Research on multi-label classification method of transformer based on DBSCAN (Density-Based Spatial Clustering of Applications with Noise) clustering algorithm

MingYu Wang¹,a, Rui Cheng¹,b *

¹ Department of Dispatching and Control, State Grid Tianjin Chengxi Electric Power Supply Company, Tianjin 300190, China
a wangmingyu74@126.com, b* meg5@163.com

Abstract. With the improvement of the intelligent level of power grid and the enhancement of the integrated characteristics of power grid, the degree of discretization of massive data of power equipment gradually increases, which brings great challenges to the safe and stable operation of power grid. How to process and analyze data effectively has become an important research content. Transformer is an important electrical equipment, therefore it is of great significance to monitor the operation status of transformer, to construct transformer operation characteristic label system based on multi-source heterogeneous data, and to realize multi-label classification function. In this paper, a transformer multi-label classification method based on DBSCAN(Density-Based Spatial Clustering of Applications with Noise) clustering algorithm is proposed, which can accurately identify outliers as Noise without input of the number of clustering to be divided, realize the key feature mining of transformer state, and to realize to provide flexible information association and historical data for dispatch and control operators.

1. Introduction

With the gradual expansion of the power grid, the improvement of the intelligent level of the power grid, the gradual formation of the new energy pattern and the strengthening of the integration characteristics of the power grid, people's demand for high efficiency of data is getting higher and higher, while the degree of quantization and discretization of monitoring information of the electric power equipment is gradually increasing. As the core element of power grid, electric power equipment bears the important mission of power generation, transmission, transformation and distribution. The quantization and discretization of monitoring information brought by the intelligence of power grid brings great challenges to the safe and stable operation of power grid. Therefore, it has become an urgent problem to improve the trend warning awareness of equipment operation status by deep integration of artificial intelligence technology and data-driven advantages, to promote the utilization of equipment assets and to improve the safety and stability of electric power equipment.

Transformer is one of the important electric power equipment of power plants and substations. Once the transformer ages or fails, and if it can not be found and dealt with as soon as possible, it will pose a great threat to the safety and reliable power supply of the power grid and cause great losses to the national economy. Therefore, it is of great significance to monitor the running state of transformer in real time or at regular time, to build an operating characteristic label system of transformer based...
on multi-source heterogeneous data, to master its operating characteristics and to realize multi-label classification.

With the quantification of monitoring data of the power grid, how to process and analyze the data effectively has become an important research content. Cluster analysis technology, as one of the important research directions in the field of data mining, has been widely used in many fields. It discussed the use of vibration fault data of hydro-generator sets and proposed an analysis and diagnosis method based on fuzzy clustering in reference [1]. It studied the classification of massive recording data collected by traveling wave transient high speed through semi-supervised clustering technology to achieve effective screening of fault data sets in multi-channel massive recording data in reference [2]. It classified insulator leakage current under different weather condition by density clustering method based on branch dimension in reference [3].

This paper proposes a multi-label classification method of transformer based on DBSCAN (Density-Based Spatial Clustering of Applications with Noise) clustering algorithm. This method does not need to input the number of clustering to be divided, can accurately identify the outlier as noise, and has no bias on the shape of clustering, so as to realize the key feature mining of transformer state.

2. Multi-source data of transformer
For transformers, combined with the data status of the dispatching and control center and the equipment department, we sum up a total of five categories of data by analyzing the data range as Figure 1 shows, including equipment model information, operation and monitoring information, equipment management information, maintenance and experiment information and external environment information.

![Figure 1. Five Multi-source of data of transformer](image)

Data generated by the operation of the power system are scattered in different business systems of the power grid. We get a total of five data sources related to transformers including OMS (intelligent dispatching management system), PMS (equipment management system), fault recording, dispatching and control cloud, D5000 (smart grid dispatching and control system), and online monitoring of power transmission and transformation. The OMS provides the information of device ledger, maintenance, and scheduling events; the PMS can provide equipment test report; the Fault recording can provide fault recording files and reports; the dispatching and control cloud can provide environment, substation information, voltage, current measurement information, alarm, remote signal shift and status information; the D5000 system provides information such as monitoring information and AVC action record; the online monitoring of power transmission and transformation provides the online monitoring status and alarm information of devices.

We standardized the multi-source data of transformer, and constructed the relation between multi-source heterogeneous data. After data processing, knowledge extraction, knowledge fusion and
knowledge update, combined with the requirements of comprehensive analysis of transformer status, we can build an accurate and adaptive operation characteristic labeling system of transformer. When abnormal transformer alarms occur, the big data label center proactively recommends flexible information association and historical data reference for the dispatcher. The transformer label architecture consists of data source acquisition, data preprocessing, label classification and data label center formation. These discrete data in the power system are summarized and stored by multidimensional labels in the data label center. Figure 2 shows the big data tagging architecture.

Figure 2. the big data tagging architecture

3. DBSCAN Clustering

With the rise of "information explosion", "big data" and other concepts, the data mining field has higher and higher requirements for clustering analysis algorithms to deal with different types of data sets. However, some traditional clustering algorithms, such as K-means and BIRCH, can only cluster spherical convex data sets, which is not widely applicable. Density-based clustering algorithm does not need to specify the number of clusters in advance, and can identify any number and shape of clusters in the data set containing noise data. DBSCAN, as a classic representative of density-based clustering algorithm, has been applied more and more in cluster analysis.

3.1. DBSCAN Basics

DBSCAN is a density-based clustering algorithm, which generally assumes that categories can be determined by the tightness of sample distribution[4]. Samples of the same category, they are closely related to each other, that is, there must be samples of the same category not far away from any sample of that category. DBSCAN creates a cluster category by grouping closely related samples. By grouping all groups of closely connected samples into different categories, we obtain the final result of all clustering categories[5].

DBSCAN algorithm requires the user to set two parameters:

MinPts: the minimum number of points in a cluster

ε: the minimum distance between points to be considered as neighbors

These two parameters are generally difficult to determine in the cube, and are usually set by users based on their experience after a preliminary understanding of the cluster data set. MinPts is
relatively easy to determine the value: when there are not many data objects in the data set, \( \text{MinPts} \) is generally 4 in the two-dimensional spatial clustering; In addition, taking the data set \( \lfloor \frac{m}{25} \rfloor \) is also an effective way (where \( m \) is the total number of data samples, and \( \lfloor \cdot \rfloor \) indicates rounding down). Related concepts in DBSCAN are as follows:

Definition 1 \( \varepsilon \)-Nearest neighbor: The region within the radius \( \varepsilon \) of an object \( p \) is called the \( \varepsilon \)-nearest neighbor of object \( p \).

Definition 2 Core object: In object set \( D \), object \( p \) is called a core object if its \( \varepsilon \)-nearest neighbors contain at least the number of objects that meet the neighborhood density threshold \( \geq \text{MinPts} \).

Definition 3 Direct density reachable: in object set \( D \), if object \( q \) is \( \varepsilon \)-nearest neighbor of another object \( p \) and \( p \) is a core object, object \( q \) is directly density reachable from object \( p \), as shown in Figure 3.

![Figure 3. Direct density reachable](image)

Definition 4 Density reachable: in object set \( D \), if there is a point chain \( p_{i_1}, p_{i_2}, \ldots, p_{i_n} \), for \( p_{i_1} \in D(1 \leq i \leq n) \), and \( p_{i+1} \) is reachable from the direct density of \( p_{i} \), then \( p_n \) is reachable from the density of \( p_{i_1} \), as shown in Figure 4.

![Figure 4. Density reachable](image)

Definition 5 Density connected: If object \( o \) exists and both object \( p \) and object \( q \) are density reachable from \( o \), the density of object \( p \) and object \( q \) are density connected, as shown in Figure 5.
3.2. DBSCAN algorithm steps
Input: data set, radius parameter, density threshold;
Output: clustering results and noise data.
Step 1: An unprocessed object is randomly selected from data set, and its -nearest neighbors meet the density threshold requirements and is called a core object;
Step 2: Traverse the whole data set to find all density-reachable objects from object and form a new cluster;
Step 3: Generate the final cluster result through density connection;
Step 4: Repeat Step 2 and Step 3 until all objects in the data set are processed.

4. Application
As the field of power grid dispatching already has the basis of real-time collection of multi-source data, the transformer related data is classified according to the basic information, operation information and status information, and the data label classification framework is put forward as shown in the figure below.

The dimensions of transformer big data labeling include: transformer basic ledger information, multi-source system transformer ID associated entry, transformer test data regularly updated, transformer unqualified time record, transformer fault type labeling, transformer defect frequency statistics, historical transformer status evaluation information, etc.

Through DBSCAN algorithm, transformer related information is calculated and analyzed, and various labels are formed as follows:
Basic ledger information labels: substation name, operation time, transformer number, transformer type, transformer maximum voltage type;
Current status label: maintenance status, four remote information, alarm information, historical overhaul times, defect location, defect cause;
Defect labeling: untreated defect, familial defect;
Computational analysis labels: probability of failure, fuzzy reasoning, clustering commonality.
Figure 6 shows an application example.
5. Conclusion
With the frequent use of large transformer equipment, its failure has become an important factor affecting the normal, safe and stable operation of power system. Timely discovery of hidden dangers of transformer operation and accurate estimation of its development trend can effectively avoid the occurrence of accidents.

In order to manage and monitor the operation status of transformers effectively, based on the monitoring of the normal operation data information of transformers, this paper constructed a multi-label classification system of big data transformer operation characteristics through DBSCAN
clustering algorithm and combined with the comprehensive analysis requirements of transformer status.

Through the analysis and calculation of transformer related data, the current transformer operation situation is summarized and analyzed, and the transformer operation and maintenance plan for the future period of time is provided with guidance and help, which provides a reference for the daily operation and maintenance of transformers and power outage repair work, and provides an important basis for the transformer fault prediction research.

References

[1] Zhang, GB, Shu HC, Yu JL. TravellingWave Field Data Contingency Screening Based on Semi-supervised Clustering Using Generalized Current Modal Components. Proceedings of the CSEE, 2012,32(10): 150-159.

[2] Chen TH, Chen QJ. Fuzzy clustering analysis based vibration fault diagnosis of hydroelectric generating unit[J]. Proceedings of the CSEE, 2002,22(3): 43-47.

[3] Zhou GL, Song YQi, Wang GL, Zhu YL. Research of Condition Monitoring Big Data Storage and Clustering[J].Transactions of China Electrotechnical Society, 2013, 28(2):337-344.

[4] Feng ZH. Research and Application of Clustering Algorithm based on DBSCAN[D]. Jiangnan University, 2016.

[5] Deng D. DBSCAN Clustering Algorithm Based on Density[C]// 2020 7th International Forum on Electrical Engineering and Automation (IFEEA). 2020.