Research on the Device Risk Value for Power Transmission Network based on k-means

Xiaorong Cheng
Computer Engineering Department
North China Electric Power University
Baoding, China
Cheng3100@sohu.com

Yun Miao
Computer Engineering Department
North China Electric Power University
Baoding, China
1029059800@qq.com

Zijun Li *
Computer Engineering Department
North China Electric Power University
Baoding, China
1193331957@qq.com
* Corresponding Author

Abstract—To evaluate the risk system of power communication transmission network (PCTN) scientifically, this paper puts forward an improved k-means clustering algorithm in allusion to the issue of the power device risk value in the PCNT. Both the rating of the device status and the importance of the device determine the power device risk value, among which the importance of device is decided by the number and type of tasks it bears. Since generated randomly, traditional initial clustering centers is easy to lead the result to a partial best instead of global. Therefore, this text uses an improved Huffman tree method based on adjusted cosine similarity to make sure the initial centre. It is verified that the improved k-means clustering algorithm can do better in clustering data, dividing the risk rank scientifically, which provides a reference frame for risk evaluation.

Keywords- Power Communication Transmission Network; Power Device Risk Value; k-means; Huffman Tree; Cosine Similarity

I. INTRODUCTION

It is practically significant to timely discover and point out the possible risk, ensure the normal operation of the business, and improve the reliability of the whole network operation by assessing the risks of electric power communication transmission network status effectively.

The methods of clustering analyzing the electric power communication transmission network equipment risk value are multitudinous, generally are: hierarchical dividing method [1-2], hierarchic method [3-4], the methods based on density [5-8]. K-average clustering method, also known as the K means clustering method is a typical method of division [9-10]. It is based on the average as the "center" of a class, and constantly iterative optimization then gets the final clustering results. This method is rapid and simple with a more intuitive geometric meaning has been applied in pattern recognition, image processing and computer vision successfully. As the initial centers are selected randomly, the traditional k-average clustering method is easy to cause the result into a local optimum rather than the global optimal. Therefore, the selection of the initial center is especially important for clustering results. In this paper we calculate the similarity between each sample to construct the Huffman tree using the adjusted cosine similarity, obtaining a more reasonable initial center, which can improve the accuracy of clustering results.

II. TRADITIONAL K-MEANS CLUSTERING METHOD AND DISADVANTAGES

Assume that the number of samples is n, the sample collection is $S_n = \{x_1, x_2, \ldots, x_n\}$. The dimension is v, and the final clustering results are divided into k classes.

(1) Selecting the initial center $C_k = \{c_1, c_2, \ldots, c_k\}$. c_i is the center vector of class i, representing the i category.

(2) Calculating the Euclidean distance between the sample x_j and c_i.

$$d(x_j, c_i) = \sqrt{\sum_{k=1}^{v} (x_{jk} - c_{ik})^2} \qquad (1)$$

Assigning x_j to the class which is the closest to it, which is the class i of $\min\{d(x_j, c_i) \mid i \in \{1,2,...,k\}\}$.

(3) After all the samples are dispensed, updating the center of each cluster c_i', m is the number of samples in each cluster.

$$c_i' = \frac{1}{m} \sum_{i=1}^{m} x_i \qquad (2)$$

(4) Repeat the above steps until convergence.

The initial center is randomly selected in this method, only the initial cluster centers divide rationally, can we get the optimal clustering results. Therefore, we made the improvement for the lack of the algorithm.

III. IMPROVEMENT OF K-MEANS ALGORITHM

A. Adjusted vector cosine similarity

1) Vector cosine similarity

Researchers calculate the degree of similarity between the vector space using cosine similarity method, which uses the cosine of the angle between two vectors vector space to measure the difference between the two individuals. For example, the cosine similarity between samples X and Y can be represented as:

$$w(X,Y) = \cos\theta = \frac{\vec{x}\cdot\vec{y}}{\|x\|\cdot\|y\|} \qquad (3)$$

The similarity among n samples is represented as similarity matrix:

$$\begin{pmatrix} 1 & & & & \\ w_{(2,1)} & 1 & & & \\ w_{(3,1)} & w_{(3,2)} & 1 & & \\ \vdots & \vdots & \vdots & \ddots & \\ w_{(n,1)} & w_{(n,2)} & \cdots & w_{(n,n-1)} & 1 \end{pmatrix}$$

$w_{(i,j)}$ represents the difference between object *i* and *j*, $w_{(i,j)} \in [-1,1]$. And $w_{(i,j)} = w_{(j,i)}$, $w_{(i,i)} = 1$. When the objects *i* and *j* are very similar with each other, $w_{(i,j)} \to 1$; in contrast, $w_{(i,j)} \to -1$.

2) Adjusted vector cosine similarity

Cosine similarity can be determined that some of the differences between individuals, mainly in individual differences between the dimensions, rather than the difference between the dimension values. To improve this situation, we will make the adjustments that all dimensions subtract its average, and then the amended results are used in cosine similarity determination. For example, two users A and B give a mark on x and y in 5-point scale, and the results are shown in Table 1. It is clear that the evaluation between users A and B on the contents x and y has a big gap. The cosine similarity calculated by traditional method is 0.970 with the result of high similarity, while the amended results are -1.000, representing the very low similarity. Obviously the latter is more in line with the facts.

TABLE I. USER RATING TABLE

	User A	User B	Cosine similarity
Original data	(1,1)	(3,5)	0.970
Revised data	(-1,-2)	(1,2)	-1.000

B. Improved k-means algorithm flow

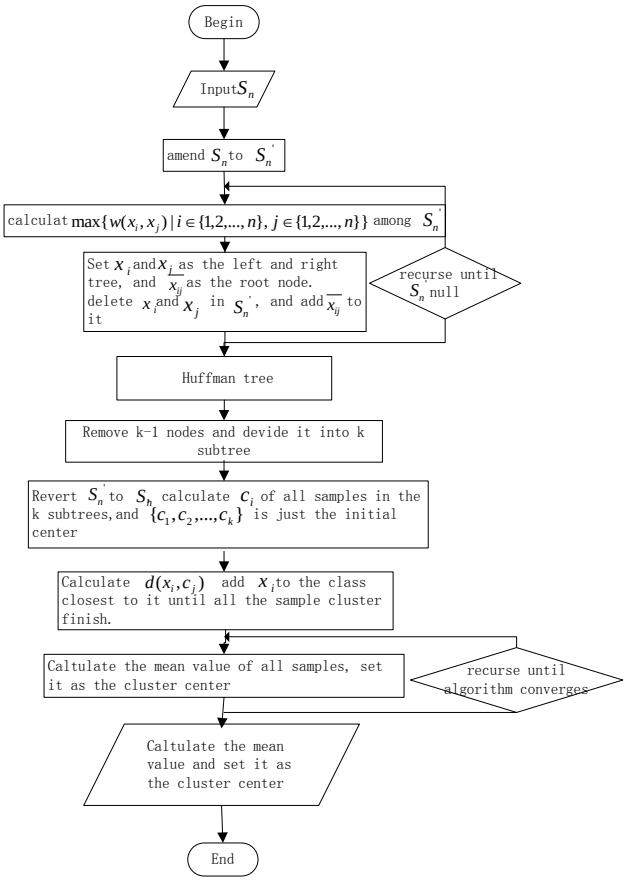

Figure 1. Algorithm flowchart

Utilizing Huffman tree structure the initial centers can avoid the instability caused by randomly generated initial center, meanwhile reduce the probability that the clustering results turn into local optimum, thereby obtaining global optimal results. It is the corrected sample when calculating the cosine similarity and structuring the Huffman tree, while it is the original data when calculating the average of all samples subtree as the initial center and step-iterating.

C. Experimental verification and results analysis of algorithm

1) Experimental verification

Assume that S_n is constituted by 30 two-dimensional objects，distributed as Fig. 2:

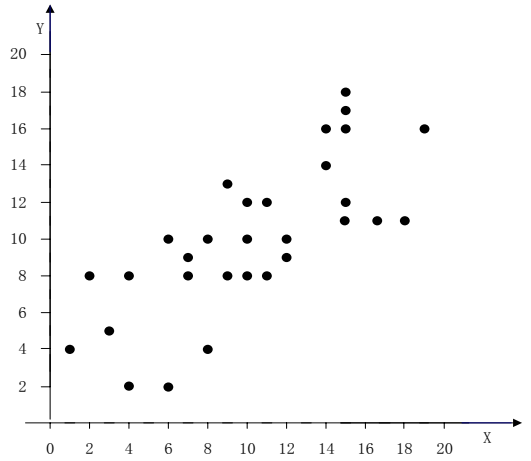

Figure 2. Data distribution

2) Results analysis

In this experiment, it can be seen from Fig. 2 that the data divided into three categories is the most reasonable. According to the improved k-means method, we need to amend the 30 data first, so one-dimensional average of the sample data is 10.23, and two-dimensional average is 10.1.For purposes of calculation, the mean value of the integer is 10, then we make the improvement in the way that both of the one-dimensional average and two-dimensional average subtract 10 calculating the cosine similarity. Then the 3 initial center vector are (10.75,8.75), (5.17,6.5), (14.21,13.5). All data are rounded in order to cluster and contrast conveniently, and it is proven that the rounded data does not affect the accuracy of the results which are (11,9), (5,7), (14,14). The optimal final clustering results are shown in Table 2 Table 3and Fig. 3. While the randomly generated initial centers are (19, 16), (15,17), (15,18), the final clustering results are shown in Table 2 Table 3and Fig. 3, Fig. 4. By contrast, the randomly generated initial center may get an optimal result rather than the best results.

TABLE II. FINAL CLUSTER CENTERS

Final cluster centers(Best)				**Final cluster centers(Random)**			
	Clustering				***Clustering***		
	1	2	3		1	2	3
x	5	14	10	**x**	12	8	16
y	4	16	9	**y**	17	8	15

TABLE III. CASE NUMBER OF EACH CLUSTER

Case number(Best)			**Case number(Random)**		
Clustering	1	7	**Clustering**	1	5
	2	10		2	20
	3	13		3	5
	Effective	30		Effective	30
	Missing	0		Missing	0

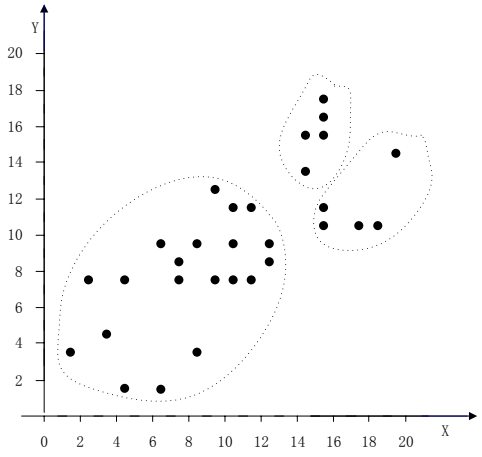

Figure 3. Traditional k-means clustering results

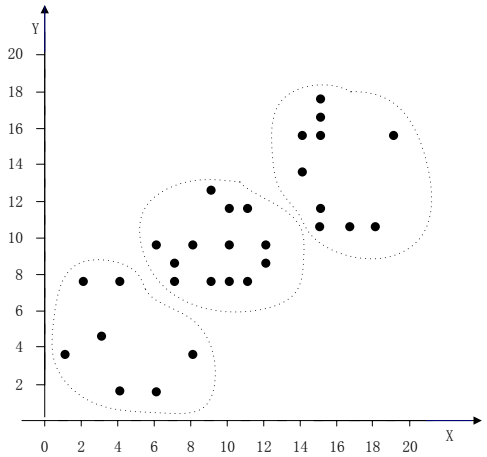

Figure 4. Improved k-means clustering results

IV. APPLICATION OF IMP ROVED K-MEANS METHOD IN DEVICE RISK VALUE CLUSTERING

Equipment risk value is determined by the device status and criticality. The risk of the device is divided into high-risk, high risk, low risk and low-risk four categories according to the power transmission network evaluation criteria. Researchers collected 30 sets of data (the data just list the first three). First, we calculate the average of each dimension of the 30 sets of data, and they are 2.7859 and 2.3621, the improved results are shown in Table 4.

TABLE IV. EQUIPMENT RISK INFLUENCE FACTORS DATA

	Original data		**Revised data**	
Equipment name	***Equipment status score***	***Equipment importance***	***Equipment status score***	***Equipment importance***
s1	2.7257	2.2424	-0.0602	-0.1197
s2	1.1534	1.9835	-1.6325	-0.3786
s3	1.3743	2.3527	-1.4116	-0.0094
...	...	...	...	...

The initial cluster centers are based on the improved k-means method in this article, as shown in Table 5, Table 6 and Fig. 5.

TABLE V. FINAL CLUSTER CENTERS

Initial cluster centers					Final cluster centers				
	Clustering					*Clustering*			
	1	2	3	4		1	2	3	4
Equipment status score	1.0503	1.5576	3.042	6.3775	**Equipment status score**	1.0503	1.3447	2.8886	6.8217
Equipment importance	3.306	1.934	1.887	3.0118	**Equipment importance**	3.306	1.9028	1.9626	3.1843

TABLE VI. NUMBER OF CASES IN EACH CLUSTER

Case number			
Clustering	*Number*	*Clustering*	*Number*
1	4	3	8
2	12	4	6

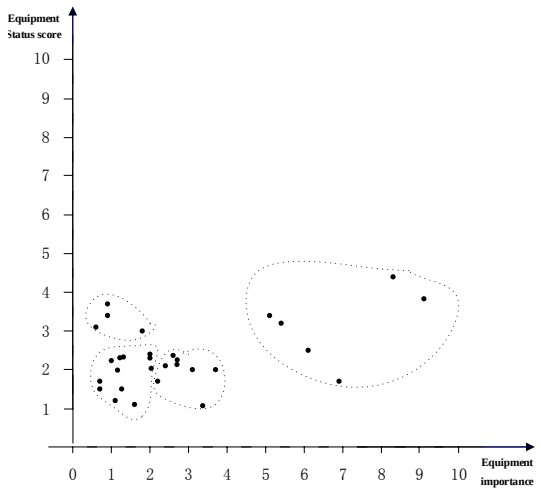

Figure 5. Clustering results

As shown in Fig. 5, X-axis represents the criticality of equipment, Y-axis represents the risk score. Equipment risk score adopts the marking scheme, so the higher the score, the higher the risk. When the equipment risk scores are equal, the higher the degree of equipment importance is, the higher the risk is. Therefore, the fourth category is the high-risk category, and the second category is the low-risk category. Although the third category is lower than the first in equipment status score but the equipment importance degree, so the first category is divided into a lower risk category, and the third category is the higher risk categories, which matches the actual.

Because of the two isolated points (8.2876,4.3719), (9.1217,3.7921), the randomly generated initial center can easily separate these two points into one category affecting the final clustering results, While the improved k-means algorithm in this articles can cluster in a more reasonable way. The risk points can be discovered in time based on the clustering results, what may provide the basis for the next step of risk management and control devices, and thus in where the power transmission network evaluation system plays a role.

V. CONCLUSIONS

K-means clustering method is improved in this article, and reasonable initial cluster centers are obtained by improving the method of selecting the initial centers which enhances the accuracy and credibility of k-means clustering method and eliminates the isolated point. The different risk levels divided based on improved k-means algorithm can cluster the equipment risk value intuitively and accurately so that make references for risk assessment system of the power transmission network.

REFERENCES

[1] Chen Li-fei, Jiang Qing-shan, Wang Sheng-rui. A hierarchical method for determining the number of clusters[J]. Journal of Software, 2008,19(1):62−72.

[2] Zheng Xiao-feng,Xu Jian-min,Lu Kai.Data Clustering of Road Transportation Information System Based on Attribute Dimension Partition and MapReduce[J]. Journal of South China University of Technology.2014,42(8)122-128,135

[3] HuWei. Improved Hierarchical k-means Clustering Algorithm[J], Computer Engineering and Applicationgs.2013,49(2),157-159.

[4] Hentila L, Alatossava M, Czink N, Kyosti P. Cluster-level parameters at 5.25 GHz indoor-to-outdoor and outdoor-to-indoor MIMO radio channels[C]. Mobile and Wireless Communications Summit, 2007:1-4.

[5] Yu Yan-wei,Wang Qin,Kuang Jun,He Jie.An On-line Density-based Clustering Algorethm for Spatial Data Stream[J].ACTA AUTOMATICA SINICA.2012,38(6),1051-1059.

[6] Xie Juan-ying,Guo Wen-juan,Xie Wei-xin,Gao Xin-bo.K-means Clustering Algorithm Based On Optimal Initial Centers Related to Pattern Distribution of Samples in Space[J].Application Research of Computers.2012,29(3),888-892.

[7] Chunsheng Hua, Sagawa R, Yagi Y. Scale-invariant density based clustering initialization algorithm and its application[J]. ICPR 2008 19th International Conference on, 2008:1-4

[8] Yongsheng Sang, Zhang Yi. Motion Determination Using Non-uniform Sampling Based Density Clustering Fuzzy Systems and Knowledge Discovery[C]. FSKD'08 Fifth International Conference on, 2008:1-4.

[9] Sahu L, Mohan B.R. An improved K-means algorithm using modified cosine distance measure for document clustering using Mahout with Hadoop[C]. Industrial and Information Systems (ICIIS), 2014 9th International Conference on, 2014:1-5.

[10] Jieming Wu, Wenhu Yu. Optimization and Improvement Based on K-MeansCluster Algorithm Knowledge Acquisition and Modeling[C]. KAM'09 Second International Symposium on, 2009:1-5.