Advanced Science and Technology Letters
Vol.143 (AST 2017), pp.169-175
http://dx.doi.org/10.14257/astl.2017.143.36

Application of Optimized GSA Algorithm on Bad-data Detection of Electric Power Dispatching System

Jianlou Lou[1]*, Jizhe Xiao[1], Hongjian Zheng[2], Zhaoyang Qu[1]

[1]School of Information and Engineering, Northeast Electric Power University, Jilin 132012, China
[2]PetroChina jilin Petrochemical Company First Power Plant, Jilin 132021, China
[1]loujianlou@qq.com, [2]jh_zhenghj@petrochina.com.cn

Abstract. The detection and identification of the bad data of the power system plays an important role in dispatching personnel to grasp the running status of the power grid in real time. In order to overcome negative effects of random selection of clustering initial values of traditional GSA bad data identification algorithm on identification precision and computation rate, this paper propose an optimized GSA algorithm based on area density statistics method. This algorithm by computing the area density of each cluster object to select k points that are farthest from each other and are at the highest area density as the initial cluster center. The experimental results show that the optimized GSA algorithm improves the accuracy of the degree of clustering dispersion and the recognition accuracy of the bad data. At the same time, the algorithm greatly reduces the computational complexity of iterative computation, improves the computing speed and saves a lot of computing time. In the case of huge system and large amount of data, this method is a rapid and efficient algorithm, and has potential of good application.

Keywords: power system; identification of bad data; area density; gap statistic algorithm; cluster

1. Introduction

The detection and identification of bad data in power system has been one of the important functions of power system state estimation [1-2]. The research on bad data identification is mainly divided into two directions. One is the identification method based on estimation and residual analysis. The other is based on data mining technology and intelligent algorithm identification method. K-Means clustering is usually applied to this kind of algorithm [3-4]. Many scholars proposed the use of GSA (Gap Statistic Algorithm) strengthen algorithm to identify bad data [5]. GSA is a kind of data mining algorithm which can improve the clustering effect, and it can estimate the best clustering number. In the bad data identification of power system, clustering of good data and the bad data can be accurately distinguished [6-7]. However, the traditional GSA algorithm does not consider the optimal selection of the initial cluster center, resulting in algorithms often terminating with local optima.

ISSN: 2287-1233 ASTL

In order to improve the computing speed and reduce the miscarriage of justice, this paper propose the optimized GSA method based on area density statistics. In the process of the algorithm, the area density of each data object is first calculated, and *k* points which are farthest from each other and at the highest area density are selected as the initial clustering center. This method provides a good clustering basis for calculating the degree of clustering dispersion for the GSA algorithm, which improves the GSA algorithm recognition accuracy, reduces the computation cost and saves a lot of computation time. The simulation results show that the method can identify bad data accurately, without misjudgment, and can significantly improve the calculation speed.

2 An Improved K-Means Algorithm Based on Area Density Statistics

In the K-Means algorithm with Euclidean distance as the measure of similarity, the *k* data objects farthest from each other are more representative than the *k* data objects randomly selected [8-9]. However, in the actual data sets often have noise data exists, if only simply take the farthest away from the *k* points to represent *k* different categories, maybe it will get to the noise point, thus affecting the clustering effect [10-11].

Generally, in a data space, a high-density data object area is divided by a low-density object area, and a point in the low-density area is generally regarded as a noise point. In order to avoid getting the noise point, we take *k* points in the high density area which are farthest away from each other as the initial clustering center.

The radius ε of the spatial area containing *n* data objects centered on a spatial point x_i is called the density parameter of the object x_i. The larger the ε, the lower the data density of the area in which the data object is located.

In space, the distance between the sample point x_i and the sample point y_i is $d(x_i, y_i)$.

$$d(\mathrm{x}_i, \mathrm{y}_i) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} \tag{1}$$

The distance between one sample point *x* and another sample set Z is defined as the shortest distance between all the distances that each point in Z to *x*. The distance formula is:

$$d(\mathrm{x}, Z) = \min\left(d(\mathrm{x}, \mathrm{y}), \mathrm{y} \in Z\right) \tag{2}$$

By calculating the density parameter of each data object x_i, the point in the high area density can be found. Thereby, resulting in a high area density point set W. The data object z_1, which is in the highest area density, is taken as the first center and added to the set Z. The distance from each sample point y_i in the W to the set Z is calculated to find another sample point that furthest from the set Z, that is, max(min(d(y_i, Z))), join the set Z as the second center z_2. In the same way, we get *k* initial clustering centers in turn.

The optimized K-Means algorithm based on area density statistics is described as follows:

Step 1: To calculate the distance between any two data objects $d(x_i, y_i)$.

Step 2: Calculating the density parameter of each object, deleting the points in the low area density, and obtaining the set W of the data objects in the high area density.

Step 3: The data object in the highest area density is taken as the first center, and added to the set Z, while removed from the W.

Step 4: From the set W, find the farthest point from the set Z, join the set Z, while removed from the D.

Step 5: Repeat (4) until the number of samples in Z reaches *k*.

Step 6: The K-Means clustering algorithm is used to select the *k* values as the initial clustering centers, and the iterative calculation is carried out to get the clustering results.

3 Optimized GSA Identification Method

Through the study of the GSA-based elbow judgment and the optimized K-Means algorithm based on the area density statistics presented in this paper, the optimized GSA algorithm for bad data identification is obtained, the following is the flow chart of the algorithm.

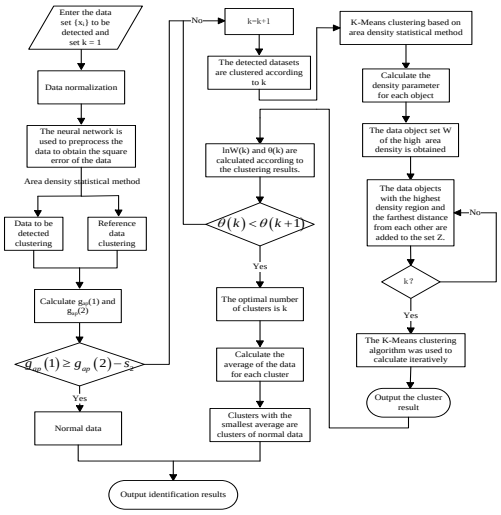

Fig. 1. Flowchart of optimized GSA algorithm

(1) Enter the data set $\{x_i\}$, and site the initial k=1.

(2) The data set is normalized, and the neural network algorithm is used to preprocess the data to get the square error data.

(3) The GSA-based elbow judgment algorithm is used to calculate the $g_{ap}(1)$ and the $g_{ap}(2)$ by the formula (3) [12].

$$g_{ap}(k) = \frac{1}{F}\sum_{j=1}^{F} \ln W_{r,j}(k) - \ln W(k) \tag{3}$$

(4) If the result satisfied formula (4), it means all the data are normal data, and output identification result.

$$g_{ap}(1) \geq g_{ap}(2) - s_2 \tag{4}$$

(5) Otherwise, let $k=k+1$. Based on the area density statistical method, determine the optimal k cluster initial centers, and the K-Means is used to iteratively calculate the data clustering results.

(6) Further, according to the GSA-based elbow judgment theory, calculate lnW(k) and θ(k), verify whether the formula (5) is met.

$$\theta(k) < \theta(k+1) \tag{5}$$

(7) If not, repeated calculation $k=k+1$. If so, the optimal number of clusters k is obtained.

(8) Finally, calculated the average of each cluster. The cluster with the smallest average is regarded as the clustering of normal data.

4 Experimental Simulation Analysis

The simulation data of the paper are taken from the real-time operating data of the D5000 system of Jilin Electric Power Company on April 26, 2016. There are 240 sets of real-time measurement data were obtained from the survey data, of which 200 groups were trained as training samples and the remaining 40 groups were used to test the neural network. After the neural network is trained, the measured value of the neural network is tested, and the square of the input and output difference is obtained as the basis of clustering analysis.

4.1 Single Bad Data Scenario

It is assumed here that the measurement data No.34 in Group 215(active power of 220kV AClineSegment in Jilin Fengman Hydropower Plant) exceeds the normal value by 25%. First of all, the 215 sets of measured data are pre-processed by the neural network algorithm.

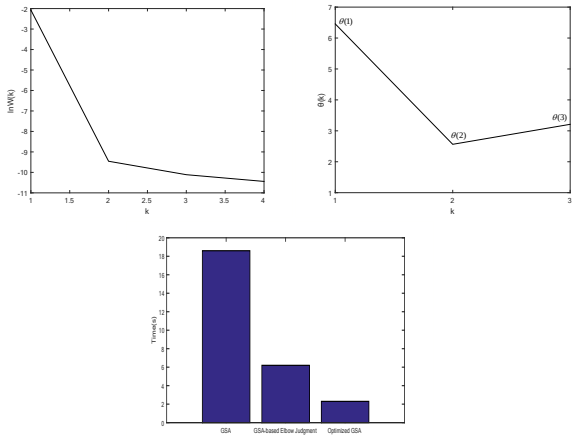

Fig. 2. The analysis of the results of single bad data scenario

After optimization GSA identification method calculate, $g_{ap}(1)=2.1521$, $g_{ap}(2)=3.6589$, and $s_2=0.0812$, it means that $g_{ap}(1)<g_{ap}(2)-s_2$, this proves that there is bad data in the data set. Then, use the elbow judgment to calculate lnW(k) and θ(k). From figure 2(left) of the cure of lnW(k)—k, it is obvious that at k=3, the curve becomes significantly flattened. And in the figure 2(middle), it can be see that $\theta(2)<\theta(3)$, so the optimal number of clusters is 2. The average value of the two internal elements was calculated, and the data of the large average clustering was the 34th measurement data. Simulation results verify the accuracy and effectiveness of the algorithm.

Most importantly, the algorithm proposed in this paper is more efficient in terms of algorithm running time. As shown in figure 2(right), the method proposed in this paper is much shorter than the other two algorithms in running time. From the above figure we can see through the method proposed in this paper, the identification of bad data accuracy and effectiveness is greatly enhanced.

4.2 Multiple Bad Data Scenario

In this case, assume that there were six bad data in the 230th sets of measurement data, numbers are 4, 15, 36, 75, 96 and 112. Its measured value exceeds the normal value of 15% to 30%. Firstly, the 230 sets of measured data are pre-processed by the neural network algorithm.

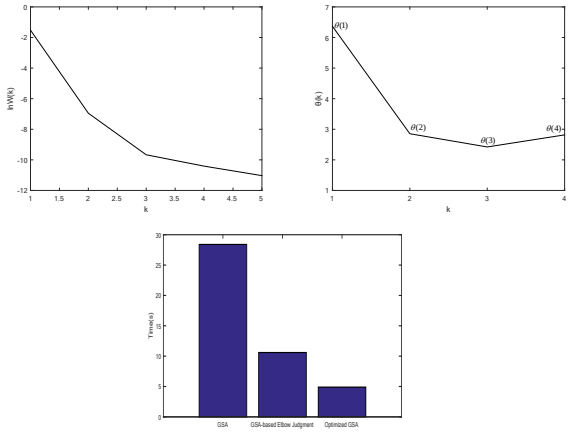

Fig. 3. The analysis of the results of multiple bad data scenario

After optimization GSA identification method calculate, we can know $g_{ap}(1)=1.2587$, $g_{ap}(2)=1.8756$, and $s_2=0.0871$, then, $g_{ap}(1)<g_{ap}(2)-s_2$, so there is bad data in the data set. Next, using the elbow judgment calculate lnW(k) and θ(k). From figure 3(left), it can be clearly seen that at k=3, the cure of lnW(k)—k becomes significantly flattened. At the same time, from figure 3(middle), we can know that $\theta(3)<\theta(4)$, so the optimal number of clusters is 3. Furthermore, the mean values of the three clusters were calculated. The two clusters with larger mean values are considered as the clustering of the bad data. Those six hypothetical bad data were accurately detected.

On the efficiency of the algorithm is concerned, it can be seen from figure 3(right) that the performance of the GSA identification method optimized by the area density statistic is more obvious when dealing with more clustering numbers. Especially when dealing with large amounts of data, the algorithm can save a lot of time and improve accuracy.

6 Conclusion

After studying the traditional GSA algorithms and the GSA-based elbow judgment theory, we found that there is no effective choice of clustering initial value in the process of determining the optimal number of clustering algorithms. Based on the above problems, this paper proposes the optimized GSA identification method based on area density statistical. By finding out the data objects of the highest area density, the optimal clustering initial value can be obtained, and the clustering accuracy and efficiency of the algorithm can be greatly improved.

In this paper, the method is applied to the local power grid real-time data of bad data detection and identification. Through several different cases of bad data simulation experiments can be found: This method is more objective and accurate, and the calculation speed is greatly improved on the basis of avoiding residual pollution and

residual submergence. Especially for the large amount of data in a large system, this method is a fast and efficient algorithm, and has a good application prospect.

Acknowledgments. This work was supported by the National Natural Science Foundation project of China (No. 151437003), Jilin province science and technology development plan project (20150204084GX) and Jilin province science and technology development plan project (21060623004TC).

References

1. Zhao Junbo, Zhang Gexiang, Huang Yanquan. Status and Prospect of State Estimation of Power System with New Energy Sources [J]. Electric Power Automation Equipment, 2014, 34(05): 7-20+34.
2. Liu Li, Di Denghui, Jiang Xinli. Current situation and development of the methods on bad-data detection and identification of power system [J]. Power System Protection and Control, 2010, 38(05): 143-147+152.
3. Zhao Li, Hou Xingzhe, Hu Jun, et al. Improved k-means algorithm based analysis on massive data of intelligent power utilization [J]. Power System Technology, 2014, 38(10):2715-2720.
4. Meng Jianliang, Liu Dechao. A new method for identifying bad data of power system based on Spark and clustering analysis [J]. Power System Protection and Control, 2016, 44(03): 85-91.
5. YuHui Luo, Jonathon C. Active source selection using gap statistic for underdetermined blind source separation[C]. Signal Processing and Its Applications 2003 Proceedings, Seventh International Symposium, Paris, France, 2003: 137-140.
6. Guo Yandong, Shen Dinghui. Bad data detection and identification based on improved GSA algorithm in grid [J]. East China Electric Power, 2013, 41(03):542-545.
7. Shi Zhiping. Study on identification of bad data in power system based on improved GSA algorithm [J]. Automation Application, 2012, 02: 57-60.
8. Di Donghai, Yu Jiang, Gao Fei, et al. K-means text clustering algorithm based on initial cluster centers selection according to maximum distance [J]. Application Research of Computers, 2014, 31(03): 712-715+719.
9. Xie Juanying, Gao Hongchao. Statistical correlation and k-means based distinguishable gene subset selection algorithms [J]. Journal of Software, 2014, 25(09): 2050-2075.
10. De Amorim R C, Mirkin B. Minkowski metric, feature weighting and anomalous cluster initializing in k-means clustering [J]. Pattern Recognition, 2012, 45(3): 1061-1075.
11. Lai Yuxia, Liu Jianping. Optimization study on initial center of K–means algorithm [J]. Computer Engineering and Applications, 2008, 44(10): 147-149.
12. Wu Junji, Yang Wei, Ge Cheng, et al. Application of GSA-based Elbow Judgment on bad-data detection of power system [J]. Proceedings of the CSEE, 2006, 226(22): 23-28.

