Resident user load classification method based on improved Gaussian mixture model clustering

Haojing Wang[1], *Yingjie* Tian[1], *An* Li[2*], *Jihai* Wu[2], and *Gaiping* Sun[2]

[1]State Grid Shanghai Municlpal Electric Power Company, 200093 Shanghai, China
[2]Shanghai University of Electric Power,200090 Shanghai, China

Abstract. In view of the limitation of "hard assignment" of clusters in traditional clustering methods and the difficulty of meeting the requirements of clustering efficiency and clustering accuracy simultaneously in regard to massive data sets, a load classification method based on a Gaussian mixture model combining clustering and principal component analysis is proposed. The load data are fed into a Gaussian mixture model clustering algorithm after principal component analysis and dimensionality reduction to achieve classification of large-scale load datasets. The method in this paper is used to classify loads in the Canadian AMPds2 public dataset and is compared with K-Means, Gaussian mixed model clustering and other methods. The results show that the proposed method can not only achieve load classification more effectively and finely, but also save computational cost and improve computational efficiency.

1 Introduction

Load classification refers to the processing of load data from a large number of power devices to extract typical load profiles [1], which can be applied to electricity consumption behavior analysis, load forecasting, tariff setting, demand-side response, etc. Accurate and effective load classification is helpful for the precise marketing of power supply departments. Therefore, the implementation of accurate load classification is of great significance to real-time dispatching, improving the economic efficiency of enterprises, and saving energy [2-3]. Load classification is a research hotspot in recent years. The existing methods can be mainly divided into artificial neural network methods [4-7] and cluster analysis methods [3-4]. Among them, load curve cluster analysis generally uses K-Means clustering . The papers [8-14] use clustering algorithms such as K-Means, K-Medoids, and Hierarchical clustering to achieve the classification and identification of daily load curves, electricity consumption trajectories, and typical electricity consumption patterns of commercial and residential customers. The above research All effectively realize the load classification. The paper [15] compared several clustering algorithms and found that the divisional clustering algorithm was more efficient but less accurate. This is also true for hierarchical clustering. Both hierarchical and partitioned clustering suffer from the problem of "hard assignment" (i.e., each point is explicitly assigned to a cluster center) and they are

* Corresponding author: spiderla97@163.com

not adaptive for large-scale data sets. Hierarchical clustering has high time complexity, and partitioning clusters may fall into local optima. Gaussian mixture model (GMM) clustering is used to assign cluster members according to the clustering probability, which is called "soft classification". It can effectively solve the problem of "hard assignment" with more information, and better clustering quality for largescale data sets. The K-Means algorithm assumes that each cluster is approximately spherical in shape and approximately equal in size. In contrast, GMM clustering has a more flexible cluster shape. The Gaussian mixture model has been widely used in speech, image recognition and other fields, but is less applied in the classification of load state.

Therefore, this paper proposes a hybrid PCA-GMM-based load state classification method by combining the advantages of "soft classification" and clustering flexibility of GMM clustering with PCA from the perspectives of improving clustering quality, clustering efficiency and saving computational cost. The dimensionality of load data is reduced by PCA, and then used as the input to GMM clustering algorithm. Thus, an accurate and effective classification of load states is achieved. To illustrate the effectiveness of the proposed method, PCA-GMM clustering and other methods are applied to the AMPds2 dataset [19]. The results show that the proposed method has better clustering quality and clustering efficiency, and it does effectively reduce the computational cost.

2 Gaussian mixture model clustering

2.1 Gaussian mixture model (GMM)

Gaussian Mixture Model (GMM) refers to a linear combination of multiple Gaussian distribution functions. It is the fastest learning probability model. GMM tries to find a mixture of multidimensional Gaussian distributions that can best simulate the input data set. The flowchart of the Gaussian mixture model clustering algorithm is as follows:

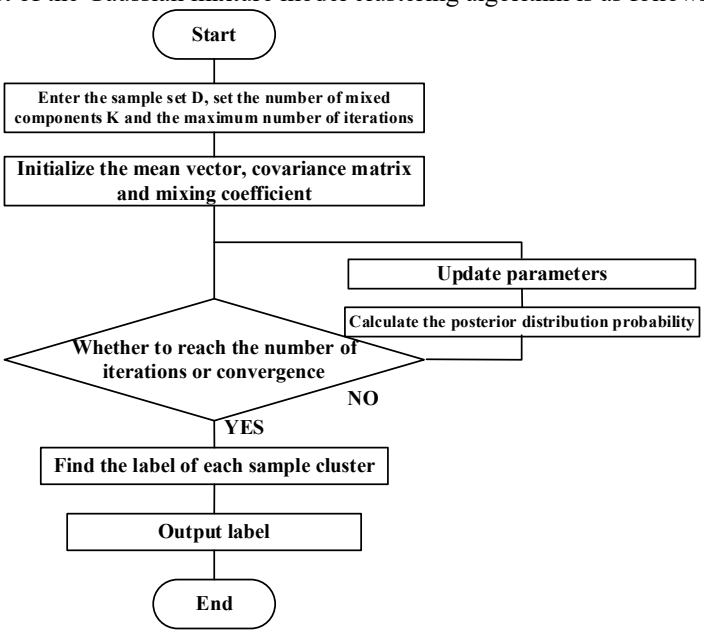


Fig. 1. Gaussian mixture model clustering algorithm.

2.2 Cluster evaluation index

2.2.1 BIC

Probabilistic estimation of the number of groups of GMM using BIC-based model selection theory. The definition of BIC is shown in the formula, and the optimal number of clusters is gradually obtained by approximation.

$$C_{\mathrm{BIC}} = -2\ln(L) + n_p \ln(m) \tag{1}$$

In Equation (7): C_{BIC} is the BIC value; n_p is the number of hyperparameters; L is the maximum value of the estimated model likelihood function.

Assuming that the errors or disturbances of the model are normally distributed, BIC can be expressed as:

$$C_{\mathrm{BIC}} = m\ln\left(\frac{S_{\mathrm{RSS}}}{m}\right) + n_{\mathrm{p}} \ln(m) \tag{2}$$

In Equation (8): S_{RSS} is the residual sum of squares of the estimated model.

C_{BIC} is an increasing function of S_{RSS} and np, that is, the introduction of residuals and unknown parameters will increase C_{BIC}. Therefore, in judging the number of load classifications, the model with a low BIC value is preferred

2.2.1 DBI

Davies-Bouldin Index (DBI), also known as classification adequacy index, is an index to evaluate the pros and cons of clustering algorithms.

Suppose there are ***m*** time series, and these time series are clustered into ***n*** clusters. The ***m*** time series are set as the input matrix ***X***, and the ***n*** cluster classes are set as ***N*** as the parameters passed into the algorithm. Use the following formula to calculate:

$$DBI = \frac{1}{N}\sum_{i=1}^{N}\max_{j\neq i}\left(\frac{S_i + S_j}{M_{ij}}\right) \tag{3}$$

The meaning of this formula is to measure the mean value of the maximum similarity of each cluster class. In the formula, $\boldsymbol{S_i}$ is the average distance between the data in the cluster and the centroid of the cluster, which represents the degree of dispersion of the time series in the cluster ***i*** , and $\boldsymbol{M_{ij}}$ is the distance between the cluster ***i*** and the cluster ***j***.

3 Principal component analysis (PCA)

PCA is a dimensionality reduction method that converts multi-dimensional data into a relatively simple spatial mapping in the simplest and most economical way. It can convey important relationships between data through highly intuitive visual output, and the dimensionality reduction quality and dimensionality reduction rate are better. Each point in the low-dimensional space obtained by PCA represents an object, and each point obtained after dimensionality reduction in this paper represents the load characteristic of each day.

The PCA algorithm flow is as follows:

Input: ***n***-dimensional sample set ***D*=(*x*(1),*x*(2),...,*x*(*m*))**, the number of dimensions to be reduced to ***n'***.

• Centralize all samples:

$$x(i) = x(i) - \frac{1}{m}\sum_{j=1}^{m} x(j) \tag{4}$$

• Calculate the covariance matrix $\boldsymbol{XX^T}$.

• Perform eigenvalue decomposition on matrix $\boldsymbol{XX^T}$.

• Take out the eigenvector $(\boldsymbol{w_1}, \boldsymbol{w_2}, \ldots, \boldsymbol{w_{n'}})$ corresponding to the largest eigenvalue. After all the eigenvectors are standardized, the eigenvector matrix $\boldsymbol{W}$ is formed.

• For each sample $\boldsymbol{x(i)}$ in the sample set, transform it into a new sample $\boldsymbol{z(i)=W^T x(i)}$

• Obtain the output sample set $\boldsymbol{D'} = (\boldsymbol{z(1)}, \boldsymbol{z(2)}, \ldots, \boldsymbol{z(m)})$.

4 Case analysis

4.1 Data set description

This article selects the AMPds2 data set released by Canadian scholars in 2016. The data is collected from a household user in Vancouver, Canada, with a total living area of 199 m^2 and a basement area of 100 m^2. The data set is collected from April 1, 2012 to 2014. On April 1, 2010, there were 1051200 records in total. The sampling interval of power data was 1min, including the total meter data and sub-metering data such as active power, reactive power, voltage and current. The sampling interval of external environment data was 1h, including temperature , Air pressure, wind speed, etc.

4.2 Data pre-processing

Due to possible errors, loss, and anomalies in the data during measurement, recording, and transmission. This requires pre-processing of the data. At first, the maximum value was estimated for the total load of the AMPds2 dataset, and values exceeding two times the estimated value were considered as outliers for deletion. Then the arithmetic mean of the two hours of data before and after was used to fill in the missing values. Finally, a sliding filtering algorithm is used for noise reduction of the data. A data buffer is created in RAM to store $\boldsymbol{N}$ sampled data in order, and for each new data read, the earliest one collected is discarded and the arithmetic mean of the N data in the buffer is calculated as the result of filtering. the overview of the AMPds2 data set is shown below.

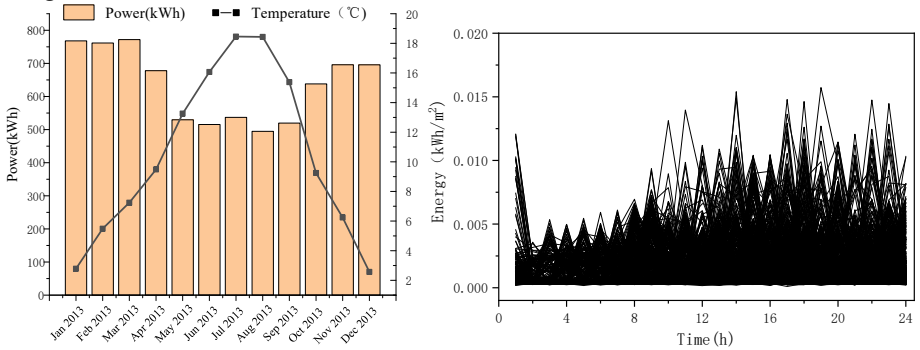


Fig. 2. Energy consumption and temperature change graph And daily load curve per unit area.

4.3 Comparison of GMM and other cluster classification results

GMM clustering belongs to "soft classification" and the clusters are flexible. Here we visualize the comparison between K-Means clustering and GMM clustering. For the determination of the optimal number of clusters, GMM clustering often uses the BIC criterion, while K-Means clustering is most commonly obtained based on the clustering effectiveness index. The DBI index is simpler and has a small range of changes, which is more widely used. . Therefore, all the clusters involved in GMM use the BIC criterion to determine the number of clusters, and other clusters use the DBI index.

Comparing Figure 4, we can see that GMM clustering divides the electricity load into 13 categories, while K-Means clustering only divides 3 categories. From the perspective of the number of classifications, GMM classification is more refined than K-Means, and contains more information. Combining the above figure and the date distribution corresponding to the clustering results of GMM and K-Means (Figure 5), it can be seen that the information obtained from the clustering results of K-Means is only that the electricity consumption of the building has small fluctuations in summer and low electricity consumption. In the spring and autumn, the electricity consumption fluctuates greatly, and the GMM classification results can be further detailed, such as the highest temperature in July and August and the lowest temperature in January and December. The electricity consumption habits are the same, and the fluctuations are relatively stable.

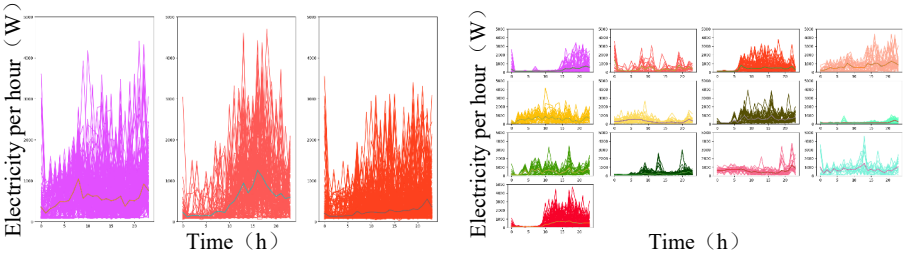
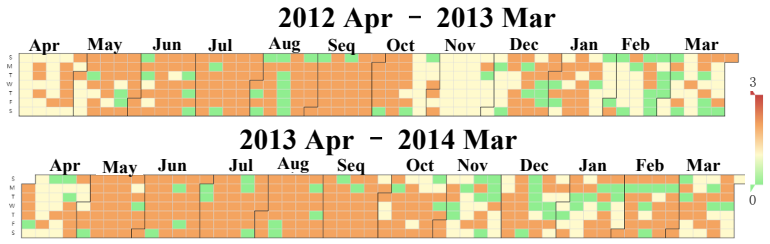
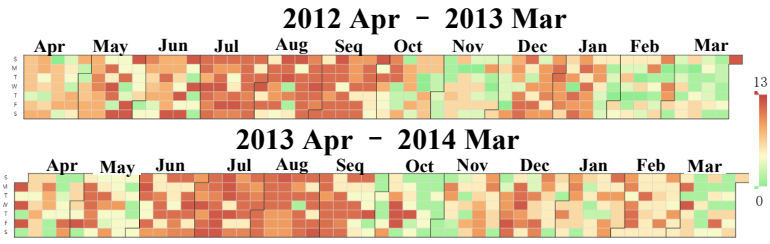


Fig. 4. Comparison of clustering results between GMM and K-Means.



(a) K-Means



(b) GMM

Fig. 5. Date distribution corresponding to K-Means and GMM clustering results.

4.4 Load classification method based on PCA-GMM

The specific process of PCA-GMM method classification is shown in Figure 6. First, the pre-processed 24-dimensional load data is mapped to a 2-dimensional space through the PCA method, as shown in Figure 7. Each point in the figure represents a data object (that is, the information of a 24-dimensional daily load curve). 728 data points. The sample distribution of data clusters after dimensionality reduction is elliptical, indicating that GMM-based clustering will be more suitable for the classification of these building loads than other clusters.

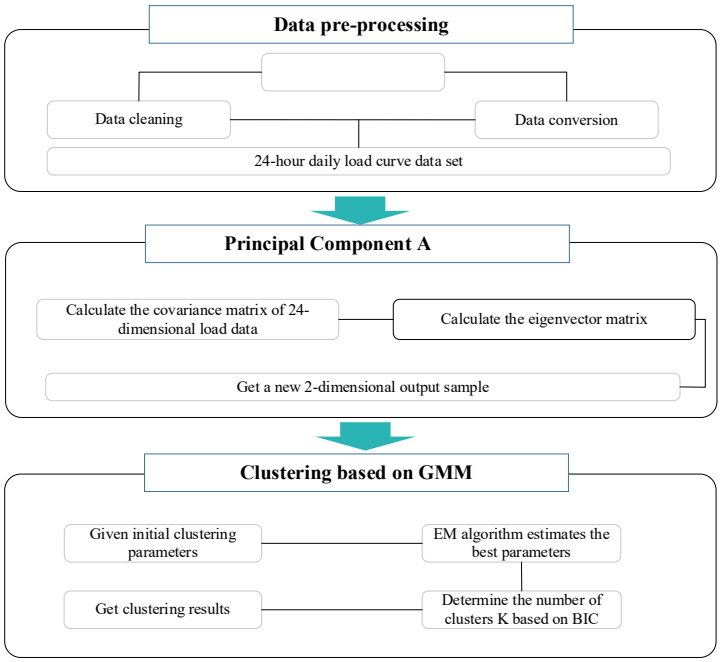


Fig. 6. Process of load classification method based on PAC and GMM clustering

Secondly, input the reduced-dimensional load data into the GMM clustering algorithm.It can be seen from the figure that the number of clusters is 6 categories. In general, the clusters identified by each building are close within the clusters and finely classified between the clusters. The clustering effect is good, and the distribution area is also related to the load on the date and time. The characteristics of seasonal distribution correspond to each other.

The following are the classification results of the PCA-GMM method.

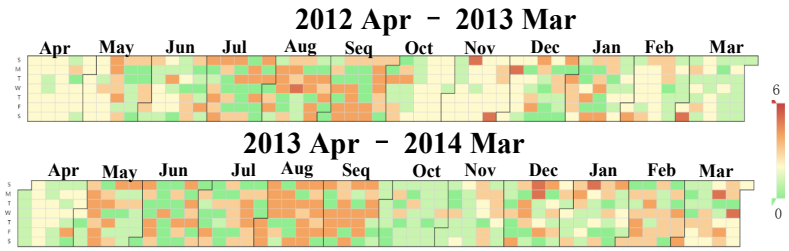


Fig. 8. Date distribution corresponding to the clustering results of PAC and GMM.

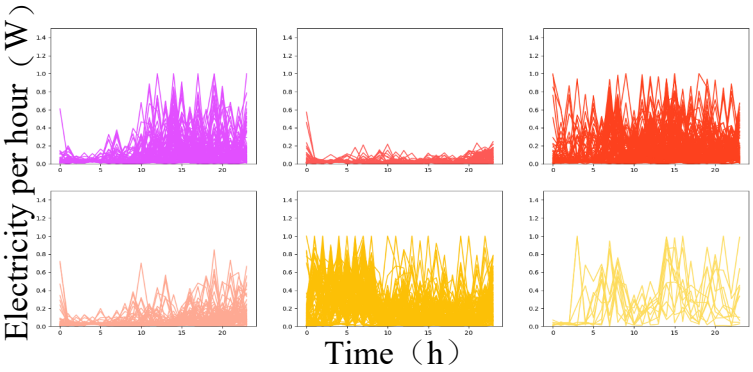


Fig. 9. PCA-GMM clustering result graph.

Through the analysis of Figure 8-9, we can get:

• From the shape of the typical daily load curve, the peak of the user's daily load curve in the past two years is mainly in the afternoon or early morning, and the shape of the class 1 and class 4 curves are relatively similar. , The peak of all-day electricity load is concentrated in the afternoon, and the magnitude of class 1 is higher; the shape of the curves of class 3 and class 5 are similar, the all-day load is at a high level, and the peak period of electricity consumption is at night; the class 2 all-day load Lower.

• From the perspective of working days and non-working days, the user did not show obvious patterns.

• From the distribution of seasons and temperature changes, category 1 is mainly distributed throughout the year with little electricity consumption at night and large electricity consumption during the day, which conforms to the normal electricity consumption law of users; category 2 with the lowest electricity load is mainly Distributed in spring and autumn when the climate is suitable and the demand for air conditioning is low; Class 3 is distributed in March, April, October, and November, with a large load throughout the day; Class 4 is similar to Class 1 distributed throughout the year, with low night load and heavy load during the day ; Class 5 electricity consumption at night is large, distributed in August and September when the temperature is higher. The above analysis shows that the classification result of the PAC-GMM method conforms to the electricity consumption characteristics of the electricity load with the seasonal and temperature changes, and it captures the correlation and dependence between the data. The information contained is better than that of K-Means. The clustering results are more refined.

5 Conclusion

In this paper, a load classification method based on Gaussian mixture model clustering and principal component analysis is proposed. Residential load classification is taken as an example. The pre-processed large-scale load dataset is reduced in dimensionality by PAC method and then input to GMM clustering algorithm to achieve load pattern classification. The results of the algorithm example show that:

- Comparing with K-Means, GMM and other methods, it is learned that the proposed PCA-GMM method achieve the classification of load patterns for each type of buildings both effectively and accurately, with more refined classification results, higher clustering efficiency and much lower storage space.

- Using this method to classify load patterns will help the power supply sector to get a better grasp of users' load characteristics, formulate reasonable tariff policies, propose energy saving strategies. It can also provide more targeted services to various users, and

motivate consumers to actively participate in the demand side management system. At the same time, it is of crucial importance to guide the rolling planning of power grid, real-time dispatching and reliability assessment of operation planning.

References

1. Y Liu, C Yuen,H Huang, et a1. Peak-to-Average ratio constrai ned demand-side management with consumer's preference in resi dential smart grid. Selected Topics in Signal Processing, IEEE Jo urnal of 8. 1084-1097. 10.1109/JSTSP.2014.2332301.
2. Y Liu, C Sun, Z Niu, et a1.Study on cl assification technology of power load characteristics based on im proved fuzzy C-Means clustering algorithm[J].Electrical Measure ment & Instrumentation,2014,51(18):5-9.
3. T Zhang, M Gu.Overview of power user load pattern extraction technology and its application[J]. Power System Techn ology,2016,40(03):804-811.
4. T Li, H Yang, Y Gao.Overview of household electric l oad identification technology for smart meters[J].Supply and Elec tricity,2011,28(06):39-42.
5. Y Liu, Y Liu, L Xu. High performance back propag ation neural network algorithm for massive load data classificatio n[J].Automation of Electric Power Systems,2018,42(21):96-105. D OI:10.7500/AEPS20171215005.
6. L Shi, R Zhou, W Zhang,et a1. Load classifica tion method using deep learning and multidimensional fuzzy C- means clustering [J]. Journal of Electric Power System and Auto mation,2019,31(07):43-50.
7. W Li, B Zhou, N Lin. Daily load characteristic curve classification and short-term load forecasting based on fuzzy clu stering and improved BP algorithm[J].Power System Protection a nd Control,2012,40(03):56-60.
8. S Lin, E Tian, Y Fu, et a1.A load classification me thod based on information entropy segmentation aggregation appr oximation and apectral clustering[J].Proceedings of the CSEE,2017,37(08):2242-2253.
9. F Bu, J Chen, Q Zhang, et a1.A controllable and r efined recognition method for load patterns based on two-layer it erative clustering analysis[J]. Power System Technology,2018,42(0 3):903-913.
10. X Wang, Z Chen, X Peng. A load pattern c ombination recognition method based on two-layer clustering analysis[J]. Power System Technology,2016,40(05):1495-1501.
11. S Wang, T Liu. Classification and recognition me thod of resident load gradient lifting tree considering power cons umption mode[J]. Proceedings of the CSU-EPSA,2017,29(09):27-3 3.
12. X Peng, J Lai, W Chen. Intelligent identification method of customer power consumption mode based on cluster a nalysis[J].Power System Protection and Control,2014,42(19):68-73 .
13. H Jia, G He, C Fang, et al. Multi-level clu stering method for hierarchical clustering and bidirectional cappin g combination for load forecasting[J]. Power System Technology, 2007(23):33-36.
14. X Li, X Jiang, J Qian, et al. Classification and com prehensive method of power industry based on user daily load c urve[J]. Automation of Electric Power Systems,2010,34(10):56-61.

15. B Zhang, C Zhuang, Hu J, et a1. Integrated clustering a lgorithm for power load curve combined with dimensionality red uction technology[J]. Proceedings of the CSEE,2015,35(15):3741- 3749.
16. K Li, Z Ma, R Duane, et a1. Identification of typical building daily electricity usage profiles using Gaussian mi xture model-based clustering and hierarchical clustering[J]. Applie d Energy,2018,231:331-342.
17. B Zeng, J Zhang, L Ding, et al. Application of improv ed adaptive fuzzy C-means algorithm in load characteristic classi fication[J]. Automation of Electric Power Systems,2011,35(12):42- 46.
18. H Yang, L Zhang, Q He, et al. A study of power load cl assification based on adaptive fuzzy C-Means algorithm[J].Power System Protection and Control,2010,38(16):111-115+122.
19. MAKONIN S, ELLERT B, BAJIC I V, et al. Electricity, water, and natural gas consumption of a residential house in Canada from 2012 to 2014[J]. Scientific Data, 2016, 3(160037):1-12.]