A clustering method of Gas load based on FCM-SMOTE

Dong Hong Huang[1], Dan Liu[2], Ming Wen[3], Xin Li Dong[4], Min Wen[5], Xing Hao Zhao[5,*]

[1]BEIJING GAS GROUP CO., LTO., ,100035 Beijing, China
[2] BEIJING GAS GROUP CO., LTO., ,100035 Beijing, China
[3] BEIJING GAS GROUP CO., LTO., ,100035 Beijing, China
[4] BEIJING GAS GROUP CO., LTO., ,100035 Beijing, China
[5]SCHOOL OF MODERN POST (SCHOOL OF AUTOMATION), Beijing University of Posts and Telecommunications,100876 Beijing, China

Abstract. For the design and planning of gas-fired boiler system, the load of gas-fired boiler is an important basic data. Load clustering analysis, combined with the application of data mining technology and gas boiler system, excavates the hidden load patterns in a large number of disordered and irregular loads, and classifies them, so as to solve many problems in gas boiler system. The current load clustering methods have more or less problems. The invention first carries out data PVA dimension reduction processing on the huge gas data, and then carries out cluster analysis. In the actual application of gas-fired boilers, the data objects we are faced with are usually unbalanced data sets. In order to solve the problem of sample imbalance, we use the FCM-SMOTE algorithm to oversample the clustered data to make the data set into a balanced data set.

1 Technical background

The research on load clustering of gas-fired boiler system is to classify gas load scientifically and effectively, and user load clustering is to dig out the relationship and composition of different types and areas of load through cluster analysis[1]. In the planning and design of the gas-fired boiler system, whether it is to estimate the economy of the project or to determine the construction scale of the gas-fired boiler system, the load of the gas-fired boiler is an important basic data in the design process[2]. The load cluster analysis combines the data mining technology and the application of the gas boiler system, through the data mining to analyze the gas load characteristics, and excavates the hidden load patterns in a large number of disordered and irregular loads, and classify it, through the typical load curve, solve many problems in the gas boiler system, such as gas load forecasting, demand side response analysis. More and more data acquisition devices, such as smart meters can collect a large number of users' gas consumption. Different types of users, such as civil, commercial, industrial, agricultural, have great differences in gas consumption patterns, even if they are the same type of users, their gas consumption patterns may be different. Based on the classification of load data, the gas consumption patterns of different gas users can not only support gas companies to carry out orderly heating, off-peak management, time-sharing gas and other market competition strategies and provide more personalized heating services. It also helps to improve the understanding of the gas consumption patterns of different gas users, so as to carry out more efficient demand side management. In addition, users can adjust consumption strategies more economically and optimally according to the problems found in load classification, which not only reduces costs, but also improves energy efficiency.

Load classification refers to the classification of load types according to different ways or different standards. China has a variety of classification standards, including functional classification, occurrence time classification, and terminal loss classification. However, users can not be refined classification, there are the following problems:

The main results are as follows:

(1) The load pattern can not be found accurately.

It is difficult to find the changes and differences of gas consumption patterns.

The general load classification affects the follow-up application of load pattern recognition.

(4) The data of gas load increases rapidly, the quality of data is difficult to be guaranteed, and it is difficult to classify.

In summary, the current load clustering methods have more or less problems. In view of the importance of load clustering in gas systems and the difficulties faced by existing traditional methods, it is necessary to study new effective load classification methods to meet the current load clustering needs.

*Xing Hao Zhao: zhxh@bupt.edu.cn

2 The invention solves the problem

At present, there are roughly two kinds of research work on the clustering of gas boiler load curves at home and abroad. If we only consider the clustering analysis of the data, because of the huge characteristics of gas load data and high data dimension, it will result in a large amount of calculation. Usually for low-dimensional data, the dataset is more complete, and each dimension has similar ability to distinguish clusters. In order to greatly reduce the storage space of the data set and shorten the time for calculating the distance of data points to improve the efficiency of the clustering algorithm, the invention first carries out data dimensionality reduction processing, and then carries out clustering analysis.

In the actual application of gas-fired boilers, the data objects we are faced with are usually unbalanced data sets. In order to solve the problem of sample imbalance, we use the smote algorithm to oversample the clustered data to make the data set into a balanced data set.

3 Research status

Based on the research status of user load curve clustering at home and abroad, the existing research work can be divided into two kinds. The first category directly uses the original load data for clustering analysis. The main algorithms are k-means (k-means) algorithm, fuzzy C-means (FCM), self-organizing map (SOM), hierarchical clustering algorithm and so on. Because the user's original daily load data is used, the disadvantage of this kind of clustering analysis is that the data dimension is high, the amount of computation is large, and it is easy to fall into dimensional disaster, which is also a test for the amount of calculation and storage. The second kind of method is indirect load clustering analysis, which uses the intermediate algorithm to preprocess the daily load curve, which is roughly divided into two processing methods. One is load dimensionality reduction, which can be applied by commonly used dimensionality reduction methods, such as singular value decomposition ((Singular Value Decomposition)). There is another method for processing the load-based time series in frequency domain and time domain, such as discrete Fourier transform ((Discrete Fourier Transfom)), and another morphological feature measurement method based on time series for clustering. These three methods can reduce the dimension of load data and reduce the amount of computation.

Liu Liqing, Ding Qiaolin [3-5]and others studied the influence of the previous numerical processing method of load clustering on the results of fuzzy C-means clustering (FCM)[6]. Firstly, various data preprocessing methods were used for the actual load measurement data of IRIS data sets and different users within the jurisdiction of a certain company, including summation standardization, maximum and minimum standardization, maximum processing standardization, etc. The accuracy and influence of different processing methods on the clustering results of FCM[7] algorithm are studied. The results show that the FCM clustering results obtained from the data processed by sum standardization and maximum standardization are the most accurate, but the clustering effect is not good for high-dimensional data sets with a large number of features, and the computation is large and inefficient. Cheng Xiang[8] selects K-means algorithm to analyze the load measurement data, and PCA, uses DBI index to select K value in dimensionality reduction technology. Aiming at substation load clustering, a weighted K-means algorithm is designed, which is more suitable for substation load clustering with classified users. Chen en and Wu Hao[9] use singular value decomposition to transform the coordinates of the characteristic values of the daily load curve. According to the contribution rate of different features under the new coordinates, the numerical values under the new coordinates are selected as the dimension reduction index of the daily load curve. Finally, the K-means algorithm is improved, and the weighted Euclidean distance is introduced as the distance function to achieve a robust load clustering method based on SVD.

Aiming at the problem that the identification of minority class samples is more difficult than that of most class samples, Weiss GM[10] deeply classifies the main reasons for the decline of traditional classification methods caused by unbalanced data sets. Oversampling is to add a small number of samples, the original data can be well retained, although oversampling also has the disadvantage of marginalization, so in the field of classification, oversampling is usually used. SMOTE (Synthetic Minority Oversampling Technique, synthesis of a small number of oversampling techniques) algorithm is a commonly used oversampling method, which well solves the problem of over-randomness in random upward sampling[11]. However, the algorithm can not overcome the problem of data distribution of unbalanced data sets, and it is easy to cause the problem of distribution marginalization. Because the distribution of negative samples determines their optional nearest neighbors, if a negative sample is at the edge of the distribution of the negative sample set, the "artificial" samples generated by the negative samples and adjacent samples will also be at this edge, and will be more and more marginalized, thus blurring the boundary between positive and negative samples, and making the boundary more and more blurred[12]. This kind of boundary fuzziness improves the balance of the data set, but increases the difficulty of the classification algorithm.

Therefore, the invention combines the FCM algorithm with the SMOTE algorithm, first uses the FCM algorithm to find out the central store of the original negative class, and then uses the SMOTE algorithm to obtain a new balanced data set.

4 Technical scheme

4.1 PCA dimensionality reduction

PCA (Principal Component Analysis, principal component analysis (PCA) is often used in machine learning and is one of the steps of data preprocessing. When PCA simplifies and reduces the dimension of data, it is mainly based on the following two factors: first, high-

dimensional feature space contains a lot of unnecessary redundant information, and features are related to each other; second, high-dimensional data calculation is more complex. The goal of PCA is to keep the information of the original data set as much as possible and simplify the high-dimensional data in the case of damage. Select the variables that contribute the most to the information of the data sample to minimize the variance of the cost function.

According to the gas load vector used in this paper, $\mathrm{x}_1, \mathrm{x}_2, K, \mathrm{x}_H$ is recorded as the gas load vector of 48 points a day. The purpose of principal component analysis is to linearly combine the original gas load vector.

In the formula:

$$\begin{cases} I_1 = v_{11}x_1 + v_{12}x_2 + K + v_{1H}x_H \\ I_2 = v_{21}x_1 + v_{22}x_2 + K + v_{2H}x_H \\ M \\ I_H = v_{H1}x_1 + v_{H2}x_2 + K + v_{HH}x_H \end{cases} \quad (1)$$

$\mathrm{x}_1, \mathrm{x}_2, K, \mathrm{x}_H$ is a daily 48-point gas load vector;

v is the gas load vector coefficient;

$I_1, I_2, K, I_{\mathrm{d}'}$ is the eigenvector corresponding to the largest d' eigenvalues.

d' is the principal component dimension under the new coordinates of the output in the process of dimensionality reduction.

And the following constraints are met:

(1) The sum of squares of the above coefficients is equal to 1: ${v_{i1}}^2 + {v_{i2}}^2 + K + {v_{iH}}^2 = 1$

(2) There is no correlation among the principal components:

$$COV(I_i, I_j) = 0, i \neq j, \quad i, j = 1, 2, K, H$$

(3) The importance of principal components is determined according to the decreasing variance.

4.2 FCM algorithm

Fuzzy C-means clustering algorithm (FCM) is one of the most widely used clustering algorithms based on objective function. Fuzzy C-means clustering algorithm is optimized by traditional hard clustering algorithm. Hard clustering adopts the principle that the membership degree is either 0 or 1, uses the mean square approximation method to construct the conditional nonlinear programming problem, and solves the clustering problem with the help of the objective function, so the clustering objective function is often expressed in the form of intra-class average error and $J(U,P) = \sum_{k=1}^{c}\sum_{i=1}^{m} d_{ik}^2$.

The principle of FCM algorithm is as follows: the gas load data sample set $X = \{x_1, x_2, ..., x_N\}$ to be clustered is divided into c classes, where $x_1, x_2, ..., x_N$ is the load clustering sample, N is the number of samples, $2 \le c < N$, and the clustering center matrix is expressed as $V = (v_1, v_2, ..., v_c)^T$, when the clustering objective function meets certain requirements. The objective function is used to represent the sum of squares of the distances from the sample points in each class to all kinds of centers, and the optimal value of the objective function is calculated by iterative optimization method. The objective function is calculated as shown in the formula:

$$\min J(X, U, v_1, v_2, ..., v_k, ..., v_c) = \sum_{k=1}^{c}\sum_{i=1}^{N} {u_{ki}}^{m_0} d_{ki}^2 \quad (2)$$

In the formula:

u_{ki} is the membership degree, indicating the gas load data I (I). The, N) samples belong to k (Kwon 1, 2, … The degree of, c) classes;

U is the membership matrix of u_{ki};

v_k is the k th gas load clustering center.

d_{ki} is the Euclidean distance between v_k and x_i, calculated as $d_{ki} = \| v_k - x_i \|$;

m_0 is a fuzzy parameter, which controls the fuzzy degree of the membership matrix U, also known as the smoothing parameter. The larger the m, the higher the fuzzy degree of the classification. Usually $m_0 = 2$.

The constraints of the objective function are as follows:

$$\begin{cases} \sum_{k=1}^{c} u_{ki} = 1, 1 \le i \le N \\ u_{ki} \in [0,1], 1 \le k \le c, 1 \le i \le N \\ \sum_{i=1}^{N} u_{ki} \in [0, N], 1 \le k \le c \end{cases} \quad (3)$$

By using Lagrange Multiplier Method to find the optimal solution, the final result can be obtained:

$$u_{ki} = \frac{1}{\sum_{j=1}^{c}\left(\frac{d_{ki}}{d_{ji}}\right)^{\frac{2}{m_0 - 1}}} \quad (4)$$

$$v_k = \frac{\sum_{i=1}^{N} (u_{ki})^{m_0} x_i}{\sum_{i=1}^{N} (u_{ki})^{m_0}} \quad (5)$$

In the formula:

λ is Lagrange multiplier;

N is the number of samples;

d_{ki} is the Euclidean distance between v_k and x_i, calculated as $d_{ki} = \| v_k - x_i \|$;

d_{ji} is the Euclidean distance between v_j and v_i, calculated as $\mathrm{d}_{ji} = \| v_j - v_i \|$;

When the clustering parameters such as the clustering number (c), the fuzzy parameter (m_0) and the maximum number of iterations are known, the clustering process of the selected samples can be realized by repeated iterations.

4.3 SMOTE algorithm

The main purpose of SMOTE algorithm is to balance the data set by increasing the number of minority samples.

Suppose an unbalanced data set, for each gas load data sample X in a small number of samples, search for its nearest neighbor samples K (these K nearest neighbor samples belong to a small number of samples). Assuming that the upward sampling rate of the data set is n, then n samples are randomly taken from the K nearest neighbor samples (there must be K > n), to mark these n samples as $y_1, y_2, ..., y_n$. The associated data samples X and y_i , carry out the corresponding random interpolation operation through the correlation formula between X and $y_i (i = 1, 2, ..., n)$, and get the interpolation samples p_i ,. For each data sample, n corresponding minority class samples are constructed.

The interpolation formula is as follows:

$$p_i = X + rand(0,1)^*(y_i - X), i = 1, 2, ..., n \quad (6)$$

In the formula:

X represents the data samples in a small number of classes in the gas load data set;

y_i represents the first of the n nearest neighbor samples of data sample X.

The sampling rate n depends on the degree of imbalance of the data set, and the calculation formula of the sampling rate n is as shown in the formula for calculating the imbalance between the majority and minority classes of the data set.

$$n = round(IL) \quad (7)$$

Where $round(IL)$ represents the value obtained by rounding IL.

Through the above interpolation operation, the majority class samples and minority class samples can be effectively balanced, thus the classification accuracy of unbalanced data sets can be improved.

We can know that in the SMOTE algorithm, because each interpolation is associated with the sample point data and its K nearest neighbors, because the SMOTE algorithm oversampling interpolation is random, if the interpolation result is not ideal, the oversampling operation will blur the positive and negative class boundary of the data type. Therefore, we preprocess the minority data sets through the clustering algorithm, and on this basis, we establish clustering to obtain the cluster center, through the cluster center as the base point for oversampling operation, we can effectively solve the deficiency of fuzzy positive and negative class boundary.

FCM-SMOTE algorithm needs to make a decision operation on the data before interpolation, then the synthesized samples can effectively avoid invalid interpolation, reduce the probability of fuzzy positive and negative class boundaries, and maintain the distribution pattern of a few classes of data.

The core of the algorithm is mainly analyzed from three parts: determining the boundary points of a few classes, judging the dangerous points and modifying the oversampling formula.

Determine the boundary points of a few classes.

For unbalanced dataset S, define it as a formula:

$$S = \{(W_1, v_1), (W_2, v_2), ..., (W_m, v_m)\} \quad (8)$$

In the formula:

W_i is an example of gas load data set S;

$v_i \in \{1.0, 0.0\}$, $1 \le i \le m$ is the class tag of W_i .

Let P be a minority class set, v_j = 1.0pm $1 \le j \le |P|$

Magi Q be a majority class set, v_l =0.0, $1 \le l \le |Q|$.

For a few class examples, the K nearest neighbors are obtained, and the categories of K nearest neighbors are judged. If there is a majority class, the minority class is a boundary minority class data sample, which is included in the boundary data sample set R, and R is expressed as shown in the formula:

$$R = \{(X_1, v_1), (X_2, v_2), ..., (X_t, v_t)\} \quad (9)$$

In the formula:

X_i represents the boundary point instance;

v_i is the class label.

At the same time of recording R, a few class instances and most of their nearest neighbors in R are recorded in the data sample set T.

Judge the danger point.

After clustering the minority clusters, it is necessary to judge the number of data samples of the boundary minority classes in the data samples of each cluster in the sample set T. If it is more than 1, the clustering c_i , needs to be judged again.

The Euclidean distance d_{wp} from the cluster center u_i , to the boundary minority class instance X and the d_{wq} of the Euclidean distance between the majority of the K nearest neighbors corresponding to the u_i to the minority class instance X are determined.

If there is $d_{wq} < d_{wp}$, and the K nearest neighbors of X are most classes, the point X is judged to be the clustering dangerous point, the point is deleted, and the cluster center is recalculated.

And so on, until this situation does not exist, and finally determine the clustering and cluster center.

Modified oversampling formula.

After preprocessing the data with clustering algorithm, we need to carry out oversampling operation, and for the improved SMOTE algorithm, we also need to modify the corresponding oversampling formula.

For the oversampling formula of FCM-SMOTE algorithm, the maximum Euclidean distance is obtained by

considering the Euclidean distance from the cluster center to each cluster data sample. Get a new interpolation formula as shown in the formula:

$$X_{new} = u_i + rand(0, H)^*(X - u_i), i = 1, 2, ..., k \quad (10)$$

In the formula:

X_{new} is the sample of new interpolation;

u_i is the cluster center;

X is the original gas load sample data in clustering with u_i as the cluster center.

Rand (0 ~ H) denotes a random number between 0 and H.

5 Case analysis

In order to verify the feasibility of the over-sampling method proposed in this section, the FCM-SMOTE over-sampling algorithm proposed in this paper is experimentally verified by using the relevant unbalanced data set on a computer with 8.00GB memory, AMD processor, 1.8GHz main frequency, 500GB hard disk and Windows7 operating system, using MATLAB R2016a software.

Firstly, dimensionality reduction is carried out for the data.The above FCM-SMOTE method was used to perform over-sampling operation on the electricity sample set, and compared with random over-sampling, basic SMOTE sampling, B-SMOTE and S1-SMOTE methods. The results are shown in Figure 1. Due to the current user to take the under-voltage method, under-current theft more, a few can also change the power factor measured by special means, and the power factor can be calculated through the voltage, current value combined with other data, power factor and voltage, current has a certain correlation. In FCM clustering in FCM-SMOTE method, cluster number C is set as 2.

Table 1 lists the parameter Settings for each comparison algorithm. M is the number of nearest neighbors of the negative class sample point, in which there are n positive class samples.

Table 1. Parameters setting of each comparison algorithm

Oversampling algorithm	Parameter Settings
SMOTE	m=20
B-SMOTE	m=20、n=5
S1-SMOTE	m=20、n=5

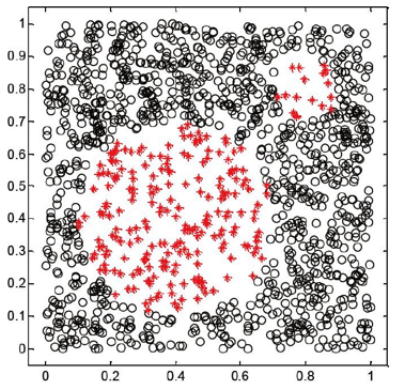

Figure 1 the original data

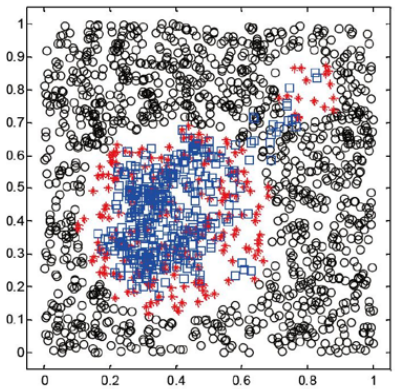

Figure 2 SMOTE

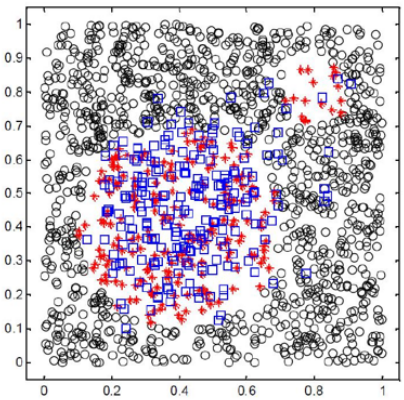

Figure 3 B-SMOTE

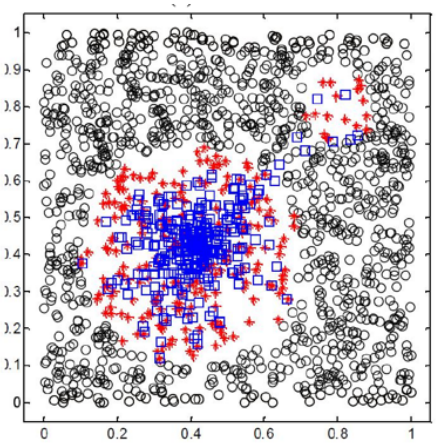

Figure 4 SI-SMOTE

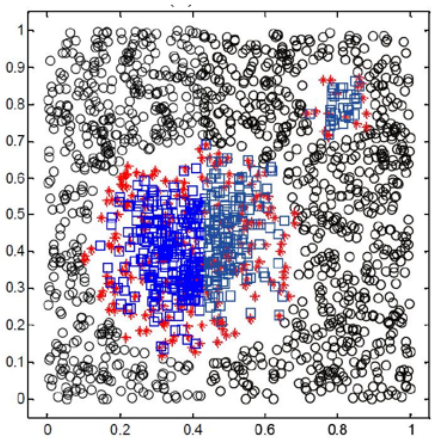

Figure 5 FCM-SMOTE

In the above figure, the black circle represents the positive sample of the original data, the red star represents the negative sample of the original data, and the blue block represents the negative sample generated by the sampling algorithm. It can be observed that the random sampling contains a large number of repeated data, but some negative data are not selected. The negative sample generated by SMOTE method is slightly better, but it also has the problem of repeating with the original data, and a lot of noise points are sampled, which is easy to confuse the boundary of positive and negative sample. The data generated by B-SMOTE generally follows the distribution trend of the original data, but the newly added negative samples overlap more with the original samples, and the boundary is fuzzy. S1-SMOTE oversampling avoids the introduction of noise to some extent, and the boundary is clear. However, as can be seen from the figure, the generated negative sample is too close to the center point. However, FCM-SMOTE method finds the center of the sample cluster through the clustering method and limits the generated negative sample within the boundary of positive and negative sample, thus avoiding the generation of noise data and the blurring of the boundary of positive and negative class.

6 Conclusion

Unbalanced data sets can have a negative impact on the performance of machine learning. The classification results of the unbalanced data sets are more likely to be biased towards the positive sample, which ignores the important information contained in the negative sample, making the decision boundary of the classifier different from the actual spur results of the positive and negative samples.

As can be seen from the above experiments, for gas data, dimensionality reduction processing is carried out first, and then cluster analysis is carried out.

Combined with FCM and SMOTE algorithm, this method has a better effect than other methods. The clustering method is used to find the center of the sample cluster and limit the generated negative sample within the boundary of positive and negative sample, thus avoiding the generation of noise data and the blurring of the boundary of positive and negative class.

7 Innovation point

In view of the problems existing in the load clustering method, the invention adopts data mining and depth learning methods to solve the problem. The innovation of this paper is as follows:

Gas load clustering method based on PCA-FCM-SMOTE.

The invention first carries out data dimensionality reduction processing, and then carries out clustering analysis, and uses the smote algorithm to oversample the clustered data so that the data set becomes a balanced data set. In the actual application of gas-fired boilers, the data objects we are faced with are usually unbalanced data sets. It solves the problem of sample imbalance, greatly reduces the storage space of the data set and shortens the time of calculating the distance of data points, so as to improve the efficiency of the clustering algorithm.

Solve the problem of non-ideal oversampling interpolation.

In the SMOTE algorithm, because each interpolation is associated with the sample point data and its K nearest neighbors, because the SMOTE algorithm oversampling interpolation is random, if the interpolation result is not ideal, the oversampling operation will blur the positive and negative class boundaries of the data type. The clustering algorithm is used to preprocess the minority data sets, and on this basis, clustering is established to obtain the cluster center. Through the oversampling operation based on the cluster center, the deficiency of fuzzy positive and negative class boundary is effectively solved. The characteristic attribute of the FCM-SMOTE algorithm determines its application significance. The algorithm is analyzed by determining the boundary points of a few classes, judging the dangerous points and modifying the oversampling formula. The synthesized samples can effectively avoid invalid interpolation, reduce the probability of fuzzy positive and negative class boundaries, and maintain the distribution pattern of a few classes of data.

The invention carries out cluster analysis on the load data of the gas-fired boiler through the PCA+FCM+SMOTE algorithm. The characteristic of gas load data is huge and the data is unbalanced. because there is a lot of unnecessary redundant information in the high-dimensional feature space and the calculation of high-dimensional data is complex, we first reduce the PCA dimension of the gas load data to reduce the storage space and computing time of the data set. Then FCM algorithm is used to classify the reduced-dimensional data. At the same time, in order to solve the problem of sample imbalance, FCM+SMOTE algorithm is used to classify the gas data more evenly.

Reference

1. Zhou K L, Yang S L, Shen C. A review of electric load classification in smart grid environment[J]. Renewable & Sustainable Energy Reviews, 2013, 24(Complete):103-110.
2. Hu Yangchun. Research on Power Load Pattern Recognition Method Based on Improved K-means Clustering Algorithm [D].University of Electronic Science and Technology of China,2018.
3. Liu Liqing. Research and design of power user load pattern recognition system [D]. North China Electric Power University, 2012. Ying Wen Chess. Energy Internet: trends and key Technologies [J]. International financing, 2020 (02): 30-32.
4. Liu Liqing, Ding Qiaolin, Zhang Tiefeng, Chen Jian. The influence of data preprocessing methods on fuzzy C-means clustering [J]. Electric Power Science and Engineering, 2011,27 (08): 24-27+46.
5. Liu Liqing, Ding Qiaolin, Zhang Tiefeng, Sun Jinbao. Research on load pattern recognition method of power users [J]. Journal of North China Electric Power University (Natural Science Edition), 2011,38 (06): 29-33
6. Ritu Agarwal and Om Prakash Verma. Robust copy-move forgery detection using modified superpixel based FCM clustering with emperor penguin optimization and block feature matching[J]. Evolving Systems, 2021, : 1-15.
7. Zhiming Cai et al. Retraction Note to: A FCM cluster: cloud networking model for intelligent transportation in the city of Macau[J]. Cluster Computing, 2021, : 1-1.
8. Cheng Xiang. Research on power load clustering method based on load measurement data [D]. Zhejiang University, 2017.
9. Chen Ye, Wu Hao, Shi Junyi, Shang Jiayi, Sun Weizhen. Application of singular value decomposition method in Dimension reduction Cluster Analysis of Daily load Curve [J]. Power system Automation, 2018, 42 (03): 105-111.
10. Qian Hongbo, he Guangnan. Summary of non-equilibrium data classification [J]. Computer Engineering and Science, 2010,32 (05): 85-88.
11. Starling James K. and Mastrangelo Christina and Choe Youngjun. Improving Weibull distribution estimation for generalized Type I censored data using modified SMOTE[J]. Reliability Engineering & System Safety, 2021, 211(prepublish) : 107505-.
12. Hongpo Zhang et al. An effective convolutional neural network based on SMOTE and Gaussian mixture model for intrusion detection in imbalanced dataset[J]. Computer Networks, 2020, 177