2019 IEEE 5th International Conference on Computer and Communications

A Hybrid Clustering Model for Lithium-ion Batteries Screening

Yudong Wang
Institute of Automation, Chinese Academy of Sciences
University of Chinese Academy of Sciences
Beijing, China
e-mail: wangyudong2018@ia.ac.cn

Jie Tan*
Institute of Automation, Chinese Academy of Sciences
Beijing, China
e-mail: jie.tan@tom.com

***Abstract*—In lithium-ion battery manufacturing, due to the variations of raw materials, manual operation and equipment, batteries performance differently from each other, which inevitably lead to a reduction in the available capacity and premature failure of a battery pack with multiple cells configured in series, parallel, and series–parallel. So it is important to screen inconsistent ones from batteries. The traditional battery screening approach is based on the capacity, voltage and internal resistance of the battery. However, this approach ignores the indicators in production manufacturing, among which battery discharge process is an important step. Discharge voltage curves are a set of time series, sensitive to values and shape, and have small fluctuations. It reflects how batteries perform during this work step. Single clustering method, density-based or shape-based, is not suitable for screening by discharge voltage curve, so we propose a hybrid model to screen inconsistent batteries. Finally, we give the experiment results, and compare with the single algorithm to prove that the model has better performance.**

***Keywords*-clustering; time series; battery screening;**

I. Introduction

In recent years, lithium-ion batteries have been widely used in various fields, such as electric vehicles, electric bicycles, and so on. Aiming to meet the demand for power and voltage, several or large numbers of lithium batteries are packed together, in series, parallel or series and parallel, to form a large battery pack. Each battery performs diversely from others, hence, consistency of batteries that make up the packs comes to be one of the most important indicators when evaluating the pack [1]. In this paper, we build a hybrid unsupervised model to screen consistent batteries from the unlabeled.

A. Lithium-ion Battery Consistency

Lithium-ion battery consistency mainly refers to the consistent characteristics of single cell performance, including the consistency of battery extrinsic characteristics (voltage, current, internal resistance), and the internal characteristics (capacity, power, energy). Battery consistency is closely related to raw materials, production process and equipment quality [2]. The first step is to mix various raw materials together with water, and then put them into manufacturing process, when the concentration and temperature are up to standard. However, in practice, the concentration of mixture is difficult to reach a certain precise value, and can only be in a range of requirements [3]. This uncertainty leads to different manufacturing technique parameters in subsequent processes.However, many of them are set manually, according to experience, which brings about the inconsistency of batteries. In addition, different aging degree of equipment and temperature change will also affect consistency of batteries [4].

B. Time-series Discharge Voltage Curve Data

Battery consistency can be represented by discharge voltage curve(DVC) during the discharge work step, and a further implication is, the higher similarity of DVC, the stronger consistency there is. Therefore, the problem of battery pack quality evaluation can be transformed into the problem of battery consistency, and then into the similarity of DVC. The DVC data is collected in real time by sensors and stored in the database during constant current discharge process of batteries, and the data is organized as time-series [5]. One of the problems is that the collected data are not labeled, therefor, time-series DVC data clustering is mainly focused on in this paper.

C. Time-series Clustering

At present, time-series clustering has been an attractive research era. A sequence of continuous, real-valued elements, is known as a time-series [6]. Time-series clustering is defined as follows: Given a time-series $D = \{D_1, D_2, \ldots, D_n\}$, the unsupervised data are partitioned into $C = \{C_1, C_2, \ldots, C_k\}$, C_i is called a cluster, and time-series in cluster are grouped together by a certain similarity measure, where $D = \cup_{i=1}^{k} C_i$ and $C_i \cap C_j = \oslash$ for $i \neq j$.

Because of feature values changing by time, an individual time-series is a series of dynamic data. Related work of time-series clustering could be divided into three categories: whole time-series clustering, subsequence clustering and time point clustering [7]. Whole time-series clustering means clustering on some amount of unsupervised, raw or processed, time-series by their similarity. Subsequence clustering differs from whole time-series clustering for it focuses on subsequences of a time-series. Time point clustering is clustering on time points, and one difference from the other

978-1-7281-4743-7/19/$31.00 ©2019 IEEE

two categories is that it is based on temporal proximity of time points and similarity of the corresponding values.

Hesam, Witold and Iqbal use DTW(dynamic time warping) distance, which is a shape-based approach, on fuzzy clustering [8]. DTW is a method that calculates similarity between two individual samples, including linear similarity and non-linear similarity. Non-linear similarity calculating is essentially a greedy algorithm, and linear similarity depends on a sliding window to scan an individual. Both approaches ignore the fact that DVC is sensitive to observations of value. Sliding and non-linear are not allowed.Bode, Gerrit, et al. talked about how unsupervised machine learning techniques can be used to tackle time-series clustering on automation and control systems. Carolina, Hernando, Joaquín gave a modified spectral merger method to process unsupervised data [9]. Rui et al. propose YADING algorithm, and it can cluster large-scale time-series data very fast.

D. Summary of Work

First, we analyze the particular of time-series DVC data and the limitations of K-means, a partitioning clustering method, and DBSCAN (Density-Based Spatial Clustering of Applications with Noise), a density based clustering method. Second, we adopt interpolation method to get new time-series representation for data alignment and easier clustering. Third, we proposed a modified DBSCAN algorithm, and create a hybrid clustering model on time-series, as is shown in figure 1.

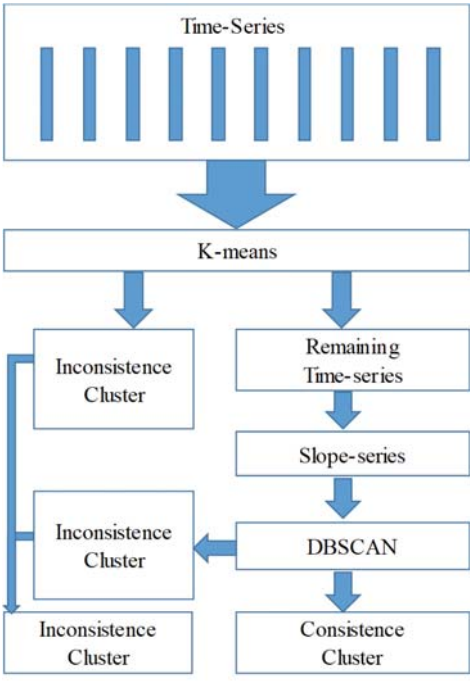

Figure 1. Hybrid clustering mode

In this model, we use K-means for pre-clustering, and give a strategy to divide the DVC time-series into 2 clusters: inconsistent cluster of K-means and the remaining time-series. Then we used an effective and simple representation approach for data processing on the remaining time-series. At last, we input processed data to DBSCAN, and divide them into inconsistent cluster of DBSCAN and consistent cluster. The inconsistent cluster of K-means and of DBSCAN are merged into inconsistent cluster. Consistent cluster consists of time-series that have similar values at the same reference time point, similar shape of curve, and similar volatility.

II. Problem Formulation

A. Unaligned Data

The collected battery time-series data consist of 9 work steps, among which DVC data can give best expression to the chemical characteristics of batteries. DVC data will decline rapidly in the initial stage, then step into a steady decline stage, and finally converge gradually. Battery data are collected from many sets of devices, and there are 512 observation points on each device, so there is no guarantee that the time-series are aligned by time points. Each of the time-series includes 800-950 observations, however, the time-series for clustering are required to be synchronized to the same reference time. Therefore, data alignment is the first problem to be addressed.

B. Limitations of K-means

The most representative partitioning clustering method is k-means. Given time-series$X = \{x_1, x_2, \ldots, x_m\}$, clusters $C = \{C_1, C_2, \ldots, C_k\}$, $center_j$ as center of cluster C_j, and $x_j \in center_j$, for $\forall x_i \notin C_j$, there is a relationship as follows:

$$dist(x_i, Center_j) > dist(x_j, Center_j)$$

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

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

Where $a_i \in a$ and $b_i \in b$.For batteries DVC data, which include a little part of time-series with volatility, it is impossible to calculate the distance accurately, and that leads to the deviation of clustering.

C. Limitations of DBSCAN

DBSCAN, as a representative of this kind of method, usually judge how to cluster an individual by its distance to an existing cluster. Given a cluster $C_i = \{x_1, x_2, \ldots, x_k\}$, and an unlabeled point x_l , $dist(x_l, C_i)$ is defined as follows:

$$dist(x_l, C_i) = min\{dist(x_l, x_i)\}$$

Where $x_i \in C_i$. Like k-means, for individual time-series with volatility, the distance from a cluster to the individual

will increase, resulting in deviation in partitioning. Another limitation is that the number of clusters is uncertain after DBSCAN, which means parameters have great implications on it.

III. Preliminaries

While collecting data with real-time database, time points and observation values are both recorded. However, the actual situation is that in different individual time-series, the recorded time points are not aligned, due to the actual network environment and different reaction time of equipment. So time-series data alignment is required to ensure the follow-up work.

For time-series, all values should be organized by the same reference time $T = t_1, t_2, \ldots, t_n$,, so data interpolation is an effective method for data alignment [10]. Specifically, c ommon i nterpolation m ethods i nclude linear interpolation, Newton interpolation, Lagrange interpolation, and so on. While selecting interpolation methods, with so many choices, the original characteristics of DVC time-series should be fully taken into consideration, to ensure that processed data will not deviate too much from the original data after interpolation. As shown in figure 2 , v alues of the great mass of time-series go down smoothly and then rapidly drop to a certain range while time accumulates. The linear interpolation, simple, with less calculations, and most important is that it barely changed trend of time-series, is chosen as the interpolation method for DVC data.

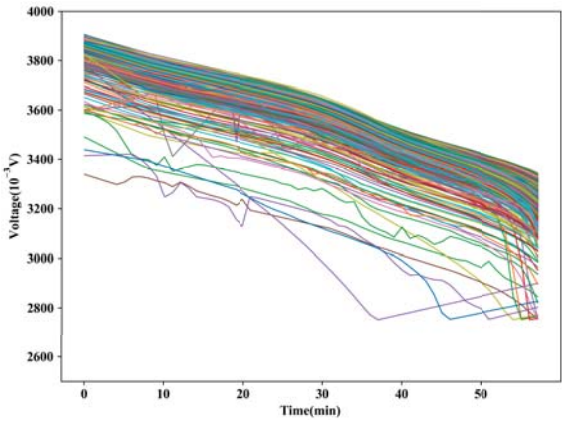

Figure 2. Discharge voltage curve

Given a raw individual discharge voltage time series values $V = \{v_1, v_2, \ldots, v_n\}$ which reference time points series is $T = \{t_1, t_2, \ldots, t_n\}$, we use linear interpolation to establish a function $V = f(T)$ to describe the relationship between V and T. It means that V is an representation of T. Suppose an adjacent interval (v_i, v_{i+1}) with reference time points interval (t_i, t_{i+1}), and $\forall t \in (t_i, t_{i+1})$, there is

$$v_t = \frac{t - t_{i+1}}{t_i - t_{i+1}} v_i - \frac{t - t_i}{t_i - t_{i+1}} v_{i+1}$$

Where v_t is a value at time $t \in (t_i, t_{i+1})$. In practice, we set time points series $T = \{1, 2, 3, \ldots, 60\}$ which ranges from 0 to 60, an get discharge voltage values $V = \{v_1, v_2, \ldots, v_60\}$ with reference time series T, by function $V = f(T)$.

IV. Methodology

The hybrid model has 3 steps. It begins with K-means algorithm for preliminary clustering, and divide time-series into several clusters. These clusters are divided into 2 parts, inconsistent cluster and the remaining time-series. Then we obtain the slope series of each individual remaining time-series, which is considered continuous derivable. At last, we use slope-series as input data of the modified DBSCAN algorithm, and then divide the slope-series into the consistent cluster and inconsistent cluster.

A. K-means for Pre-clustering

Generally, K-means algorithm is highly sensitive to the value of k and initial clustering centers. That means different k and initial cluster centers will generate different clusters. So Kd-tree is usually used as an initialization method, for a much better choice of initial clustering centers than random centers [11]. In fact, if every individual time series is presents a same shape(shape of curve, and values are ignored), it is easy to prove that the K-means clustering results are related to the means of each curve. Therefore, as a part of the hybrid model, the task of K-means is to cluster the time-series initially in this paper, and we don't care about the clustering deviation. Results are shown in figure 3.

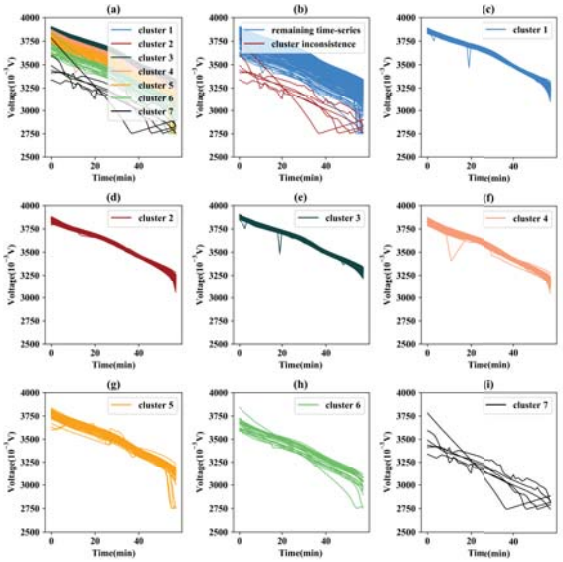

Figure 3. K-means for pre-clustering

After K-means algorithm we get inconsistent cluster and the remaining time-series. The inconsistent cluster consists of clusters that have relatively low cluster mean values

and less time-series samples. The clustering strategy is as algorithm 1.

Algorithm 1 Clustering Strategy

Input:

$Data_{raw}$: The raw DVC time-series data.

Output:

C_A , C_B :

1: $data_{aligned} = interpolation(Data_{raw}, T)$

2: $C, Centers = KMEANS(Data_{aligned})$

3: Calculate each clusters mean value,

$$Mean_k = \frac{1}{sum(Mean_k)T} \sum_{x_i \in C_k} \sum_{x_{i,j} \in x_i} x_{i,j}$$

4: $C_A = C - C_B = \{C_i | Means_i = min(Means)\}$

5: **return** C_A, C_B

Among 7 clusters, cluster 7 and cluster 6 have least samples, meanwhile, cluster 7 has minimum mean value. So, cluster 7 is chosen to compose the inconsistent cluster.

B. Slope-series Representation

There are two problems when directly using DBSCAN for time-series clustering. Firstly, we cannot always figure out how the time-series goes when time accumulate, especially while encountering data with little range volatility (In fact, that is the truth). In this case, calculating the distance between two individual time series accurately becomes difficult. Secondly, because each observation of a time point is treated as a dimension, it is very difficult to find the appropriate parameters. At the same time, even though the parameters are determined, when the reference time axis changes, the parameters will fail.

Given an individual time-series $x_{raw} = \{x_1, x_2, \ldots, x_m\}$ with reference time points $t_raw = \{t_1, t_2, \ldots, t_m\}$, we can get $x = \{x^{(1),x^{(2)},\ldots,x^{(n)}}\}$, with corresponding reference time points $t = \{t^{(1)}, t^{(2)}, \ldots, t^{(n)}\}$by linear interpolation. Then, define slope-series of x as s,

$$s = \{\frac{x^{(2)} - x^{(1)}}{t^{(2)} - t^{(1)}}, \frac{x^{(3)} - x^{(2)}}{t^{(3)} - t^{(2)}}, \ldots, \frac{x^{(n)} - x^{(n-1)}}{t^{(n)} - t^{(n-1)}}\}$$

When using linear interpolation for data alignment, we set all each individual time series with same reference time points, and for any two adjacent time points, there is a same interval length. That means,

$$t^{(k)} - t^{(k-1)} = t^{(i)} - t^{(i-1)}$$

where $2 <= k, i <= n$, therefore, the series s can be defined as,

$$s = \{x^{(2)} - x^{(1)}, x^{(3)} - x^{(2)}, \ldots, x^{(n)} - x^{(n-1)}, \}$$

In this way, we get all slope-series, $S = \{s_1, s_2, \ldots, s_m\}$, where m is the size of the whole time-series.The slope-series is shown in figure 4.

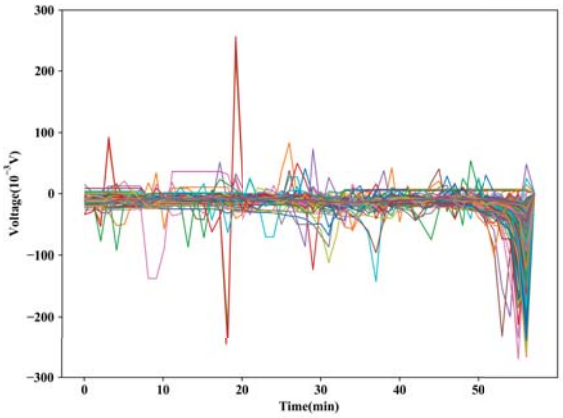

Figure 4. Slope-series

C. Modified DBSCAN for Final Clustering

The DBSCAN algorithm is based on density, and it performs well on dataset with noise. DBSCAN works with 2 important parameters: ϵ and $min - samples$. Given a set of slope-series, ϵ is the maximum distance between two samples for one to be considered as in the neighborhood of the other. This is not a maximum bound on the distances of points within a cluster. This is the most important DBSCAN parameter to choose appropriately for a data set and distance. Experiment results are shown in figure 5.

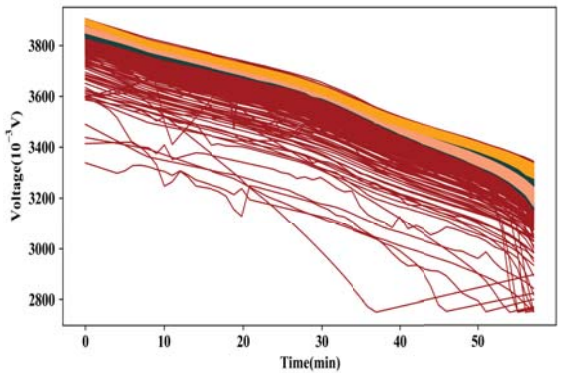

Figure 5. Results of DBSCAN

In slope-series, noise means that fluctuation o f some individuals differs from others too much. In the modified DBSCAN, we divided the slope-series in 2 clusters, C_A and C_B.We define C_A as cluster of inconsistent slope-series; C_B as cluster of consistent ones. For a set of slope-series, observation values are ignored, that means C_A consists of DVC time-series with different shapes from others in C_B. In DBSCAN, it calculates the $\epsilon - nei$ as neighborhood of a sample. If the size of $\epsilon - nei$ is too small, samples in $\epsilon - nei$ are treated as noise. Actually, size of an $\epsilon - nei$ changes dynamicly with another parameter $min - dist$. If the distance between 2 samples is greater than $min - dist$, the 2 samples

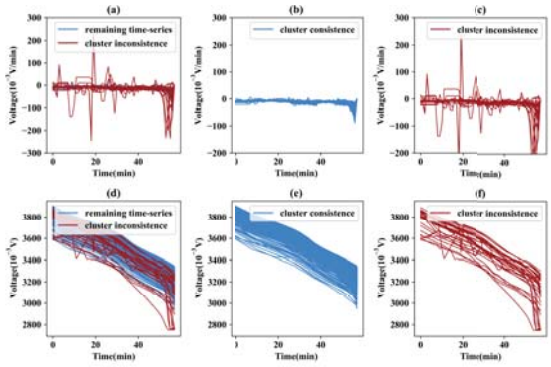

Figure 6. As a hybrid model, at first we use K-means for pre-clustering, and filter a part of inconsistent series. Then transform the remaining to slope-series as input data to the modified DBSCAN. This figure shows results of the modified DBSCAN.

can will not be in one $\epsilon-nei$. Given $\epsilon = 1$, we set a threshold for $min-dist$, to ensure that if one sample's distances from others are greater than $min - dist$, it will be clustered into C_A. At last, we set $C_A = C_{AbyK-means} \cup C_{Abym-DBSCAN}$ Results of un-modified DBSCAN are shown in figure 5, and of modified DBSCAN are shown in figure 6 . Clustering strategy is as algorithm 2.

Algorithm 2 Modified DBSCAN

Input:
$Data_{slope}$
Output:
C_A , C_B

Mark all $p \in Data_{slope}$ as unvisited
do
    Visit p, and set p visited.
    For each $p_i \in p$, calculate $\epsilon - nei$,
    and get set N_i, size $s_i = size(N_i)$.
    If
        $min(s_i) = 1$, mark p as a member of C_A.
    else
        mark p as a member of C_B.
until $visited(\forall slope \in Data_{slope}) = true$.
return C_A, C_B

V. Conclusion

In this paper, we give a hybrid clustering model for lithium-ion batteries screening ,and in this model we propose a modified DBSCAN algorithm. Output of the model is C_A and C_B, and batteries of C_B are consistent. As experiment results show that our model clearly gives the boundaries of consistent batteries and inconsistent ones. For Lithium-ion Batteries screening, it performs better than single clustering method.

Acknowledgment

This work is supported by the National Natural Science Foundation of China under Grant U1701262 and U1801263.

References

[1] K. Liu, Y. Liu, D. Lin, A. Pei, and Y. Cui, "Materials for lithium-ion battery safety," *Science advances*, vol. 4, no. 6, p. eaas9820, 2018.

[2] X. Feng, M. Ouyang, X. Liu, L. Lu, Y. Xia, and X. He, "Thermal runaway mechanism of lithium ion battery for electric vehicles: A review," *Energy Storage Materials*, vol. 10, pp. 246–267, 2018.

[3] T. W. Liao, "Clustering of time series data—a survey," *Pattern recognition*, vol. 38, no. 11, pp. 1857–1874, 2005.

[4] K. Turcheniuk, D. Bondarev, V. Singhal, and G. Yushin, "Ten years left to redesign lithium-ion batteries," 2018.

[5] C.-Y. Wang, G. Zhang, S. Ge, T. Xu, Y. Ji, X.-G. Yang, and Y. Leng, "Lithium-ion battery structure that self-heats at low temperatures," *Nature*, vol. 529, no. 7587, p. 515, 2016.

[6] A. Manthiram, "An outlook on lithium ion battery technology," *ACS central science*, vol. 3, no. 10, pp. 1063–1069, 2017.

[7] S. Aghabozorgi, A. S. Shirkhorshidi, and T. Y. Wah, "Time-series clustering–a decade review," *Information Systems*, vol. 53, pp. 16–38, 2015.

[8] H. Izakian and W. Pedrycz, "Anomaly detection in time series data using a fuzzy c-means clustering," in *2013 Joint IFSA World Congress and NAFIPS Annual Meeting (IFSA/NAFIPS)*. IEEE, 2013, pp. 1513–1518.

[9] J. Paparrizos and L. Gravano, "Fast and accurate time-series clustering," *ACM Transactions on Database Systems (TODS)*, vol. 42, no. 2, p. 8, 2017.

[10] P. Upchurch, J. Gardner, G. Pleiss, R. Pless, N. Snavely, K. Bala, and K. Weinberger, "Deep feature interpolation for image content changes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7064–7073.

[11] X.-y. Zhang, Q. Shen, H.-y. Gao, Z. Zhao, and S. Ci, "A density-based method for initializing the k-means clustering algorithm," in *Proceedings of International Conference on Network and Computational Intelligence (ICNCI 2012), IPCSIT*, vol. 46, 2012, pp. 46–53.