

PAPER • OPEN ACCESS

Transmission Line Fault Analysis Method Based on Fuzzy Knn

To cite this article: Guorong Zhang *et al* 2019 *IOP Conf. Ser.: Earth Environ. Sci.* **300** 042110

View the [article online](#) for updates and enhancements.

Transmission Line Fault Analysis Method Based on Fuzzy Knn

Guorong Zhang¹, Yu Du¹, Zhiguo Wang¹, Junni Li², Liting Zhai¹, Li Xu¹, Yuanlong Ruan³, Jing Shang^{3,*}

¹State Grid Gansu Provincial Electric Power Company Jinchang Power Supply Company, Gansu 737100, China

²Beijing Guotong Network Technology Co., Ltd., Beijing 100000, China

³College of computer science and Information Engineering, Tianjin University of Science & Technology, Tianjin 300457, China

*Corresponding author e-mail: shangjing48@qq.com

Abstract. The improvement of the quality of life of the masses has led to more stringent requirements on the safety of the power system. Aiming at the problem that knn has low accuracy when dealing with the unconstrained data of the category and the complexity of the power system and the diversification of the structure, the Spark+Hadoop+Hive is used as the technical framework to establish a real-time fault analysis model of the transmission line based on the Spark computing platform. The problem of low timeliness in transmission line analysis further improves the accuracy of the algorithm in fault analysis of transmission lines.

1. Introduction

The current power system has become an important foundation for social development, and the issue of big data has also been upgraded to the national strategic level. With the continuous innovation and development of emerging data mining and analysis technologies such as big data and artificial intelligence, it provides unlimited space for the development of business innovation and intelligent decision-making in the power industry, and provides technology for the safe and reliable operation of the smart grid. Support. When the data is large, the reliability and sensitivity of the previous power grid fault diagnosis analysis model need to be improved; the intelligent power grid fault analysis model has not been constructed yet, and the analysis of fault automation cannot be achieved.

At present, for the fault diagnosis analysis of power grids, the expert system uses the way of reasoning and judgment of the problem by simulating the logic analysis process of the person in decision-making, and dealing with complex problems that only experts can solve. However, with the increase of the amount of data in the knowledge base, it is too time-consuming in terms of information search, update and expansion, which leads to the failure of the actual application effect, and the poor adaptability and poor fault tolerance. Higher requirements for the widespread use of technology. In addition, the offline processing of massive data represented by Hadoop requires frequent I/O operations on the disk, resulting in low computational efficiency and often failing to meet the requirements for online status monitoring and evaluation in the power system.



This paper combines the fuzzy theory and the knn algorithm, and develops a fuzzy knn algorithm to analyze the transmission line data to determine whether it is faulty. The Spark+Hadoop+Hive (full distributed environment) technology architecture is used to construct a real-time fault analysis model for transmission lines based on Spark computing platform, which meets the requirements of real-time fault analysis of power grids.

2. Design of real-time fault analysis model for transmission lines

Real-time analysis of transmission line fault data is an important part of real-time fault monitoring of power systems. Accurate and rapid detection and analysis of real-time data is of great significance in power systems. The invention adopts the method of Spark+Hadoop+Hive, and the technical framework is as shown in Figure 1:

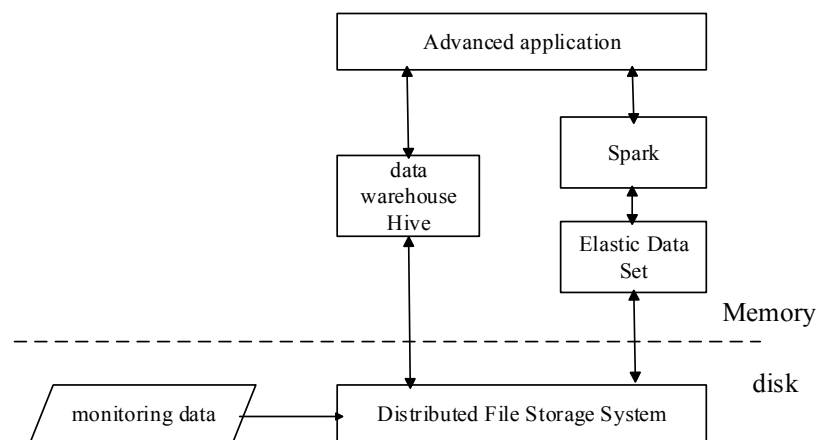


Figure 1. Digital electric line real-time fault monitoring system framework diagram.

The content of the data acquisition module in the model contains the historical data and real-time data of the power grid, and the three-phase voltage and three-phase current signal models of the transmission line simulation results. For historical data, it is especially important to pay attention to the data at the time of the failure, and the changes and fluctuations of the data before the failure, which are of great help to the prediction and judgment. After obtaining real-time data, the data is analyzed by batch processing or stream processing. The key is to compare with the fault data and then make correct classification of the data. The present invention solves the problem in detail. However, due to fewer failures in real life, some of the data required for faults and non-fault situations needs to be obtained through simulation software. The data obtained by the simulation is used to train with the data stored in the power transmission and transformation system database, and then combined with the data prototype obtained from the real system through the fuzzy knn classifier for online fault analysis. The specific process is shown in Figure 2:

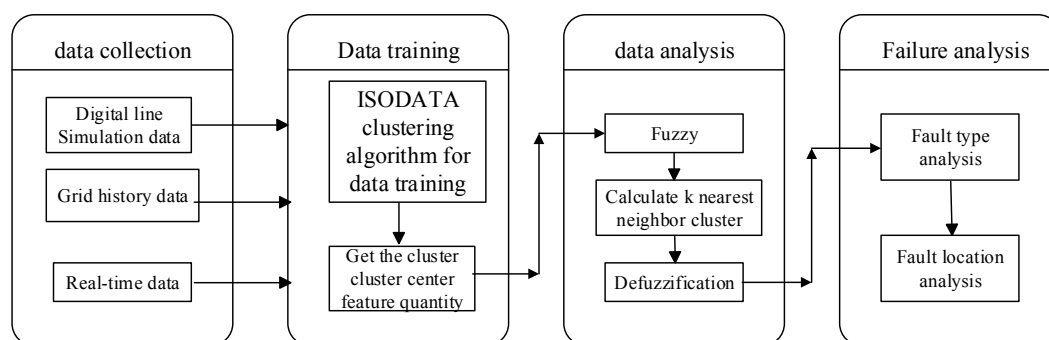


Figure 2. Fault Analysis Model of Transmission Line Based on Fuzzy Knn.

3. Data training based on ISODATA algorithm

ISODATA algorithm, also known as dynamic clustering algorithm, is a commonly used clustering analysis algorithm, which belongs to the unsupervised classification method and is used to obtain representative central points in the cluster. In the data training phase based on ISODATA (iterative self-organizing data analysis) algorithm, the supervised learning method is adopted to use the ISODATA clustering algorithm to reduce the amount of data, so as to obtain the center point of representative clusters. The clustering radius ρ of each one. One of the design goals is to get a more accurate cluster center point and its characteristics, so you need to label the training data to control the merging and splitting of the cluster. The ISODATA algorithm flow is shown in Figure 3.

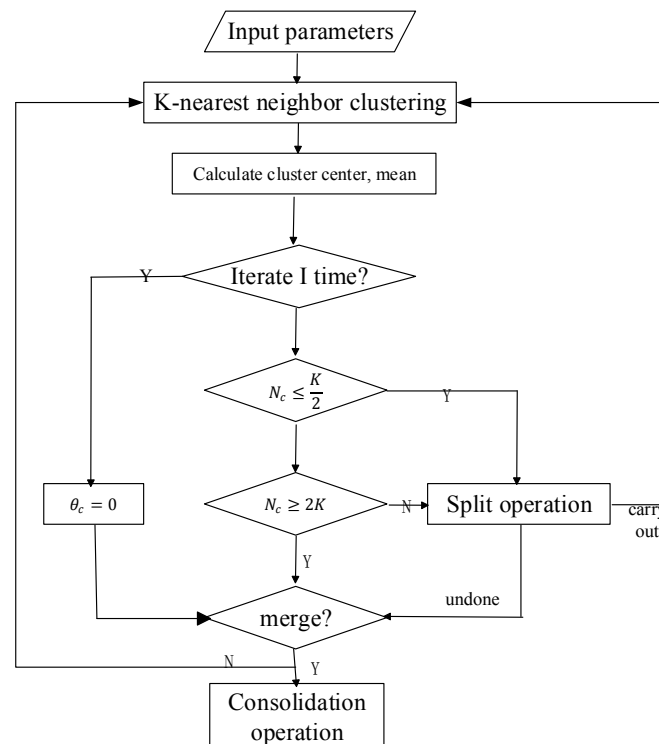


Figure 3. ISODATA algorithm flow chart.

In the calculation of the membership degree of the test data, according to the k samples selected in the previous step, the neighbors of the pattern samples are sequentially given reasonable weights, and the categories are determined according to the membership function of the pattern samples x_i . The exact classification of the samples is guaranteed to determine the type of failure.

3.1. Establish the initial center

Input N pattern samples $\{x_i, i = 1, 2, \dots, N\}$ that collect historical data and simulation results, preselect N_c initial cluster centers $\{z_1, z_2, \dots, z_{N_c}\}$, which may not equal to the number of cluster centers required, the initial position can be arbitrarily selected from the sample.

3.2. Streamlined sample

The N pattern samples are assigned to the nearest cluster S_j , if $D_j = \min\{\|x - z_i\|, i = 1, 2, \dots, N_c\}$, the distance of $\|x - z_j\|$ is the smallest, then $x \in S_j$. If the number of samples in S_j is $S_j < \theta_N$, the sample subset is canceled, at which time N_c is decremented by 1.

3.3. Fix each cluster center

$$z_j = \frac{1}{N_j} \sum_{x \in S_j} x, \quad j = 1, 2, \dots, N_c \quad (1)$$

Calculate the average distance between the pattern samples in each cluster domain S_j and each cluster center:

$$\bar{D}_j = \frac{1}{N_j} \sum_{x \in S_j} \|x - z_j\|, \quad j = 1, 2, \dots, N_c \quad (2)$$

Calculate the total average distance of all pattern samples and their corresponding cluster centers:

$$\bar{D} = \frac{1}{N} \sum_{j=1}^N N_j \bar{D}_j \quad (3)$$

3.4. Discriminate splitting, merging, and iterative operations

- (1) If the number of iteration operations has reached I , that is, the last iteration, then $\theta_c = 0$;
- (2) If $N_c \leq \frac{K}{2}$, that is, the number of cluster centers is less than or equal to half of the specified value, the existing cluster is split;
 - a. Calculate the standard deviation vector of the sample distance in each cluster

$$\sigma_j = (\sigma_{1j}, \sigma_{2j}, \dots, \sigma_{nj})^T \quad (4)$$

Where the components of the vector are

$$\sigma_{ij} = \sqrt{\frac{1}{N_j} \sum_{k=1}^{N_j} (x_{ik} - z_{ij})^2} \quad (5)$$

Where $i = 1, 2, \dots, n$ is the dimension of the sample feature vector, $j = 1, 2, \dots, N_c$ is the number of clusters, and N_j is the number of samples in S_j .

- b. Find the largest component of each standard deviation vector $\{\sigma_j, j = 1, 2, \dots, N_c\}$, represented by $\{\sigma_{j_{max}}, j = 1, 2, \dots, N_c\}$.

- c. In any of the largest component sets $\{\sigma_{j_{max}}, j = 1, 2, \dots, N_c\}$, if $\sigma_{j_{max}} > \theta_s$, one of the following two conditions is satisfied: $\bar{D}_j > \bar{D}$ and $N_j > 2(\theta_N + 1)$, that is, the total number of samples in S_j is more than double the specified value or $N_c \leq \frac{K}{2}$. Then split Z_j into two new cluster center sums, and $N_c + 1$. among them:

$$\begin{aligned} Z_j^+ &= \sigma_{j_{max}} \text{ corresponding component } + k\sigma_{j_{max}}; \\ Z_j^- &= \sigma_{j_{max}} \text{ corresponding component } - k\sigma_{j_{max}}; \end{aligned}$$

- (3) If the number of iteration operations is an even number of times or $N_c \geq 2K$, the existing clusters are combined.

3.5. Update cluster center z_j

If it is the last iteration, the algorithm ends, and outputs the radius ρ of each cluster of the final cluster center z_j ; otherwise, it goes to the first step and goes to the next iteration until the result converges.

4. Test data membership

In this paper, the clustering center feature data z_j after training is fuzzified, the membership function of the cluster center is obtained, and the membership degree of the cluster center z_j subordinate to a certain category c is calculated. The difference from the knn algorithm is that the discriminant of the data type

depends not only on its nearest cluster center category, but also sets the weight of each cluster center according to the size of the distance, and finally determines the type of the data according to the weight.

In each cluster, firstly, according to the Euclidean distance between the test sample X and the cluster center W obtained during the training phase, k samples with the smallest distance from the sample to be classified in the cluster center are selected; then the neighbors of the pattern samples are given. A large weight ρ_k is given, that is, the larger the distance is, the smaller the weight is; the membership degree of the test sample in the cluster is calculated, and finally the category of x_i is determined according to the membership function of the test sample x_i . It concludes that the pattern sample x_i belongs to the membership function $u_c(x_i)$ of the category c , which is expressed as an expression:

$$u_c(x_i) = \frac{\sum_{k=1}^k \frac{u_c(w_k)}{\|x_i - w_k\|^2}}{\sum_{k=1}^k \frac{\rho}{\|x_i - w_k\|^2}} \quad (6)$$

In the formula, the membership function $u_c(w_k)$ of the cluster center w_k takes the value of the radius ρ of the cluster center. According to the membership function $u_c(x_i)$ of the test data, the membership data of the test data x_i subordinate to the k clusters is calculated, and the category label of the test data x_i is the category label of the cluster cluster with the largest membership degree.

5. Conclusion

In the data processing stage, a clustering algorithm based on ISODATA is designed. By comparing the number of samples in the original sample, the purpose of reducing the amount of data is realized. The average distance between the sample and each cluster center is used to realize the cluster. The central point W and the selection of the clustering radius ρ of each one are determined by the clustering radius ρ and the reasonable weights are assigned to the neighbors of the model samples, so that the accurate classification of the fault categories is realized. By writing the fault detection data into the Hadoop system in the form of streaming data, the data is read as an elastic distributed data set, which satisfies the requirements of real-time analysis of the integrated transmission line fault monitoring data.

References

- [1] Pasi Luukka, Feature selection using fuzzy entropy measures with similarity classifier, J. Expert Systems with Applications. 2010 (4)
- [2] J M Keller, M R Gray, J A Givens, et al. A fuzzy k-nearest neighbor algorithm. IEEE Transactions on Systems Man and Cybernetics. 1985
- [3] MA Wei, WU Tao, DUAN Mengya. Research on Clustering Algorithm Based on K Nearest Neighbor Membership, J. Computer Engineering and Applications. 2016 (10)
- [4] Zhou Zhiyang, Feng Baiming, Yang Penglin, Wen Xianghui. Research and Implementation of Stream Data KNN Classification Algorithm Based on Storm, J. Computer Engineering and Applications. 2017 (19)
- [5] Yunlong Gao, Feng Gao. Edited AdaBoost by weighted kNN, J. Neurocomputing. 2010 (16)
- [6] Hong Cheng, Rongchao Yu, Zicheng Liu, Lu Yang, Xue-wen Chen. Kernelized pyramid nearest-neighbor search for object categorization, J. Machine Vision and Applications. 2014 (4)
- [7] Cheng Debo, Su Yijuan, Zong Ming, Zhu Yonghua. Adaptive Nearest Neighbor Classification Algorithm Based on Sparse Learning, J. Computer Engineering and Design. 2015 (07)
- [8] Wang Wei, Zhang Wenbo, Xu Jiwei, Wei Jun, Zhong Hua. Research on Fault Detection Technology of Distributed Software System Based on Statistical Monitoring in Cloud Environment, J. Journal of Computers. 2017 (02)
- [9] Zhou Donghua, Hu Yanyan. Fault Diagnosis Technology of Dynamic System, J. Acta Automatica Sinica. 2009 (06)
- [10] Yan Lijuan, Li Xingyi. Research on KNN Algorithm for Big Data Classification, J. Computer Application Research. 2014 (05)