Lithium-Ion Battery Screening by K-Means with DBSCAN for Denoising

Yudong Wang[1, 2], Jie Tan[1, *], Zhenjie Liu[1] and Allah Ditta[3]

Abstract: Batteries are often packed together to meet voltage and capability needs. However, due to variations in raw materials, different ages of equipment, and manual operation, there is inconsistency between batteries, which leads to reduced available capacity, variability of resistance, and premature failure. Therefore, it is crucial to pack similar batteries together. The conventional approach to screening batteries is based on their capacity, voltage and internal resistance, which disregards how batteries perform during manufacturing. In the battery discharge process, real time discharge voltage curves (DVCs) are collected as a set of unlabeled time series, which reflect how the battery voltage changes. However, few studies have focused on DVC based battery screening. In this paper, we provide an effective approach for battery screening. First, we apply interpolation on DVCs and give a method to transform them into slope sequences. Then, we use density-based spatial clustering of applications with noise (DBSCAN) for denoising and treat the remaining data as input to the K-means algorithm for screening. Finally, we provide the experimental results and give our evaluation. It is proved that our method is effective.

Keywords: Lithium-ion battery, battery screening, K-means, denoising.

1 Introduction

In recent years, lithium-ion batteries have been widely used in various applications, such as electric vehicles, industrial energy storage equipment, and automobile starting devices. In the majority of circumstances, batteries are packed together to satisfy voltage and capability needs [Mathew, Kong, Mcgrory et al. (2017); Al-Zareer, Dincer and Rosen (2017); Liu, Tan and Wang (2018)]. However, due to variations in raw materials, manual operation, and differing ages of the equipment, there is inconsistency among a pack of batteries, which usually manifests as inconsistent voltages and internal resistances [Zhang, Cheng, Ju et al. (2017)]. When a pack is in operation, unbalanced voltages and internal resistances will lead to partial heating, which can accelerate battery aging and aggravate the imbalance. Over

[1] Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China.

[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, 100049, China.

[3] University of Education, Township, College Rd Lahore Punjab, Lahore, 54770, Pakistan.

[*]Corresponding Author: Jie Tan. Email: tan.jie@tom.com.

Received: 19 April 2020; Accepted: 20 June 2020.

CMC. doi:10.32604/cmc.2020.011098

time, the battery life will be greatly reduced [Liu, Liu, Lin et al. (2018)].

1.1 Battery consistency and screening

Battery consistency refers to consistency in the characteristics of cell performance, including voltage, current, internal resistance (which can be measured by instruments), and dynamic changes during charging and discharging [Zhang, Jiang, Jiang et al. (2017)]. Consistency cannot be measured directly; therefore, batteries are screened according to the similarity of their voltages, internal resistances and capabilities [Li and Pan (2005)]. At the same time, batteries are chosen from the same batch, which means they contain similar materials [Liu, Liu, Lin et al. (2018)]. However, this approach neglects the way batteries perform during charging and discharging [Wang, Zhang, Ge et al. (2016)].

1.2 Discharge voltage curve and time series

The discharge voltage curve (DVC) is a series of data collected by sensors during the discharging process in battery manufacturing. Observed values are sent to databases every few seconds to record real-time voltage. We use V to represent observations and T for reference time points to describe DVC records, and we have

$$(V,T)=\{(v_1,t_1),(v_2,t_2),\dots,(v_n,t_n)\} \tag{1}$$

where an observation v_i is always paired with its reference time point t_i. This kind of data is called a time series, and it consists of a sequence of observation values and corresponding reference time points. In general, we place more emphasis on the DVC than on the charge voltage curve (CVC). This is because charging and discharging are reversible chemical reactions, and when batteries work as power supplies, they are always discharging during operation. While batteries are screened, there is no supervision, so battery screening is a time-series clustering problem. Raw data are shown in Fig. 1.

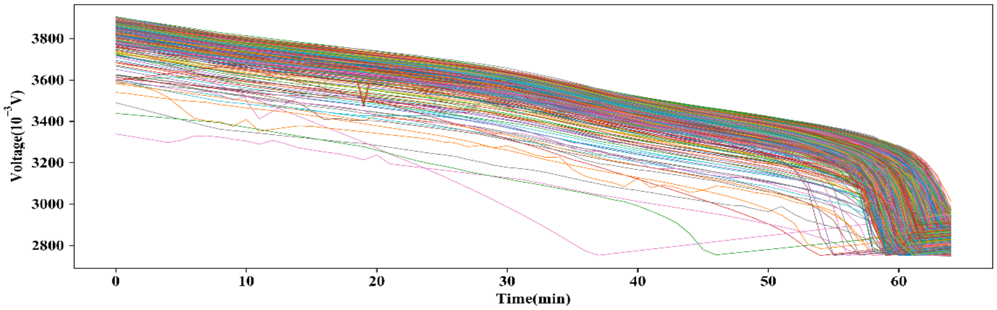

Figure 1: The DVC data are a set of time series with observation values and corresponding reference time points. The discharging process lasts 60 minutes

1.3 Time-series clustering

Clustering belongs to the category of unsupervised machine learning, and it aims to divide unlabeled data into several clusters. Specifically, if data are represented as a set of series consisting of values and time points, we call this task time-series clustering [Liao (2005)]. Among conventional clustering methods, there are shape-based, density-based, and distance-based approaches. It should be noted that there is no clear boundary between these three types of approaches.

Distance-based Method: When using distance-based methods, we calculate the similarity of two samples by the distance between them. There are many ways to determine similarity, such as the Euclidean distance and Manhattan distance. Among them, the Euclidean distance is the most widely used [Aghabozorgi, Shirkhorshidi and Wah (2015)]. For example, given a sample x and existing clusters A and B with centers $Center_A$ and $Center_B$, K-means calculates the distances $d(x, A) = dist(x, Center_A)$ and $d(x, B) = dist(x, Center_B)$. If $d(x, A) < d(x, B)$, x is classified into cluster A ; otherwise, it is classified into B [Gan and Ng (2017)].

Density-based Method: Unlike distance-based methods, density-based methods calculate the density den_i of each sample. Given a threshold ε, if $den_i < \varepsilon$, $sample_i$ is treated as a noise sample. Density-based spatial clustering of applications with noise (DBSCAN), for instance, generates a cluster by judging the density of a neighborhood and the distance from other samples or clusters.

Shape-based Method: Among shape-based methods, the dynamic time warping (DTW) method is well known. DTW is a measure to calculate the distance between time series. The basic problem that DTW attempts to solve is how to align two sequences to generate the most representative distance measure of their overall difference. If there is any discrepancy in the alignment of time series, the DTW algorithm uses a dynamic programming technique to solve this problem. The first step is to compare each point in one sequence with every point in the second, generating a matrix. The second step is to work through this matrix, starting at the bottom-left corner and ending at the top-right.

2 Problem formulation

2.1 Unaligned data

In a lithium-ion battery charging and discharging unit, there are tens of thousands of sensors for data collection. Equipment is divided into tens to hundreds of work areas that charge and discharge simultaneously. Theoretically, sensors collect voltage observations every t seconds. Although values are stored in a database (a real-time database, generally) at the same time intervals, in actual situations of networks and computer scheduling, reference time points cannot be aligned.

2.2 Data noise

In this paper, we aim to divide batteries that are consistent with each other into several groups. However, when collecting data, noise is inevitable. Some voltage curves show fluctuation, which is caused by the instability of the electrolyte in batteries, and this makes it difficult to distinguish these curves with the fluctuations from the rest of the data. Some noise is generated by the sensors. For instance, when errors occur, inaccurate values are collected and stored in the database.

2.3 K-means limitations

Given the time series $X = \{x_1, x_2, \ldots, x_m\}$, which is aligned by a reference time sequence with the same time interval, and clusters $C = \{C_1, C_2, \dots, C_k\}$, where $center_j$ is the center of cluster C_j and $x_j \in C_j$. for $\forall x_i \notin C_j$ [Krishna and Murty (1999)]. we have the following relationship:

$$dist(x_i, center_j) > dist(x_j, center_j) \tag{2}$$

Here, $dist(a, b)$ denotes the distance between a and b. For example, if the function $dist(a, b)$ is defined as Euclidean distance, we have

$$dist(a, b) = \sqrt{\sum(a_i - b_i)^2} \tag{3}$$

where $a_i \in a$ and $b_i \in b$.

An example: In the following experiment, we individually use K-means on raw data, and the results are shown in Fig. 2.

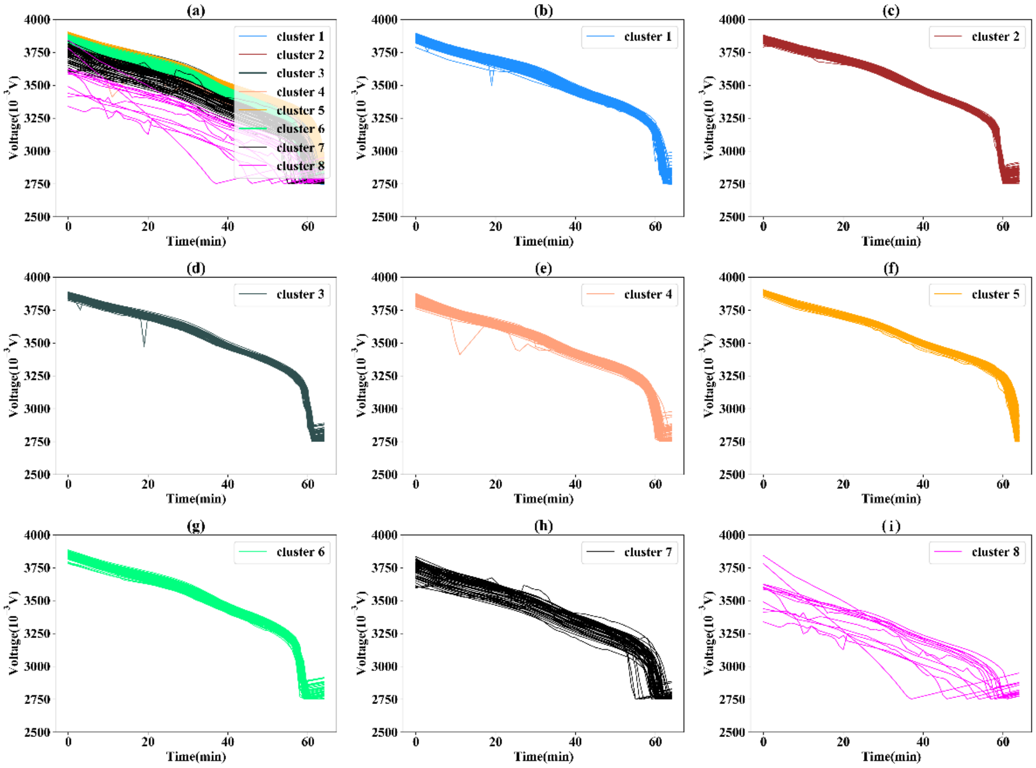

Figure 2: K-means clustering results without denoising are shown in subgraphs (a), (b) to (i), which show the details of each cluster

Fig. 2 shows the results of K-means. For battery DVC data, which include the part of the time series with high volatility (as shown in the subplot), it is impossible to calculate the distance accurately, which leads to deviation in clustering.

Why DTW should not be used. The DTW distance seems to be a reasonable method, but it performs poorly in practice. It is very time consuming, making it unsuitable for industrial applications. However, this is not the main reason why it should not be used. Generally, DTW works well on time series, but DVC data are more discrete than continuous. Notably, during the discharging process, the observation value of the voltage at each time point is important. However, DTW always aims to find a minimum distance between two curves, which leads to the problem that some points are ignored when calculating distance.

As shown in Fig. 3, we have 2 discharging sequences named DVC 1 and DVC 2, and there are 2 DTW intermediate steps named DTW steps a and b. When calculating the distance between DVC 1 and 2, DTW always finds a minimum distance, such as in steps a and b, but screening compares each voltage observation at the same time point. Therefore, DTW does not meet our needs.

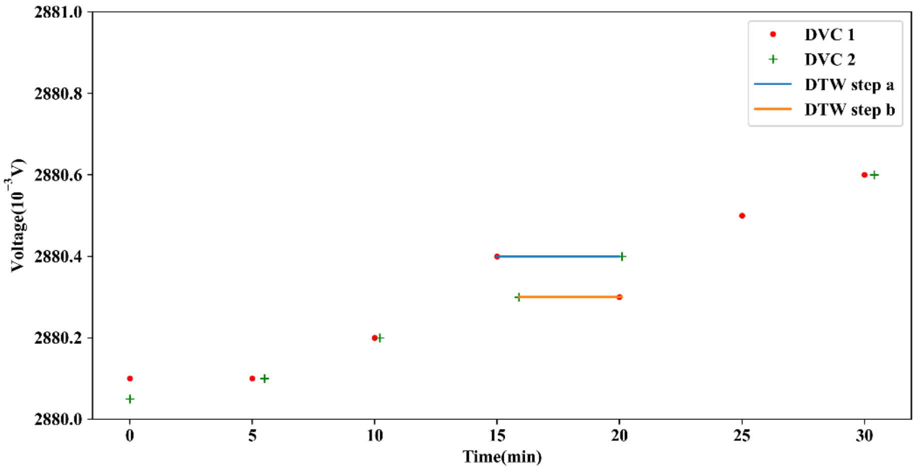

Figure 3: DVC 1 and 2 are discharging voltage and time sequences. DTW steps a and b show how to calculate the distance between 2 sequences

2.4 Main work

In this paper, our goal is to divide batteries that are consistent into several groups. We use an unsupervised approach to obtain them from unlabeled data. However, the DVC time series are not aligned, which makes it difficult for the clustering method to work properly. The first step of our work is to align the data to the same reference time sequence. We use the interpolation method to obtain a new time series. Second, we transform the DVC to slope sequences as a representation of the raw data. Then, the DBSCAN method is used for denoising. Finally, we use K-means for clustering and provide an evaluation of the results. Our main work is shown in Fig. 4.

3 Data processing

We process the data in 2 steps. The first step is data aligning, which aims to align the data to the same reference time sequence. The second is data representation, the goal of which is to obtain new data slope sequences.

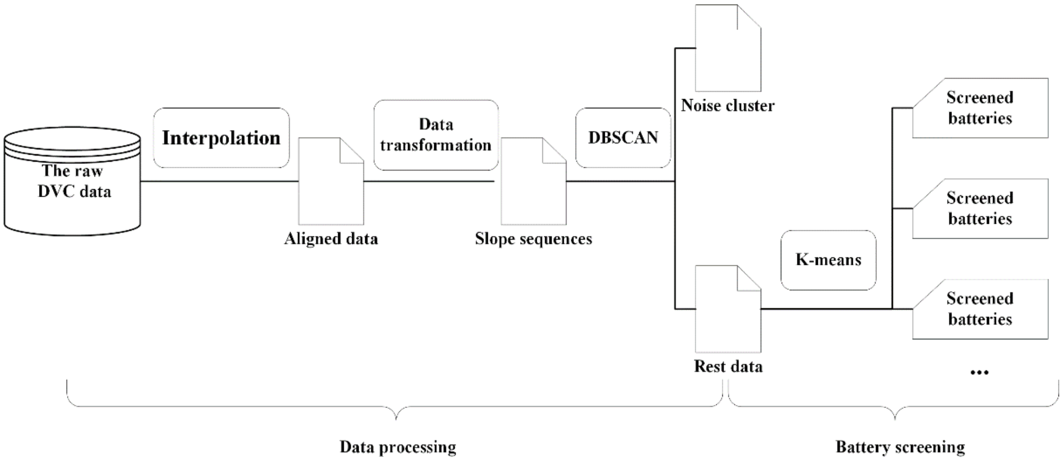

Figure 4: Our main work

3.1 Alignment

A set of battery DVC data consists of both observation values and reference time-point sequences. Because of the limitations of the network situation and the collection mechanism, data are not always aligned. as shown in Fig. 5.

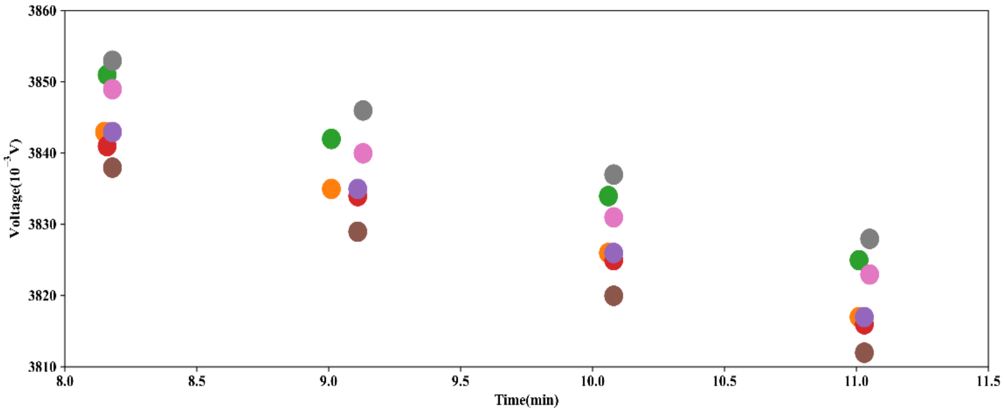

Figure 5: The 7 collected samples, for which data collection was performed 4 times. The sample collection times are not aligned, with each being different from one another

For the raw DVC time series, given v_i as observations and t_i as reference time points, we have

$$v_i = \{v_i^{(1)}, v_i^{(2)}, \dots, v_i^{(n)}\} \tag{4}$$

$$t_i = \{t_i^{(1)}, t_i^{(2)}, \dots, t_i^{(n)}\} \tag{5}$$

where (n) denotes the n-th collection and v_i, t_i are always paired together. Unaligned sequences are defined as follows: for $i, j \in M$, if $i \neq j$, then $len(v_i) \neq len(v_j)$ and

$len(t_i) \neq len(t_j)$, where $len(v)$ is the size of a sequence v. However, when we use a set of time series, we need it to be aligned, so interpolation is applied to the data.

Common methods include linear interpolation, Newton interpolation, and Lagrange interpolation. In selecting interpolation methods, the most important principle is to preserve the characteristics of the original data to ensure that the new data set will not deviate too much from the original. As shown in Fig. 1, the values of the DVC time series decrease smoothly and then rapidly drop to a certain range as time passes. Linear interpolation is simple, with few calculations, and most importantly, it barely changes the trend of the time series. Therefore, we choose it as the interpolation method for DVC data.

Suppose we have an interval of adjacent voltage values $(v_h^{(i)}, v_h^{(i+1)})$ with reference times $(t_h^{(i)}, t_h^{(i+1)})$; then,

$$v_h^{(j)} = \frac{t_h^{(j)} - t_h^{(j+1)}}{t_h^{(i)} - t_h^{(i+1)}} v_h^{(i)} - \frac{t_h^{(j)} - t_h^{(i)}}{t_h^{(i)} - t_h^{(i+1)}} v_h^{(i+1)} \tag{6}$$

where $j \in (i, i+1)$ and $v_h^{(j)}$ is the new value of sample v_h at time point $t_h^{(j)}$ after interpolation. The process is shown in Fig. 6.

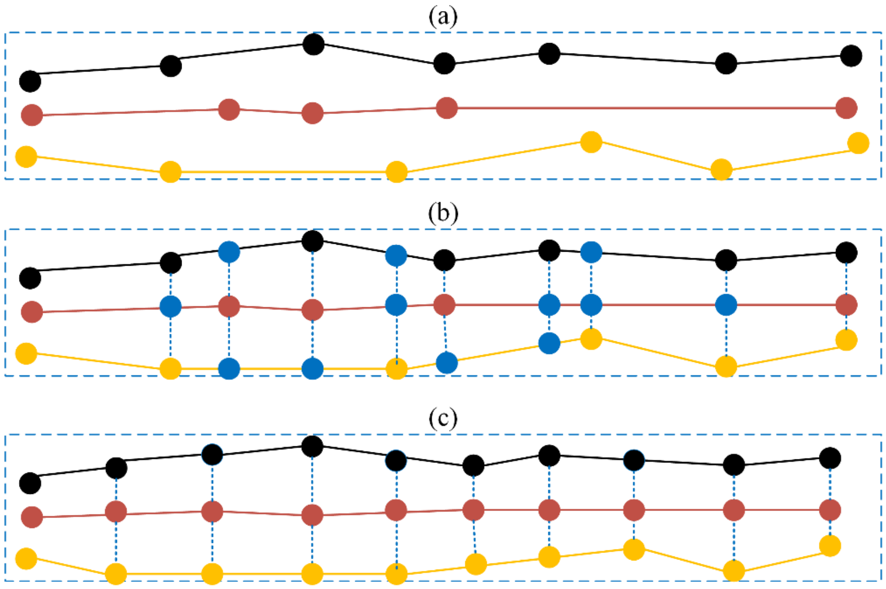

Figure 6: (a) shows how raw data appear without alignment. Samples do not have the same reference time sequence. (b) shows that when we align samples to the same time sequence, the time intervals in the series are not always equal. (c) shows the results after average interpolation

3.2 Representation

The DBSCAN algorithm is sensitive to parameters; therefore, using DBSCAN directly is not the best choice. In this paper, we transform time series into a set of slope sequences. Given an interpolated data set $X = \{x_1, x_2, \ldots, x_m\}$, each $x_i \in X$ is paired with the same reference time $T = \{t_1, t_2, \dots, t_n\}$. We have the slope sequence

$$s_i = \left\{x_i^{(1)}, \frac{x_i^{(2)} - x_i^{(1)}}{t_2 - t_1}, \frac{x_i^{(3)} - x_i^{(2)}}{t_3 - t_2}, \dots, \frac{x_i^{(n)} - x_i^{(n-1)}}{t_n - t_{n-1}}\right\} \tag{7}$$

For the same time interval, and letting $s_i^{(j)} = x_i^{(j+1)} - x_i^{(j)}$, we have

$$s_i = \{x_i^{(1)}, s_i^{(1)}, s_i^{(2)}, \dots, s_i^{(n-1)}\} \tag{8}$$

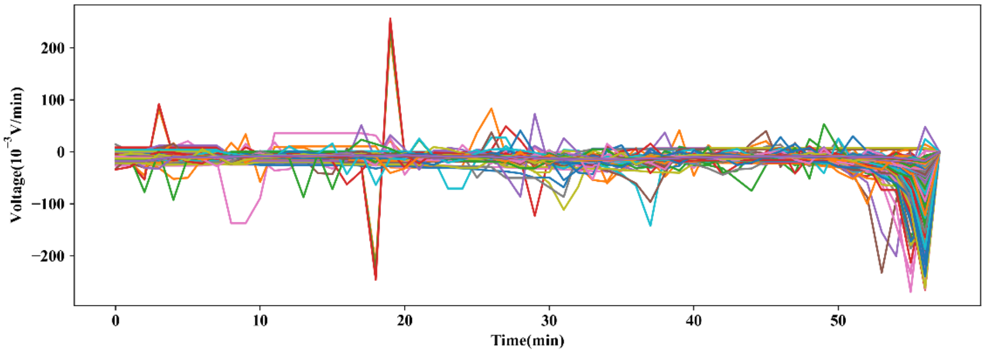

Figure 7: After interpolation and representation, the DVC data are transformed to slope sequences

In this way, s_i is a representation of x_i. Each $s_i^{(j)} \in s_i$ shows how the DVC changes at time point t_i. Ignoring the first item of each slope sequence, the whole sample representation is shown in Fig. 7.

4 DBSCAN for denoising

4.1 Why slope sequences are used

The DBSCAN algorithm is based on density, and it performs well on data with noise. DBSCAN uses 2 important parameters: ε and $min - samples$. Given a set of slope sequences, ε is the maximum distance there can be between two samples for one to be considered to be in the neighborhood of the other. This is the most important DBSCAN parameter. Without transforming data into slope sequences, DBSCAN would not be able to denoise well. This is because DBSCAN is sensitive to parameters, and DVC values are large (2700 to 4000), so choosing suitable parameters hard. At the same time, normalization does not work well, as very low values make it more difficult to initialize a DBSCAN. The experimental results are shown in Fig. 8.

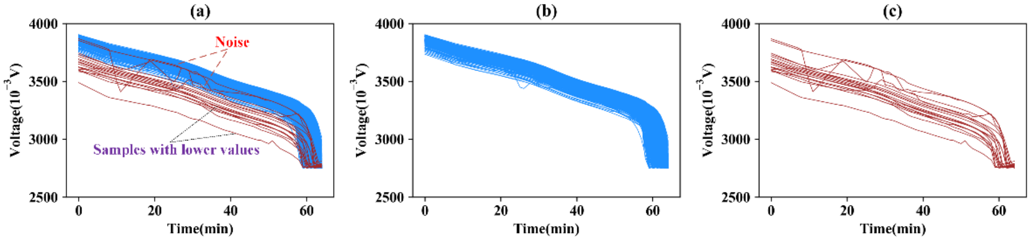

Figure 8: Experimental results of DBSCAN with aligned data.

Noise is clearly shown in subgraph (c). In fact, if data were not transformed into slope sequences, samples with low values and high volatility would be divided into noise clusters [Schubert, Sander, Ester et al. (2017)]. Thus, it is difficult to find a suitable value threshold to divide only samples with abnormal volatility into noise clusters. This is because DBSCAN focuses on sample values but not on how DVCs change in the reference time sequences.

4.2 Experiment on slope sequences

Our aim is to divide the data into $Cluster_{noise}$ and $Cluster_{normal}$. In this experiment (as shown in Fig. 9), we use DBSCAN on slope sequences, and the results are shown as (a), (b) and (c). Then, we transform the slope sequences back to aligned time series, and we can see how noise is detected in (d), (e) and (f).

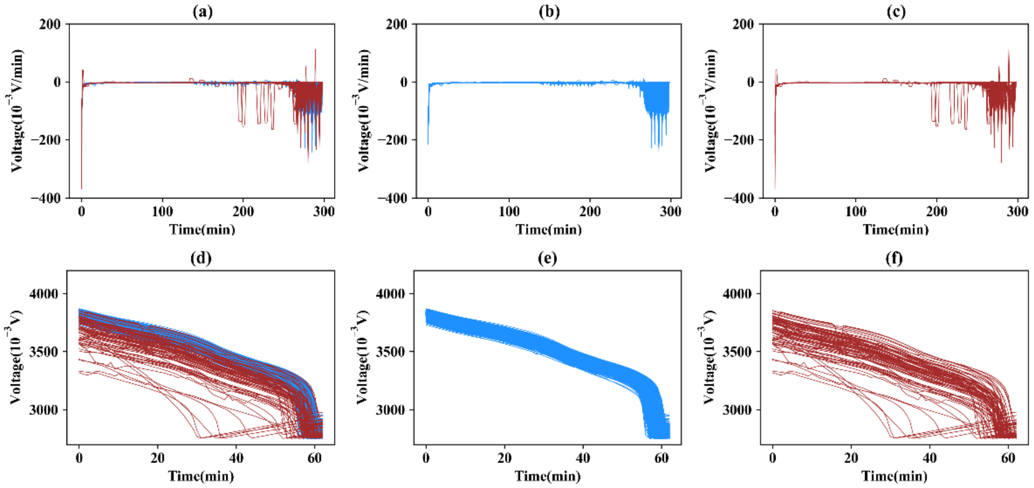

Figure 9: (a), (b) and (c) show how we perform denoising with DBSCAN on slope sequences. (d), (e) and (f) show the results with aligned DVC samples

5 Battery screening and evaluation

5.1 Screening by K-means clustering

The DVC time series data are divided into a noise cluster and a normal cluster after DBSCAN. The normal cluster will be used as the input data for the K-means algorithm, which is one of the best-known clustering methods; it is simple, effective, and the most widely used. Given a data set X, K-means divides it into several clusters, as determined by the parameter k defined by the user. The objective function is to minimize the loss function E, and E is given as follows:

$$E = \sum_{j=1}^{k} \sum_{i=1}^{n} \left|\left|x_i^{(j)} - c_j\right|\right|^2 \qquad (9)$$

where $x_i \in X$, c_j is denoted as the cluster center, and $\left|\left|x_i^{(j)} - c_j\right|\right|^2$ is the distance from x_i to the cluster center c_j. In this experiment, we divide the normal cluster into 8 new clusters to screen batteries that are consistent.

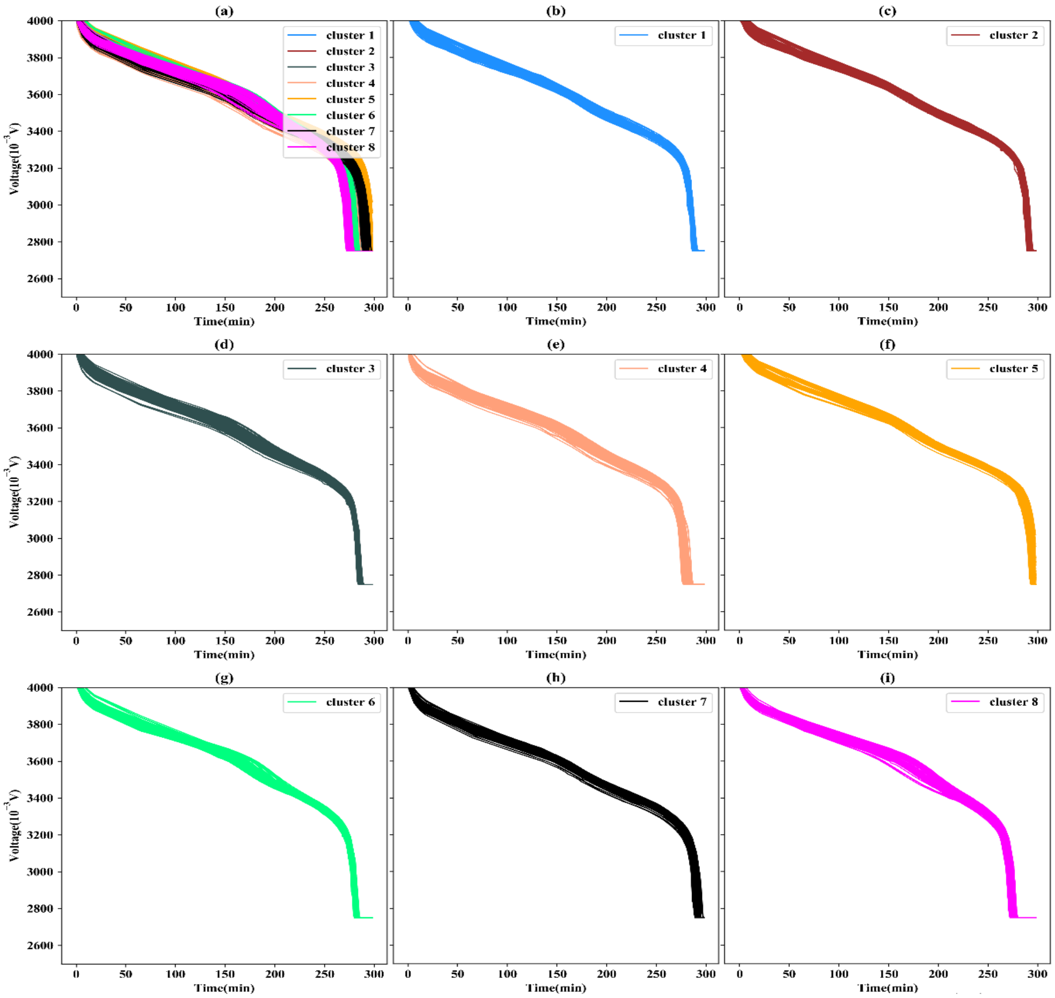

Figure 10: (a) shows whole clusters and their distribution, and (b) to (i) give the details of each cluster

Fig. 10 shows the results of K-means on slope sequences. Compared with Fig. 2, we can see that the results shown in figure 10 are better, with clearer borders. In subgraph (a), a total of 8 clusters are displayed, but there is partial overlap between them. Subgraphs (b) to (i) give the details of each cluster. Batteries belonging to one cluster have more consistency and better performance when they are packed together. Finally, we provide an essential-criteria principle to evaluate our experimental results.

5.2 Evaluation

In general, evaluation criteria are divided into internal categories and external categories. For K-means with Euclidean distance, we prefer to use the SSE (sum of squared error) to evaluate our experiment. The SSE is one of the most widely used criteria for clustering [Saxena, Prasad, Gupta et al. (2017)]. It is defined as

$$SSE = \sum_{k=1}^{K} \sum_{\forall x_i \in C_k} \left|\left| x_i - \mu_k \right|\right|^2 \tag{10}$$

where C_k is one of the clusters and μ_k is the vector mean of cluster k, as shown in Fig. 11. Here, the SSE indicates the inconsistency of a battery group, and the more inconsistent the batteries are, the lower their SSE.

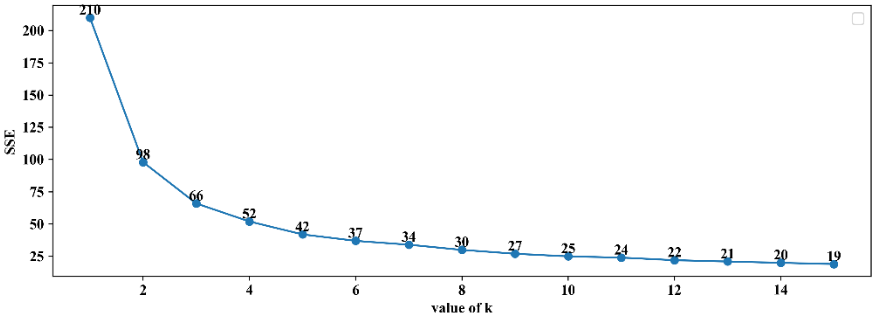

Figure 11: SSE changes with the k value

Here, the value K means how many clusters we classify batteries into. Batteries in the same cluster have a high degree of similarity, and that means they have similar DVCs. The voltage curves of the cells directly indicate the voltage variation law under working condition. Furthermore, they reflect the variation law of the battery capacity, internal resistance, and temperature. Therefore, cells with similar discharge voltage curves are considered largely to have similar electrochemical characteristics, and that means they are more consistence.

6 Conclusion

In this paper, we apply interpolation and representation to raw DVC time-series data to obtain aligned slope sequences. Then, DBSCAN is used for denoising and removing noise samples with irregular fluctuations. Finally, we use the K-means algorithm to divide the slope sequences into 8 clusters and provide an evaluation. Our experimental results show that performing K-means on aligned slope sequences with DBSCAN for denoising is an effective approach to battery screening.

Funding Statement: This work is supported by the National Key Research and Development Program under Grant 2018YFB1703400 and the National Natural Science Foundation of China under Grants U1801263 and U1701262.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

Aghabozorgi, S.; Shirkhorshidi, A. S.; Wah, T. Y. (2015): Time-series clustering-a decade review. *Information Systems*, vol. 53, pp. 16-38.

Al-Zareer, M.; Dincer, I.; Rosen, M. A. (2017): Novel thermal management system using boiling cooling for high-powered lithium-ion battery packs for hybrid electric vehicles. *Journal of Power Sources*, vol. 363, pp. 291-303.

Gan, G.; Ng, M. K. P. (2017): K-means clustering with outlier removal. *Pattern Recognition Letters*, vol. 90, pp. 8-14.

Krishna, K.; Murty, M. N. (1999): Genetic K-means algorithm. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 29, no. 3, pp. 433-439.

Li, X. Z.; Pan, H. B. (2005): Study on the uniformity of storage batteries. *Chinese Battery Industry*, vol. 10, no. 5, pp. 285-289.

Liao, T. W. (2005): Clustering of time series data-a survey. *Pattern recognition*, vol. 38, no. 11, pp. 1857-1874.

Liu, C.; Tan, J.; Shi, H.; Wang, X. (2018): Lithium-ion cell screening with convolutional neural networks based on two-step time-series clustering and hybrid resampling for imbalanced data. *IEEE Access*, vol. 6, pp. 59001-59014.

Liu, K.; Liu, Y.; Lin, D.; Pei, A.; Cui, Y. (2018): Materials for lithium-ion battery safety. *Science advances*, vol. 4, no. 6, eaas9820.

Mathew, M.; Kong, Q. H.; Mcgrory, J.; Fowler, M. (2017): Simulation of lithium ion battery replacement in a battery pack for application in electric vehicles. *Journal of Power Sources*, vol. 349, pp. 94-104.

Saxena, A.; Prasad, M.; Gupta, A.; Bharill, N.; Patel, O. P. et al. (2017): A review of clustering techniques and developments. *Neurocomputing*, vol. 267, pp. 664-681.

Schubert, E.; Sander, J.; Ester, M.; Kriegel, H. P.; Xu, X. (2017): DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems*, vol. 42, no. 3, pp. 1-21.

Schuster, S. F.; Brand, M. J.; Berg, P.; Gleissenberger, M.; Jossen, A. (2015): Lithium-ion cell-to-cell variation during battery electric vehicle operation. *Journal of Power Sources*, vol. 297, pp. 242-251.

Wang, C. Y.; Zhang, G.; Ge, S.; Xu, T.; Ji, Y. et al. (2016): Lithium-ion battery structure that self-heats at low temperatures. *Nature*, vol. 529, no. 7587, pp. 515-518.

Zhang, C.; Cheng, G.; Ju, Q.; Zhang, W.; Jiang, J. et al. (2017): Study on battery pack consistency evolutions during electric vehicle operation with statistical method. *Energy Procedia*, vol. 105, pp. 3551-3556.

Zhang, C.; Jiang, Y.; Jiang, J.; Cheng, G.; Diao, W. et al. (2017): Study on battery pack consistency evolutions and equilibrium diagnosis for serial-connected lithium-ion batteries. *Applied Energy*, vol. 207, pp. 510-519.