ISSN 2616-5767 Vol.6, Issue 7: 11-17, DOI: 10.25236/AJETS.2023.060703

Spatial Distance-based Uniformity Evaluation Method for Power Flow Datasets

Meng Xianbo, Li Yalou, Wang Zigan, Hu Shanhua

China Electric Power Research Institute, Haidian District, Beijing, China

Abstract: *Research on the application of artificial intelligence (AI) methods in power grid analysis has been widely conducted. The power flow dataset required for the training of AI models is not uniformly distributed, and to improve the quality of the dataset, related research has generally applied tuning methods that tend to have a more uniform sample distribution, resulting in improved performance of AI models. However, the lack of research on methods to evaluate the uniformity of distribution of the dataset poses an obstacle to the judgment of the validity of uniformity adjustment and the analysis of the impact of uniformly distributed datasets on the performance of AI models. This paper proposes a method to evaluate the uniformity of power flow datasets by using the calculation of distances in the high-dimensional feature space of the flow datasets and plotting the minimum distance statistics as a way to evaluate the uniformity of the flow datasets. It was also tested on the power flow dataset of 36-node grid and evaluated separately for two uniformity levels, which can clearly represent the uniformity of distribution of the dataset in the high-dimensional feature space.*

Keywords: *power flow dataset; high-dimensional feature space; uniformity evaluation*

1. Introduction

At present, research on the application of artificial intelligence technology in power system analysis has been gradually carried out [1-3]. Since the performance of AI models is highly dependent on the quality of the training dataset [4], to improve the performance of AI techniques in grid analysis applications, researchers have conducted research on methods for generating, tuning, and expanding datasets [5-10]. The existing power flow dataset suffers from the problem that there are more similar samples for stable operation and fewer samples for extreme operation states, a situation similar to the sample class imbalance problem [11,12]. Therefore, the trend of adjustment of the dataset is all about eliminating some similar samples and supplementing the samples of rare operating states to meet the requirements of uniformity and adequacy of the power flow dataset [13]. The adjusted dataset does improve the performance of the AI model, but the state of the dataset itself before and after the adjustment is not clear due to the lack of evaluation methods for the characteristics of the power flow dataset. Therefore, it is impossible to judge whether the adjustment method really achieves the adjustment of the corresponding distribution characteristics, and whether the corresponding distribution characteristics dataset is beneficial to the performance of the AI model.

Uniformity of distribution of fower flow datasets is a common requirement, but there are difficulties in evaluating uniformity in high-dimensional feature spaces, and some methods that work in low-dimensional spaces are difficult to generalize to high-dimensional spaces. As in computer graphics, the uniformity of distribution of point sets is evaluated using spectral analysis [14-16], which allows the uniformity of point sets to be clearly observed in the graph of amplitude-frequency characteristics. However, the Fourier transform required for spectral analysis is difficult to implement in high-dimensional space.

In view of this, this paper investigates a distance-based method for evaluating the uniformity of power flow datasets, which treats power flow datasets as point sets in a high-dimensional feature space, and achieves the evaluation of distribution uniformity by using the relationship between the distance between points. This paper is organized as follows: Section 2 introduces the content and format of the grid power flow dataset, which is the basis for the subsequent discussion of uniformity in high-dimensional feature space; Section 3 introduces the distribution uniformity evaluation method of the power flow dataset, and verifies the effectiveness of the method by the uniformity evaluation results of the power flow dataset of the CEPRI 36-node grid model with different distribution cases in Section 4.

ISSN 2616-5767 Vol.6, Issue 7: 11-17, DOI: 10.25236/AJETS.2023.060703

2. Power Flow Dataset

2.1. Contents of Power Flow Dataset

The power flow dataset consists of ***n*** power flow samples, each sample represents a corresponding power system state. For a given power system state, often only part of the physical quantities is known, i.e., the definite solution conditions in the power flow calculation, and the other power flow state quantities are to be found, and the known physical quantities do not reflect the full picture of the power flow. If the power flow calculation converges, the complete power flow state data can be calculated.

An operating point of a power system in a certain steady state can be represented by the combination of power and voltage of all buses when the system is in steady state.

$$\{\{P_i,Q_i,U_i,\theta_i\}|i=0,1,\ldots,N\}_{N\times 4} \tag{1}$$

Or complex numbers form:

$$\{\{S_i,\dot{U}_i\}|i=0,1,\ldots,N\}_{N\times 2} \tag{2}$$

Where N is the number of active buses of the grid, including generators and loads. The relationship between the above variables is given by a set of nonlinear power flow equations, so that only the known condition in the power flow calculation can also represent an operating point of the power system. Usually this known condition consists of the active power P and reactive power Q at the PQ node, the active power P and voltage magnitude V at the PV node and the voltage magnitude and phase at the balance node, so that the amount of data to be stored is reduced by half and the real number The representation is changed from an $N \times 4$ matrix to an $N \times 2$ matrix. This representation may suffer from a non-convergence of the power flow calculation compared to the complete combination, which is needed for some tasks related to power flow convergence discrimination or adjustment.

2.2. Format of Power Flow Dataset

To match the dataset format of the data-driven method, it also needs to be adjusted on top of the description form. The datasets required for the data-driven approach can be divided into labeled and unlabeled datasets, matching supervised learning tasks and unsupervised learning tasks, respectively. If part of the dataset is labeled and the other part is unlabeled, such datasets match semi-supervised learning. Since there is a mature grid simulation software as the basis, it is easy to obtain the corresponding labeled information for the tasks when the data of the grid state is known, so this paper focuses on the labeled dataset. The labeled dataset takes the form of a feature-label pair, where the feature $\boldsymbol{x}$ is usually represented as a d-dimensional row vector, and is typically written as follows

$$\boldsymbol{x}=(x_1,\ldots,x_d), \tag{3}$$

and can be abbreviated to $(x_i)_{1\le i\le d}$. The label y is usually an element of a finite set $\mathcal{Y}$. The value of y can also be expressed as an integer when encoded by its index in the elements of $\mathcal{Y}$. A sample in the form of a feature-label pair can be represented as

$$s:=\{(\boldsymbol{x},y)|\boldsymbol{x}\in\mathcal{X},y\in\mathcal{Y}\} \quad \text{or} \quad \{((x_i)_{1\le i\le d},y)|x_i\in\mathbb{R},y\in\mathbb{R}\} \tag{4}$$

where $\mathcal{X}\subseteq\mathbb{R}^d$ is the feature space when the constraints of $\boldsymbol{x}$ are not considered $\mathcal{X}=\mathbb{R}^d$. In practice, the feature vector $\boldsymbol{x}$ is constrained by physical properties, such as the upper and lower limits of node power and node voltage, resulting in the feature space $\mathcal{X}$ as a high-dimensional finite space. To eliminate the effect of dimensional differences between the eigenvalues, the data are also usually normalized, i.e., x_i is mapped linearly between $[0,1]$, resulting in a hypercube with side length 1. The discussion of the data distribution in this paper is after normalization, so the feature space $\mathcal{X}$ is also the feature space of the unit hypercube. The $N \times 2$ matrix of inputs to the power flow equation is expanded into a d-dimensional row vector in a fixed order of node placement, at which point the grid state is presented in the standard sample format (feature-label pair) in labeled data, where the dimension d of the feature vector takes a value equal to twice the number of nodes N.

ISSN 2616-5767 Vol.6, Issue 7: 11-17, DOI: 10.25236/AJETS.2023.060703

With the above settings, the definition of a power flow sample is given as follows:

Definition 1 (power flow sample): a sample containing known conditions for power flow calculation at a run point, expressed in the form of feature-label pairs, is called a power flow sample.

The training samples required for the specific task are formed by adapting feature $\boldsymbol{x}$ to the specific grid analysis task and assigning the corresponding physically meaningful values to label *y*. The resulting samples remain in a consistent format.

Definition 2 (Power flow dataset): A dataset consisting of power flow samples is called a power flow dataset and can also be expressed as a pairwise combination of the form of the feature matrix $\boldsymbol{X}$ of $n \times d$ and the label matrix $\boldsymbol{Y}$ of $n \times 1$.

$$D_{\text{flow}} := \{s_j\}_{1 \le j \le n} = \{\boldsymbol{X}_{n \times d}, \boldsymbol{Y}_{n \times 1}\} \tag{5}$$

By definition, a power flow dataset is a set of labeled data sets, where the features represent the grid operating points and the labels represent the state information that this operating point has in a specific task, and the format is also consistent with supervised learning. If the dataset feature matrix $\boldsymbol{X}$ is considered as the coordinates of *n* points on a *d*-dimensional feature space and these points are plotted in the feature space, a point cloud will be formed. This perspective transformation gives the dataset geometric properties, and the associated geometric properties are only related to $\boldsymbol{X}$, which makes the labeled dataset consistent with the unlabeled dataset in subsequent modeling.

3. Analysis of Blue Noise Distribution Evaluation Method

3.1. Blue Noise Distribution

Uniformity sampling of the sample space is a basic method to ensure comprehensive coverage of the data set. In computer graphics, uniformity sampling has been studied in more depth, and the blue noise distribution is recognized as the best distribution characteristic to cover the sampling domain to the maximum extent.

3.2. Blue Noise Distribution Evaluation Method

The power spectral density of noise (the spectral distribution of power) is often used to distinguish between different types of noise. In fields such as acoustics and physics, this classification of noise is often given a different "color" designation for different power spectral densities, i.e., different types of noise are named different colors. The color classification of noise comes from a formal analogy between a noise spectral density function in the frequency domain and a light wave signal in the frequency domain, i.e., if a light wave has the same spectral density pattern as blue noise in the frequency domain, the light wave will appear blue, and so on.

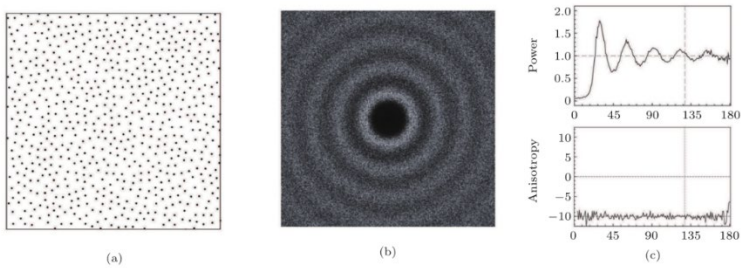

Figure 1: Example of blue noise point set and its spectrum analysis.

In computer graphics, the concept of blue noise is sometimes also used generically to refer to any noise that has a minimal low-frequency component and no significant peaks appear in the spectrum. As shown in Figure 1, (a) is the sampled blue noise point set, (b) is the power spectrum of the point set, and (c) is a further processing of (b) for the radial mean and normal anisotropy of the power spectrum. The upper plot in figure (c) shows that the power spectrum of the planar point set has a small low-frequency component and a large high-frequency component, which is consistent with the blue-noise characteristic

ISSN 2616-5767 Vol.6, Issue 7: 11-17, DOI: 10.25236/AJETS.2023.060703

[17]. The blue noise characteristic is determined by the power spectrum function curve in the frequency domain and is found to be related to the uniformity and homogeneity of the distribution of the planar point set, i.e., whether the spectral curve of the planar point set conforms to the blue noise characteristic indicates whether the distribution of the point set is uniform, and the degree of conformity of the spectral characteristic to the blue noise characteristic also reflects the homogeneity of the point set.

3.3. Blue Noise Sampling Versus Random Sampling

Many studies default random sampling as a kind of uniformity sampling, in fact, pure randomness does not bring the most desirable results. As shown in Figure 2, under uniform distribution, the left panel shows the set of points formed by blue-noise sampling, and the right panel shows the set of points formed by random sampling. Sampling points tend to be chaotic and leave blank space in the region, while blue-noise sampling will make the sample points as uniform as possible. If each sample point can represent the information in a certain surrounding area, the blue noise distribution can cover a larger feature space, while the distribution formed by random sampling shows the characteristics of some regions with blank space and others with denser points.

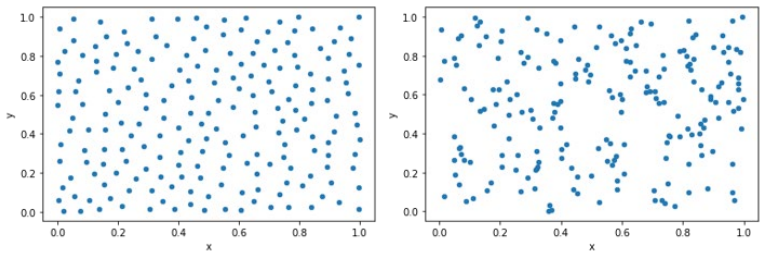

Figure 2: Blue noise sampling and stochastic sampling in the two-dimensional plane

The online data of power flow presents a more non-uniform distribution characteristic, and a large number of duplicate and similar samples are represented in the feature space as a dense cloud of points in a region, while there are no non-converging samples in the online data, that is, the non-converging region in the feature space is blank. The concept of high-dimensional blue noise characteristic was proposed in the literature [18]. Since the number of configurable samples in the feature space is much larger than the usually used sample set capacity as the dimensionality grows, forming a significant sparsity, the high-dimensional blue noise characteristic cannot cover the entire feature space uniformly and can only ensure that no two sample points are too close to each other. The high-dimensional blue-noise distribution is still a distribution characteristic that covers the largest range of feature space with the same sample capacity. Therefore, it is still important to evaluate the uniformity of the sample distribution in the high-dimensional feature space.

4. A Method for Evaluating the Uniformity of Power Flow Datasets

4.1. Distance in High-dimensional Feature Space

The distance between the sample input features $\boldsymbol{x}$ in vector form, such as sample $\boldsymbol{x}_i[x_{i1}, x_{i2}, ..., x_{in}]$ and sample $\boldsymbol{x}_j[x_{j1}, x_{j2}, ..., x_{jn}]$ is calculated as

$$d_{ij} = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2} \tag{6}$$

Therefore, for the sample point $\boldsymbol{x}_i[x_{i1}, x_{i2}, ..., x_{in}]$, the distance from its minimum point can be expressed as

$$d_{i-\min} = \min\{d_{ij}, j \neq i\} \tag{7}$$

The n-dimensional array $\boldsymbol{d}_{\min}$ of the minimum distance of the sample points is obtained when i is taken over all the points in the point set, and the uniformity of the sample can be judged by counting the number of $\boldsymbol{d}_{\min}$ in each of the taken segments.

ISSN 2616-5767 Vol.6, Issue 7: 11-17, DOI: 10.25236/AJETS.2023.060703

4.2. Method Flow

According to the definition of the blue noise distribution in the literature [18], the statistic of the minimum distance between sample points and points is used as the evaluation index.

The specific steps are as follows:

(1) Calculate the distance d_{ij} between two points in the sample set according to Equation(6), where $i \neq j$;

(2) Calculate the minimum distance between each point and other sample points in the data set, i.e., the distance from each point to the nearest adjacent point $d_{i-\min}$, according to Equation(7), and calculate the $d_{i-\min}$ of all sample points to form an array $\boldsymbol{d}_{\min}$.

(3) Statistical analysis of $\boldsymbol{d}_{\min}$ is performed to form a histogram.

When the statistic of $\boldsymbol{d}_{\min}$ forms a peak at a small distance value, it indicates that there are a large number of sample points clustered together, or a large number of similar or repeated samples, then the uniformity is poor; when the statistic shows a Gaussian distribution and the peak is high, it indicates that each sample point stays close to the nearest sample point, i.e., the uniformity is good.

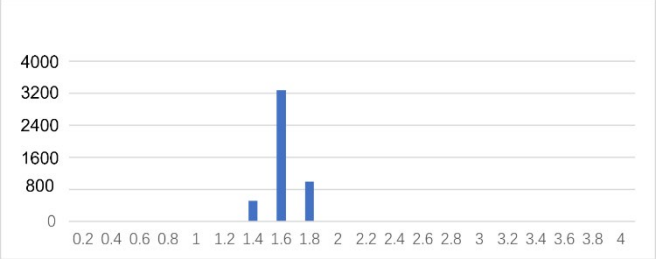

Figure 3: Example of uniformity evaluation results graph

As shown in Figure 3, the results will be presented in the form of such a bar chart, with the horizontal coordinate indicating the value of $\boldsymbol{d}_{\min}$ and the vertical coordinate indicating the number of sample points in this interval. The results shown in the figure are that most of the points are concentrated within a relatively small interval from their nearest neighbors, and the data set has a good uniformity.

4.3. Algorithm Implementation Considering Performance

In the implementation of the program, calculating the minimum distance between a sample and all other sample points is a computationally intensive task. In this paper, we use kd-tree to implement the minimum distance $\boldsymbol{d}_{\min}$ to improve the computational efficiency. Kd-tree, as a data structure for accessing high-dimensional data, is highly efficient in static queries and is commonly used to achieve fast k-nearest neighbor search of data, and the function implemented in this paper using kd-tree is to quickly search the distance of the nearest point.

5. Experimental Validation

5.1. Example Introduction

The samples in the power flow dataset of this paper describe various modes of operation of the grid model CEPRI36, and the grid structure is shown in Figure 4, where some nodes are connected to capacitors or reactors that are not involved in regulation, and there are 18 nodes of generating units or loads involved in regulation, with the nodes injecting power as the input feature values, for a total of 36 variables, i.e., the sample contains a feature dimension of 36 dimensions.

ISSN 2616-5767 Vol.6, Issue 7: 11-17, DOI: 10.25236/AJETS.2023.060703

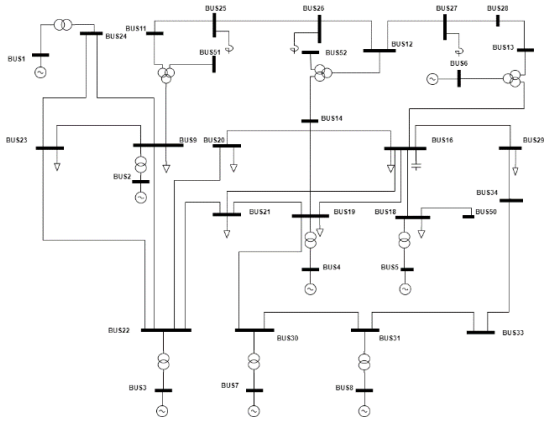

Figure 4: CEPRI36 grid model topology connection diagram

In order to demonstrate the results of the evaluation of the dataset with or without uniformity by the method in this paper, two datasets were generated using different methods, respectively. Dataset 1 with uniformity is a blue-noise dataset generated using the method in the literature [18]; Dataset 2 without uniformity is designed with reference to the distribution characteristics of the dataset in actual runs, where the dataset has a high sample similarity and a low number of sample points for the extreme run cases. Both datasets have 30,000 samples.

5.2. Results and Discussion

The experimental results are shown in Figure 5. The upper bar graph is the minimum distance statistics of the blue noise data set, while the lower bar graph is the minimum distance statistics of the inhomogeneous data set made as a comparison group. The horizontal coordinates are the different values of the minimum distances, and the vertical coordinates are the number of samples whose distances from the nearest neighbors fall in the corresponding value interval.

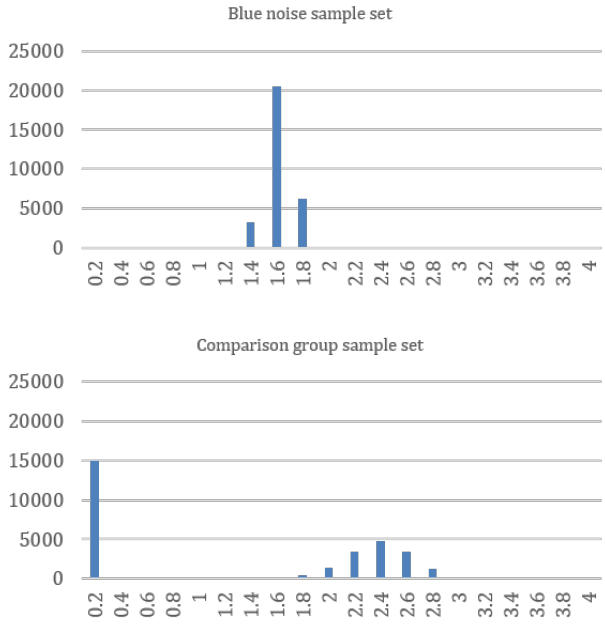

Figure 5: Experimental results graph

The blue noise dataset has better uniformity, so the minimum distance statistic plot presents a Gaussian distribution with higher peaks, and higher peaks indicate a more uniform distribution of the dataset. In contrast, the inhomogeneity of the comparison group is reflected by a part of similar samples gathered together, and other samples are more sparsely distributed, which is presented as two peaks in the minimum distance statistical graph, the peak in the range of 0~0.2 corresponds to a large number of similar samples, while the second more gentle peak corresponds to other samples, due to the existence of a large number of similar samples, the minimum distance of other samples is also larger and the

ISSN 2616-5767 Vol.6, Issue 7: 11-17, DOI: 10.25236/AJETS.2023.060703

distribution is more sparse.

6. Conclusions

This paper presents a method for evaluating the distribution uniformity of a power flow dataset. The method treats the power flow dataset as a set of points in a high-dimensional space and achieves the distribution uniformity assessment by using the relationship between the distance between points. By testing different distribution datasets, the uniformity of the dataset can be clearly reflected, which is of positive significance for the subsequent research on the need of generating power flow datasets and the influence of power flow dataset distribution characteristics on AI models.

Acknowledgements

This work was supported by State Grid Corporation of China Science & Technology Project - Research on rapid exploration and learning method of key information in high proportion new energy power system - under Grant 5100-202155429A-0-0-00.

References

[1] Wang Z, Zhou Y, Guo Q, and Sun H (2021). Transient Stability Assessment of Power System Considering Topological Change: a Message Passing Neural Network-based Approach. Proceedings of the CSEE 44.07: 2341-2350(in Chinese).
[2] Su T, Liu Y, Shen X, et al (2020). Deep Learning-driven Evolutionary Algorithm for Preventive Control of Power System Transient Stability[J] Proc. CSEE, 40.12. (in Chinese)
[3] Shi Z, Yao W, Zeng L, et al (2020). Convolutional neural network-based power system transient stability assessment and instability mode prediction[J]. Applied Energy, 263: 114586.
[4] Sun C, Shrivastava A, Singh S, et al (2017). Revisiting unreasonable effectiveness of data in deep learning era[C]. Proceedings of the IEEE international conference on computer vision. 843-852.
[5] Chen J., Chen Y., Tian F., Guo Z., and Li T. (2019). The Method of Sample Generation for Power Grid Simulation Based on LSTM. Proc. CSEE, 39, 4129-4134.(in Chinese)
[6] Tan B, Yang J, Lai Q, Xie P, Li J, and Xu J. (2019) Data augment method for power system transient stability assessment based on improved conditional generative adversarial network. Automation of Electric Power Systems 43.1: 149-157(in Chinese).
[7] Batista G E A P A, Prati R C, Monard M C (2004). A study of the behavior of several methods for balancing machine learning training data [J]. ACM SIGKDD Explorations Newsletter, 6(1): 20-29.
[8] Chawla N V, Bowyer K W, Hall L O, et al (2002). SMOTE: synthetic minority over-sampling technique [J]. Journal of artificial intelligence research, 16: 321-357.
[9] Zhang H, Cisse M, Dauphin Y N, et al (2017). Mixup: Beyond empirical risk minimization [J]. arXiv preprint arXiv:1710.09412.
[10] He H, Bai Y, Garcia E A, et al. ADASYN: Adaptive synthetic sampling approach for imbalanced learning [C]. 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence). IEEE, 2008: 1322-1328.
[11] Japkowicz N, Stephen S (2002). The class imbalance problem: A systematic study[J]. Intelligent data analysis, 6(5): 429-449.
[12] He H, Garcia E A (2009). Learning from imbalanced data[J]. IEEE Transactions on knowledge and data engineering, 21(9): 1263-1284.
[13] Zhang Y, Zhang H, Li C, and Pu T. (2021) Review on deep learning applications in power system frequency analysis and control. Proceedings of the CSEE 41.10: 3392-3406+3665(in Chinese).
[14] Dippe M A, Wold E H. (1985) Antialiasing through stochastic sampling [C]. Proceedings of the 12th annual conference on Computer graphics and interactive techniques. 69-78.
[15] Cook R L (1986). Stochastic sampling in computer graphics[J]. ACM Transactions on Graphics, 5(1): 51-72.
[16] Yuksel C. (2015) Sample elimination for generating Poisson disk sample sets[C]. Computer Graphics Forum: Vol. 34. Wiley Online Library, 25-32.
[17] Yan D M, Guo J W, Wang B, et al (2015). A Survey of Blue-Noise Sampling and Its Applications [J]. Journal of Computer Science and Technology, 30(3): 439-452.
[18] Meng X, Li Y, Shi D, et al (2022). A Method of Power Flow Database Generation Base on Weighted Sample Elimination Algorithm[J]. Frontiers in Energy Research, 10: 919842.