Energy Proceedings
Vol 39, 2024

ISSN 2004-2965

The Influence of Data Distribution Diversity on Prediction Model for SOC of Lithium-ion Battery

Lin He[1,2], Bin Liu[1,2*], Jiangyan Liu[1,2], Kuining Li[1,2]

1 Key Laboratory of Low-grade Energy Utilization Technologies and Systems, Chongqing University, Ministry of Education, Chongqing 400044, China

2 School of Energy and Power Engineering, Chongqing University, Chongqing 400044, China
(*Bin Liu: liubin0921@cqu.edu.cn)

ABSTRACT

The working condition data of lithium-ion battery can vary significantly due to factors such as battery type, production processes, and usage conditions. These data differences pose a challenge to accurately predicting the state of charge (SOC) , leading to various scenarios where the model exhibits low training accuracy, high training accuracy and low prediction accuracy, and so on. To investigate the impact of data differences on the training results, it is crucial to study the influence of distribution diversity of large-scale data on the generalization of the prediction model of SOC. Therefore, 32 operational data sets of actual lithium batteries were studied in this paper. Considering the demand of advanced battery management technology, random forest (RF) was combined with MIMO strategy to predict multi-step SOC, and prediction models were established for 32 operational data sets respectively. The application effect of RF is studied and the effect of data set properties on multi-step prediction model of SOC is analyzed. The results indicate that, for large-scale lithium-ion battery data, excluding a small amount of data, the RF-MIMO model achieves an R^2 training accuracy of approximately 0.95 or higher for predicting future SOC with a time step of 180 intervals. The median R^2 accuracy of each model to predict other data sets remains about 0.9. When the dataset meets the requirements of a wide distribution range of SOC, a left-skewed tendency in the kernel density curve, and a relatively uniform distribution, the model training can obtain high precision.

Keywords: lithium-ion battery, large-scale data, state of charge, multistep prediction

NONMENCLATURE

Abbreviations	
SOC	State of Charge
RF	Random forest
MIMO	*Multi-input Multi-output*
Symbols	
Φ_k	the battery parameters at time k
k	Time
G_D	Gini index
h_k	the decision tree

1. INTRODUCTION

Lithium-ion batteries are widely used as energy storage devices for electric vehicles due to their high energy density and low self-discharge rate [1][2]. In order to ensure the safe and stable operation of lithium-ion batteries, an efficient and intelligent battery management system is particularly important, and the state of charge estimation of batteries is one of the key technologies. SOC is generally defined as the ratio of available capacity to reference capacity [3-5]. Accurate SOC estimation can prevent the battery from overcharging and discharging and extend the service life of the battery. SOC cannot be measured directly [6] and is often estimated indirectly based on voltage, current, temperature and other data. SOC estimation methods can be divided into traditional estimation methods, model-based estimation methods and data-driven estimation methods [7]. Among them, the data-driven method is to map the relationship between SOC and voltage, current, temperature, etc. directly into the data-driven model based on measurement data, which is

# This is a paper for 15th International Conference on Applied Energy (ICAE2023), Dec. 3-7, 2023, Doha, Qatar.

simple to establish and can be combined with cloud data in the future, and has potential in SOC estimation [8-10].

At present, most studies using data-driven method to predict the state of charge are conducted on individual working condition data of the same type or different types of lithium batteries. However, different types of lithium batteries will lead to great differences in battery working condition data due to differences in chemical composition, charge and discharge performance, cycle life, etc. Even for the same type of lithium batteries, the data of battery working condition will be different due to different production processes, aging conditions and user habits. The difference in data significantly increases the difficulty of accurate prediction of the state of charge, resulting in low training accuracy, high training accuracy and low prediction accuracy of the model and so on. In order to explore how data differences affect the training results of the model, etc., It is necessary to study the impact of the distribution diversity of large-scale data on the generalization of charge state prediction model.

Therefore, this paper studied the actual operation data of multiple lithium batteries and established models respectively for SOC prediction. Considering the demand of advanced battery management technology, random forest algorithm was combined with multi-input, multi-output and multi-step prediction strategy to predict SOC in the future period and the application effect is studied. On this basis, the influence of data set distribution on the multi-step SOC prediction model is analyzed, which provides a reference for how to select data sets for SOC prediction in actual working conditions in the future.

2. THEORY AND METHOD

2.1 Research approach

The main framework of the study consists of four parts. First, 32 actual data sets of lithium battery packs were collected under different temperatures, different driving speeds and different usage habits and so on. We selected the total voltage, total current, SOC and temperature of the battery pack as features, pre-processed the data, and then input it into the multi-step prediction model of SOC with optimal parameters, respectively trained each data and predicted other data sets besides itself. Finally, the training and prediction results of each method are counted, and the influence of distribution diversity on model generalization is analyzed from the perspective of feature distribution. The analysis results are helpful to evaluate whether the lithium battery data set is conducive to training the prediction model of SOC.

2.2 multi-step prediction method of SOC

The SOC at time k is a function of the battery parameters, which can be expressed by formula (1), where Φ_k represents the battery parameters at time k, and k=1,2,... t_E. t_E stands for last moment. In order to perform SOC estimation of time series, obviously, it is necessary to determine the range of input time t_w, that is, the input step length. SOC can be expressed as formula (2), where $k \geqslant t_w > 0$. If multi-time estimation is required and the output step length is more than one moment, SOC can be expressed by formula (3). Multi-input Multi-output (MIMO) strategy is to establish a multi-output model to predict the multi-step SOC value at one time, which not only takes advantage of the correlation of the input multi-time battery parameters, but also considers the correlation of the output multi-time SOC, effectively reducing the accumulation of errors. The general schematic diagram is shown in Figure 1.

$$SOC_k=f(\varphi_k,\varphi_{k-1},\ldots\varphi 1) \tag{1}$$

$$SOC_k=f(\varphi_k,\varphi_{k-1},\ldots\varphi_{k-tw+1}) \tag{2}$$

$$(SOC_k,SOC_{k+1}\ldots SOC_{k+h})=f(\varphi_k,\varphi_{k-1},\ldots\varphi_{k-tw+1}) \tag{3}$$

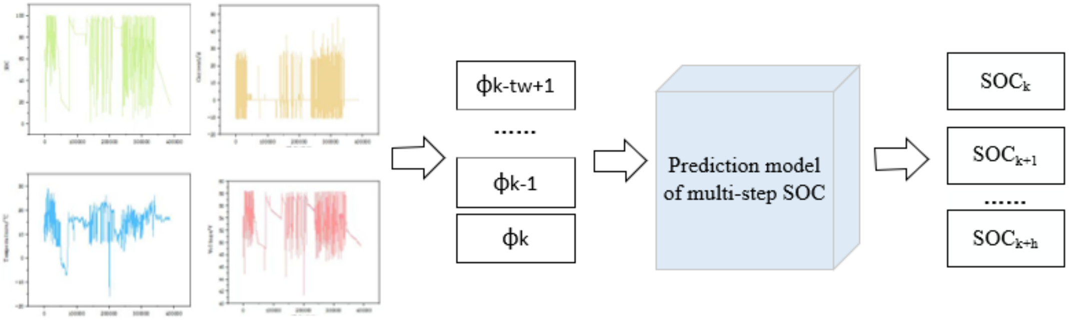

Figure 1 Schematic diagram of the method

2.3 Random forest algorithm

Random forest is an integrated learning algorithm based on decision tree, which adds the idea of bagging and random subspace to solve the problem of low accuracy and easy overfitting of decision tree model [11-13]. Random forest adopts bootstrap technology to form a self-service sample set from random samples returned from the data, and builds decision trees according to CART algorithm [14]. Each tree has root nodes, intermediate nodes and leaf nodes, as shown in the figure. The attribute selection measure of CART algorithm is Gini index. Assuming data set D contains m categories, the calculation formula of Gini index G_D is as follows:

$$G_D=1-\sum_{j=1}^{m} p_j^2 \tag{4}$$

Where, p_j is the frequency of occurrence of class j elements.

For each attribute, consider every possible binary partition, select the subset of the smallest Gini index generated by the attribute as its split subset, and under this rule, split continuously from top to bottom until the decision tree is generated, and finally take the average of the results of each tree as the predicted value, that is

$$y^{'}=\frac{1}{K}\sum_{k=1}^{K} h_k(x_i) \tag{5}$$

Where h_k represents the decision tree and K is the number of trees.

3. PREDICTION PROCESS

3.1 Data sources and preprocessing

The data in this paper come from the historical data of the actual operation of a shared electric bicycle. The electric bicycle battery pack consists of 14 battery cells in series, and its basic parameters are shown in Table 1. A total of 32 operating condition data sets are collected, which contain sensing information such as battery temperature, total voltage of battery pack, voltage of battery unit, battery capacity, actual SOC, etc. The operating state of the battery reflects charge and discharge through positive and negative current.

Table 1 Main parameters of lithium-ion batteries

Main parameters of lithium-ion battery	
Number of cells	14
Connection mode	series
Nominal voltage(V)	58
Nominal capacity(Ah)	32
Charging mode	CC-CV
Cooling method	passive air cooling
Heating method	none

Each dataset is sampled at a interval of 10 seconds, denoted as a time step. In this paper, battery temperature, total voltage of battery pack, current and actual SOC are selected as characteristics for analysis.

According to statistics, the total number of missing values in each data set accounts for a small proportion of the total data set, so a simple linear interpolation method is used to fill the missing values in the data set, and the outliers are filtered by the quartile range rule. Data standardization[15] is mainly to conduct standardized data processing. In this paper, Min-Max method is used for data standardization:

$$x_{scale}=\frac{x-x_{min}}{x_{max}-x_{min}} \tag{6}$$

Where x is the initial data, x_{max} and x_{min} respectively represent maximum the and minimum values in the data, and x_{scale} is the result of data standardization.

3.2 Model training and optimization

The pre-processed battery data set was divided into a training set and a test set according to the ratio of 8:2, and the multi-step prediction model was trained using the training set. In order to reduce the accumulation of errors, a small amount of battery information is used to predict the SOC for a long period of time, and the input step is set to 10 and the output step is set to 180. On this basis, multiple data sets of actual working conditions are used to optimize the model hyperparameters by grid search method.

3.3 Model Evaluation

The evaluation indexes used are Mean Absolute Percentage Error (MAPE) and goodness of fit R^2. The mean absolute percentage error can be expressed as:

$$\text{MAPE}=\frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y^{'}_i-y_i}{y_i}\right| \tag{7}$$

Where $y_i^{'}$represents the i th predicted value, y_i represents the i th true value, and n represents the number of samples.

Goodness of fit represents the fitting effect between the predicted value and the true value of the model, namely:

$$R^2=1-\frac{\sum_{i=0}^{n}(y_i-y^{'}_i)^2}{\sum_{i=0}^{n}(y_i-\bar{y})^2} \tag{8}$$

Generally speaking, the smaller the MAPE, the smaller the model prediction error and the higher the accuracy; The closer R² is to 1, the better the model fits and the higher the accuracy.

4. RESULTS

4.1 Training performance of the model

The training results of 32 data sets of RF model combined with MIMO strategy are shown in Figure 2. The R^2 of many data sets is above 0.95, indicating that this method can accurately predict multi-step SOC.

The accuracy of some data sets in the model is also very low, which may be affected by the data distribution. In terms of MAPE index, the RF-MIMO model's errors are basically below 0.05.

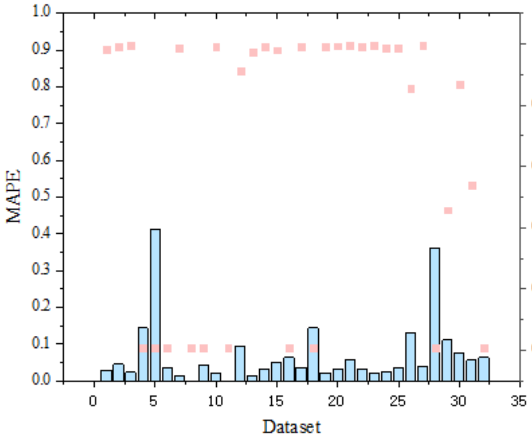

Figure 2 Training results of the model

4.2 Predictive performance of the model

In order to comprehensively observe the SOC multi-step prediction effect of each data set and facilitate analysis, all R^2 results of training and prediction of each data set were drawn into heat maps, as shown in Figure 3. According to the figure, the results of RF-MIMO model are basically above 0.8. In general, accuracy results above 0.9 account for 1/2 of all prediction results. For some datasets, R^2 is even below 0.3, indicating that the current model has poor generalization in these data sets.

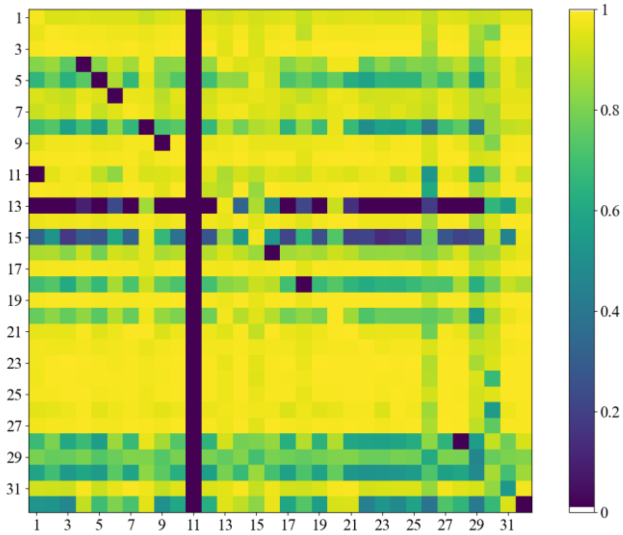

Figure 3 Prediction results of the model

4.3 Influence of data distribution properties on model

According to the results in Figure 3, data sets with high and low training accuracy are found for research according to the indexes with R^2 greater than 0.90 and R^2 less than 0.55.

Considering the influence of SOC distribution on model training, kernel density curve was used to explore SOC distribution of common data sets with high and low training accuracy.

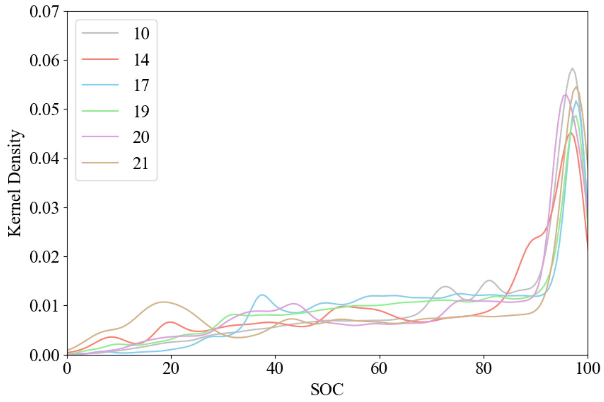

Figure 4(a) SOC distribution of common data sets with high training accuracy

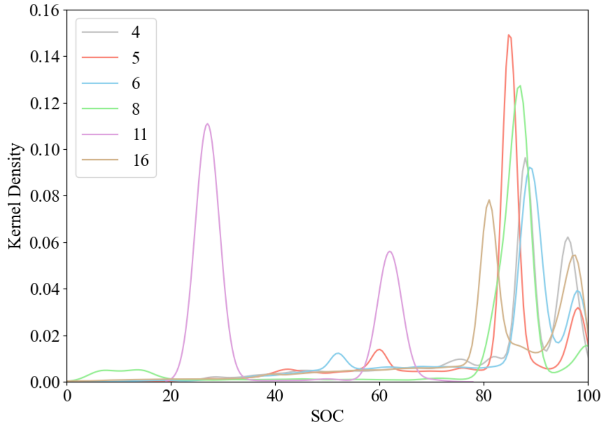

Figure 4(b) SOC distribution of common data sets with low training accuracy

Figure 4 shows the distribution of common data sets SOC with high and low training accuracy. The SOC core density curve of data sets with high training accuracy has a high density when SOC is 95-100, while the density of other intervals is about 0.01, showing a left-leaning trend and an overall uniform distribution. However, the SOC distribution range of data sets with low training accuracy is narrower, and the kernel density is very large in some locations, exceeding 0.1, and the overall is not uniform.

5. RESULTS

Studying the impact of the distribution diversity of large-scale data on model generalization is helpful to evaluate whether the lithium battery data set is conducive to training SOC prediction models. In this paper, the random forest and MIMO strategy are combined to predict multi-step SOC and we built models respectively for SOC prediction based on 32 actual data of lithium batteries.

The output step length is 180, that is, the prediction time is half an hour. The application effect of the algorithm is studied and the influence of data distribution on the multi-step SOC prediction model is explored. Specific conclusions are as follows:

(1) RF-MIMO model training accuracy R^2 is mostly above 0.95, MAPE is mostly below 0.05, so the model can accurately predict SOC;

(2) RF-MIMO model has excellent prediction performance when predicting data sets other than its own, and the median R^2 result is basically above 0.9;

(3) When the data set meets the following requirements, the model can be trained with high precision:

The SOC distribution range is wide, and the nuclear density curve tends to be left, and the overall distribution is uniform.

REFERENCE

[1] G.L. Plett, Extended Kalman filtering for battery management systems of LiPB based HEV battery packs, J. Power Sources 134 (2) (2004)252–261.

[2] C. Liu, W. Liu, L. Wang, G. Hu, L. Ma, B. Ren, A new method of modeling and state of charge estimation of the battery, J. Power Sources 320 (2016) 1–12.

[3] Li, J. Huang, B.Y. Liaw, J. Zhang, On state-of-charge determination for lithiumion batteries, J. Power Sources 348 (2017) 281–301.

[4] W. Waag, C. Fleischer, D.U. Sauer, Critical review of the methods for monitoring of lithium-ion batteries in electric and hybrid vehicles, J. Power Sources 258(2014)321–339.

[5] J. Kalawoun, K. Biletska, F. Suard, M. Montaru, From a novel classification of the battery state of charge estimators toward a conception of an ideal one, J. Power Sources 279 (2015) 694–706.

[6] CHENG K W E, DIVAKAR B P, WU H, et al. Battery-Management System (BMS) and SOC Development for Electrical Vehicles[J]. IEEE Transactions on Vehicular Technology, 2011,60(1):76-88.

[7] X. Hu, F. Feng, K. Liu, L. Zhang, J. Xie, B. Liu, State estimation for advanced battery management: key challenges and future trends, Renew. Sust. Energ. Rev. 114 (2019).

[8] Shu X, Li G, Zhang Y, Shen S, et al. Stage of charge estimation of lithium-ion battery packs based on improved cubature kalman filter with long short-term memory model. IEEE Trans Transport Electrif 2021;7(3):1271–84.

[9] Sun H, Sun J, Zhao K, Wang L, et al. Data-driven ICA-Bi-LSTM-Combined lithium battery SOH estimation. Math Probl Eng 2022;2022:9645892.

[10] Li D, Wang L, Duan C, Li Q, et al. Temperature prediction of lithium-ion batteries based on electrochemical impedance spectrum: a review. Int J Energy Res 2022;46.

[11] BREIMAN L, CUTLER A. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.

[12] BREIMAN L. Bagging predictors[J]. Machine Learning, 1996,24(2): 123-140.

[13] HO T K. The random subspace method for constructing decision forests[J]. IEEE Transactions on Pattern Analysisand Machine Intelligence, 1998, 20(8): 832-844.

[14] BREIMAN L，FREIDMANJ H，OLSHEN R A，etal.Classification and regression trees[M]. Chapman & Hau/CRC，1984.

[15] Dr. K. D S, Dr. C. S B. Data Analytics: Why Data Normalization[J]. International Journal of Engineering & Technology. 2018, 7(4.6).