A Density-Based Spatial Clustering of Application with Noise Algorithm and its Empirical Research

Jianfeng Liu [1], Huaqin Qin [1], Ziyi Liu [2, *], Sheng Wang [2], Qiang Zhang [1], Zhengmin He [1]

[1] Beijing Kedong Electric Power Control System Co., Ltd., No. 15, Xiaoying East Road, Qinghe, Haidian District, Beijing 100089, China

[2] School of Economics and Management, North China Electric Power University, Beijing 102206, China

*Corresponding author e-mail:liuziyi@ncepu.edu.cn

Abstract. With the rapid development and wide popularization of information technology, a large amount of data also follows. It is very important to use data mining tools to screen valuable information from complex data. As one of the widely used density clustering algorithms, density-based spatial clustering of application with noisy (DBSCAN) algorithm is an important data mining method. It can find the multi-dimensional relationship between data elements from the data set, complete the clustering of arbitrary shape and noisy data sets when the number of cluster classes is unknown, and support spatial database. Therefore, based on the example of judging the correctness of the relationship between the user's meter and the substation transformer, and supported by the clustering technology of DBSCAN, this paper finally verifies that the density-based clustering method has a good classification effect on the data with high complexity.

Keywords: Station area; Household change relationship; DBSCAN algorithm.

1. Introduction

The line loss rate is a comprehensive core economic and technical indicator of the power company. Strengthening the line loss management in Taiwan is a long-term strategic task and system engineering of the power company. The correct and reliable relationship between households is the basis and prerequisite for the accurate calculation of line losses in the station area. For this reason, the identification of the relationship between households and changes is of great significance to the planning and operation of the power grid and the measurement management of marketing. It can be said that the effect and degree of informatization of a region's regional management is a direct reflection of the region's marketing management level.

Since the low-voltage station area is the final link of the distribution network power supply, there are many types of loads and weak wiring rules. During the operation of the electric power company, technical reasons and irregular management will cause the electric energy meter files to be inconsistent with the actual situation, and the identification of the household change relationship has been Become a difficult problem in line loss management. In the traditional front-line work, the verification of the relationship between Taiwan and households is mainly conducted manually, and on-site power outages and carrier communication are two methods of investigation. Carrier communication is a widely used verification method. This method uses the carrier communication terminal on the transformer and the handheld receiver on the user's meter side to perform carrier communication, and analyzes the characteristics of the communication message to determine the user's ownership. Since it is difficult to ensure the stability of users' power consumption due to power failure verification, carrier communication has many noise sources and is susceptible to interference. Obviously, manual methods should not be used as conventional verification methods. For this reason, a practical method is urgently needed that can quickly and accurately identify the relationship between households in Taiwan and districts. Since the identification of household change relationship is a typical classification problem, common clustering algorithms can be divided into five categories, namely, partition-based clustering, hierarchical clustering, density-based clustering, and grid-based

clustering. As well as model-based clustering, the clustering algorithm has low requirements for the expression and continuity of the model, and can effectively extract and classify system features, and has strong implementation. There are many researches on the application of clustering algorithm to the identification of household change relations. The literature [1] researches and experiments on several platform recognition algorithms such as clustering and deep learning, and analyzes and concludes that the optimal path method is used for platform recognition. , Is a current station recognition algorithm with high recognition accuracy and low cost; Literature [2] analyzes the characteristics of user voltage clusters based on the similarity of user voltage time series data to further detect the relationship between users; Literature [3] After the massive voltage data is preprocessed, the K-means clustering method is used to cluster the data after feature extraction to realize the recognition of the relationship between households; literature [4] extracts the voltage data features by improving the K-means clustering algorithm, and Applying correlation coefficients to realize the diagnosis of the correct station area for users with abnormal characteristics; Literature [5] uses the MDS algorithm to reduce the dimensionality of the voltage data, and uses the improved K-means method to achieve user clustering, thereby identifying the relationship between users; Literature [6] Use principal component analysis to extract features from data collected by smart meters, and simulate different objects to perform fuzzy C-means classification to identify different categories of users; literature [7] uses quantum genetic and fuzzy clustering The zero-crossing offset of the voltage is classified, and the household-change relationship identification is realized by comparing with the zero-crossing offset of the transformer terminal. The clustering methods used in the above studies mostly focus on the partitioned clustering algorithm, and the algorithm has been optimized and improved. However, the partition-based clustering is suitable for small and medium-sized data and the clustering effect has a large relationship with the choice of initial parameters. It is easy to fall into the local optimum. For this reason, the density-based clustering method has begun to be used in the identification of household change relationships. The algorithm is less sensitive to the distribution of data points, and is suitable for the investigation of users who do not match the household relationship. Literature [8] uses an adaptive segmentation aggregation approximation method to extract voltage data features, and applies DBSCAN clustering to further identify users who do not match the user relationship; Literature [9] proposes a derivative-based dynamic time warping algorithm and a density-based algorithm. Noisy spatial clustering application algorithm of household change relationship recognition method. The density-based clustering algorithm is more flexible and convenient when applied to the identification of Taiwan household relationship. In addition, other studies have tried to use model-based clustering algorithm to explore the recognition method of Taiwan household relationship. Literature [10] proposed the two-stage clustering method combining the self-organizing feature mapping algorithm and the k-means algorithm completes the intelligent identification of the relationship between households.

To sum up, the clustering algorithm provides the idea of classification and clustering for the recognition of household change relationships, and relies on the collected big data to improve the efficiency of recognition. At present, research and analysis based on voltage data are more extensive, but there are few related studies that make full use of user power consumption and station line loss data. In order to realize the accurate identification of the relationship between Taiwanese users, this paper applies the density-based DBSCAN clustering algorithm to extract the characteristics of user power consumption to obtain abnormal user clusters to identify the wrong users.

2. DBSCAN algorithm

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm is a density-based clustering algorithm proposed by Ester in 1996. The advantage of this method is that there is no need to determine the number of clusters, and it can be used for clusters of any shape. Make processing and can effectively find noise points and outliers [11]. Different from the general clustering method, the traditional clustering method is only suitable for convex sample sets, and

DBSCAN can not only adapt to the non-convex sample set, but also has a very good degree of fitting to the convex sample set. The key to the DBSCAN clustering algorithm is the determination of the two parameters Eps and MinPts, where Eps represents the neighborhood distance threshold, and MinPts represents the critical value of the number of samples in the neighborhood [12]. If there is a data set $D = X_1, X_2, \ldots, X_n$. The basic definition involved in DBSCAN is as follows:

Definition 1: Core object. For any sample $X_j \in D$, if it contains at least MinPts samples in its Eps neighborhood, then Xj is called the core object.

Definition 2: Direct density. If Xi is located in the Eps neighborhood of Xj, and Xj is the core object, then Xi is directly reached by the density of Xj.

Definition 3: The density is reachable. If there is a data set sequence $p_1, p_2, \ldots, p_T$ that satisfies $p_1 = X_i, p_T = X_j$, and pT+1 is directly reached by the density of pT , then it is said that Xj is reachable by the density of Xi , that is, the reachability of the density is transitive.

Definition 4: Density is connected. If there is a core object sample Xk, both Xi and Xj can be reached by the density of Xk, and the density of Xi and Xj is said to be connected.

As shown in the figure below, X1 and X2 are the core points, X3 and X4 are the boundary points, X2 is directly reached by the density of X1, X3 is directly reached by the density of X2, X4 is reachable by the density of X1, and the density of X4 and X3 are connected.

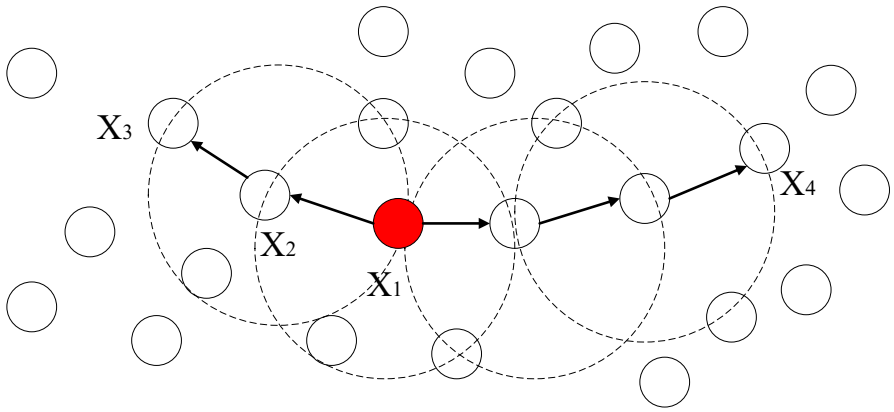

Fig. 1 Schematic diagram of DBSCAN

Based on the above definition, the basic principle of DBSCAN clustering is: choose a point m from the entire data set to determine whether the point is the core point. If it is a core point, search for all the points with reachable density starting from the m point, and these points will generate a cluster. Then the points in the cluster are added to the seed table, and the cluster is expanded by looking for points with the density of the seed points, until no points are added, a complete cluster is formed. If m is not a core point, mark m temporarily as a noise point. Repeat the above steps for the next unprocessed point in the data set to expand the next cluster. All points in the data set have been processed, and the clustering ends [13].

3. Calculation example analysis

3.1 Recognition of the relationship between households

Take the station area with ID 201069 as an example, select the time period from October 1st to October 31st, 2019, cluster the single-day electricity consumption data of each user in the station area through the DBSCAN algorithm, and accumulate 1302 histories Data, which contains equivalent data. To realize DBSCAN clustering, two important parameters Eps and MinPts need to be set. Take any point from the sample data set, and if the point meets the judgment of the core object, form a cluster of all the data points whose density can be reached, and mark the data points that do not belong to any cluster as noise points. Since the settings of Eps and MinPts have a greater impact on the

clustering effect, Table 1 draws up 4 parameter combinations, and performs cluster analysis based on the different parameter combinations set forth in Table 1 to select parameters.

Table 1. Selection of parameters for the DBSCAN algorithm

Label	Eps	MinPts	Number of clusters
a	1.13	2.79	4
b	5.05	2.79	3
c	5.05	10.11	2
d	10.24	10.11	2

The comparison of the clustering results of the four parameters of the DBSCAN algorithm is shown in Figure 2. Each sub-graph shows the relationship between the user's single-day power consumption and the amount of change in the power consumption. The vertical axis represents the single-day power consumption of each user, and the horizontal axis represents the amount of change in electricity consumption on two adjacent days. Various types of users are represented in different colors.

According to Figure 2, the increase in Eps increases the number of points that fall in the neighborhood of the core object, and the number of clusters will decrease, resulting in data that is not of the same type will also be classified as one type. On the contrary, the number of categories will increase, and the data that was originally a category will be fragmented. At the same time, another key parameter MinPts and Eps are adjusted together. When Eps is unchanged, the increase of MinPts reduces the number of core objects. At this time, the data in the cluster that is originally a type will be regarded as noise point markers. Based on the discussion of the clustering effect produced by different parameter combinations, Eps=10.24 and MinPts=10.11 are finally selected as the input parameters of the DBSCAN algorithm to extract user power consumption characteristics, and obtain users with abnormal power consumption characteristics as the users to be detected. To this end, for the first type of users in sub-figure d, the household change relationship identification is performed, and the problematic user and the target station area of the user are found.

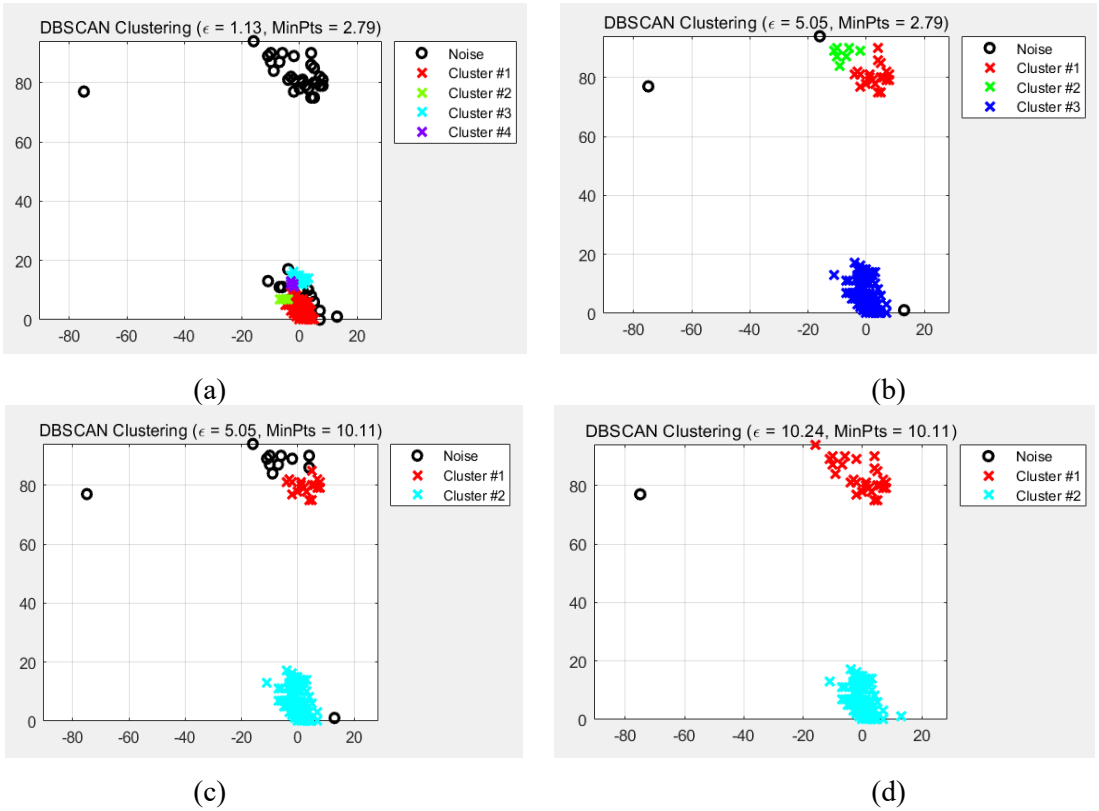

Fig. 2 Comparison of clustering effect with different parameters

3.2 Recognition result output

Based on the above calculations, data with the following household change relationship errors are obtained. Among them, the first group of output data is the result calculated through the calculation examples in this paper. In order to verify that the proposed method is still feasible for the identification of other stations, another 3 stations are randomly selected for verification, the second group of data household change calculation period is still selected in October, and the third and fourth groups of data are selected in December. . Finally, the output results are summarized to obtain the household change relationship output data shown in Table 2.

Table 2. Output data of household transformer relationship

Group	Original station area IDv	Target zone ID	user ID	Time point of household change calculation
1	201069	203846	157547335	2019.10.1~2019.10.31
2	4074594	4148882	158460871	2019.10.1~2019.10.31
3	3478123	201692	183447177	2019.12.1~2019.12.31
4	201692	5484531	174550093	2019.12.1~2019.12.31

According to the data in Table 9, the user number should be adjusted from the original station area to the target station area. Later, through specific verification, the specific time point of the household change can be obtained, and the relationship between the households can be adjusted. Obviously, the analysis of using three correlation coefficients at the same time is more convincing than using a single correlation coefficient to identify the effect, and it can enhance the reliability of the recognition of the relationship between the households to a certain extent. However, there are still some subjective factors. When identifying the situation of multiple users in a station area, there is a deviation from the actual situation. How to further improve the detection accuracy requires further research and discussion.

4. Conclusion

Based on the importance and necessity of the recognition of the relationship between Taiwan and households, this paper proposes a new method based on the DBSCAN clustering algorithm to determine whether the relationship between Taiwan and households is abnormal, and conducts an empirical study on this method. The conclusions drawn are as follows: (1) Clustering algorithm is an important method to analyze the relationship between Taiwanese users, and abnormal users can be classified through clustering. (2) The density-based DBSCAN clustering algorithm can accurately and efficiently classify users and realize the extraction of abnormal users with different characteristics. The density-based clustering method has certain advantages in the recognition of the relationship between users. (3) Accurate and reliable verification methods for the relationship between Taiwan and users can improve work efficiency and guide the analysis of line loss. The clustering algorithm provides theoretical guidance for further application of data mining to identify the relationship between Taiwan and households, improves the efficiency of the recognition of the relationship between Taiwan and households, and has high feasibility in practical engineering applications.

References

[1] Li Guochang, Song Weiqiong, Xian Huizhu, Han Liu, Hu Xiaoye, Yuan Juan. Research and application of electric energy meter area recognition method based on zero-crossing time and SNR algorithm[J]. Electrical Measurement and Instrumentation, 2019, 56(07)):148-152.

[2] Cai Yongzhi, Tang Jie, Que Huakun, Li Jian, Guo Wenchong. A method for verifying the relationship between households and transformers in a station based on the analysis of voltage cluster characteristics[J]. Guangdong Electric Power, 2021, 34(08): 50-60.

[3] Huang Xu, Wang Weiheng, Wu Shuang, Hu Wei. Research on the Recognition Method of Relations Between Taiwan and Households Based on Big Data of Electricity Consumption[J]. Power Supply and Consumption, 2019, 36(10): 22-29.

[4] Zhou Gang, Huang Rui, Liu Dudu, Zhang Zhimin, Hu Junhua, Gao Yunpeng. Diagnosis of abnormal household change relationship based on improved K-means clustering and Pearson correlation coefficient [J/OL]. Electrical measurement and instrumentation: 1-13 [2021-11-02].

[5] Wang Jiaju, Wan Zhongbing, He Zhongxiao, Wang Jia, Xie Zhi, Wang Xiao. Recognition method of household relationship based on multidimensional scale analysis and improved K-means[J]. Electrical Automation, 2020, 42(02): 56-59.

[6] Zeng Shunqi, Wu Jiekang, Li Xin, Cai Zhihong. Identification of the relationship between station voltage and users based on fuzzy C-means clustering algorithm[J]. Sichuan Electric Power Technology, 2021, 44(03): 69-75+87.

[7] Yao Li, Zhang Jiangming, Ni Linna. Recognition of household change relations in low-voltage stations based on quantum genetics and nuclear fuzzy clustering[J]. Electrical Measurement & Instrumentation, 2020, 57(20): 106-113.

[8] Cui Xueyuan, Liu Shengyuan, Jin Weichao, Lin Zhenzhi, Xuan Yuhua, Wang Haibo. Household change relationship and phase identification method based on APAA and improved DBSCAN algorithm [J]. Power System Technology, 2021, 45(08): 3034-3043.

[9] Liu Su, Huang Chun, Hou Shuaishuai, Huang Shifu, Li Jianqi. A method of household change relationship recognition based on DDTW distance and DBSCAN algorithm [J]. Automation of Electric Power Systems, 2021, 45(18): 71-77.

[10] Song Weiqiong, Guo Shuai, Li Ji, Liu Heng, Guo Qiuting, Hu Wei. Intelligent identification of household change relations in distribution stations based on voltage time series data [J/OL]. Journal of Electric Power System and Automation: 1-8[2021-11-02].

[11] Ma Liangyu, Sun Jiaming, Yu Shilei, Zhao Shangyu. Research on abnormal working conditions of wind turbines based on DBSCAN and SDAE [J]. Chinese Journal of Power Engineering, 2021, 41(09): 786-793+808.

[12] Jiang Qijia, Jiang Zhongming, Tang Dong, Zeng Jingming. Research on the method of detecting gross errors in slope safety monitoring data based on SSA-DBSCAN [J/OL]. Journal of Yangtze River Scientific Research Institute: 1-8 [2021-10-13].

[13] Guo Yanjie. Improved DBSCAN algorithm and its application research [D]. North University of China, 2020.