Power Transformer Condition Assessment Method based on Improved Rough Fuzzy K-means Clustering

Chunpeng Li1,a, Xinyuan Hu2,b, Guofei Guan1, Zhongwen Li2, Qiqi Luan1, Feng Jiang1

1Jiangsu Fangtian Power Technology Co., Ltd., Nanjing, Jiangsu, China
2Nanjing University of Posts and Telecommunications, Nanjing, Jiangsu, China
aemail: licp@js.sgcc.com.cn, bemail: 1219053930@njupt.edu.cn

Abstract: As an important equipment of power transmission and transformation, power transformer plays a direct role on the reliability of power supply. Considering the low proportion of power transformer fault data and the overlap boundary with normal data, an imbalance measurement of cluster size is introduced and a power transformer state assessment based on rough fuzzy k-means clustering is proposed. The developed method can not only judge the type of fault that has occurred, but also predict the possible future faults of power transformers. The effectiveness of the designed method is verified by experimental comparative analysis.

1. Introduction

Power transformer status assessment is based on the monitored transformer status parameters, using a variety of methods to analyze and judge the status of the equipment. Initially, the monitored characteristic data is used to judge the status of the power transformer. And then, with the increase of monitoring equipment, the single monitoring feature can no longer reflect the state of power transformer very well. In order to make comprehensive use of monitoring information while avoid the one-sidedness of single feature, the literature [1] use fuzzy mathematical theory. The uncertainty and fuzziness information of transformer are well calculated by the membership vector of each feature.

Cluster analysis is a very important unsupervised learning algorithm in the field of machine learning and data mining. At present, the application of cluster analysis in transformer data processing can improve the processing of unbalanced data sets, improve the evaluation level of transformer state, and more accurately reflect the real operation state of transformer. K-means clustering divides each sample into a clear cluster, and requires the sample data to have a clear boundary. However, for transformer monitoring data, cluster boundaries are fuzzy or even overlapped, and abnormal data account for the ratio is very low, that is, the scale of clusters is extremely unbalanced, and it is difficult for traditional “hard clustering” methods to achieve ideal clustering results. For the cluster analysis of cross clusters, scholars have also proposed a series of “soft clustering” algorithms [2-5]. What’s more, a few scholars have tried to improve the processing effect with rough k-means algorithm on unbalanced data sets [6-8]. For example, Noordam JC proposed a fuzzy K-means (Cluster size insensitive FCM, csiFCM) which is not sensitive to cluster size; Lin PL improved siibFCM (Size-insensitive Integrity-based FCM, based on csiFCM, siibFCM); a rough K-means clustering algorithm based on the imbalance measurement within a cluster was proposed [9].

Although the above algorithms improve the processing effect of the rough K-means algorithm on imbalanced data sets to a certain extent, the clustering performance of the algorithm on imbalanced data...
sets is still not ideal. In this paper, for the analysis and processing of monitoring data with unbalanced cluster sizes and cross boundaries, an improved fuzzy membership function considering the unbalanced cluster size measurement is introduced to measure the data samples in the cross boundary area. The improved clustering algorithm can analyze power Unba lanced monitoring data for common faults of transformers to evaluate the operating status of power transformers. The effectiveness of the algorithm is verified by the practical experimental data set.

2. Cluster analysis algorithm

2.1 Fuzzy Measure of Cross-boundary Samples of Clusters

Aiming at the impact of the unbalanced cluster size on the clustering results of the RFKM algorithm, this paper first proposes a measurement method for the unbalanced cluster size. Since the upper approximation set contains all the data samples that may belong to the cluster, it describes the size of the cluster to a certain extent, so the measurement formula for the imbalance of the cluster size is proposed as follows:

$$f_i = \frac{|\bar{C}_i|}{\sum_{j \in C^u} |C_j|}$$

where, $|\bar{C}_i|$ represents the number of samples that fall into the approximate set on the i-th cluster, and $C_z$ represents the cross cluster where the data sample $x_j$ is currently located.

The fuzzy membership degree of the cluster size imbalance measure is designed as follows:

$$\mu = \left\{ \begin{array}{ll}
\frac{1}{\sum_{j \in C^u}(f_j, d_j)^{v-1}} X_j \in \bar{C}_i \\
1 & X_j \not\in \bar{C}_i
\end{array} \right.$$

It improves the measurement accuracy of samples in the cross-boundary area. Moreover, it also greatly reduces the computational costs of the algorithm.

2.2 Improved RFKM algorithm based on measurement of sample imbalance at cross-boundary

The iterative calculation formula of the traditional cluster center is as follows:

$$\hat{v}_i = \frac{\sum_{j \in C^u}(f_j, d_j)^{v-1} X_j \in \bar{C}_i}{\sum_{j \in C^u}(f_j, d_j)^{v-1}}$$

It can be seen from the above formula that if the scale of clusters is more severely imbalanced, for minority clusters, the average proportion of the boundary area data samples is relatively large, while for most clusters, the boundary area data the proportion of the average of the sample is relatively small. In order to weaken the influence of the center iteration formula, this paper adopts the iterative calculation formula of cluster center mean as shown in formula (4).

$$\hat{v}_i = \frac{\sum_{j \in C^u} X_j}{\sum_{j \in C^u}}$$

Combining the above-mentioned cluster scale imbalance measurement and the iterative calculation formula of the cluster center mean value, the following improved RFKM algorithm based on the cross-boundary sample imbalance measurement is further given. The steps of the algorithm are as follows.

**Input:** Data set $U$: $U=\{X_j\}_{j=1}^N$.

**Output:** The data object set $U$ is divided into $k$ clusters.

**Step 1:** The number of clusters is $k$, initialize the cluster center $v_i$, distance judgment threshold $\zeta$. 

fuzzy coefficient $m$.

Step 2: For each object $X_j$, calculate the Euclidean distance $d_{ij}$ from $X_j$ to the center point of each cluster $v_i$. Choose $o=(i|d_{ij}=\max(d_{ij}))$, $i=1,...,k$, if $\exists \omega=(i'|d_{ij} \leq \zeta, i' \neq i)$, then $x_j \in C_i$ and $x_j \notin C_{i'}$. Otherwise, $x_j \in C_{i'}$. For all clusters, $C=C_1 \cup C_2$.

Step 3: Calculate the number of samples $|C_j|$ in the approximate set $\overline{C_j}$ on each cluster, and use formula (1) to calculate the imbalance between clusters. Use formula (2) to calculate the membership degree of each data object $x_j$ to its belonging cluster. Use formula (4) to iteratively update the center point of the cluster.

Step 4: If the cluster no longer changes, the algorithm terminates, otherwise it returns to Step 2.

In the iterative process of the above algorithm, the scale of clusters can be adaptively measured according to the changes in the classification of data samples into different clusters, and the contribution of data objects in the boundary area to the calculation of the center points of the cross clusters can be adjusted to reduce the cluster size. The unbalanced cluster size has adverse effects on minority clusters.

On the other hand, when the cluster size is roughly balanced, when the two clusters are $C_1=C_2$, and $f_1=f_2$ at this time, formula (2) degenerates into the traditional fuzzy metric formula. Therefore, the traditional rough fuzzy K-means algorithm can be seen as a special case of the algorithm in this chapter, and the algorithm proposed in this chapter has better adaptability.

3. State Evaluation of Power Transformer Based on Cluster Analysis

The research of emerging artificial intelligence methods is mainly concentrated in the field of supervised learning. It requires the use of a large amount of historical data of known labels to train the model and then use it for state evaluation. Its workload is large and the density of abnormal data is low. Supervised algorithms appears powerless. Through the cluster analysis of power transformer monitoring data, the status changes and existing safety hazards can be found in time, which provides a basis for power transformer status maintenance. The clustering algorithm can quickly and accurately pre-process massive data, and at the same time, it can further evaluate the state of the power transformer and give the evaluation result.

This paper will use the proposed improved RFKM clustering algorithm to deal with the unbalanced data of power transformers. The algorithm introduces the concept of rough set, puts data that is difficult to distinguish between states into the “boundary” area, reminds operation and maintenance personnel to pay attention to these boundary data, and improves the clustering accuracy and state evaluation accuracy.

The evaluation method divides the transformer state grade into two states: normal, abnormal / fault. Firstly, the corresponding indexes of each fault type in the above index system are filtered out from the original data, and then the improved RFKM algorithm based on cross boundary sample imbalance measurement is used to cluster analysis. The specific evaluation process is shown in figure 1.
Input transformer monitoring data set

According to the index system, cluster and analyze the index data corresponding to each failure

Count the lower approximate set samples of the "normal" clusters and mark them as "1"
Generate the state matrix of the transformer

Count the sample of the border area and mark it as "0"

Count the lower approximate set samples of the "fault" cluster and mark them as "-1"

End

Generate evaluation model

Training sample recognition rate >90%
Adjustment parameters m, c

Yes

No

Start

Training sample and model

Adjustment parameters m, c

Training sample recognition rate >90%

Generate evaluation model

According to the index system, cluster and analyze the index data corresponding to each failure

Count the lower approximate set samples of the "normal" clusters and mark them as "1"

Count the sample of the border area and mark it as "0"

Count the lower approximate set samples of the "fault" cluster and mark them as "-1"

End

Figure 1 Evaluation flowchart

The transformer state evaluation method proposed in this paper makes full use of the idea of rough set, clustering the monitoring data of transformers to the lower approximate set that is determined to belong to the cluster and the boundary area where the cluster cannot be determined. If the group of monitoring data is determined to belong to the lower approximate set of the "abnormal/fault" cluster, then this group of data is determined to belong to the fault data; if the group of monitoring data belongs to the boundary area, the group of data belongs to the "abnormal data", that is, it is still running, but may malfunction in the future.

4. Experiment analysis

In order to verify the effectiveness of the evaluation method, 200 sets of power transformer monitoring data were extracted, including 40 sets of normal data and 160 sets of abnormal/fault data, with an imbalance ratio of 1:4. 180 groups as training samples and 20 groups was taken as test samples.

Firstly, the training data set is used to train the classic RFKM algorithm and the improved RFKM algorithm respectively. Then the test data (part) as shown in Table 1 is used to test the accuracy and evaluation effect of the proposed method.

| H2  | CH4 | C2H6  | C2H4  | C2H2 | Types      |
|-----|-----|-------|-------|------|------------|
| 7.5 | 5.7 | 3.4   | 2.6   | 3.2  | Normal     |
| 14.7| 3.8 | 10.5  | 2.7   | 0.2  | Normal     |
| 63.1| 16.7| 4.3   | 9.3   | 10.1 | Normal     |
| 56.3| 22.5| 15.5  | 9.9   | 8.3  | Normal     |
| 19.6| 320.7|279.2 | 574.7 | 0    | Malfunction|
| 147 | 766 | 506   | 2078  | 3.1  | Malfunction|
| 41.6| 25.1| 124   | 15.7  | 206  | Malfunction|

Each group of monitoring data in Table 1 is expressed as a $1 \times 5$ matrix, that is, $x_i = [x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5}]$. Then, the data in Table 1 is normalized according to the formula (5). After the normalization is completed, the data shows in Table 2.

| H2  | CH4 | C2H6  | C2H4  | C2H2 | Types      |
|-----|-----|-------|-------|------|------------|
| 0.3348|0.2545|0.1518|0.1161|0.1429|Normal     |
| 0.4608|0.1191|0.3392|0.0846|0.0063|Normal     |
With the classical RFKM algorithm: divide \{4, 8, 12\} three groups of data samples originally belonging to "fault" into "normal" clusters, and \{7, 9, 10, 19\} four groups of data samples into the boundary area. With the Improved RFKM algorithm: divide a group of \{15\} data samples that originally belonged to "faults" into "normal" clusters, and a group of data samples \{8\} into the boundary area.

In order to show the effect of the algorithm more intuitively, the principal component analysis (PCA) is used to find the "main" components in the data, remove noise and redundancy, and project the above 5-dimensional test data into a 3-dimensional space. The results are shown in Figure 2. Where, the blue "*" and "o" indicate samples that are clustered correctly; the red "o" indicates that samples originally belonging to the majority cluster have been incorrectly divided into at least several clusters; the black "□" indicates samples divided into border areas.

![Figure 2](image)

From the experimental results of the above 20 groups of test data, the traditional RFKM algorithm misclassified three groups of data samples and divided four groups of data samples into the boundary region. The power transformer state evaluation method based on the clustering algorithm proposed in this paper introduces the concept of rough set upper and lower approximation, and considers the impact of cluster size imbalance on the clustering results in the algorithm. The algorithm is more suitable for analyzing power transformers. The improved RFKM algorithm only misclassified 1 set of data samples and 1 set of data samples into the boundary set. Compared with the traditional RFKM clustering algorithm, it achieved the best evaluation results.

5. Conclusion

The improved RFKM algorithm proposed by this paper introduces an improved fuzzy membership function that takes into account the cluster size imbalance measurement to analyze and process the data samples in the cross-boundary area, which improves the clustering effect of the RFKM algorithm on the power transformer imbalance monitoring data. This method can not only judge the fault type of the transformer that has failed, but also predict the possible future fault type of the transformer.

References

[1] Ruijin Liao, Fanjin Meng, Nianrong Zhou, etc. Evaluation Method of Transformer Insulation State Based on Fusion of Set Pair Analysis and Evidence Theory [J]. High voltage engineering, 2014, 40(02): 474-481.

[2] Peters G, Crespo F, Lingras P, et al. Soft clustering - Fuzzy and rough approaches and their extensions and derivatives [J]. International Journal of Approximate Reasoning, 2013, 54(2): 307-322.

[3] Fumin Ma, Ruiqiang Lu, Tengfei Zhang, πRFKM clustering algorithm based on local fuzzy enhancement in boundary region[J]. Control and Decision, 2017, 32(11): 1949-1956.

[4] Ruiqiang Lu, Fumin Ma, Tengfei Zhang. Rough k-means clustering algorithm based on interval 2-type fuzzy metric [J]. Pattern Recognition and Artificial Intelligence, 2018, 31(3): 265-274.
[5] Bicego M, Figueiredo M A T. Soft clustering using weighted one-class support vector machines [J]. Pattern Recognition, 2009, 42(1): 27-32.

[6] Liu Y, Hou T, Liu F. Improving fuzzy c-means method for unbalanced dataset [J]. Electronics Letters, 2015, 51(23): 1880-1882.

[7] Aksoylar C, Qian J, Saligrama V. Clustering and Community Detection with Imbalanced Clusters [J]. IEEE Transactions on Signal & Information Processing Over Networks, 2017, 3(1): 61-76.

[8] Ofek N, Rokach L, Stern R, et al. Fast-CBUS: A fast clustering-based undersampling method for addressing the class imbalance problem [J]. Neurocomputing, 2017, 243: 88-102.

[9] Tengfei Zhang, Long Chen, Yun Li. Rough K-means clustering algorithm based on intra-cluster imbalance measurement [J]. Control and Decision, 2013, 28(10): 1479-1484.