Life prediction of battery based on random forest optimized by genetic algorithm

1st LI Cailian
School of Computer and Information Technology
Beijing Jiaotong University
Beijing , China
18125213@bjtu.edu.cn

2nd ZHANG chun
School of Computer and Information Technology
Beijing Jiaotong University
Beijing , China
chzhang1@bjtu.edu.cn

***Abstract*—When the random forest algorithm is used for battery life prediction, the prediction result is unstable, and it is difficult to ensure the accuracy of the model. In view of the above problems, this study proposes to use genetic algorithms to optimize the random forest prediction model. While ensuring the prediction accuracy, the depth and number of decision trees in the random forest are optimized, and the optimal combination of decision tree depth and number is used Life prediction. Using the lithium-ion battery data published by NASA to conduct simulation experiments and evaluate the prediction performance of the model, and then compare with the prediction results of the random forest prediction model and lasso prediction model.**

Keywords—random forest algorithm, lithium-ion battery, life prediction, genetic algorithm

I. INTRODUCTION

As a key component in the EMU, the performance of the battery plays an important role in ensuring the safety and reliability of the entire system. Battery failures can cause performance degradation or failure of power equipment, increasing costs. An effective battery management system will help to run the battery efficiently while extending its life. Therefore, the accurate prediction of the remaining life of the battery plays an increasingly important role in the estimation of the health status of the battery.

The health status of a battery is usually defined as SOH. The most typical definition of SOH is based on battery capacity. The changing trend of battery capacity can directly reflect the health status of the battery, that is, the battery life. When the battery capacity gradually decreases, it represents the degradation of battery performance. When the capacity reaches a given threshold (typically 70% to 80% of the factory capacity), it means that the battery has failed. This threshold is hereinafter referred to as the failure threshold. Therefore, if we can predict the battery capacity trend based on the battery's historical usage data, we can understand the battery life in advance, and then make adjustments and maintenance strategies in a timely manner to avoid accidents and reduce maintenance costs.

In recent years, there have been more and more methods for battery life prediction. The common prediction methods are summarized into three categories, namely statistical distribution-based, model-based and data-driven prediction methods. The method based on statistical distribution is to make predictions by fitting the distribution using historical failure data. This type of method is difficult to accurately predict and describe the life status of the battery. The model-driven prediction method uses the system's physical model and regression data model to make predictions. Such methods involve too many physical and chemical characteristics, and modeling is difficult and complicated. LYU C et al. Used the internal SOH parameters reflecting the aging of the battery as state variables, and used the fitting equations of the capacity decay simulation and known state variables as the measurement model and process model, respectively, and combined the electrochemical model and particle filter to establish a prediction model [1]; YANG F et al. Proposed a two-logarithmic empirical capacity degradation model based on Bayes' rule, using an improved particle filter algorithm to update the model parameters to predict the battery [2] ; ZHANG X et al. Improved unscented particle filter prediction method based on Markov chain Monte Carlo[3]. Data-driven prediction methods use pattern recognition and machine learning to detect changes in parameters to make predictions. This method does not consider internal chemical reactions and failure mechanisms, and directly uses highly relevant data for predictive analysis. Therefore, some people at home and abroad have used data-driven methods to predict battery life and perform algorithm optimization. Ding Yangzheng et al. Proposed an improved PSO-optimized ELM to predict the remaining life of lithium ion batteries[4]; Wu Haiyang et al. Proposed a genetic algorithm-based BP neural network battery life prediction metho[5]; YI WU et al. A method for predicting the remaining life of lithium-ion batteries based on neural networks and bat particle filters is proposed[6].

It should be noted that most of the indicators estimated in the existing literature for SOH use capacity and internal resistance that can directly reflect battery life, and ignore the indirect health indicators (HIS) estimated by SOH, such as voltage and current, in the battery It is easier to collect during use, but the attention is not high. Therefore, the battery data collected in this paper extracts indirect health indicators for life prediction.

II. DATA PREPROCESSIN

Considering that when estimating the remaining battery life, it will be affected by various factors, so the data can be collected together. However, because the data dimension is

978-1-7281-6286-7/20/$31.00 ©2020 IEEE

relatively large, there may be factors with low correlation with battery life, and the collected data needs to be subjected to dimensionality reduction processing.

A. Data Set

We use the NASA Ames Center lithium-ion battery data set. The battery data set is from NASA PCoE. This article uses the data of four lithium-ion batteries B5, B6, B7 and B18. In this data set, the battery accelerates charging, discharging, and electrochemical impedance measurement in three different operating modes. The experiment is performed at 25 degrees Celsius and the experimental results are analyzed. Record observations. Specific steps are as follows:

- Charging process: charge in the constant current (CC) mode of 1.5A until the battery voltage reaches 4.2V, then continue charging in the constant voltage (CV) mode until the charging current drops to 20mA.
- Discharging process: Perform at a constant current (CC) level of 2A until the battery voltage drops to 2.7V, 2.5V, 2.2V, and 2.5V, respectively.
- Impedance measurement: It is performed by an electrochemical impedance spectroscopy (EIS) frequency sweep from 0.1 Hz to 5 kHz.

Repeated charging and discharging leads to accelerated battery aging, and impedance measurements provide insight into internal battery parameters that change with the aging process. When the battery capacity drops to the failure threshold, the battery life ends and the experiment is terminated. Fig.1 below shows the battery capacity change trend during the entire life cycle of the battery.

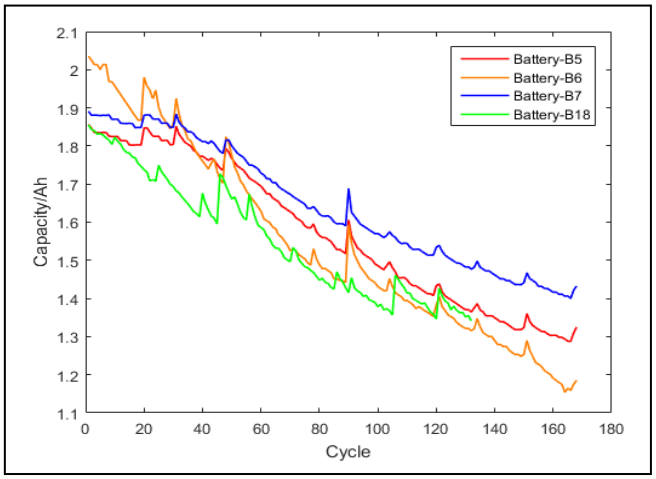

Fig. 1. Four battery capacity trends.

B. Feature Extraction

According to the introduction of the above data set, the following 7 characteristics are selected for prediction.

- Number of battery cycles C.The average current $\bar{I}_i$ at both ends of the battery at the i-th discharge.
- The average voltage $\overline{U}_i$ at both ends of the battery at the i-th discharge.
- The roughly calculated internal resistance R_i of the battery at the i-th discharge.

$$R_i = \frac{\overline{U}_i}{\bar{I}_i} \tag{1}$$

In the above formula, R_i represents the internal resistance of the battery estimated at the i-th time, while $\overline{U}_i$ and $\bar{I}_i$ represent the average voltage and average current at the i-th discharge, respectively.

- The average temperature $\overline{T}_i$ of the battery during the i-th discharge.
- The discharge depth D_i of the battery during the i-th discharge.
- The Isobaric discharge time V_i of the battery during the i-th discharge,calculate the time from when the battery starts to discharge to the lowest voltage at each discharge.

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \tag{2}$$

Among them, x_i and y_i represent the value of the i-th variable, $\bar{x}$ and $\bar{y}$ represent the average value of the i-th variable.

The Spearman correlation coefficient is used to calculate the correlation between the above 7 characteristics and the battery capacity Q^* measured during discharge, and the contents in TABLE I are obtained (using the B00005 battery as an example):

TABLE I. COMPARISON OF CHARACTERISTICS AND CAPACITY

	C	$\bar{I}_i$	$\overline{U}_i$	R_i	$\overline{T}_i$	D_i	V_i
Q^*	0.991	0.895	0.953	0.875	0.765	0.997	0.997
seq	3	5	4	6	7	1	2

C. Data Normalization

Due to the difference in the units of each feature, the values have different sizes. For the convenience of calculation,

the normalization principle is used to perform simple processing operations on the data to reduce the differences between the data and control it to [0,1] Within range.

D. Feature Combination

The normalized features are combined into the same matrix, each column represents the collected data of a feature, and each row represents the value of 7 features during one discharge.

C	I	U	R	T	D	V
0	0.55724	0.65753	0.42951	0.41677	1	1
0.005988	0.55304	0.73239	0.45139	0.46806	0.98007	0.98242
0.011976	0.54908	0.7965	0.47057	0.44043	0.95701	0.96358
0.017964	0.58262	0.7958	0.43056	0.39749	0.93465	0.96387
0.023952	0.58455	0.78258	0.42531	0.35303	0.93198	0.96189
0.02994	0.55028	0.7725	0.46372	0.37042	0.95772	0.96341
0.035928	0.5502	0.76941	0.46311	0.38593	0.95611	0.96261
0.041916	0.6517	0.90038	0.37297	0.36246	0.8688	0.94616
0.047904	0.68719	0.88842	0.32932	0.34089	0.84446	0.94462
0.053892	0.68578	0.87113	0.32712	0.31762	0.84519	0.9449
0.05988	0.72184	0.85201	0.28168	0.27747	0.82179	0.94495
0.065868	0.75717	0.9566	0.26463	0.3469	0.77809	0.92629
0.071856	0.79269	0.93984	0.22112	0.33151	0.75499	0.92609
0.077844	0.79242	0.91596	0.21622	0.27747	0.75468	0.92561
0.083832	0.78968	0.98645	0.23463	0.23955	0.73273	0.90692
0.08982	0.82797	0.97351	0.18923	0.20182	0.70947	0.90587
0.095808	0.86693	0.96242	0.14399	0.20519	0.68727	0.90631

Fig. 2. Normalized feature data table.

III. Random Forest Algorithm And Optimization

A. Random Forest Algorithm

Random forest is a highly flexible machine learning algorithm. It is an algorithm that integrates multiple trees through the idea of ensemble learning. Its basic unit is a decision tree, but its essence is a large branch of machine learning, that is, an ensemble learning method[7]. The "forest" in the random forest is the forest composed of the multiple decision trees we explained above, and the "random" is the part we want to optimize this time.

Random forest is more commonly used in classification problems. Each decision tree is considered as a classifier, and then a statistical analysis is performed on the classification results of multiple decision trees. The larger probability is the result of this classification. However, the prediction capability of random forests is also very strong. It predicts the results through each decision tree, then calculates the average of the prediction results of all decision trees, and then obtains the final prediction result.

B. Random Forest Algorithm based on Genetic Algorithm Optimization

When a random forest generates a decision tree, the depth of the decision tree (mtry) is the number of split attributes selected when the decision tree is generated. The larger mtry is, the more accurate the decision tree prediction result is. But the larger mtry will directly affect the calculation amount of the algorithm;The number of randomly combined decision trees (ntree) will also affect the accuracy of the model. Therefore, this paper uses genetic algorithms to optimize random forests, determine the optimal number of decision trees and the optimal depth of decision trees, and achieve the goal of efficient prediction.

1) Parameter optimization algorithms

TABLE II. Comparison of parameter optimization algorithms

Algorithm	***Concept***	***Pros***	***Cons***
Grid search	Each variable is a row or a column to form a grid for exhaustive search	Easy to implement	Sampling at equal intervals is required, which is easy to cause the value to be locally optimal; large data volume is easy to cause a combination explosion
Random search	Generate random points in the interval	Random sampling is adopted and the calculation amount is small	Unstable optimization
Genetic algorithm	Based on genetics to simulate biological evolution	Based on bionics, you can quickly and randomly search	Difficult to optimize when the dimension is high

According to the above comparison, it is concluded that for the range of mtry and ntree, the calculation amount will be larger when selecting the grid search, which affects the optimization speed. The optimization effect of random search is extremely unstable, so the genetic algorithm is selected for parameter optimization.

2) Genetic Algorithm

Genetic Algorithm (GA) is a computational model based on biological evolution that searches for the optimal solution by simulating the evolutionary process in nature[8]. The main steps of GA are as follows:

a) Initialize the population: Choose a coding scheme (usually binary coding) to encode the randomly generated matrix in the solution space to form the initial population of GA.

b) Calculating fitness function: Used to convert the objective function to the corresponding fitness value.

c) Select operation: According to the fitness of the individual, the individual with higher fitness is selected from the current population through the method of roulette.

d) Cross operation: The individual selected in c) is controlled by the probability threshold Pc whether to use single-point crossover or multi-point crossover to generate new individuals.

e) Mutation operation.: Use the probability threshold Pm to control whether the individuals in d) are subjected to

single-point mutation or multi-point mutation operations on part of the genes of the individual.

f) Iterate in turn: Update the population, find the optimal solution in the new population obtained after e), and then continue to repeat the operation of b) until the number of iterations is satisfied and output the optimal solution.

3) Optimization of Random Forest Algorithm

Objective function:

$$\begin{aligned} \min \quad & \mathrm{reg}(\mathrm{ntree}, \mathrm{mtry}) \\ \text{s.t.} \quad & 1 \le \mathrm{ntree} \le 500 \\ & 1 \le \mathrm{mtry} \le 7 \end{aligned} \tag{3}$$

Which represents the number of randomly selected decision trees in a random forest, and the depth of each decision tree, which is the number of attributes of the split decision tree.

$\mathrm{reg}(\mathrm{ntree}, \mathrm{mtry})$ represents the root mean square error between the predicted value Y_hat and the true test value Y_tst when using ntree decision trees and the depth of each decision tree is mtry ,The formula is:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (Y\_tst - Y\_hat)^2}{n}} \tag{4}$$

Because when the genetic algorithm is used to optimize the solution space, the model calculates the maximum value of the function, a preliminary conversion of the objective function is needed to obtain the fitness function:

$$\begin{aligned} \max \quad & -\mathrm{reg}(\mathrm{ntree}, \mathrm{mtry}) + 1 \\ \text{s.t.} \quad & 1 \le \mathrm{ntree} \le 500 \\ & 1 \le \mathrm{mtry} \le 7 \end{aligned} \tag{5}$$

The GA-RF algorithm steps are as follows:

a) Initialize the population: Select a coding scheme (usually binary coding) to encode the randomly generated matrix in the solution space to form the initial population of GA.

b) Calculate fitness function -reg(ntree,mtry)+1: Calculate the fitness value corresponding to each group (ntree.mtry) in the initial population and save it to the matrix..

c) Select operation: According to the individual's fitness, the individual with higher fitness is selected from the current population by roulette.

d) Cross operation: Use the single-point cross to generate new individuals using the individuals selected in c). The probability threshold pc = 0.6.

e) Mutation operation: Perform a single-point mutation operation on part of the genes of the individuals crossed in Step 4 with a probability threshold of pm = 0.005.

f) Iterate in turn: Update the population, find the optimal solution in the new population obtained after e), and then continue to repeat the operation of b) until the number of iterations reached 100 and output the optimal solution.

IV. EXPERIMENT

In this paper, two experiments are conducted to predict the remaining battery life:

A. Experiment One

For the data of two batteries (B5, B6), their lifespans are predicted respectively. 60% of the data of each battery is selected as the training data set, and 30% is used as the verification data set. The training data set adopts a random combination method. After the data set is divided, the lasso algorithm, the random forest algorithm and the GERF algorithm are used to establish the prediction model, and the remaining life of the battery is predicted by the battery current, voltage, depth of discharge, and constant voltage drop discharge time. Finally, the predicted error rate is calculated to evaluate the model.

- B0005 battery

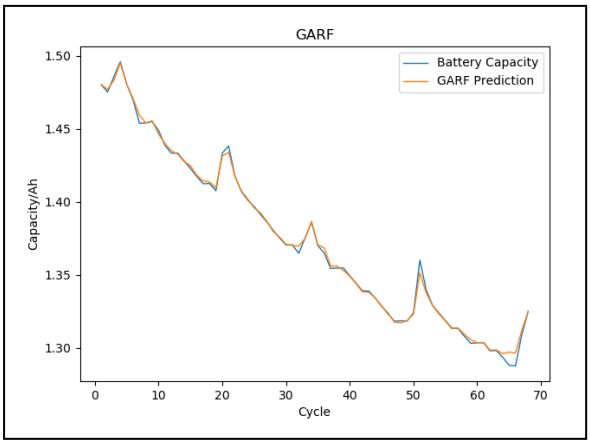

Fig. 3. Graph of GA-RF algorithm prediction (ntree=413,mtry=5)

- B0006 battery

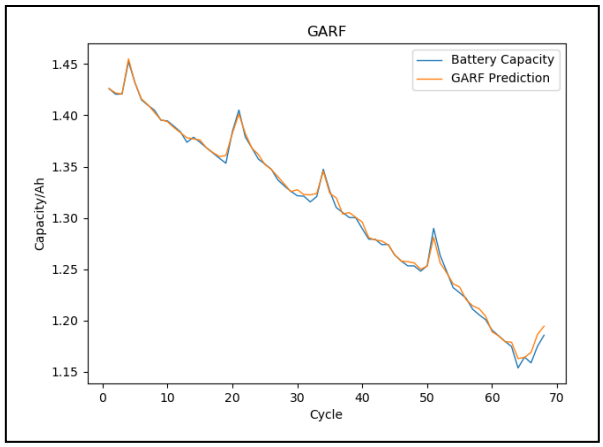

Fig. 4. Graph of GA-RF algorithm prediction (ntree=459,mtry=5)

TABLE III. RMSE OF MODEL PREDICTION OF EXPERIMENT ONE

Type of Battery	*RMES(RF)*	*RMES(GA-RF)*	*RMSE(Lasso)*
B0005	0.0036	0.0024	0.0022
B0006	0.0055	0.0040	0.0038

B. *Experiment Two*

For the data of four batteries (B5, B6, B7, B8), three battery data were selected as the training set, and the other battery data was used as the test set. The three algorithms of the same experiment one were used to build a battery life prediction model. The current, voltage, depth of discharge, and constant voltage drop discharge time of the battery predict the remaining life of the battery. Finally, the predicted error rate is calculated to evaluate the model.

- B0005 battery

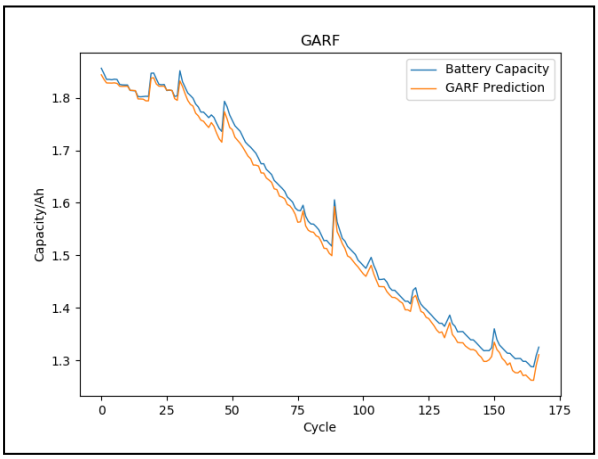

Fig. 5. Graph of GA-RF algorithm prediction (ntree=379,mtry=2)

- B0006 battery

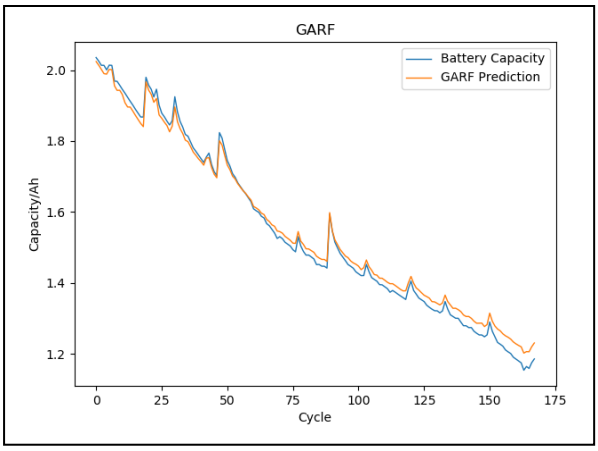

Fig. 6. Graph of GA-RF algorithm prediction (ntree=163,mtry=4)

- B0007 battery

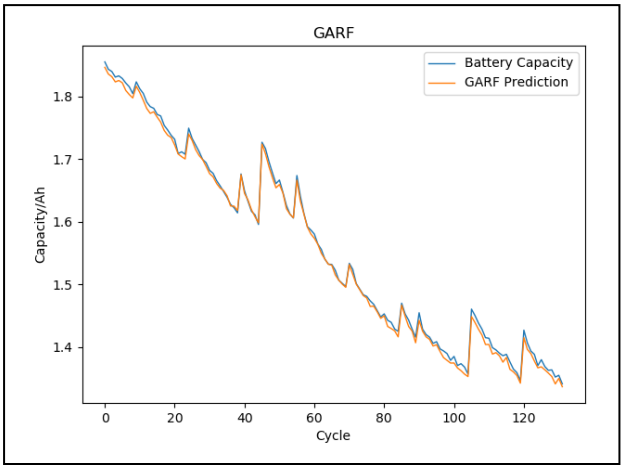

Fig. 7. Graph of GA-RF algorithm prediction (ntree=264,mtry=3)

- B0018 battery

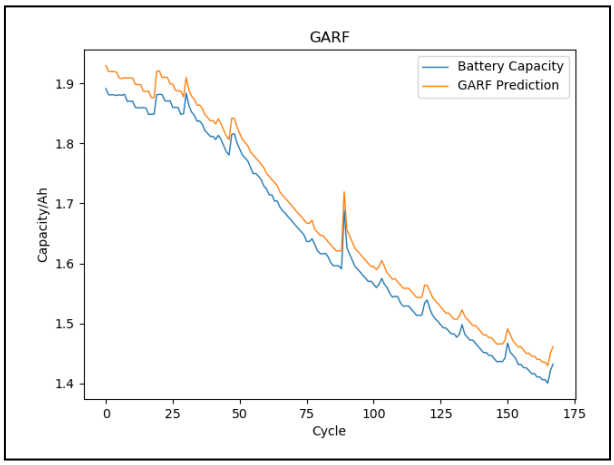

Fig. 8. Graph of GA-RF algorithm prediction (ntree=323,mtry=5)

TABLE IV. RMSE OF MODEL PREDICTION OF EXPERIMENT TWO

Test set	*RMES(RF)*	*RMES(GA-RF)*	*RMSE(Lasso)*
B0005	0.029	0.016	0.018
B0006	0.048	0.021	0.025
B0006	0.052	0.029	0.033
B0006	0.0093	0.0068	0.0086

According to the error rate analysis of the above two experiments, the prediction accuracy of the Lasso prediction model for the same battery is high, and the prediction accuracy of the GARF model for different batteries is high. Combining the error analysis of the two experiments, the overall accuracy of the GARF model is higher than that of the other two models.

V. SUMMARY

In this study, genetic algorithms were used to optimize the random forest prediction model to improve the accuracy of the model prediction. The experiment compares the predictions of Lasso model, random forest model and optimization model. The rms errors of the three models in battery prediction are calculated. The accuracy of GARF optimization model is high. However, the prediction speed after model optimization is slower than before optimization, which requires further experimental optimization in subsequent model improvements.

REFERENCES

[1] LYU C,LAI Q,GE T,et al. A lead-acid battery's remaning useful life prediction by using electrochemical model in the particle filtering framework[J]. Energy,2017,120:975-984.

[2] YANG F,WANG D,XING Y,et al. Prognostics of Li(NiMnCo) O2-based lithium-ion batteries using a novel battery degradation model[J]. Microelectronics Reliability,2017,70:70-78.

[3] ZHANG X,MIAO Q,LIU Z. Remaining useful life prediction of lithium-ion battery using an improved UPF method based on MCMC[J]. Microelectronics Reliability,2017.

[4] Ding Yangzheng,Jia Jianfang.Improved PSO optimized extreme learning machine predicts remaining useful life of lithium-ion battery[J].JOURNAL OF ELECTRONIC MEASUREMENT AND INSTRUMENTATION,2019,33(02):72-79.

[5] 5.WU Haiyang,MIAO Weiwei,GUO Bo,LV Shunli,WU Hao,TENG Xinyuan.Research on Battery Life Prediction of BP Neural Network Based on Genetic Algorithm[J].Computer & Digital Engineering,2019,47(05):1275-1278.

[6] Zhang, Yongzhi & Xiong, Rui & Hongwen, he & Pecht, Michael. (2018). Long short-term memory recurrent neural network for remaining useful life prediction of lithium-ion batteries. IEEE Transactions on Vehicular Technology. PP. 1-1. 10.1109/TVT.2018.2805189.

[7] WEI Zhenhan,SONG Shuxiang,XIA Haiying.State-of-charge Estimation Using Random Forest for Lithium Ion Battery[J].Journal of Guangxi Normal University(Natural Science Edition),2018,36(04):27-33.

[8] XI Weijie,LI Donghui.BP Neural Network based Genetic Algorithm Optimization for Prediction of OCS Wear[J].Electric Railway,2019,30(S1):47-49