Big-data Analysis and Knowledge Discovery of Battery Fault in Numerous Real-world Electric Vehicles

Guangyu Zhao[1], Qingle Sun[2*]

[1]guangy_zhao@163.com, [2*]756907958@qq.com

[1]School of Public Administration, Huazhong Agricultural University, Hubei, 430070, China
[2]The National Engineering Research Center of Electric Vehicles, Beijing Institute of Technology, Beijing 100081, China

Abstract—Battery safety and aging have been considered as most import issues restricting the further deployment and development of electric vehicles in real-world applications. Due to differences in technology, materials, groups and physical and chemical reactions inside the battery, lithium-ion concentration gradient will be formed inside the battery, which is directly reflected in the inconsistency of external parameters. To study the correlation between battery faults and external parameters, in this study, all-life-cycle big-data of 41 electric vehicles are analysed by data mining. Firstly, the characteristic parameters representing voltage consistency are studied, and then the selected related parameters are statistically analysed in each dimension to explore the factors affecting the safety and reliability of batteries. The interesting knowledge discovered by this study can provide follow-up support for battery safety.

Keywords-Big-data analysis, knowledge discovery, electric vehicles (EVs), correlation analysis, hypothesis testing

1 Introduction

With the increasing demand for power and energy of EVs, hundreds of the cells are connected in series and parallel to form battery packs. However, the inconsistency brought about by battery formation can affect the energy density, durability, and safety performance of the battery pack [1]. The inconsistency of a battery pack has two sources, one is the inherent inconsistency generated during the manufacturing process, and the other is the inconsistency that becomes progressively larger due to the different temperatures and currents of the cells in the battery pack [2]. Accurately identifying the inconsistency characteristics of a battery pack can contribute to accurate state-of-charge (SOC) predictions and battery safety [3].

Many studies have been deployed to explore battery inconsistency issue based on data mining methods and the history operation data of batteries. B. Duan et al. argued that the inconsistency of a battery pack needs to be evaluated using a combination of indicators, including capacity, internal resistance, and specific value of the constant current charge capacity and constant voltage charge capacity, and presented a comprehensive analysis of the

ICIDC 2022, June 17-19, Qingdao, People's Republic of China

DOI 10.4108/eai.17-6-2022.2322760

inconsistency of a battery pack based on information entropy theory [4]. X. Feng et al. proposed a method of inconsistency assessment for battery packs based on a clustering quality evaluation index applied to time-series data. The method directly used the cell voltage as an assessment factor with low complexity [5]. M. Ouyang et al. proposed an internal short circuit detection method based on the internal battery consistency of the battery pack. This method used the recursive least squares (RLS) method of the Mean Difference Model (MDM) to estimate the model parameters. First, the basic parameters of the MDM were identified, and then the characteristic parameters, such as the voltage difference and the fluctuation function of the internal resistance, were calculated according to the basic parameters of the MDM. When an internal short circuit occurs, the internal short circuit can be detected based on the significant changes in these characteristic parameters [6]. Differential voltage analysis [7] and capacity increment analysis [8, 9] can indicate the deterioration of the battery. The width and height of the peak curve reflect the aging characteristics of the battery, and each battery has a unique curve attribute that represents its state of health (SOH). The charge-discharge characteristic curve describes the relationship between the battery parameters reflecting the dynamic internal reaction, and the difference between the curves indicates the consistency of the power battery pack.

Some studies analyzed big data of numerous real-world vehicle operating data and proposed fault diagnosis method. P. Liu et al. proposed a method for mapping the cell voltage as a high-dimensional vector and quantifying the magnitude of voltage inconsistency using the variance of cosine similarity, furthermore, used a clustering algorithm to improve computational efficiency [10]. Z. Wang et al. used the electric vehicle monitoring platform to obtain voltage data. Based on the comparison of the sample entropy and Shannon entropy on the time series, they believed that Shannon entropy had a better effect on voltage fault diagnosis. Based on the modified Shannon entropy, a Z-score-based safety early warning strategy was proposed to provide early warnings [11]. J. Tian et al. proposed battery consistency evaluation methods based on multi-feature weighting and clustering analysis. They found that the open-circuit voltage (OCV), ohmic resistance and polarization resistance have different effects on the consistency of the power battery pack, and the experimental results showed that the consistency variation is negatively correlated with the mileage, which can be approximately fitted by a first-order function [12].

With the rapid development of EVs, the EV data will grow by leaps and bounds. However, current methods are based on data of single EV, yet has disadvantages of discovering the knowledge by numerous real-world electric vehicles. In addition, the large amount of computation, and online implement of existing methods are also great challenges. To cope with the issue, all-life-cycle big-data of 41 electric vehicles are analysed by data mining. Firstly, 7 voltage characteristic parameters that measure the voltage dispersion of the cells at a moment were proposed. Then, all the data of a faulty electric logistics vehicle were used to carry out the chi-square tests, and the chi-square values of 7 parameters were obtained. Finally, the selected parameters are analysed by big-data of 41 electric vehicles from season and mileage dimension and interesting knowledge are discovered to provide follow-up support for battery safety.

2 Data Acquisition

The studied data is of real-world EVs is stored and used as time-series data, which includes vehicle data, drive motor data, pole data, alarm data, vehicle location, fuel cell data, engine data, etc. The vehicle data parameters are shown in Table 1.

Table1: The vehicle data parameters

Items	Vehicle data					
Parameters	Vehicle state		Charging state	Gear	Speed	Mileage
	Total voltage	Total current	State of Charge	DC-DC state	Operational mode	Insulation resistance

This paper makes a study of voltage inconsistency at the same moment. Since most of the vehicles operate normally in the real world, there are very few vehicles with failures. Therefore, the number of vehicles with poor voltage consistency is also very small. The data of alarm for poor cell consistency and cell voltages are selected according to the requirements of the study. In this paper, an electric logistics vehicle that has failed is selected to study the relationship between cell inconsistency and characteristic parameters. The data of 41 passenger cars are analysed by big-data method, while the patterns of the characteristic parameters are investigated. Some data examples are shown in Table 2.

Table2: Data examples

Index	Time	Speed (km/h)	Total voltage (V)	Cell voltages (V)	Alarm for poor cell consistency	…
0	2019-01-02 10:41:03	37	346.1	4.010_..._4.006	1	…
1	2019-01-02 10:41:13	45.9	351.8	4.010_..._3.974	1	…
2	2019-01-02 10:41:23	55.8	344.9	3.961_..._3.974	1	…
…	…	…	…	…	…	…

3 Characteristic Parameters and Correlation Analysis

In this paper, seven descriptive statistics commonly used to measure data dispersion, including range, relative range, interquartile range, variance, standard deviation, mean deviation, and coefficient of variation, were selected as parameters to characterize consistency, and parameter values were calculated for the poor consistency alarm and non-alarm data. The chi-square test was used to test whether each parameter was correlated with the poor consistency alarm. The larger the chi-squared value of a parameter, the more likely it is that the parameter is correlated with the poor consistency alarm.

3.1 Characteristic Parameters

The seven parameters have similarities, but each has its own focus. The range is the difference between the maximum and minimum values of all data, while the interquartile range is only used for 50% of the data. The relative range is a relative value. The variance is usually used to determine the stability of a set of data and is the square of the standard deviation. The mean deviation can measure the difference between the data and the arithmetic mean. The coefficient of variation is sensitive to small changes in the mean.

The range of voltages V_R refers to the difference between the maximum cell voltage and the minimum cell voltage within the battery pack, which is susceptible to extreme values. The calculation formula is as follows

$$V_R = V_{\max} - V_{\min} \tag{1}$$

Where $V_{\max}$ is the maximum cell voltage and $V_{\min}$ is the minimum cell voltage.

The relative range of voltages V_{RR} refers to the ratio of V_R to the mean of voltages $\overline{V}$. The calculation formula is as follows

$$V_{RR} = \frac{V_R}{\overline{V}} \tag{2}$$

The interquartile range of voltages V_Q refers to the range of the middle 50% of the sorted voltages. V_Q is less susceptible to extreme values compared to V_R. The calculation formula is as follows

$$V_Q = V_{Q3} - V_{Q1} \tag{3}$$

Where V_{Q3} is the third quartile of the sorted voltages, and V_{Q1} is the first quartile of the sorted voltages.

The variance of voltages V_{s^2} is the arithmetic mean of the square of the difference between the cell voltage and the mean. The standard deviation of voltages V_s is the arithmetic square root of V_{s^2}. V_{s^2} and V_s reflect the overall magnitude of the dispersion of the voltages. The formula is as follows

$$V_{s^2} = \frac{1}{n}\sum_{i=1}^{n}(V_i - \overline{V})^2 \tag{4}$$

$$V_s = \sqrt{V_{s^2}} \tag{5}$$

Where n is the number of cells, V_i is the voltage of the i-th cell.

The mean difference of voltages V_{MD} is the arithmetic mean of the absolute value of the difference between the cell voltage and the mean. V_{MD} uses the absolute value to erase the effect of positive and negative signs, and the variance of voltages uses the square. The calculation formula is as follows

$$V_{MD} = \frac{1}{n}\sum_{i=1}^{n} | V_i - \overline{V} | \tag{6}$$

The coefficient of variation of voltages V_{CV} is the ratio of V_s to $\overline{V}$. The effect of measurement scale and dimension can be eliminated by V_{CV} relative to V_s. The calculation formula is as follows

$$V_{CV} = \frac{V_s}{\overline{V}} \tag{7}$$

The above parameters are formed into parameter characteristic matrix F

$$F = \begin{bmatrix} V_{R1} & V_{RR1} & V_{Q1} & V_{s^2 1} & V_{s1} & V_{MD1} & V_{CV1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ V_{Rt} & V_{RRt} & V_{Qt} & V_{s^2 t} & V_{st} & V_{MDt} & V_{CVt} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix} \tag{8}$$

Where t is any time.

3.2 Correlation Analysis

The voltage inconsistency characteristic parameters are derived mathematical and statistical methods, and each parameter has a different correlation with the cell inconsistency. In order to select parameters with potentially high correlation for the next part of the statistical analysis, the statistical chi-square test was used to assess the correlation of each parameter. The chi-square test is often used to quantify the difference between the actual observed value and the theoretical inferred value, widely used to test the independence between two variables. The larger the chi-square value, the greater the degree of deviation of the two variables. In general, the chi-square test is divided into four steps.

(1) Calculating the chi-square value

A parameter can be regarded as a feature, and establish the matrix F as in (8). Whether a poor consistency alarm occurs can be regarded as a label, and establish the label matrix L as in (9). The label matrix L ([1 0] for no alarm, [0 1] for alarm]) is processed with one-hot coding.

$$L=\begin{bmatrix} y_1 \\ \vdots \\ y_t \end{bmatrix} \tag{9}$$

Dot product the transpose of matrix L by matrix F, and sum values of all features by class to obtain the observation matrix A. Since it is a binary classification problem, the number of rows of matrix A is 2. The first row is the sum of the 7 parameters for all non-alarm times, and the second row is the sum of all alarm times.

$$A=\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} & a_{16} & a_{17} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} & a_{26} & a_{27} \end{bmatrix} \tag{10}$$

The sum of each feature and the class frequency are calculated. The expectation matrix E is calculated analogously to the above steps.

$$E=\begin{bmatrix} e_{11} & e_{12} & e_{13} & e_{14} & e_{15} & e_{16} & e_{17} \\ e_{21} & e_{22} & e_{23} & e_{24} & e_{25} & e_{26} & e_{27} \end{bmatrix} \tag{11}$$

Calculate the chi-square value.

$$\chi^2=\sum\frac{(A-E)^2}{E} \tag{12}$$

(2) Solving for degrees of freedom

The degrees of freedom represent the total number of variables minus the number of constraints. The degrees of freedom for each feature are the number of categories minus 1, so the degree of freedom is 1.

(3) Setting significance level value

The significance level refers to the probability of rejecting the null hypothesis when it is correct, representing the probability of a small probability event. The significance level is usually determined before the statistical test, and is usually taken as 0.05 or 0.01. In this paper, the significance level is 0.05.

(4) Looking up the table

A part of the chi-square distribution table with degrees of freedom equal to 1 is shown in Table 3. It can be finally determined whether the two variables are correlated by looking up the table.

Table3: Chi-square distribution table

α	χ^2
0.10	2.71
0.05	3.84
0.02	5.41
0.01	6.64

4 Results and Discussion

An electric logistics vehicle that has failed is selected for the experiment. The battery type of this logistics vehicle is ternary lithium-ion battery and the number of data pieces exceeds 390,000. The results of calculating 7 parameters by randomly selecting one piece of data in two cases respectively are shown in Figure1. It is obvious that the value of each parameter is higher than when there is no alarm. It seems that all these parameters and the poor consistency alarms are related to some extent, but it could also be an accidental result due to random sampling. Therefore, the chi-square tests are used to test whether they are really correlated to reduce the random errors.

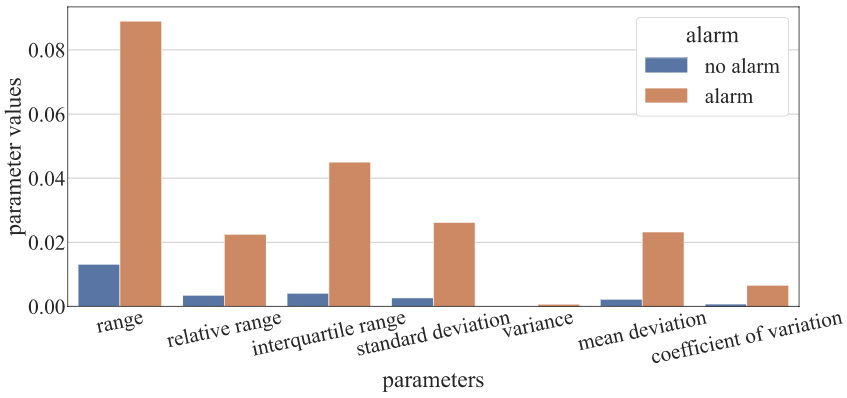

Fig.1: Comparison of parameter values in case of alarm and no alarm.

The chi-square values of all the parameters were calculated as shown in Figure2. When the chi-square value of a parameter is greater than 3.84, it means that there is more than 95% confidence in rejecting the null hypothesis. That is to say it is believed that the parameter and the poor consistency alarm are correlated. When the chi-square value is larger, it means that the degree of deviation between the theoretical inferred value and the actual observed value is

greater. The order of the probability of correlation of the parameters according to the test results is range, interquartile range, relative range, standard deviation, mean deviation, variance and coefficient of variation.

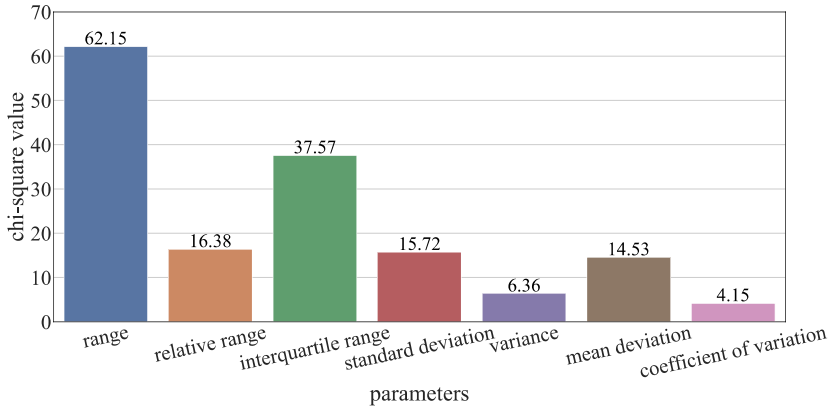

Fig.2: The result of the chi-square values of parameters.

The top two parameters, the range and the interquartile range, with their chi-square values of 62.15 and 37.57, were chosen to perform the analysis of the influencing factors. The data for the analysis experiment are 41 passenger cars belonging to the same model.

Temperature is an important factor affecting the consistency of power battery pack. In this paper, a large amount of data of 41 passenger cars are used to perform statistical analysis of vehicles in different seasons. The curves of the mean range and the mean interquartile range with seasons are shown in Figure3. It is obvious that the mean range in winter is the largest, reaching more than 30mV, followed by autumn, and lower in spring and summer. This indicates that the consistency of the power battery pack of this model will become gradually worse as the ambient temperature decreases. The pattern presented by the mean interquartile range is basically the same as the mean range. From the principle point of view, the higher temperature in summer provides a warm working environment for the power battery pack, which results in a smoother insertion and extraction of lithium-ions compared to the cold winter. Therefore, the consistency of the power battery pack is better in summer.

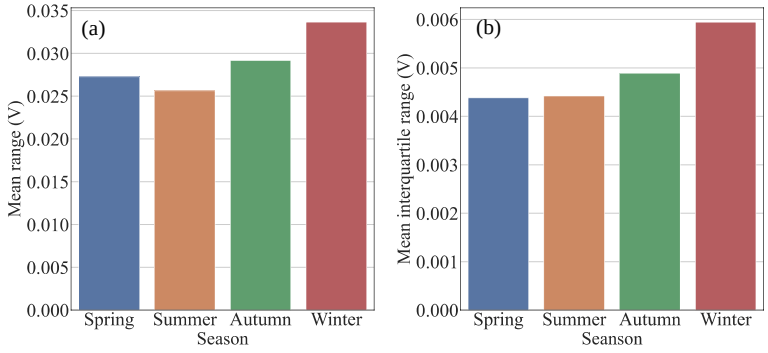

Fig.3: The result of two characteristic parameters in different seasons. (a) Mean range; (b) Mean interquartile range.

When the electric vehicles leave the factory, the batteries are in the initial stage of their life cycle. As the mileage of the electric vehicles increases, whether it is the original inconsistency caused by manufacturing differences, or caused by a micro-short circuiting during the operation of the vehicle, or the inconsistency caused by impurities mixed into the power battery pack will be revealed. This paper makes a statistical analysis of vehicles with different mileage. The curves of the mean range and the mean interquartile range with mileage in four seasons are shown in Figure4, corresponding to four sub-graphs respectively. The mileage is divided into 4 intervals.

It can be seen that regardless of the season, the mean range shows a U-shaped curve with the increase in mileage, indicating that in each season, the voltage consistency gradually becomes better and then worse with the increase in mileage. Due to manufacturing differences, the consistency of the power battery pack may not be optimal in the initial stage. Since the structure of the power battery pack is in series and parallel, there is a situation where one cell charges another cell. As the mileage of electric vehicles increases, the consistency gradually becomes better. In the later stage, the influence of mileage gradually increases, the battery may be degraded or short-circuited, and the consistency becomes worse. It can be seen that 25,000 to 35,000 kilometers is a demarcation point.

Compared with the mean interquartile range, the mean range has a larger variation. The range measures the difference between the maximum and minimum voltages in the power battery pack, and the interquartile measures 50% of the cells, which shows that the poor consistency of the power battery pack is mainly reflected in the individual cells. The abnormality of individual cells causes the maximum or minimum voltage to change drastically, which makes the range more obvious than the interquartile range.

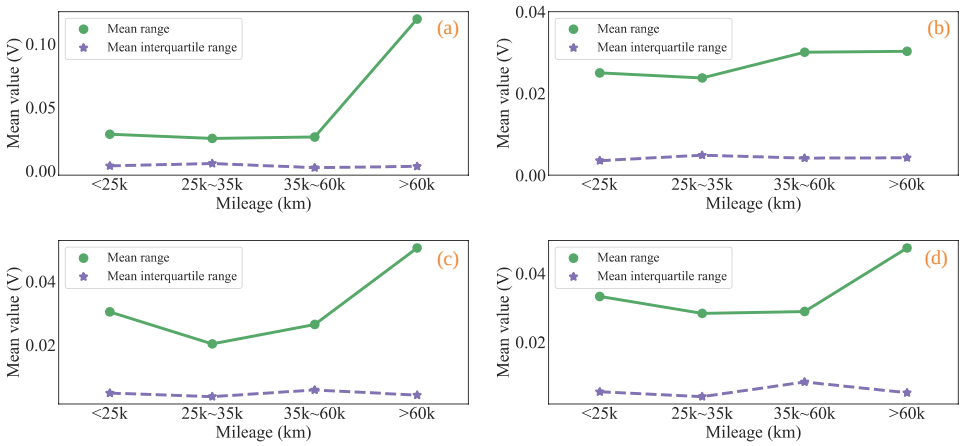

Fig.4: The result of two characteristic parameter values under different mileage in each season. (a) Spring; (b) Summer; (c) Autumn; (d) Winter.

5 Conclusions

In this paper, all-life-cycle big-data of 41 electric vehicles are analysed by data mining, 7 voltage characteristic parameters that measure the voltage dispersion of the cells at a moment were proposed, including voltage range, voltage relative range, voltage interquartile range, voltage standard deviation, voltage variance, voltage mean deviation and voltage coefficient of variation. All the data of a faulty electric logistics vehicle were used to carry out the chi-square tests, and the chi-square values of 7 parameters were obtained. Among them, the voltage range reaches 62.15 and the voltage interquartile range reaches 37.57, ranking the top two. Therefore, they were selected as statistical parameters. Based on the two parameters and all the data of 41 passenger cars of the same model, the relationship and difference of the above two parameters were analyzed in the seasonal dimension and mileage dimension. We can draw the conclusion that voltage range and the voltage interquartile range are high in winter and low in summer. The voltage range has a U-shaped curve with the mileage, which is more obvious than the voltage interquartile range. The voltage range can be used to quantify the inconsistency of the power battery pack instead of relying on alarm information. The interesting knowledge discovered by this study can provide follow-up support for battery safety.

Acknowledgments. This work is supported by National Natural Science Foundation of China (51775042).

References

[1] Zhou, L., Zheng Y., Ouyang M. et. al. (2017) A study on parameter variation effects on battery packs for electric vehicles. Journal of Power Sources, 364: 242-252.

[2] Feng, F., Hu, X., Hu L. et. al. (2019) Propagation mechanisms and diagnosis of parameter inconsistency within Li-Ion battery packs. Renewable and Sustainable Energy Reviews, 112: 102-113.
[3] She, C., Wang, Z., Sun, F. et. al. (2020) Battery aging assessment for real-world electric buses based on incremental capacity analysis and radial basis function neural network. IEEE Transactions on Industrial Informatics, 16: 3345-3354.
[4] Duan, B., Li, Z., Gu, P. et. al. (2018) Evaluation of battery inconsistency based on information entropy. Journal of Energy Storage, 16: 160-166.
[5] Feng, X., Zhang, X., Xiang, Y. (2020) An inconsistency assessment method for backup battery packs based on time-series clustering. Journal of Energy Storage, 31:101666.
[6] Ouyang, M., Zhang, M., Feng, X., et. al. (2015) Internal short circuit detection for battery pack using equivalent parameter and consistency method. Journal of Power Sources, 294: 272-283.
[7] Wang, L., Pan, C., Liu, L., et. al. (2016) On-board state of health estimation of LiFePO4 battery pack through differential voltage analysis. Applied Energy, 168: 465-472.
[8] Jiang, Y., Jiang, J., Zhang, C., et. al. (2017) Recognition of battery aging variations for LiFePO4 batteries in 2nd use applications combining incremental capacity analysis and statistical approaches. Journal of Power Sources, 360: 180-188.
[9] Weng, C., Cui, Y., Sun, J., et. al. (2013) On-board state of health monitoring of lithium-ion batteries using incremental capacity analysis with support vector regression. Journal of Power Sources, 235: 36-44.
[10] Liu, P., Wang, J., Wang, Z. et. al. (2019) High-dimensional data abnormity detection based on improved Variance-of-Angle (VOA) algorithm for electric vehicles battery. 2019 IEEE Energy Conversion Congress and Exposition (ECCE), 5072-5077.
[11] Wang, Z., Hong, J., Liu, P. et. al. (2017) Voltage fault diagnosis and prognosis of battery systems based on entropy and Z-score for electric vehicles. Applied Energy, 196: 289-302.
[12] Tian, J., Wang, Y., Liu, C. et. al. (2020) Consistency evaluation and cluster analysis for lithium-ion battery pack in electric vehicles, Energy, 194: 116944.