The Lithium-ion Battery State-of-Charge Estimation using Random Forest Regression

Chuanjiang Li, Zewang Chen, Jiang Cui,Youren Wang, Feng Zou
College of Automation Engineering
Nanjing University of Aeronautics and Astronautics
Nanjing, China

***Abstract*—The battery state-of-charge (SOC) is a very important part of battery prognostics and health management (PHM). For the problem that the SOC of lithium-ion battery cannot be measured directly, a kind of battery SOC estimation method using random forest regression is proposed in this paper. Firstly, a training set was constructed which used the battery current, battery voltage, battery temperature and other correlation factors as the model's training input and the corresponding battery SOC as the model's training output. Then, the model was trained with random forest algorithm. Finally, the trained model was applied to the battery SOC estimation. In this paper, this method is applied separately to the battery steady discharge process and dynamic discharge process for the estimation of battery SOC. Experimental results show that, the proposed method can effectively estimate the battery SOC and has a higher estimation precision than BP neural network estimation method.**

Keywords-lithium-ion battery; random forest regression; state-of-charge (SOC) estimation

I. INTRODUCTION

Compared with the rechargeable Nickel-cadmium battery and Ni-MH battery, lithium-ion battery has the advantages of high density, long service life, high working voltage and no pollution etc. With these advantages, lithium-ion battery has been widely used in various portable information processing terminals, electric vehicles and other fields. With the wide application of lithium-ion battery, the battery SOC estimation has become a more and more important issue, because accurate SOC estimation can display the battery's usable charge to the user before recharging and prevent the battery from being over or under-charged to extend the lifetime. However, the SOC cannot be measured directly, it must be estimated by measuring such parameters as the battery voltage, current, temperature etc.

There are several methods to estimate the SOC[1~5].Ah counting method is simple and easy to use, but it has the initial error and accumulated error problem. The open circuit voltage method can accurately estimate the battery SOC, but open circuit voltage can only be measured a long time after the battery stops working. So it is not suitable for online estimation. The artificial neural network is a kind of black box method based on the historical data of battery. It builds the correlation between the SOC affecting factors and SOC to estimate the battery SOC, but there are many affecting factors, such as the battery voltage, current, temperature, battery chemistry and battery history etc. If too many factors are chosen, it will lead to huge computation. If too few factors are chosen, the artificial neural network cannot correctly reflect the SOC. Currently, the principal component analysis(PCA) is usually used to analyse the importance of the battery affecting factors in order to select appropriate affecting factors to train the model ,but this process requires a lot of time. Moreover, the artificial neural network method also has some problems. It is sensitive to the affecting factor dimension difference which lead to the input variables should be normalized. It also has the over-fitting and easily fall into local minimum value problem.

To solve the problems of the artificial neural network method, this paper presents a method to estimate the SOC using random forest regression. Random forest has the ability to analyse how important the affecting factors are. The phenomenon of over-fitting rarely happens when the method is used and can explain hundreds of affecting factors without increasing computation compared with artificial neural network method.

This paper is consisted of five parts as following. The first part is the introduction. The second part gives a brief introduction about random forest regression. The third part presents the experimental source of the data of the steady discharge process and dynamic discharge process. The fourth part displays the simulation result and compares it with that of BP neural network. The fifth part provides the conclusion of this paper.

II. RANDOM FOREST

A. Random Forest Regression

Random forest regression was proposed by Leo Breiman[7~8]. Random forest for regression are formed by regression trees $\{h(X,\theta_k), k=1,...,K\}$,where X is the observed input vector, X and θ_k are independent and identically distributed random vectors. The output prediction values of random forest regression are numerical. The output true values are Y. The training data is assumed to be independently drawn from the joint distribution of (X,Y) ,the random forest prediction is unweighted average of all regression tree prediction:

$$\bar{h}(X) = (1/K)\sum_1^K h(X,\theta_k) \quad (1)$$

This work was supported by the Fundamental Research Funds for the Central Universities (grant no. NS2014028).

978-1-4799-7958-5/14/$31.00 ©2014 IEEE

The mean-squared generalization error for any numerical predictor $h(X)$:

$$E_{X,Y}(Y-h(X))^2 \tag{2}$$

Random forest has the following features:

1) As $K \to \infty$,the law of Large Numbers ensures:

$$E_{X,Y}(Y-\bar{h}(x))^2 \to E_{X,Y}(Y-E_\theta h(X,\theta)) \tag{3}$$

The quantity on the right is the generalization error of random forest, designated $PE^*(forest)$.The convergence in (3) implies that random forest has not over-fitting.

2) Assume that for all $\theta, E(Y)=E_X h(X,\theta)$,then

$$PE^*(forest) \le \bar{\rho} E_\theta E_{X,Y}(Y-h(X,\theta))^2 \tag{4}$$

$E_\theta E_{X,Y}(Y-h(X,\theta))^2$ in (4) is the average generalization error of a tree, designated $PE^*(tree)$, $\bar{\rho}$ is the weighted correlation between residuals $Y-h(X,\theta)$ and $Y-h(X,\theta')$ where θ and θ' are independent. It can be seen from (4) that what is required to ensure the prediction accuracy of random forest: low correlation between residuals of different tree members of random forest and low generalization error for the individual trees. Low correlation and generalization error are obtained by adjusting $\bar{\rho}$.

In order to ensure that the random forest to reach the standard, it can be achieved by using the following strategies:

1) N samples are randomly selected from the whole dataset using bootstrap method and N regression trees are constructed, the out-of-bag(OOB) data that are not selected are used as testing set.

2) If the number of initial variables is M, at each tree node, m variables are chosen(m<M) and the best split is selected based on these variable to split the node. The m usually is M/3.

3) Each tree is fully grown without pruning.

4) The predicted result of the random forest can be obtained by averaging the predicted results of all regression trees.

B. *Variable Importance Evaluation*

Variable importance evaluation for individual feature is based on a kind of idea which is that when a relative feature (the feature may play an important role in prediction accuracy)is changed, the prediction accuracy of random forest will significantly reduce. It can be realized by the following strategies:

1) Each regression tree which has generated is tested using OOB data for a OOB accuracy.

2) A feature V in OOB data is randomly changed with noise, then this regression tree is tested by OOB data with noise. A new OOB accuracy is obtained.

3) The testing error between the original OOB data and OOB data with noise can be seen as the importance evaluation of the feature V.

4) The sequencing of the variable importance evaluation of all the features is based on initial OOB accuracy and OOB accuracy with noise.

C. *Software*

The "randomForest" toolbox in "R" statistic analysis software is chosen to construct the random forest model. The "R" is widely used and free to download for the public(http://www.r-project.org/). There are two main parameters, ntree and mtree. The ntree represents the number of trees in random forest. The value of ntree should not be too small for ensuring the convergence of the model. The parameter, mtree, governs how many variables are drawn.

III. THE EXPERIMENT DATA SET

The SOC estimator based on random forest regression is applied separately to the battery steady discharge process and dynamic discharge process. The experiment data of steady discharge process are collected from a group of lithium-ion battery which was tested at Idaho National Laboratory for free download in the NASA Ames excellence database prediction center[9]. A set of four lithium-ion batteries(#5,#6,#7,#18) were tested through three different operation(charge, discharge and impedance) at room temperature. In this paper, the discharge data of the battery #5 is used as the source of experiment data.

The charge process of the battery #5 was carried out in a constant current(CC) mode at 1.5A until the battery voltage reached 4.2V and then continued in a constant voltage(CV) mode until the charge current dropped to 20mA. After a rest time, discharge was carried out at a constant current(CC) level of 2A until the battery voltage fell to 2.7V. During the rest time, impedance measurement was carried out. The experiments were stopped when the batteries reached end-of-life criteria, which was a 30% fade in rated capacity(from 2Ahr to 1.4Ahr). The discharge data of the battery #5 contains the battery voltage , current , temperature and other parameters. As the entire discharge process is recorded, the SOC can be calculated by (5).

$$SOC = SOC_0 + \frac{1}{C_R}\int_{t_0}^{t} I_{cm} d_\tau, C_R = \int_0^{t_{EDV}} I_{cm} d_t \tag{5}$$

In (5), SOC_0 is the initial state of charge, C_R is rated capacity, I_{cm} is discharge current, t_{EDV} is discharge time when the discharge voltage reach the cut-off voltage.

The data collected from the first four discharge process are used to train the model, the data collected from the seventh discharge process are used to test the model accuracy. The discharge data (voltage , current , temperature and SOC) collected from the first discharge process are shown in figure 1.

The experiment data of the dynamic discharge process in the estimation process are collected from the electric vehicle analysis software ADVISOR[10],which contains the battery voltage , current, temperature and SOC, as shown in Figure 2. The data are obtained from a hybrid drive cycle, which covers the main operation modes of the vehicle. This cycle is comprised of UDDS and US06. They are widely used to test

the performance of electric vehicle. The UDDS cycle simulates the city road condition and the US06 cycle simulates the high speed and acceleration conditions. The data is collected every second and the total of the data is 1971. The model is trained by the data which are randomly drawn 60% from experiment data set and other 40% data are used as test data to test the model.

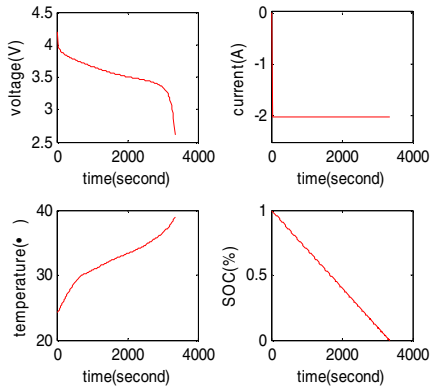

Figure 1 The steady discharge data collected from the first discharge process

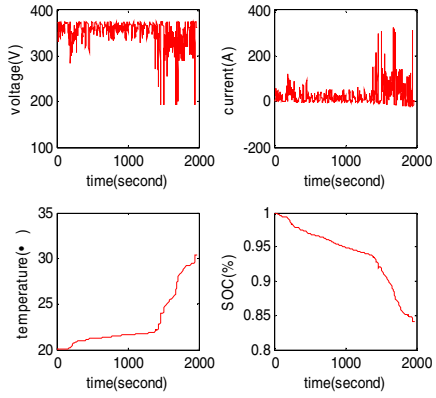

Figure 2 The dynamic discharge data

IV. Experimental Results and Analysis

The root mean square error(RMSE) is used to evaluate the accuracy of the random forest regression estimator. The RMSE is defined as follows:

$$\delta_{RMSE} = \sqrt{\sum_{i=1}^{n} \left| f(x_i) - y_i \right|^2 / n} \qquad (6)$$

In (6), $f(x_i)$ is the SOC estimation based on the random forest model, y_i is the true SOC value, n is the number of the estimation points.

In the steady discharge process, the battery voltage(V2),current(V3),temperature(V4),discharge time(V5), voltage change(V6) and temperature change(V7) are used as the model input, the corresponding SOC(V1) is used as the model output. This change is mainly compared with the previous sampling point. The ntree is 500(the test result shows that the model is able to ensure the convergence),the mtree is 2. To evaluate the performance of the random regression algorithm in battery SOC estimation process, the BP neural network method is simultaneously adopted in the estimation process as a baseline. The BP neural network takes a three-layer structure: six(steady state) or seven(dynamic) nodes in the input layer, six in the hidden layer and one in the output layer. The estimation results as shown in Figure 3~Figure 4.

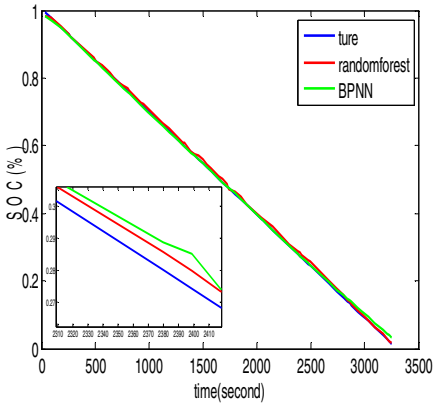

Figure 3 The steady state experimental result

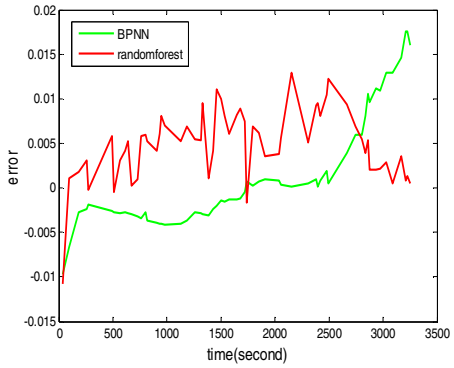

Figure 4 The steady state estimation error

Figure 3 and Figure 4 indicate that both random forest and BP neural network have high estimation accuracy, but the estimation error of BP neural network become more and more bigger, on the contrary, the random forest can ensure the error within a certain range. The maximum error estimation of random forest and neural network are respectively 0.013 and 0.018.

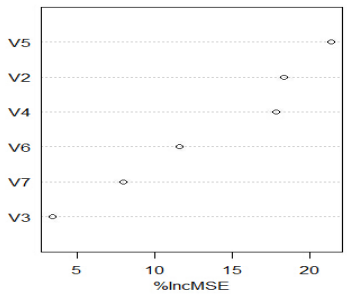

Figure 5 The variable importance analysis of steady state

Figure 5 is the result of variable importance analysis based on random forest. It indicates that the battery SOC mainly depends on the discharge time, voltage and temperature in this kind of steady discharge mode. So, when the model is trained next time, we can only consider the effect of the three factors in order to simplify the model input and reduce computation quantity.

In the dynamic discharge process, the battery voltage(V2),current(V3),temperature(V4),discharge time(V5), voltage change(V6) , current change(V7) and temperature change(V8) are used as the model input, the corresponding SOC(V1) is used as the model output. The ntree is 500(the test result shows that the model is able to ensure the convergence),the mtree is 2.

Figure 6 and Figure 7 indicate that the estimation result of random forest is more close to the true value, but its early and late estimation accuracy is not good. The maximum error estimation of random forest and neural network are respectively 0.0115 and 0.0175.

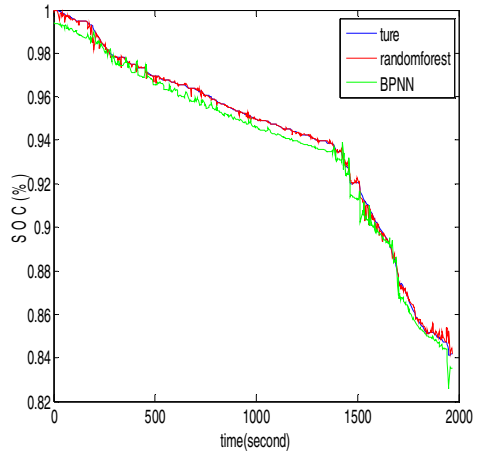

Figure 6 The dynamic state experimental result

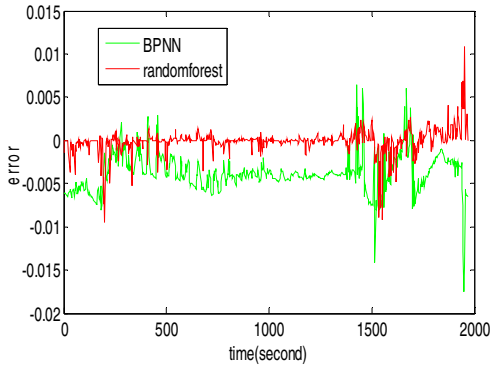

Figure 7 The dynamic state estimation error

Figure 8 indicates that the battery temperature and discharge time are the main influence factors in this kind of the dynamic discharge mode, which provides a reference for future model simplification.

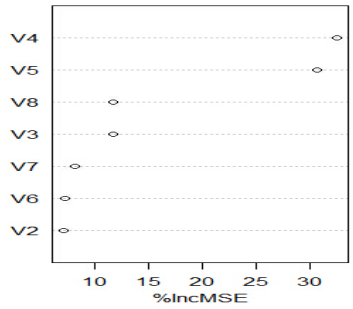

Figure 8 The variable importance analysis of dynamic state

Table 1 is the RMSE of estimation results in the two discharge modes. By comparing with BP neural network , we find that the random forest has a higher accuracy whether dynamic discharge process or steady discharge process.

TABLE I. THE RMSE COMPARISON

RMSE	*Random forest*		*BP neural network*	
δ_{RMSE}	*The steady state*	*The dynamic state*	*The steady state*	*The dynamic state*
	0.0059	0.0014	0.0063	0.0044

V. CONCLUSION

The SOC estimator based on random forest regression is applied separately to the battery steady discharge process and dynamic discharge process. We can conclude that (1) Random forest regression can be used to estimate the battery SOC. (2) By comparing with BP neural network , the random forest has a higher accuracy whether dynamic discharge process or steady discharge process. (3) The variable importance analysis of random forest can be used to analyse the importance of the SOC influence factors in order to simplify the model input and reduce the computation quantity.

ACKNOWLEDGMENT

This work was supported by the Fundamental Research Funds for the Central Universities (grant no. NS2014028).

REFERENCES

[1] W. Waag, C. Fleischer, D. U. Sauer, "Critical review of the methods for monitoring of lithium-ion batteries in electric and hybrid vehicles," Journal of Power Source, 2014,vol.258,pp.321-339.

[2] S. Lee, J. Kim, J. Lee, B.H.Cho, "State-of-charge and capacity estimation of lithium-ion battery using a new open-circuit voltage versus state-of-charge," Journal of Power Source,2008,vol.185, pp.1367-1373.

[3] S. Grewal, D. A. grant. "A novel technique for modeling the state of charge of lithium ion batteries using artificial neural networks," INTELEC 2001, pp.174-179. 14-18 October 2001.

[4] J. C.Peng, Y.B.Chen. R.Eberhart. "Battery pack state of charge estimator design using computational intelligence approaches," Battery Conference on Applications and Advances, 2000,pp.173-177.

[5] C.C.Chan,E.W.C.Lo,W.X.Shen, "The available capacity computation model based on artificial neural network for lead–acid batteries in electric vehicles," Journal of Power Source, 2000, vol. 87 ,pp .201-204.

[6] Y.S.Lee, W.Y.Wang, T.Y. Kuo, "Soft computing for battery state-of-charge(BSOC)estimation in battery string systems," IEEE TRANSACTION ON INDUSTRIAL ELECTRONICS,pp.229-239. January 2008.

[7] L.Breiman,"Random Forest," Machine Learning,2001,vol.45,pp.5-32.

[8] M. Segal,"Machine Learning Benchmarks and Random Forest Regression," Center for Bioinformatics and Molecular Biostatistics, UC San Francisco,2004.

[9] B.Saha, K.Goebel. "Battery Data Set, " NASA Ames Prognostics Data Repository[http://ti.arc.nasa.gov/project/prognostic-data-repository], NASA Ames, Moffett Field, CA. 2007.

[10] Q.S. Shi,C.H. Zhang, N.X. Cui, "Estimation of battery state-of-charge using v-support vector regression algorithm," International Journal of Automotive Technology,2008,vol.9,pp.759-764.