Research on Unreasonable Parameter Identification Model of Overhead Line Based on k-means Clustering Algorithm

Yifei Zhang\textsuperscript{1, a}, Yanfeng Gong\textsuperscript{2, b} and Jin Luo\textsuperscript{2, c}

\textsuperscript{1}North China Electric Power University, Beijing, China.
\textsuperscript{2}State Grid Shanghai Construction Company, Shanghai, China
\textsuperscript{a}zhangyifei6685@yeah.net, \textsuperscript{b}yanfeng.gong@ncepu.edu.cn, \textsuperscript{c}luojinlxn@sina.com

Abstract. The existence of unreasonable in the planning of grid data may influence the economy and rationality of the power grid planning and so on, which may lead to unreasonable structure or potential risk in planning power grid. And it may even threaten the safe operation of the entire power system. This paper summarizes the general characteristics and methods of power grid planning, introduces the theory of data mining, and establishes an unreasonable parameter identification model for overhead transmission lines based on k-means clustering algorithm. Finally, an example is given to illustrate the validity of the model and method.

1. Introduction
Electric power industry is an energy power industry. Its basic task is to provide cheap and reliable electricity to users and improve social benefits. Thus, it is very important to make a good electric power planning and strengthen the construction of power grid for promoting the national economic level. Due to the increasingly complex structure of the modern power grid, the operation mode is increasingly complicated. However, there are many factors affecting the power grid planning parameters. The existence of unreasonable data in power grid planning may affect the economic and rationality of power grid planning, which may lead to unreasonable structure or potential risks in the planned power grid. Seriously, it may even threaten the safe operation of the entire power system. It is precisely for this reason that it is extremely important to detect and identify unreasonable data in order to meet the requirements for safe and reliable operation of the power grid.

Power system state estimation method can be used to identify unreasonable parameters of power system\cite{6}\cite{7}. For the detection and identification of unreasonable data in high voltage distribution system, the power flow calculation is carried out by using the forward and backward substitution method by making use of the correlation between telemetry and telemetry and the relationship between telemetry and quantity. Then the unreasonable data are processed by the improved genetic algorithm to obtain the correct operation state of the power grid\cite{6}. The method of state estimation based on the highest reliability takes into account all kinds of constraints and types of measurement points in practice, especially suspicious ones which are easy to ignore. It has high superiority and sensitivity\cite{13}. Data mining technology is widely used in the identification of unreasonable parameters in power network planning\cite{5}\cite{9}\cite{10}. In recent years, data mining technology can be applied in five aspects: prediction, classification, association rule analysis, clustering and outlier analysis\cite{5}. The tree method can be used to classify the data, and then the nearest neighbor rule can
be used to estimate the unreasonable data. This algorithm not only meets the accuracy requirements of fault data estimation, but also greatly improves the calculation speed [10].

In general, the detection and identification of unreasonable parameters play an important role in the accuracy of state estimation results. There are many measuring devices in the transmission network, which leads to more actual measurement and high data redundancy. Therefore, the unreasonable data detection and identification in the power network planning are mainly concentrated in the transmission network. However, due to the effect of various uncertainties, residual pollution and residual submergence are prominent in the state estimation. Hence, it is particularly important to detect and identify the unreasonable parameters in the transmission network as accurately as possible. As the structure and operation mode of modern power grid are becoming more and more complex, the data volume of power grid planning is also showing an explosive growth trend. Therefore, the combination of unreasonable data identification and data mining technology can be used to retrieve and identify unreasonable power flow data of overhead lines in power grid planning, thus reducing the tedious work of manual checking input data, and also improving the convergence and accuracy of power flow. Not only that, but it will also greatly improve the efficiency of large-scale power grid.

2. Power grid planning theory

Power grid structure is very important for maintaining system stability, improving system economy, ensuring good power quality and flexibility of power grid dispatching control. Power grid planning should not only consider the requirements of national economic development and power load growth, but also the requirements of rational allocation of resources and environmental protection. Therefore, new technologies are adopted to solve a series of problems such as power supply reliability, power quality, unreasonable distribution of power grid and so on. Generally speaking, power network planning is to determine the new power lines to meet the demand of load growth in the planning period by considering the existing power network structure, power planning, load forecasting and other factors. Generally speaking, power grid planning is to determine the new power lines to meet the demand of load growth in the planning period by considering the existing power network structure, power planning, load forecasting and other factors. At the same time, various constraints should be taken into account to minimize the economic cost and maximize the reliability of the power grid.

The purpose of power grid planning is to pursue the most reasonable investment plan for power grid construction, so as to ensure the long-term optimal development of power system operation. The earliest grid planning was usually carried out in the form of scheme comparison. This method is often selected from several well-designed alternatives, and then the optimal one is selected through technical and economic comparison. Since this method is generally based on experience by grid planners, there are inevitably some problems. At present, the common methods of power grid planning can be divided into two categories: heuristic optimization method and mathematical optimization method. The similarities are based on the prediction results, and the relevant mathematical models are created to finally obtain the optimal grid planning scheme.

2.1. Heuristic Optimization Method

Based on the direct research, the heuristic optimization method carries out sensitivity analysis on some related parameters of one of the performance indicators of the system, and iterates successively according to the relevant criteria until the schemes that meet the requirements are obtained, so as to select the hypothetical routes. The heuristic optimization method includes three parts: overload verification, sensitivity analysis, and scheme formation. The advantage of heuristic method is simplicity and sensitivity. Due to the small amount of data processing, the required calculation time is short, and it is easy to combine with the long-term accumulated work experience of the grid staff. Heuristic optimization method can simulate power behavior more accurately, which is a unique advantage over other mathematical methods.

However, it is not a very strict optimization method, and it cannot guarantee the optimality of the solution. Therefore, it is difficult to select relevant indicators that are simple and can reflect the power
grid planning problem. Moreover, this method cannot fully take into account the mutual interference and influence of decision-making in each stage of power grid planning. Especially when the grid planning project is large in scale, the planning period is long, and the number of lines to be selected is complicated, the impact of performance indicators on a group of programs is not much different, and the results may be greatly deviated from the real optimal solution. It is thus difficult to obtain an optimal solution.

2.2. Mathematical Optimization Method
Mathematical optimization method is to express the power grid planning in a concise mathematical language, transform the practical problem into the extreme value problem with limited conditions in mathematics, and then use the optimization method to calculate the results. Compared with heuristic optimization method, mathematical optimization method has theoretical superiority, so it has been widely studied and developed. Mathematical optimization methods mainly include linear programming, dynamic programming and multi-objective programming.

In theory, this method can ensure the optimality of the solution, but there are many factors involved in practical application, which cannot consider the constraints comprehensively. As a result, the dimension of data variables is very large, the amount of calculation is too large and the calculation time is too long. Therefore, it is very difficult to establish a reasonable mathematical model. Moreover, even if the mathematical model is established, the next solution is difficult to carry out. Because many factors in power grid planning cannot be formalized by mathematical expression, the mathematical model of the original problem is often simplified, which may lead to the problem of losing the optimal solution. For these reasons, although the method has been applied quickly, there are still many problems to be solved in the actual power grid planning.

Although many uncertain factors have been taken into account to interfere with the final results, power grid planning still faces many problems such as local optimum, difficult constraints to deal with, dimension disaster and so on. Therefore, different solutions are adopted to deal with many uncertain factors, and many planning methods are developed, such as tabu search algorithm, ant colony algorithm, simulated annealing algorithm, genetic algorithm, particle swarm optimization and so on.

3. Data Mining and K-means Clustering Algorithms
Data mining is the process of discovering potentially useful information from a large number of incomplete, noisy, ambiguous, and accidental actual operational data. As a complex system, the power system has many data sources, mixed data types, and is prone to some problems such as poor data quality. Therefore, the application of data mining technology in the power field has received more and more attention. There are many data mining algorithms, such as C4.5 algorithm, K-means clustering algorithm, support vector machine, maximum expectation algorithm, AdaBoost algorithm, Aprior algorithm, PageRank algorithm, K nearest neighbor classification algorithm and classification regression tree algorithm.

The data mining process is generally divided into three main stages: data preparation, data mining and interpretation evaluation. Data mining begins with data preparation, which is crucial because data preparation will have a direct impact on the whole process. A good data preparation is an inevitable requirement to ensure the accuracy of the entire process and the validity of the final data model. The process of data preparation can be further subdivided into four parts: data integration, data selection, data preprocessing and data conversion. The final stage of interpretation and evaluation is based on the final purpose of decision-making. By evaluating the knowledge mined, the most meaningful information is selected and interacted with users and received feedback from users.

Clustering is to divide a data set into groups or classes in a certain way, and classify data objects with higher similarity into a group, and distinguish and group data objects with low similarity. In short, the process of clustering is the process of classifying massive data into different groups.

The criterion of classification and grouping of data objects in clustering process is determined by the value of the characteristics of the expressed data objects. The general method is to describe a group
of concrete or abstract data objects according to the distance between data objects. By calculating the similarity between them mathematically, the high similarity can be grouped into a group, so as to find out the value relationship between data objects. In many applications, such as statistics, data mining, computer technology and other practices have clustering applications.

K-means clustering algorithm is a clustering algorithm based on the mean of clustering. As the most famous and widely used method of partitioning and clustering, it has the characteristics of simplicity and high efficiency. The algorithm is to specify a database including n objects and the number of clusters needed k. k is specified by the user. K-means clustering algorithm divides the data into k clusters repeatedly according to a distance function. The objects with high similarity are classified into different clusters while those with low similarity are classified into different clusters. Finally, k clusters that meet the minimum variance criteria are output.

Cluster similarity is calculated by using a cluster of objects in each cluster to obtain a "central object". According to the distance between the cluster object and each cluster center, it is sequentially assigned to the cluster in which the closest "center object" is located. In the cluster analysis, the distance calculation uses the Euclidean distance.

The processing flow of cluster analysis is as follows:

⚫ K objects are arbitrarily selected from n data objects as initial cluster centers.
⚫ According to the "central object" of each clustering object, the similarity (distance) of each object with each cluster center is calculated, and the related objects are distinguished according to the closest distance.
⚫ Recalculate each "central object" with a variable cluster.
⚫ Continuously loop the process until a certain criterion function no longer changes significantly or the clustered data object stops changing.

The criterion function of the general k-means algorithm adopts the square error criterion, which is defined as:

\[ E = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2 \]  

Among them, \( E \) is the sum of mean square deviation of all objects in the database and corresponding clustering centers. \( p \) is a point in the space representing the object. \( m_i \) is the mean of cluster (both \( p \) and \( m_i \) are multidimensional). The purpose of the clustering criteria represented in the above formula is to make the obtained clusters have the following characteristics:

Each cluster object itself is as compact as possible, and is most clearly distinguished from other cluster objects.

4. Example analysis
The example data is used to plan grid parameters in a certain area of Northwest China. The main voltage levels are 330kV and 220kV, for a total of 101 overhead lines. The identification of unreasonable parameters in the overhead lines is mainly for line resistance, reactance, and susceptance. Since the parameters of the given line are all standard values, before the parameter identification model is established, the conversion of the standard value and the nominal value is performed, and then the resistance value, the reactance value and the susceptance value of the line unit length are calculated. Then the K-means clustering algorithm is used to cluster the overhead lines in the example and identify the unreasonable parameters. The distribution of overhead line parameters is shown in Figure. 1:
According to the distribution of parameters, the distribution of overhead lines can be divided into five groups. Therefore, it can be initially determined that the initial k value of the k-means algorithm is 5, and iterative calculation is performed until the obtained result is not changed.

After several iterations, it is found that the result of the fourth iteration is basically the same as that of the third iteration, so the final clustering center and results are as follows:

![Figure 1. Scatter plot of overhead line parameter distribution](image1)

**Figure 1.** Scatter plot of overhead line parameter distribution

![Figure 2. Cluster center distribution scatter plot](image2)

**Figure 2.** Cluster center distribution scatter plot
The final cluster centers are shown in the table below. The first column represents the resistance value of the cluster centers, the second column represents the reactance value of the cluster centers, and the third column represents the susceptance value of the cluster centers.

Table 1. Distribution values of cluster center parameters

| Cluster number | resistance value | reactance value | susceptance value |
|----------------|------------------|-----------------|-------------------|
| 1              | 0.0584           | 0.3416          | 0.00017841        |
| 2              | 0.4059           | 3.0963          | 0.00075277        |
| 3              | 0.0643           | 0.5012          | 0.00015263        |
| 4              | 0.1247           | 0.8342          | 0.00022802        |
| 5              | 0.0176           | 0.1200          | 0.000094831       |

Finally, the relative errors between the resistance, reactance and susceptance of the line and the cluster center resistance, reactance and susceptance are compared to identify the unreasonable data.

5. Conclusion

In this paper, the theory of power grid planning is expounded, and the general steps and methods of power grid planning are briefly described. In addition, this paper outlines the application of data mining in power system, and introduces the related algorithms of data mining. The basic idea of clustering algorithm is elaborated, and the K-means clustering algorithm is introduced emphatically. Finally, aiming at the problem of unreasonable data of power flow in large-scale power grid planning, a model for identifying unreasonable parameters of transmission network is established by using data mining technology, and the validity of the model is proved by an example.

References

[1] E Handschin, F C Schwepp, J Kohlas, et al. Bad Data Analysis for Power System State Estimation, IEEE Transactions on Power Apparatus and Systems. 1975, Pas-94 (@):329-337.
[2] Abur A. A bad data identification method for linear programming state estimation [J], IEEE Transactions on Power Systems, 1990, 5 (3):894-901.
[3] Zhang B M, Lo K L. A recursive measurement error estimation identification method for bad data analysis in power system state estimation [J], IEEE Transactions on Power Systems, 1991, 6 (1):191-198.
[4] Fasong Wang, Hongwei Li, Rui Li. Data Mining with Independent Component Analysis[J], IEEE Computer Application in Power, 2004, 15 (3): 6043-6047.
[5] Li Zhiyong. Data Mining of Power System Operation Information [D], Zhejiang University, 2009.
[6] Zhou Xiaobao. Research and application of bad data detection and identification methods for power system state estimation [D], Hunan University, 2013.
[7] Huang Shyh-Jier, Lin Jeu-Min. Enhancement of Power System Data Debugging Using GSA-Based Data-Mining Technique. IEEE Transactions on Power Systems. 2002, 17 (4): 1022-1029.
[8] Merlin, A., and F. Brousso. Fast Method for Bad Data Identification in Power System State Estimation, IFAC Proceedings Volumes 10.1 (1977):449-453.
[9] Du Zhijia, Wu Junji. Research on Power System State Estimation Based on Data Mining [C], Annual Academic Conference of Electric Power System and Automation Specialty in Chinese Colleges and Universities. 2008.
[10] Zhao Jie, Qin Yi, Li Jing. State Estimation of Bad Data in SCADA System Based on Data Mining Technology [J], Science and Technology Information, 2007 (30): 6-7.
[11] Yanjun, Zhang, et al. Bad Data Identification Based on Innovation Graph Technique in Wide Area Measurement System, Automation of Electric Power Systems 34.23 (2010):25-29.
[12] Abur, Ali, and Y. Lin. A Highly Efficient Bad Data Identification Approach for Very Large Scale Power Systems, IEEE Transactions on Power Systems (2018):1-1.

[13] Ren Yiyi. Research on State Estimation Algorithms for Transmission and Distribution Networks [D], Hefei University of Technology, 2013.

[14] Xindong Wu and et. Al., 2008. Top 10 Algorithms in Data Mining, Journal of Knowledge and Information Systems, vol. 14. Issues 1-37, DOI: 10.1007/s10115-007-0114-2.

[15] Liao Zhiwei, Sun Yaming. Data Mining Technology and Its Application in Power System [J], Power System Automation, 2001, 25 (11): 62-66.

[16] Chen Zheng. Identification of Overhead Line Parameters in Power Network Planning Data [J], China High-tech Enterprise, 2016 (22): 62-64.

[17] Chen Zhen. Identification and correction of bad data in power network planning based on data mining [D], North China Electric Power University (Beijing), 2016.