ACTIVE LEARNING ENTROPY SAMPLING BASED CLUSTERING OPTIMIZATION METHOD FOR ELECTRICITY DATA

Wang Qingnan and Zhang Zhaogong

School of Computer Science and Technology, Heilongjiang University, Hei Long Jiang, China

ABSTRACT

Clustering is a crucial part in the field of data mining, and common clustering methods include division-based methods, hierarchy-based methods, density-based methods, and grid-based methods. In order to improve the accuracy of clustering, an optimization study is made mainly for the division-based method FCM clustering, and an FCM clustering method that integrates active learning and principal component analysis (PCA) is proposed. The method first uses principal component analysis to reduce the dimensionality of the data to reduce the computation of electricity data, then trains the sample model by active learning, and introduces the entropy (Entropy) method in the uncertainty sampling method, the larger the entropy means the greater the uncertainty of the sample, and the smaller the entropy means the smaller the uncertainty of the sample, so as to filter the electricity data, and finally the electricity data are clustered by FCM clustering The power data is finally categorized by FCM clustering, and with the proliferation of power data, the power data can be more accurately categorized using this method to achieve the stability of the power grid as well as the utilization rate. Experimental results on three datasets show that this method improves the accuracy of power data clustering by up to 2 percentage points compared to the traditional clustering method without active learning, and achieves good results in each dataset compared to other methods.

KEYWORD

Active Learning, Data Mining, FCM Clustering, Principal Component Analysis, Unsupervised Learning

1. INTRODUCTION

In recent years, China's electricity market reform has made great progress, China's electricity market construction continues to advance in depth, the initial establishment of a 'unified and open, competitive and orderly' power market system, effectively promote the optimal allocation of power resources, energy clean and low-carbon transformation. But at the same time there are still a series of problems such as immaturity of new energy participation in the market trading mechanism, the price of power auxiliary services and compensation mechanism is not sound, the lack of technical standards system in the power market, the professional quality of power trading personnel still needs to be improved.

With the increasing scale of power users, the electricity consumption data has risen sharply, but the classification of power users has been unreasonable and not detailed. This is due to the fact that the process of classification is not combined with reality and the real power consumption load attributes of electricity users are not considered. Along with the continuous progress of data mining technology, the continuously improved clustering algorithms have been applied to the

David C. Wyld et al. (Eds): AIAA, DPPR, DSA, ICSS, IOTE, NLPTA, WEST, CACIT, BMLI- 2022
pp. 209-220, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.122316

classification of electricity users one after another. The literature [1] classifies electricity users by user load curve through FCM algorithm for the demand side of electricity users, and evaluates the effect of clustering using internal evaluation index DBI. The literature [2] classifies the daily load into weekday electricity load, weekend electricity load, and holiday electricity load by using daily load characteristic analysis, and clusters the electricity users by five characteristic indexes from maximum power, peak total ratio, flat total ratio, valley total ratio, and load factor, respectively, using k-mean algorithm. The literature [3] proposed hierarchical clustering to classify electricity users, firstly by Markov model to collect customer electricity load, and then by the first layer of meticulous collection for the second layer of clustering analysis, using K-means algorithm for hierarchical analysis. In the literature [4], based on the analysis of the characteristics of electric power big data, a classification method of electric power users that fuses downscaling and clustering is proposed, and the fusion of downscaling and clustering is realized by using Spark platform, which has a great improvement in accuracy and running time than before. In the literature [5], an evaluation index of electric load clustering independent of the normalization method is proposed, and the nearest neighbor propagation algorithm is applied to load curve clustering, and the results show that the peak normalization method has a better clustering effect and performs better compared to the traditional load curve clustering. The purpose of power user classification is to maintain the stability of the power grid and improve the utilization of power grid resources so as to alleviate the growing contradiction between power supply and demand in China. Wu [6] published a study on smart meter user classification based on deep learning in 2020, and proposed a deep neural network model with a simpler structure by optimizing the structure of convolutional neural network. This model identifies and classifies users based on the features automatically extracted by the convolutional neural network, which reduces the computational complexity in the calculation of high-dimensional samples, reduces the overfitting problem, and has obvious advantages in prediction accuracy. Active learning is to query the most useful unlabeled samples by a certain algorithm and leave it to manual labeling, and then train the classification model with the queried samples to achieve a better model with fewer labeled samples. Active learning enables effective power data filtering by labeling the most useful sample data for classification to train the classification model, thus filtering the data with abnormal or highly deviated power data. In the literature [7], active learning and data classification algorithms are integrated and studied, and the principles and execution process of active learning and classification algorithms are analyzed in detail, and an active learning data classification algorithm for big data is proposed, which not only reduces the cost of manual labeling but also reduces the size of the training set, which is a good solution for the bottlenecks encountered in data classification. In the literature [8], a KNN algorithm based on group active learning is proposed to combine with electricity market transaction data for example analysis, which has better classification effect and is suitable for electricity market transaction behavior analysis and supply and demand decision. In the literature [9], a clustering adaptive active learning selection strategy is proposed in order to reduce the redundancy of selected samples in the active learning process. The initial samples are selected by clustering, which makes the initial samples representative and accelerates the active learning process.

Combined with the above discussion, this paper proposes an optimization method for power data clustering based on active learning entropy sampling. The ultimate goal of power data classification is to achieve the stability of the power grid and to improve the utilization of power resources while significantly improving power transactions.

Contribution

(1) Because of the large scale of power data, there are some power data abnormalities or large deviations from most of the data, so the introduction of active learning in the screening of power data and the introduction of entropy (Entropy) [10]method in uncertainty sampling can effectively

screen the data and greatly save the cost of manual labeling.
(2) The power trading users are categorized by FCM clustering method, and the users are divided into five categories based on electricity load, daily load curve, and maximum power, which achieves the stability of power grid and improves the utilization rate of power resources.
In this paper, we first select unlabeled samples by using principal component analysis, then filter electricity data by active learning, and finally categorize electricity data by FCM clustering, and compare the proposed method with other existing methods, and the results show that the proposed method improves the accuracy of electricity data categorization.

The first part of the paper describes the progress of work on electricity user clustering and active learning, and proposes an optimization method for electricity data clustering based on active learning entropy sampling. The second part details the method used in this paper, the third part presents the implementation of the method and the clustering results, the fourth part presents the experiments conducted on three datasets and the experimental results, and the fifth part draws conclusions and provides an outlook on future work.

2. METHODS

2.1. Data Preprocessing

In this paper, 96 points of data [11] obtained from 200 electricity users taken every 15 min from 0 to 24 were selected as the feature vector for cluster analysis. Since the initial data are usually missing or wrongly omitted, the data need to be cleaned. If the data point is suddenly high or low, the data is deleted; if the data point is missing more than 20%, the data point is deleted; if the data point is missing less than 20%, the mean value of the data point's neighboring data is used to supplement; if the neighboring data is also missing, the search for non-empty data continues backward.

In order to improve the efficiency of the algorithm, feature extraction is required for the selected data. Different load characteristic indexes are affected by the load components, and in order to eliminate the influence of weather and period, the average value of 96 points of load data of consecutive working days of sample users is selected as the daily load curve data in this paper.

The value domain of the original data may have large differences, and if the original data are processed directly, it will make the data with large values and small values very different and cannot be analyzed effectively, so the data need to be normalized. In this paper, the maximum value method is used for normalization, and the values are normalized to the interval [0,1].

2.2. Model Flow

Firstly, we use principal component analysis algorithm to reduce the dimensionality of user load data and extract user data features, and then use FCM algorithm to cluster the processed load data. The 96 data points of the daily load curve represent the user load data of a day, and 96 data points of a day are 96 principal components. Due to the time cost factor caused by the excessive amount of data, the principal component analysis method is used, which can firstly reduce the data dimensionality and represent most of the user features through fewer dimensions; secondly, the data less than 85% can be directly classified into one category, without the next clustering, reducing the time required for clustering and improving the efficiency of the algorithm. Therefore, this paper introduces principal component analysis as a way to improve the efficiency of the algorithm[12].

Active learning is used to query the most useful unlabeled samples through certain algorithms and hand over to manual labeling, and then the queried samples are used to train the classification model to achieve a better model with fewer labeled samples. Active learning enables effective power data screening by labeling the most useful sample data for classification to train the classification model, thus screening the data with abnormal or highly deviated power data.
Using the FCM algorithm, the power user classification results are obtained by clustering the load data greater than 85%. The typical electricity consumption curves are extracted for each category of electricity user data to obtain the electricity consumption curves of each category, and the electricity consumption curves are used to analyze the category to which each user belongs[13].

3. IMPLEMENTATION

3.1. Implementation of Principal Component Analysis

In this paper, 96 points of data obtained every 15 minutes from 0 to 24 are selected as the data of cluster analysis. The principal component analysis method is used to reduce the dimensions of power data[14]. The first three principal components are selected for clustering if their cumulative contribution rate is greater than 85%, and if their cumulative contribution rate is less than 85%, they are not clustered. The selected data is used as the manual annotation sample for active learning, so as to reduce the cost of manual annotation. As shown in Figure 1:

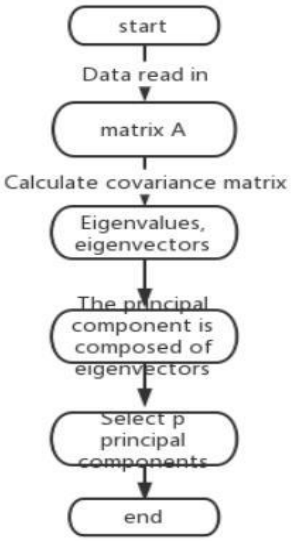

Figure 1. PCA Flow Chart

First, the input data sets are standardized so that each of them can be analyzed roughly in proportion. There are m samples (x1, x2,..., xm) for principal component analysis, and each sample is described by n features. The formula for standardization is as follows[15]:

$$S = \frac{1}{n}\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad (1)$$

$$\tilde{x}_i = \frac{x_i - \bar{x}_i}{s_i} \quad (2)$$

Next, calculate the correlation coefficient matrix. Rij is the correlation coefficient between the ith index and the jth index. The formula is as follows[16]:

$$r_{ij} = \frac{\sum_{k=1}^{n} \tilde{x}_{ki} \cdot \tilde{x}_{kj}}{n-1} \quad (3)$$

Then calculate the eigenvalues and corresponding eigenvectors of matrix R, and m new index variables are composed of eigenvectors. Finally, the cumulative contribution rate of the eigenvalue is calculated as follows[17]:

$$b_j = \frac{\lambda_j}{\sum_{k=1}^{m} \lambda_k} (j = 1, 2, \cdots, m) \quad (4)$$

$$\alpha_p = \frac{\sum_{k=1}^{p} \lambda_k}{\sum_{k=1}^{m} \lambda_k} \quad (5)$$

The samples whose cumulative contribution rate is greater than 85% are selected as sample labels for active learning of artificial marks.

3.2. Implementation of Active Learning

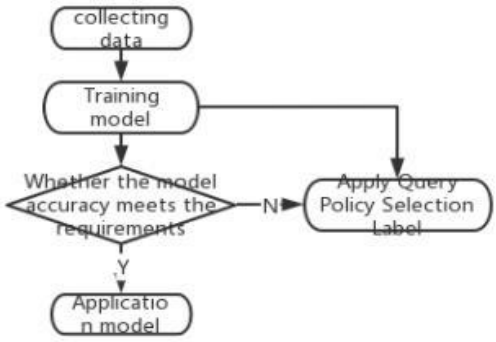

Figure 2. Active learning Flow Chart

First, the samples obtained by principal component analysis are taken as active learning samples. Active learning includes the following processes, sample selection, model training, model prediction and model updating, as shown in Figure 2.

In the process of sample selection, the query method of uncertainty sampling is used to extract the sample data that is difficult to distinguish from the model and provide it to business experts or labeling personnel for labeling, so as to achieve the ability to improve the effect of the algorithm

at a faster speed. Entropy is introduced to describe the uncertainty of a sample. Entropy can be used to measure the uncertainty of a sample. The larger the entropy, the greater the uncertainty of the sample. The smaller the entropy, the smaller the uncertainty of the sample. Compared with the least confidence and margin sample, the entropy method takes into account all categories of the model for an x. While the least confidence only considers the maximum probability, the margin sample considers the maximum probability and the secondary probability. Therefore, the sample data with large entropy can be selected as the pending label data. The mathematical formula is[18]:

$$x_H^* = \text{argmax}_x - \sum_i P_\theta(y_i \mid x) \log P_\theta(y_i \mid x) \tag{6}$$

3.3. Implementation of Clustering Algorithm

FCM is a fuzzy clustering algorithm, which provides more flexible clustering results than other clustering algorithms. First initialize the membership matrix, then judge whether the maximum number of iterations is exceeded, calculate the cluster center, update the membership matrix, and finally judge whether the iteration termination conditions are met, as shown in Figure 3.

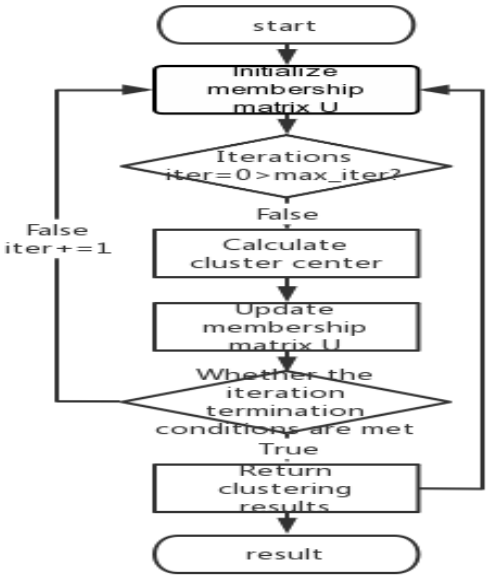

Figure 3. Cluster Flow Chart

Objective function of FCM[16]:

$$J_m(U,V) = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^m d_{ij}^2 \tag{7}$$

U_{ij} refers to the membership value, the membership degree of element j to category i, d_{ij} square refers to the distance between element j and center point i, and the whole represents the sum of weighted distances from each point to each category.

The final effect of clustering is that the intra class similarity is the minimum, and the inter class similarity is the maximum. At this time, the sum of weighted distances between the point and the center is the minimum. Therefore, it is enough to minimize the objective function.
Therefore, the expression of the optimal solution[19]:

$$\min(J_m(U,V)) = \min\left(\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^m d_{ij}^2\right) \tag{8}$$

Selection of cluster number c:

$$\mathrm{L}(c) = \frac{\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^m \|v_i - \bar{x}\|^2/(c-1)}{\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^m \|x_j - v_i\|^2/(n-c)} \tag{9}$$

Cluster center:

$$p_i = \frac{\sum_{i=1}^{n} (u_{ij})^m x_i}{\sum_{i=1}^{n} (u_{ij})^m} \tag{10}$$

The load clustering based on FCM algorithm is measured by the load curve. After the above data processing, the FCM clustering algorithm divides the power users into five categories according to the idea of maximum membership. The daily load curve drawn according to the characteristics is shown in Figure 4.

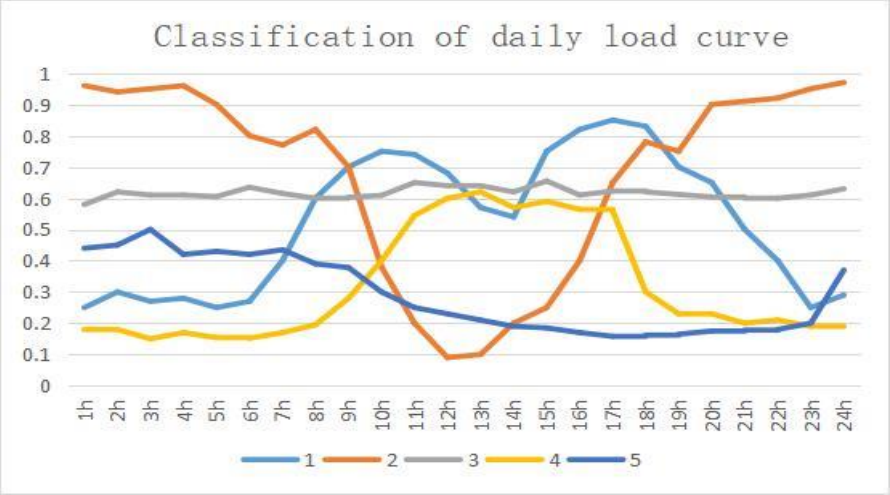

Figure 4. Daily load curve

After clustering, the first type of users accounts for about 30% of the total number of samples. The load of the first type of curve is small from 1 o'clock to 6 o'clock, rising rapidly from 7 o'clock until the peak load is reached around 11 o'clock. From 11:13, the load decreased, and from 13:00 to 18:00, the load increased, reaching the second peak of power consumption. After 18:00, the load dropped sharply. The curve generally shows that the power consumption is large in the daytime and small at night, with two peaks at 11:00 and 18:00. Most of the curves correspond to state-owned enterprises and public institutions.

The second type of users after clustering accounts for about 15% of the total number of samples. This type of curve has a high power consumption from 1 to 8 points, and a sharp decline from 8 points to 13 points. It rose sharply again from 13:00 to 24:00. Such users generally have low power consumption in the morning and high power consumption from afternoon to evening. Most of the curve corresponded to heavy industrial metal enterprises.

The third category of users accounts for about 24% of the total number of samples after clustering. This kind of curve is stable from 1 point to 24 points, with small fluctuations. This kind of users generally have the highest electricity consumption in the morning and night, and most of them are energy consuming enterprises and manufacturing industries.

After clustering, the fourth type of users accounts for about 18% of the total number of samples. This type of curve rises sharply from 8 points to 12 points, tends to be stable from 12 points to 17 points, and then drops sharply after 17 points. In general, power consumption is high during the day, low at night, and average power load is medium. Most of these curves correspond to banks or shopping malls.

The fifth category of users accounts for about 13% of the total number of samples after clustering. The curve tends to be stable from 1 point to 8 points, tends to decline after 8 points, and slowly rises again after 22 points. Most of these curves correspond to light industries, hospitals and other enterprises.

4. EXPERIMENT

4.1. Baseline Method

1. K-means: Given a K value and K initial class cluster centroids, each point is assigned to the class cluster represented by the nearest class cluster centroid, and after all points are assigned, the centroids of the class cluster are recalculated based on all points within a class cluster, and then the steps of assigning points and updating class cluster centroids are performed iteratively until the change in class cluster centroids is small or the specified number of iterations is reached. number of iterations, through this method to categorize the power data[20].
2. Random: Randomly select the unlabeled sample data for active learning, and then categorize the power data by clustering algorithm.
3. FCM: FCM algorithm is a division-based clustering algorithm, its idea is to make the same cluster is divided into the maximum similarity between objects, and the minimum similarity between different clusters, through the degree of fuzziness to divide the data, so as to categorize the power data.
4. BVSB: The idea of uncertainty sampling is to focus the selection on those samples that cannot be classified with certainty by the current classifier, and the BvSB value of the sample is defined as the difference between the two highest conditional probabilities, by which the electricity data are categorized.

4.2. Experimental Result

We tested the proposed method on three image data sets: COIL-20, Caltech101, and MNIST. And compared with other methods, first is K-means clustering algorithm, second is the method of randomly screening data, FCM clustering algorithm and BVSB (Best versus Second Best) method.

4.2.1. COIL-20 Dataset

We tested the proposed method on the COIL-20 dataset[21]. The dataset contains 1500 images with 20 different objects. For each image, we resize it to 32 × 32. We iterated all the methods for 140 times, and showed the accuracy in figures and tables. The results show that the proposed method has the highest accuracy.

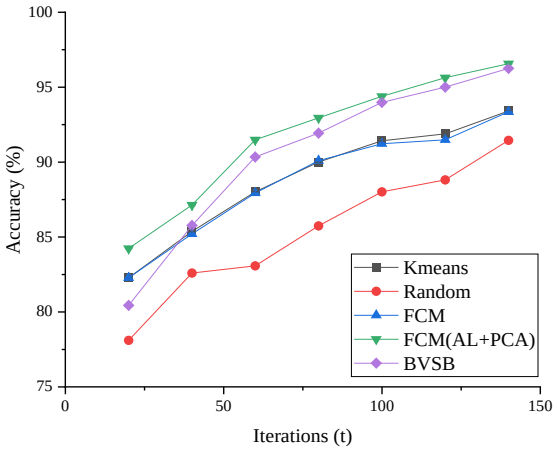

Figure 5. Accuracy on COIL-20

Table 1. Accuracy on COIL-20 Dataset

Iterations	K-means	Random	FCM	FCM(AL+PCA)	BVSB
20	82.27	78.10	82.27	**84.23**	80.43
40	85.37	82.59	85.20	**87.13**	85.77
60	88.01	83.07	87.93	**91.48**	90.34
80	90.01	85.74	90.09	**92.95**	91.93
100	91.42	88.01	91.22	**94.38**	93.98
120	91.88	88.81	91.48	**95.63**	95.05
140	93.41	91.45	93.35	**96.58**	96.20

4.2.2. Caltech101 Dataset

We also tested the performance of the proposed scheme on the Caltech101 dataset. Since the number of some categories is very small, we only selected 10 categories, including aircraft, bonsai, car side, chandelier, face, motorcycle and watch, among which each category contains more than 100 images. There are 3379 pictures in total. We iterated all the methods for 140 times, and showed the accuracy in figures and tables. The results show that the proposed method has the highest accuracy[22].

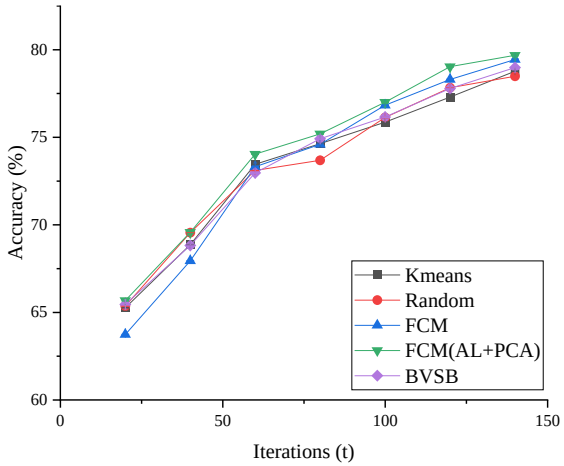

Figure 6. Accuracy on Caltech101

Table 2. Accuracy on Caltech101 Dataset

Iterations	K-means	Random	FCM	FCM(AL+PCA)	BVSB
20	65.26	64.50	65.36	**65.67**	63.73
40	68.86	67.19	69.55	**69.57**	67.94
60	73.45	69.92	73.10	**74.03**	73.31
80	74.63	74.69	73.68	**75.19**	74.59
100	75.86	75.32	76.13	**77.02**	76.83
120	77.29	76.82	77.83	**79.03**	78.30
140	78.77	78.48	78.94	**79.68**	79.22

4.2.3. MNIST Dataset

We also tested this result on a subset of the MNIST dataset. The dataset contains 300 images for each number (0-9), a total of 3000 images. We iterated all the methods for 140 times, and showed the accuracy in figures and tables. The results show that the proposed method has the highest accuracy[23].

Table 3. Accuracy on MNIST Dataset

Iterations	K-means	Random	FCM	FCM(AL+PCA)	BVSB
20	65.71	64.65	65.86	**67.33**	65.64
40	69.58	68.61	70.55	**72.18**	69.94
60	71.65	71.10	71.41	**73.48**	71.70
80	72.85	73.31	73.85	**75.18**	72.51
100	74.26	74.95	75.19	**76.12**	75.24
120	75.12	75.13	75.38	**76.88**	76.17
140	76.52	76.11	76.77	**78.11**	76.85

4.3. Result Analysis

For each method, we ran 140 simulations on each dataset, and showed the classification accuracy under different iterations in the figure[24]. We observed that the proposed method has obvious performance advantages over other methods (K-means, Random, FCM, BVSB) in three datasets (COIL-20, Caltech101, MNIST), especially in COIL-20 and MNIST datasets. Although the advantages of different datasets are different, the performance of the proposed method on all three datasets is generally better than that of other methods. From the above results, the proposed method performs better than K-means and Random in almost all cases. On COIL-20 Dataset, although the accuracy of the proposed method is slightly higher than BVSB, it is much higher than the other three methods[25]. On the Caltech 101 dataset, although the accuracy of the proposed method is similar to the other four methods, it is also higher than the other four methods. The method proposed on MNIST dataset also has higher accuracy than the other four methods. We find that by using this method, the accuracy of classification is improved and the cost of manual annotation is saved.

5. CONCLUSIONS

In this paper, we propose a method that combines principal component analysis and active learning into clustering algorithm. The introduction of active learning in power data screening and the introduction of entropy method in uncertainty sampling can effectively screen data and greatly save the cost of manual labeling. The power trading users are classified by FCM clustering method. Based on the power load, daily load curve and maximum power, the users are divided into five categories, realizing the stability of the power grid and improving the utilization rate of power resources. In this paper, the principal component analysis method is used to reduce the dimensions of the data, and then the power data is filtered through active learning. Finally, the power data is classified through FCM clustering. Comparing the proposed method with other existing methods, we demonstrate the effectiveness of the proposed method on three data sets. From the experimental results, we can see that the proposed method has achieved good results on three image reference data sets, and the results show that the proposed method improves the accuracy of power data classification.

In the future, we will investigate active learning in more depth, integrating it with other practical applications, and will continue to optimize and improve performance in our work on active learning combined with clustering.

REFERENCES

[1] Research and Markets; Global Electricity Trading Markets to 2023 with Sales Analysis on both Day-ahead Trading & Intraday Trading - ResearchAndMarkets.com[J]. Energy Weekly News, 2019, 20(9).
[2] Da Xu and Jinting Bai. A Preliminary Study on Edge Computing[J]. Journal of Research in Science and Engineering, 2022, 4(2).
[3] Wan Shaohua et al. Special Issue on Optimization of Cross-layer Collaborative Resource Allocation for Mobile Edge Computing, Caching and Communication[J]. Computer Communications, 2022, 181 : 472-473.
[4] Lin Yu et al. Secure Deduplication Schemes for Content Delivery in Mobile Edge Computing[J]. Computers & Security, 2022, : 102-602.
[5] M.A.Rahman et al., “Blockchain-Based Mobile Edge Computing Framework for Secure Therapy Applications,” IEEE Access, 2018.
[6] Zeng Siming, Li Tiecheng, Li Shun, Liang Jifeng, Fan Hui, Yang Jun, Wu Fuzhang. Power load clustering analysis based on improved density peak algorithm [J]. Science and Technology and Engineering, 2022,22 (25): 11032-11040.

[7] Xiaoqin Gao, Design and implementation of power user classification and demand side management platform based on FCM clustering [D] Nanchang University, 2021 DOI:10.27232/d.cnki. gnchu. 2021.002557.
[8] L. Huang, X. Feng, C. Zhang, L. Qian, and Y. Wu, "Deep reinforcement learning-based joint task offlfloading and bandwidth allocation formulti-user mobile edge computing," Digit. Commun. Netw., vol. 5, no. 1, pp. 10–17, 2019.
[9] Han Ruobing Research on short-term load forecasting based on cluster analysis and feature learning [D]. China University of Mining and Technology, 2022. DOI: 10.27623/d.cnki. gzkyu. 2022.001334.
[10] XU Yuanbin, LI Guohui, GUO Kun, et al. Power load clustering based on improved parallel K-Means algorithm[J]. Computer Engineering and Applications, 2017, 53(17): 260-265.
[11] ZHANG Bin, ZHUANG Chijie, HU Jun, et al. Integrated clustering algorithm for power load curve based on dimensionality reduction technology[J]. Proceedings of the CSEE, 2015, 35(15): 3741-3749.
[12] Xie bin Research on peak shaving and valley filling evaluation method of distribution network based on power consumption behavior of power users [D] Beijing University of Posts and telecommunications, 2021 DOI:10.26969/d.cnki. gbydu. 2021.000306.
[13] Sun Yi, Mao Yehua, Li Zekun, Zhang Xudong, Li Fei Comprehensive clustering method of user load characteristics and adjustable potential for power big data [J] Chinese Journal of electrical engineering, 2021,41 (18): 6259-6271 DOI:10.13334/j.0258-8013. pcsee. 201928.
[14] Li Yujiao, Huang Qingping, Liu Song, Chen Yu, Liu Peng Power user load pattern extraction method based on Clustering Fusion Technology [J] Electrical measurement and instrumentation, 2018,55 (16): 137-141 + 152.
[15] Li Zhiyong, Wu Jingying, Wu Weilin, song Baoming Power user load curve clustering based on self-organizing mapping neural network [J] Power system automation, 2008 (15): 66-70 + 78.
[16] Luan Le, Ma Zhiyuan, Mo Wenxiong, Xu Zhong, Zhou Kai, Guo Qianwen Classification method of high-quality power users considering the needs of both power suppliers and consumers [J] Journal of electric power science and technology, 2021,36 (06): 171-181 DOI:10.19781/j.issn. 1673-9140.2021.06.021.
[17] Duan rubidium, Zhang Caiqing, Liu aifang Application of fuzzy clustering in power user classification [J] DSM, 2005 (05): 18-20.
[18] Wu Xiaoxiang Research on user classification of smart meter based on deep learning [D] Nanchang University, 2021 DOI:10.27232/d.cnki. gnchu. 2021.003162.
[19] Zhang Yue, Ren Chunlei. Simplified equivalence of active distribution network based on data mining and cluster analysis [J]. China Test, 2022,48 (03): 163-168.
[20] Liu Hongkai, Zhang Jifu. A DBSCAN clustering analysis algorithm based on inverse nearest eighbor and influence space [J]. Computer Application and Software, 2022,39 (02): 287-293+349.
[21] Hou Qing, Yang Rongxin, Zhang Yingjie, Li Wei. Adaptive image clustering with deep learning and clustering analysis [J]. Computer Technology and Development, 2022,32 (01): 98-103.
[22] Zhao Zhongqi, Chang Xiqiang, Fan Yanfang, Xu Sen, Fan Mao. Cluster analysis of power load based on self encoder [J]. Science and Technology and Engineering, 2021,21 (32): 13737-13743.
[23] Cao Duanxi Improvement of clustering algorithm and research on clustering effectiveness index [D]. Nanjing University of Posts and Telecommunications,2021.DOI:10.27251/d.cnki.gnjdc.2021.000738.
[24] Wu Qingxiao, Wang Hening, Qiu Haoyu, Xie Yitian, Dong Junfeng. Power load pattern recognition of residential users based on depth clustering [J]. Science and Technology Innovation and Application, 2022,12 (24): 29-33+37. DOI: 10.19981/j.CN23-1581/G3.2022.24.008.
[25] Chang Le, Wang Qingnian. Ultra short term power load forecasting based on optimized clustering decomposition and XGBOOST [J]. Foreign Electronic Measurement Technology, 2022,41 (05): 46-51. DOI: 10.19,652/j.cnki.femt.2203706.

AUTHORS

Wang Qingnan: Male, born in Harbin, Heilongjiang Province, China, Master, his main research direction is data mining.1159925124@qq.com

Zhang Zhaogong: Male, born in Qingdao, Shandong Province, China, professor and master's supervisor of Heilongjiang University. His main research interests are datamining and biological information. zhangzhaogong@hlju.edu.cn