the conventional model by more than double in accuracy and precision.

Across all four temperatures, the data-enhanced XGBoost model consistently exhibited lower estimation errors, RMSE, and MAE than conventional XGBoost. The maximum estimation error for the conventional XGBoost, covering the entire validation dataset across all four

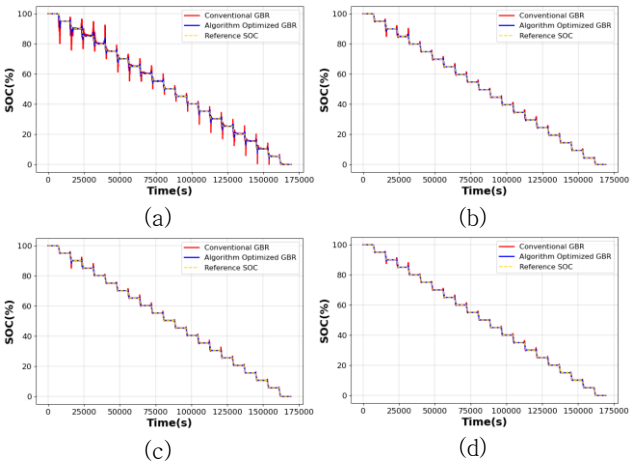

Fig.3 SOC estimation results of the testing dataset for each temperature: (a) -10℃; (b) 25℃; (c) 35℃; (d) 45℃.

temperatures stood at 9.9%, whereas the data-enhanced XGBoost notably decreased to 4.24%. Additionally, the average RMSE for the entire validation dataset was 0.28% for the conventional XGBoost and 0.10% for the data-enhanced XGBoost, with corresponding average MAE values of 0.04% and 0.02%.

4. Conclusion

A data-enhancing technique, feature engineering, was introduced in this study to address the issue of limited data or battery features by creating an entirely new battery variable. The effectiveness of this approach was validated using two XGBoost models. The models were developed, and the training process utilized a dataset comprising three driving profiles at six different temperatures to evaluate the algorithms' ability to generalize across various scenarios. The data-enhanced model resulted in exceptional SOC estimation outcomes, two times better than the conventional model. The average RMSE and MAE were notably low, at 0.103% and 0.015%, respectively.

This paper was supported by the Korea Institute of Industrial Technology Evaluation and Management (No. 20015572, battery pack thermal management for rapid charging and high power operation of electric vehicles) and SNS (Battery Management System Development and New Species Business R&D).

참 고 문 헌

[1] M. Korkmaz, " SoC estimation of lithium-ion batteries based on machine learning techniques: A filtered approach, " J. Energy Storage, vol. 72, no. PA, p. 108268, 2023, November.

[2] Center for Advanced Life Cycle Engineering (CALCE): Lithium-ion Battery Experimental Data. Available: https://www.calce.umd.edu/battery-data