A dimensionality reduction method of power load data based on the combination of VMD-OMP-Kmeans

Li Tong¹, Song Jinliang¹, Chen Jian¹, Tong Haosong¹, Cui Jia²*, Wang Anni³

¹ Electric Power Research Institute State Grid Liaoning Electric Power Co, Ltd.
Shenyang, Liaoning, 110870, China
² The school of electrical engineering Shenyang University of Technology, Shenyang, Liaoning, 110870, China
³ Shenyang Power Supply Company of State Grid Liaoning Electric Power Co., Ltd., Shenyang 110031, Liaoning Province, China;
*cujiayouxiang@sut.edu.cn

Abstract: With the power data becoming more and more complex, it is of great significance to reduce the dimension and reconstruct the power data for the subsequent processing and application of power big data. This paper presents an algorithm based on variational mode decomposition (VMD) decomposition and orthogonal matching pursuit (OMP) reconstruction to reduce the dimension of power load data. Firstly, VMD method is used to decompose and filter the power load data; Secondly, K-means clustering algorithm is used to synthesize the decomposed data mode into a cluster center data set. Then, the data set is used as a dictionary; Thirdly, an OMP algorithm is proposed to select the elements to be reconstructed in the dictionary. And the corresponding data weights are given to reconstruct the load data with high-precision. As a result, the dimensionality of massive load data is reduced automatically. Finally, this paper builds a flexible load data dimensionality reduction model based on VMD method to verify the proposed method.

1. Introduction
Data reduction and reconstruction can reduce the cost of data acquisition, greatly improve the calculation accuracy of data classification, which is of great significance to the operation of power system automation. The research on dimensionality reduction of power data is still in the development stage in China, and the research results are not mature. Reference [1] proposed the idea of reducing the dimension of big data in smart grid for technical problems such as large amount of data, high dimension and difficult to identify, and designed a smart grid big data management system based on random forest algorithm to reduce the big data in smart grid from high latitude to low latitude. But the accuracy of data processing is relatively low. In the environment of power big data, the research of applying OMP algorithm to power data reconstruction is just rising. In reference [2], OMP algorithm was used to optimize the image reconstruction, and better image reconstruction effect was obtained, and the resolution and quality were improved. This paper introduces a compressed sensing algorithm based on matching pursuit. The advantages of OMP algorithm in data reconstruction and optimization processing can well explain the feasibility of combining the VMD based dimensionality reduction method with OMP algorithm.

Through the analysis of domestic and foreign literature, the research on power load dimensionality
reduction and reconstruction based on VMD method and OMP algorithm is still in the exploratory stage. In this paper, VMD method and OMP method are combined for the first time, and a dimensionality reduction reconstruction algorithm with high efficiency and strong versatility is obtained. The algorithm is not easy to be affected by noise and has high stability, and is more suitable for the power load big data processing derived from the increasingly prosperous information age.

2. The research on data dimension reduction reconstruction and clustering algorithm

2.1 Load decomposition based on VMD algorithm

VMD is a new adaptive and accurate orthogonal frequency decomposition method. It can decompose the central mode of a redundant component signal into several intrinsic modes in a limited bandwidth. The VMD decomposition process mainly includes the decomposition and constraint of variables. The specific operation process of establishing the constraint mode problem of variational function is described as follows:

i) The analytical signal associated with each mode is calculated by the following formula:

\[ \left( \delta(t)+\frac{j\pi t}{\tau} \right) * u_k(t) \]  

(1)

ii) The center frequency of each mode is estimated by exponential \( e^{j\omega_c t} \), and the previously estimated baseband center frequency is adjusted by adding an exponential term operator to introduce a spectrum bandwidth conversion of this mode to baseband;

iii) The spectrum bandwidth of each mode integral function is obtained by calculating the standard of Gaussian smoothness and frequency gradient squareness of high frequency signal;

iv) A new variable integral function constrained modal problem is obtained. Then, by using quadratic Lagrange penalty function term and quadratic Lagrange multiplier term, a new problem without variational constraint mode can be obtained. Finally, the constrained problem is solved according to equation (2).

\[
\begin{aligned}
\min_{\{u_k\}/\{i_{nk}\}} & \left\{ \sum_k \left\| \delta(t) + \frac{j\pi t}{\tau} u_k(t) \right\|^2 \right. \\
\text{s.t.} & \sum_k u_k = f
\end{aligned}
\]

(2)

2.2 The principle and flow of K-means clustering algorithm

K-means data set belongs to unsupervised learning and does not need to prepare training set. Based on the basic principle of K-means clustering algorithm, this paper can be divided into the following six parts:

Step i: First, we need to determine the value of K and input it into the algorithm, and then divide each dataset into k classes according to the data category.

Step ii: K data points are randomly selected from the data as the initial centroid, and other data points are used as auxiliary data.

Step iii: According to formula (3), the distance from each initial centroid is calculated for each subsidiary data in the dataset. The satellite data is assigned to the nearest initial centroid.

\[
d(i, j) = \sqrt{\left(x_{i1} - x_{j1}\right)^2 + \left(x_{i2} - x_{j2}\right)^2 \cdots + \left(x_{in} - x_{jn}\right)^2}
\]

(3)

Step iv: A batch of data is collected under each initial centroid, and then a new centroid is selected according to equation (4).
Step v: If the distance between the new centroid and the initial centroid is very small or 0, it indicates that the new centroid is reasonable and the clustering has reached the expected goal. At this time, the algorithm ends.

Step vi: If the distance between the new centroid and the initial centroid is large, it is necessary to re select the new centroid and allocate the auxiliary data. That's to repeat steps 3-5.

2.3 The reconstruction based on OMP algorithm

Orthogonal matching pursuit algorithm is mainly realized by using the least square method in the selected column. In the aspect of atom selection standard algorithm, OMP algorithm greatly optimizes the iteration times by introducing atomic orthogonalization for many times, thus reducing the running time of the algorithm, which is the core content of OMP algorithm compared with MP algorithm. When the iteration stops, there are enough atoms to reconstruct the signal.

OMP algorithm starts with zero solution. In the iteration process, the column matrix with the highest similarity to the residual signal is selected from the sensor matrix, and added to the index set. The index term of vector x is solved by least square method:

$$\theta_t(I_t) = \arg \min_\theta \| y_t - A_{I_t} \theta_t \|_2^2$$

(5)

According to the knowledge of linear algebra, the solution can be solved by formula (6).

$$A_{I_t}^T A_{I_t} x_t (I_t) = A_{I_t}^T y_t$$

(6)

By repeating the above steps, and the iteration ends when the norm of the remaining signal r is lower than $\varepsilon$ that is a small value. At this time, the index set composed of a finite number of column vectors is obtained.

3. The improvement of power data dimension reduction method based on VMD-K means-OMP

3.1 The adaptive improvement based on VMD method

In this study, the method of finding the best value of variation or model variable in the signal constraint is used to better remove the noise problem in the original load data. By finding the function $\alpha_k(t)$ of K modes, the sum of each mode is maximized and the sum of bandwidth of each mode function is minimized to make it the same as the signal $f(t)$ of each input terminal. In order to solve the optimal value of the constrained penalty variational factor problem mentioned above, a Lagrange multiplier and a second constrained penalty variable factor are introduced to form an extensible Lagrange multiplier formula, $L(\{u_k\}, \{f(t)\}, \lambda)$:

$$\alpha \sum \left[ \int \left( \frac{\partial}{\partial u_k(t)} \right) u_k(t) \right] e^{2i\alpha x t} \left( f(t) - \sum_k u_k(t) \right)^2 + \int \left( \frac{\partial}{\partial \lambda(t)} \right) \lambda(t) \left( f(t) - \sum_k u_k(t) \right)$$

(7)

The optimization problem of formula (8) is divided into two different subproblems: solving the minimization problem of (mode) and (center frequency). These two optimization problems can be solved by formula (9) (10). Finally, the original signal f is decomposed into K-IMF components.
3.2 The improved OMP reconstruction optimization algorithm

In this study, the power load data processed by VMD method for dimension reduction are used to form a data center by k-means algorithm clustering, which serves as a dictionary for OMP algorithm to select data from. This method realizes the reconstruction of power load data, and then realizes the optimization of OMP algorithm.

3.2.1 The principle of data center dictionary based on K-means clustering

K-means is a special case of sparse dictionary learning. The data set formed by this algorithm belongs to unsupervised learning and does not need to prepare training set. Moreover, the principle is simple, the implementation is relatively easy, the results are relatively reliable and the interpretation is relatively good. The general algorithm of dictionary construction is divided into two steps: sparse representation and dictionary update.

First of all, we need an initialization dictionary D, and represent signal Y sparsely by using OMP algorithm.

\[ u_k = \arg \min_{u_k} \left\{ \alpha \left\| j \omega \left( 1 + \text{sgn}(\omega + \omega_k) \right) u_k (\omega + \omega_k) \right\|_2^2 + \left\| f(\omega) - \sum u_i (\omega) + \frac{\hat{\lambda}(\omega)}{2} \right\|_2^2 \right\} \]  

(8)

\[ u_k (\omega) = \frac{\hat{f}(\omega) - \sum u_i (\omega) + \frac{\hat{\lambda}(\omega)}{2}}{1 + 2\alpha (\omega - \omega_k)^2} \]  

(9)

\[ BC(D(X), D(Y)) = \sum_{x \in X, y \in Y} \sqrt{D(X), D(Y)} \]  

(10)

3.2.2 Problems to be solved by OMP algorithm

It is specifically expressed as: The sparse decomposed signal b is represented by a linear combination of K columns selected from the dictionary: \( b = A\alpha \), and the value of \( \alpha \) is calculated. A dictionary whose signal length is much smaller than the number of atoms is an over-complete dictionary. In
formula \( \mathbf{b} = \mathbf{A}\alpha \), the size of \( \mathbf{A} \) is \( n \times m \), the length of the signal \( \mathbf{b} \) is \( n \times 1 \), \( \alpha \) is \( m \times 1 \).

In MP algorithm, dictionary atoms are not orthogonal to each other and need many iterations. In OMP algorithm, residuals and selected atoms are always orthogonal, and the results converge in limited steps. However, for the signal with large amount of data, OMP algorithm will expose its disadvantages of tedious orthogonalization. Therefore, this study uses K-means data set to concentrate the data patterns decomposed by VMD algorithm into data sets. Then OMP algorithm is used to select the data in K-means data set and give the corresponding weight, which greatly reduces the signal reconstruction time, improves the data processing efficiency, achieves the purpose of obtaining high-precision data, and reduces the data dimension.

4. Analysis of examples

According to the above scheme and principle, this paper combines K-means clustering algorithm and OMP data reconstruction method on MATLAB software to build a flexible load data dimension reduction model based on VMD method. The data used in this study are measured data in a certain area, and 500 pieces of data are selected to build the model. The sampling time is 15 minutes, and the whole day is divided into 96 points. According to the empirical value, the number of clusters is \( K = 5 \).

4.1 Comparative analysis of simulation results of VMD method

Figure 1 shows the comparison of spectrum analysis results of power data decomposed by various methods. Figure 2 is the simulation results of dimensionality reduction and denoising by VMD method. It can be seen from the figure that the frequency dimension of power data is obviously reduced after the VMD method is used in this paper, and the data processing process takes a short time. It can realize the dimensionality reduction of complex and large amount of power load data, and the requirements for hardware are relatively low.

4.2 K-means clustering simulation results

The above three figures (Fig. 3, 4, and 5) are the simulation results of K-means clustering processing
electric heat storage data and load data of power enterprises. According to the classification results, a redundant dictionary is formed, and the sum of each category is selected as the dictionary element. So far, the data classification has been completed. According to the classification results, a redundant dictionary is formed. And the sum of each category is selected as the dictionary element, which provides conditions for the next step of orthogonal matching pursuit.

4.3 Comparative analysis of OMP reconstruction algorithm simulation results

According to the comparison between the reconstructed curve and the initial target curve in Figure 6, the reconstructed curve after OMP algorithm tends to a stable range with the change of time, and the change amplitude is significantly lower than that before reconstruction, and the distance between adjacent peaks or troughs is smaller. Figure 7 shows the comparison curves before and after the data is processed by three reconstruction algorithms. Compared with convex tracking method and internal focus tracking method, the difference between peak and valley values of power load data reconstructed by OMP algorithm in this paper is more stable and has a fixed relative stable interval, which makes it easier to reconstruct and optimize the data, and the efficiency of reconstruction is improved by 18.75% and 43.75% respectively.

5. Conclusion

In order to solve the problem of complex data in power automation system, this paper proposes a dimensionality reduction and reconstruction algorithm of power load data based on VMD decomposition and OMP reconstruction. The algorithm is suitable for any dimension of power load data samples, and can process and analyze complex and large amounts of data quickly and accurately.

The conclusions are as follows:

Firstly, VMD method, K-means algorithm and OMP algorithm are combined for the first time. Compared with fast Fourier transform, EMD and its improved method and wavelet transform denoising method, the dimensionality reduction speed of power load data is increased by 66%, 50% and 24% respectively. At the same time, the efficiency of dimensionality reduction is about 60%.

Secondly, the clustering center is formed by collecting the decomposed data mode as a dictionary. The k-means algorithm matches with the signal mode of power load data, and it will have a more sparse representation when combined with this study, which is more suitable for power load data and more flexible for model data.

Thirdly, the OMP method is used to select data elements from the dictionary composed of K-means clustering algorithm, so as to achieve high-precision reconstruction. Compared with the convex tracking method and the internal focus tracking method, the reconstruction efficiency is improved by 18.75% and 43.75% respectively. Compared with the target data, the error rate of simulation results is 0.523, and the effect is better.
Reference

[1] Jingyi Lu, Wenping Ma, Dong Ye. Research on Sound Signal Enhancement Algorithm Based on VMD [J]. Journal of Mechanical Engineering, 2018, 54(10): 10-15.

[2] Yifeng Zhang, Yingying Li, Yibo Sun. Digital Watermarking Based on Joint DWT–DCT and OMP Reconstruction[J]. Circuits, Systems, and Signal Processing, 2019, Vol.38 (11):5135-5148

[3] OPES M E. Unknown sparsity in compressed sensing: Denoising and inference[J]. IEEE Transactions on Information Theory, 2016, 62(9):5145-5166.

[4] Chen Wan, You Jian, Chen Bo, et al. A sparse representation and dictionary learning based algorithm for image restoration [J]. Neurocomputing, 2018, 286: 130-140.

[5] Guo Jian, Song Bo, He Yi, et al. A survey on compressed sensing in vehicular infotainment systems[J]. IEEE Communications Surveys & Tutorials, 2017, 19(4): 2662-2680.

[6] SCHNASS K. Average performance of Orthogonal Matching Pursuit (OMP) for sparse approximation[J]. IEEE Signal Processing Letters, 2018, 25(12): 1865-1869.

[7] Jia Cui, Renzhe Yu, Dongbo Zhao, et al. Intelligent load pattern modeling and denoising using improved variational mode decomposition, Applied Energy, 2019 (247):480–491.