Research on feature extraction method of power grid online data based on big data

Ji lin Chen1*, Yanhao Huang1, Weijiang Qiu1, Xinglei Chen1, Ning An1, and Kaiyu Zhang2

1 China Electric Power Research Institute, Beijing, East Road of Xiaoying 15#
2 State Grid ShangHai Municipal Electric Power Company, 1122 Yuanshen Road, Pudong New Area, Shanghai
*Corresponding author’s e-mail: chenjilin@epri.sgcc.com.cn

Abstract. The development of power simulation artificial intelligence technology needs massive open sample data. It is a trend to use the characteristics of online data to construct sample data. In order to solve the problem that the online data of power grid is rich in information, but the utilization rate of features is not high, aiming at the information features of generators, the LTTB dimension reduction and DBSCAN + L2 clustering methods are proposed, which reduce the complexity of feature extraction of time series data. The method is verified by the actual power grid data, and has achieved certain results.

1. Introduction
With the wide application of online security analysis and early warning system, online calculation data and stable results form a set of cross-section data every 5 minutes, which is about 10M, and the data of a year is about 1T. Massive online data is rich in information, but the utilization rate of data features is not high. In this paper, in order to solve the problem of rich online data and low utilization rate of features in power grid, different dimension reduction and clustering methods are proposed for feature extraction of generator and load in online data. Because online data of power grid is time series data, online data feature extraction is feature extraction of time series data. Firstly, the common technical methods of data dimension reduction are introduced, and the advantages and disadvantages of different dimension reduction methods are analyzed; Secondly, this paper introduces the common technical methods of time series data clustering, analyzes the advantages and disadvantages of different clustering methods, and finally it achieves preliminary results by actual power grid data[1].

2. Dimension reduction of time series data
Time series data is the distributed data in time domain. The commonly used time series data dimension reduction technologies include PAA (Piecewise aggregate approximation), LTTB (Largest Triangle Three Buckets) etc[2].

2.1. PAA
The sliding window is used to reduce the dimension of time series. The mean value method is used to aggregate the time series in each window. The calculated mean value is used to represent a series of values in a window.
PAA calculation: set time series dimension reduction to, where the results are as follows:

$$\bar{x}_j = \frac{m}{n} \sum_{j=(i-1)+1}^{n} x_j$$  

(1)

2.2. LTTB

LTTB is based on the maximum effective area of time series data dimension reduction and fitting. The three segment maximum triangle algorithm divides the original time series data into equal segments (buckets). For each segment, the algorithm selects the most important point (denoted as f point) to represent all the points in the current segment, so as to achieve the effect of dimension reduction. PAA uses the mean method to select f points, and lttb algorithm selects f points based on the maximum effective area (MEA). EA (effect area) of a point is defined as the area of the triangle formed by two adjacent points. Just as figure 2.

3. Clustering method for time series data

Clustering pattern is an abstract algorithm framework. Clustering pattern algorithm obtains the similarity between samples. Through comparison, the sample groups with higher similarity are divided into one category and iterated repeatedly until all the categories are divided, so as to complete the clustering task. Therefore, the clustering process can be divided into the combination of similarity calculation and clustering mode operation.
3.1. K-means clustering
In the data set, k points are selected as the initial center of each cluster according to a certain strategy, and then the remaining data is observed, and the data is divided into the clusters nearest to the k points. That is to say, the data is divided into K clusters to complete a partition, but the new cluster formed is not necessarily the best partition. Therefore, in the new cluster generated, the center point of each cluster is recalculated, then the partition is repeated until the result of each partition remains unchanged[3].

3.2. DBSCAN (Density-Based Spatial Clustering of Applications with Noise) clustering algorithm
DBSCAN is a typical density based clustering algorithm with noise. Compared with K-means, DBSCAN can be applied to both convex and non convex samples, it receives two parameters: neighborhood threshold and minpts, which are the minimum number of points needed to form core points (high-density regions / clustering centers). It is advisable to set up a set of unaccepted points $P = \{p_1, \ldots, p_n\}$, The algorithm performs alternately according to the following two steps, find and broadcast.

3.3. Distance calculation
Similarity reflects the degree of similarity between samples. In the field of clustering, the concept of distance is usually used to represent the dissimilarity between samples. The distance of sample points in space can be calculated in different ways. Distance is inversely proportional to similarity, and zero distance means the same sample, the distance calculation are $L_1$-Norm (Manhattan distance, (2)) and $L_2$-Norm (Euclidean distance, (3)).

$$d(a, b) = \sum_{i=1}^{k} |x_{ai} - x_{bi}|$$

(2)

$$d(p, q) = d(q, p) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

(3)

4. The verification of real power grid data
4.1. The Clustering methods verification of real power grid data
Based on the QS data of 40000 current nodes of the state grid dispatching, the data of DC voltage (VDC) and AC active power (PAC) in North China are extracted for dimension reduction of PAA and LTTB respectively. The dimension reduction scale will be nearly 100 times reduced dimension and nearly 200 times dimension reduction dimension, and the comparison and analysis will be made. In the experiment, the annual data of DC voltage and AC active power are nearly 80000-dimensional, and the lttb algorithm and PAA algorithm are used to reduce to 365 and 1095 dimensions. The VDC results are as follows[4]:

\[\text{VDC results are as follows[4]}:\]
PAA and LLTB are used to reduce dimensions by nearly 200 times, 100 times, 30 times, 20 times and 10 times respectively. The average loss degree of PAA and LLTB is obtained by repeated experiments, and the following Table 1 is obtained (the smaller the average loss degree, the better)[5,6]. Practice has proved that LLTB algorithm can capture the overall and detailed characteristics of DC line voltage and power.

| The multiple | ×200  | ×100  | ×30   | ×20   | ×10   |
|--------------|-------|-------|-------|-------|-------|
| PAA          | 29.680| 16.335| 6.521 | 4.197 | 1.886 |
| LLTB         | 29.593| 13.905| 3.136 | 1.835 | 0.773 |

4.2. The Dimension reduction verification of real power grid data

Based on more than 10000 load data of 40 000 nodes in North China, after dimension reduction of time series data and before clustering, we should analyze the selection strategy of super parameters. In the process of clustering, different clustering patterns are used to combine different similarity
measurement methods for clustering. Finally, the clustering results are compared to test the clustering tasks of different combination methods, 500 random sampling time series curves are used for elbow rule measurement, and the best K value of each combination is selected for the experiment, and the contour coefficient corresponding to the best cohesion is recorded. The contour coefficient table (generally more than 0.5 means good clustering) is shown in the following table2.

It can be seen from the table that the effect of DBSCAN + L2 is better than that of other combinations[7,8]. Taking 10495 load curves as an example, time series curves with different change patterns (original load data, black) are divided, and characteristic curves (cluster center, red) are extracted from these similar time series curves to represent the overall change. The effect of DBSCAN + L2 mode is shown in the figure 5.

![Figure 5. The Dimension reduction of DBSCAN+L2](image)

|            | L1       | L2       |
|------------|----------|----------|
| K-means    | 0.312    | 0.5013   |
| DBSCAN     | 0.4258   | 0.6093   |

5. Conclusion
In recent years, with the gradual development of artificial intelligence technology in the field of power grid simulation analysis, the demand for sample data is more and more urgent. How to use online data to make sample data. Aiming at the problem of feature extraction of massive online data, a feature extraction method of online data based on Dimension Reduction Clustering of time series data is proposed. Different dimension reduction and clustering methods are used to verify the grid data, and the preliminary results are obtained. In the future, we can use the feature extraction method in this paper to construct a virtual grid, and use the structure of the virtual grid and online data features to
generate massive simulation samples, which provides data support for the application of artificial intelligence technology in power system. However, there are still some shortcomings. Before dimension reduction and clustering, the processing of missing data needs to be explored. In the next step, more clustering and dimension reduction methods can be used to deal with different characteristics of online data.

Acknowledgments
This work is supported by the Science and Technology Project of State Grid: Research on Key Technologies of Power Grid Simulation Test Examples Generation.

References
[1] LIU Li, WANG Gang, ZHAI Denghui. Application of k-means clustering algorithm in load curve classification[J]. Power System Protection and Control, 2011, 39(23): 65-68.
[2] SU Haozun. Research on power load classification based on K-means clustering algorithm[J]. Electronics World, 2017(23): 78-79.
[3] SU Haozun. Research on power load classification based on K-means clustering algorithm[J]. Electronics World, 2017(23): 78-79.
[4] HU Feng, ZHU Chengzhi, WANG Zhihua. Research on the power load classification based on improved K-means algorithm[J]. Electronic Measurement Technology, 2018, 41(12): 44-48.
[5] Zhang Rui, Du Tao, Qu Shouning et al. Adaptive density-based clustering algorithm with shared KNN conflict game[J]. Information Sciences, 2021, 565.
[6] Bin Xiao, Chang Liu. A Multi-angle Improved Small Sample Clustering Algorithm[J]. Journal of Physics: Conference Series, 2021, 1848(1).
[7] LIAN Zikuan, YAO Li, LIU Shengyuan, et al. Phase and meter box identification for single-phase users based on t-SNE dimension reduction and BIRCH clustering[J]. Automation of Electric Power Systems, 2020, 44(8): 176-184.
[8] TANG Zeyang, ZHOU Kunpeng, CAO Kan, et al. Substation area topology verification method based on distribution network operation data[J]. High Voltage Engineering, 2018, 44(4): 1059-1068.