ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 1 - 2026

Research on Probabilistic Prediction Model for Power Load Forecasting in Uncertain Scenarios Based on Clustering Algorithm for Imbalanced Data

Lin Guo[1,*], *Gang Wu*[2], *Jieying Liu*[2], *Yuxin Xiao*[2], *Wei Wang*[2] *and Ruiguang Ma*[2]

[1] State Grid Sichuan Electric Power Co., Ltd., Chengdu, Sichuan, 610041, China
[2] State Grid Sichuan Electric Power Company Economic and Technological Research Institute, Chengdu,Sichuan, 610041,China

SUMMARY: *To solve the problem of unbalanced dataset clustering, this paper proposes an unbalanced data clustering method based on adaptive competitive learning. By optimizing competitive learning, new centroids are added adaptively to update the number of subclass centroids and integrate the structural features in the dataset. The two metrics of compactness and divisibility are combined to calculate the subclass merging difficulty coefficient to obtain the final clustering results. And the MCCL algorithm is tested for accuracy on the dataset characterized by imbalance, and the short-term power load forecasting step based on the combined MCCL-BILSTM model is designed. The PCA and K-means clustering algorithms are utilized to screen out the similar days for power load forecasting in response to the realistic demand of uncertain scenarios as well as multiple influencing factors of the combined power load. The MCCL-BILSTM model is utilized to forecast the electricity load for the four categories of similar days. In the comparison of short-term power load forecasting, the forecast curves of LSTM model, RNN model, BP model and GRNN model for similar days can roughly reflect the trend of the peak power load, but the MCCL-BILSTM model using clustering to screen the similar days is able to show a more superior forecasting effect.*

KEYWORDS: *competitive learning; power load forecasting; PCA; K-means clustering; similar days; unbalanced data*

1 Introduction

The electric power industry is an important basic energy industry in the development of the national economy, and is likewise the key to ensuring the sustained, stable and healthy development of the national economy and society [1, 2]. With the rapid development of the global economy and the demand for sustainable energy development, the demand for electricity in various fields is surging, driving the power system to digitalization and intelligent transformation and development [3-5]. According to relevant data, global electricity demand will increase by more than 30% from 2020 to 2030 [6]. Moreover, the extreme weather in several regions of the world starting in 2023 highlights the need to strengthen the security of power supply [7].

In order to ensure the stability and reliability of power supply, the power system needs to rationally dispatch generation, transmission and distribution [8, 9]. And power load forecasting, as the basis of power system scheduling and operation, is a crucial part of power system

*becautious22@163.com
https://doi.org/10.65102/is2026006
Published: 30 April, 2026

operation and planning [10, 11]. By accurately predicting the future trend of power load, it can not only guide the stable operation of the power system, optimize resource allocation, reasonably arrange the input and output of power generation equipment, reduce energy waste and environmental pollution, but also help optimize the balance of supply and demand in the energy market and predict the changes in power prices, which is of great significance for achieving sustainable development of the power system and improving the efficiency of the use of power resources [12-16].

Power load forecasting needs to be based on historical load data and its influencing factors, taking into account the external conditions of the power load forecasting time period and customer demand, establishing relevant forecasting models and performing model optimization to achieve reliable forecasting of power system loads [17]. With the access of large-scale renewable energy sources and changes in power market demand, as well as under various other related influences (such as weather conditions, human activities, types of industrial processes, time and seasonal characteristics, etc.), the uncertainty characteristics of power load data are significant [18-21]. Therefore, researchers have modeled power load forecasting under uncertain scenarios through various approaches.

Electricity load forecasting can be divided into three stages, i.e., traditional forecasting methods, intelligent forecasting methods, and probabilistic forecasting methods. For traditional forecasting methods, regression analysis models, time series methods, and gray models are often used. Literature [22] constructed three robust power load forecasting models with two iterative reweighted least squares regression models and a LASSO regression model, which provide a guarantee of data integrity that is invalidated by attacks. Literature [23] estimates the knowledge uncertainty and stochastic uncertainty in power load forecasting by diffusion-based Seq2seq structure and robust additive Cauchy distribution respectively, which effectively guarantees the accuracy of power load forecasting results. Literature [24] combined the autoregressive integral sliding average model and error gradient sampling to forecast the source load of the distribution network, and obtained the forecast interval of unit output and load demand of a region over a 10-year period under rolling forecast. Literature [25] used the exponential smoothing method to smooth the power load data and established a gray prediction model based on the smoothed series to form an improved exponential smoothing gray model for predicting short-term power loads, which shortened the prediction period while improving the prediction accuracy.

Intelligent forecasting methods are mainly used for electricity load forecasting based on machine learning, deep learning methods, i.e., Bayesian networks, Support Vector Machines (SVMs), Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), Long and Short-Term Memory Networks (LSTMs), and Gated Recurrent Units (GRUs). Literature [26] relied on historical consumption, temperature, socio-economic, and electricity usage data in multifamily load forecasting and used Bayesian networks to predict ultra-short-term electricity loads in uncertainty and variability environments. Literature [27] with the help of K-mean clustering method for selecting similar days and dividing the load data into weekdays and holidays, introduced SVM prediction model to predict short-term electricity loads, whose prediction accuracy (39.75%) and runtime (128.89%) were significantly improved compared to the traditional methods. Literature [28] performed a step-based sliding window approach for time-series generation of power transmission operations data, with inputs of multivariate time-series data, combined with a GRU calibrated in order to generate multistep forecasts, which outperformed the traditional approach. Literature [29] used LSTM model for ultra-short-term power load forecasting and introduced time series decomposition-reconstruction model for component series superposition to reduce the forecasting error as a way to improve the accuracy of forecasting results. Literature [30] provides a CNN-based method for extracting nonlinear

relationships between load values for residential power load forecasting, where the extraction results form a load-temperature cube, and another CNN model is used to capture the hidden features in the cube, and then the load is predicted with a support vector regression model, which improves the utilization of data features and also narrows the prediction error. Literature [31] used the scenario prediction model to simulate the stochastic behavior in power loads, fully considered the uncertainty in power, proposed a scenario prediction method for power loads based on raw pixel CNN, and introduced an optimization model, which made the power load prediction to obtain higher prediction results from the scenario simulation. Literature [32] uses CNN model to extract features of electricity data, and incorporates forward LSTM and inverse LSTM models to fully utilize the forward and inverse temporal features of the load data, and the CNN-bidirectional LSTM model has a more efficient and accurate forecasting performance in the context of uncertainty in generation and demand.

However, due to the integration of renewable energy, electric vehicles, and microgrids, the inherent uncertainty of load demand has been further emphasized, and the poor dynamic adaptability of traditional models and insufficient quantification of uncertainty prediction risks have led to the difficulty of forecasting power loads under uncertainty scenarios. In addition, intelligent algorithm-based power load forecasting methods hinder the accuracy of forecasting in uncertain scenarios due to the need for a large amount of training data and the “black box” problem of intelligent algorithms. In contrast, the output of probabilistic forecasting methods is presented as a probability density function or confidence interval, which provides the possible distribution of future electricity consumption and can effectively assess the uncertainty of electricity load forecasting [33-35]. Literature [36] addresses the uncertainty of customer demand, uses Bayesian deep learning to construct a multi-task probabilistic load forecasting architecture to quantify the uncertainty commonality and difference characteristics among customer groups, and proposes a clustering-based pooling method to compensate for the diversity of electricity load data, forming a probabilistic load forecasting model with high performance. Literature [37] developed three Bayesian deep neural networks based on RNN, LSTM, and GRU for quantifying attendant uncertainty and knowledge uncertainty for probabilistic building electricity load forecasting, among which the Bayesian LSTM model has the best performance with a 15.4% reduction in the prediction error, and it can rely only on 10 hours of lagged electricity load data to construct an effective forecasting model. Literature [38] designed an ultrashort-term probabilistic prediction interval predictor to successfully predict the user's energy demand for the next 15 minutes with an AI-driven predictor based on artificial intelligence, while calculating the user's choice of upper and lower confidence levels with a probabilistic prediction interval algorithm. Literature [39] constructed a probabilistic load forecasting method using artificial neural networks and association rules in order to forecast the electric load for the next 2 hours, estimated the electric load forecast and its difference from the observed values through artificial neural networks, used the difference to obtain an accurate forecast interval, introduced adjustments to this interval to improve the accuracy of the forecast. Literature [40] proposed a new probabilistic load forecasting method by constructing a long and short-term pattern network with two-stage attention, combining Monte Carlo noise reduction and soft-thresholding techniques for constructing uncertainty models and reducing data noise under consideration of weather factors.

Often there is an imbalance problem in power load data, coupled with the lack of uncertainty prediction, which leads to a large deviation in the final prediction results. Therefore, for the data imbalance problem in power load forecasting, scholars have given solutions based on clustering algorithms. Literature [41] used K-mean algorithm in short-term load forecasting model to classify scenario-based power load data, while load scenario imbalance was classified using balanced K-nearest neighbor method, and introduced locally weighted linear regression

algorithm and Apache Hadoop programming framework to improve the performance of load forecasting model for massive and high-dimensional data processing, which has a better performance compared to the traditional forecasting model. Literature [42] used clustered decision tree algorithm to detect building operating conditions, explored the impact of data imbalance based on changes in operating conditions, and created a multi-model forecasting method to solve the data imbalance problem, which reduces the average absolute error of load forecasting by 1.32-9.83%. Literature [43] designed an improved fuzzy C-mean clustering algorithm based on normalization, affiliation matrix updating, and dynamic time regularization for dealing with the problem of cluster size inhomogeneity that occurs with unbalanced datasets in power loads. Literature [44] proposes a density-based clustering algorithm, a combination of random undersampling and oversampling techniques for unbalanced data processing, which can effectively solve the data imbalance problem and still performs well in highly unbalanced datasets.

In this paper, we propose an adaptive competitive learning based multi-center clustering algorithm for unbalanced data to optimize the processing of power unbalanced dataset clustering problem.The MCCL algorithm re-selects the winning point and failure point strategies, and updates the subclass generation and subclass merging methods respectively. The MCCL algorithm is tested on multiple datasets for clustering analysis to verify the feasibility of the MCCL algorithm. Combined with the basic steps of power load forecasting, the -BILSTM model is proposed and the short-term power load forecasting process based on the combined MCCL-BILSTM model is designed. PCA and K-means clustering are applied to screen out the similar days for power load forecasting, and the MCCL-BILSTM model is applied to forecast power load on different categories of similar days.

2 Optimization of unbalanced data sets and clustering algorithms

2.1 Overview of unbalanced data sets

Unbalanced data, which refers to certain data sets in which the sample points are unevenly distributed. That is, the number of samples of one class in such a data set is significantly less or greater than the number of samples of another class. However, in reality, classification predictions for data information in a given situation are generally more concerned with the case of a few classes of samples. Learning unbalanced data means finding valuable information in a few classes in an unevenly distributed dataset, and its imbalance is mainly manifested in the following two situations: inter-class imbalance and intra-class imbalance.

(1) Inter-class imbalance: in an unbalanced dataset, there is a large gap between the amount of data in two classes, the data in the minority class is much less than the data in the majority class, and the boundaries between classes are often unclear, which makes it difficult to categorize the data.

(2) Intra-class imbalance: if in a dataset, the positions within several sample classes are unevenly distributed or the size of the samples within the sample space is different, this can lead to problems such as fragmentation of the data in the samples of the minority class. In this case, traditional classification algorithms may misclassify the few samples in classes that do not belong to the multi-sample set or have a low density of sample distribution as noise in order to maximize the global classification accuracy.

2.2 Multi-center clustering algorithm for unbalanced data

2.2.1 Selection of Strategies and Strategy Updates

The core of the competitive learning algorithm is the selection strategy of winning and losing centroids and their update strategy. If the selection strategies of the winning and losing centroids are unreasonable, some centroids have no chance to win, i.e., the "dead cell" problem.

Aiming at the above problems, this paper proposes a new update strategy for the selection of winning and losing points. It is divided into the winning and losing point selection strategy, the rewarding strategy for winning points and the punishing strategy for losing points. These two strategies will be described in detail next.

For ease of narration, let the dataset be $X=\{x_1,x_2,\cdots,x_N\}$, N is the number of data points, x_t is the data point under discussion at this time, and $M=\{m_1,m_2,\ldots,m_K\}$ is the initial set of center points at this point.

(1) Winning and Failing Points Selection Strategy

The selection strategy of winning and losing points, the strategy can be expressed as:

$$I_{j,x_t}=\begin{cases}1 & j=v=\arg\min\limits_{1\le i\le K} d_{it}\\ -1 & j=f=\left\{i\middle| d_{it}\le \mu+\sigma, 1\le i\le K \,\&\, i\ne v\right\}\\ 0 & \text{otherwise}\end{cases} \tag{1}$$

where $\mu=\sum_{i=1}^{K} d_{it}/K$ is the mean of the distances between all centroids to the data point x_t. $\sigma=\sqrt{\sum_{i=1}^{K}\left(d_{it}-\mu\right)^2/K}$ denotes the standard deviation of the distances between all centroids to the data point x_t. The d_{it} is the Euclidean distance between the center point m_i and the data point x_t. The winning point m_v is the closest center point to the data point x_t.

The update strategy for the selection of winning and losing points is shown in Fig. 1. In the figure, the winning point is m_v, the set of failure points includes m_{f_1} and m_{f_2}, and m_o is the remaining invalid center point.

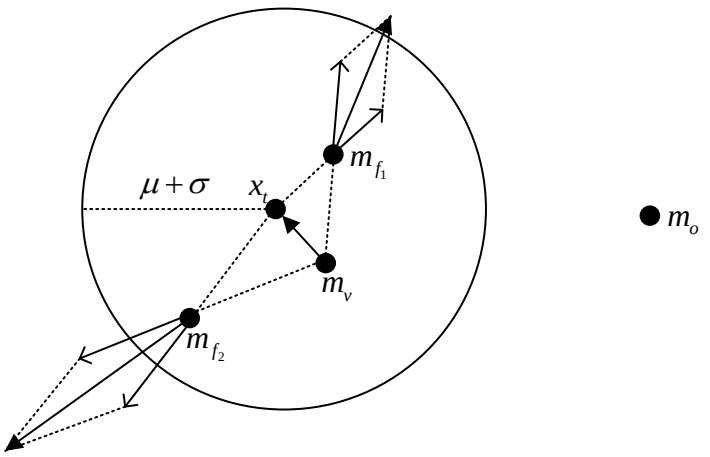

Figure 1: Select update strategy for winning points and failure points

(2) Center point update strategy

The algorithm selects winning and losing points based on Eq. (1) and then uses Eq. (2) to update the location of the center point. Eq:

$$m_j(t+1)=\begin{cases} m_j(t)+K\alpha_v\left(x_t-m_j(t)\right) & I_{j,x_t}=1 \\ m_j(t)+K\alpha_v\eta\beta_j & \\ \left(\left(m_j(t)-x_t\right)+\left(m_j(t)-m_v(t)\right)\right) & I_{j,x_t}=-1 \\ m_j(t) & \text{otherwise} \end{cases} \tag{2}$$

where α_v is the learning rate, which indicates the degree of reward for winning points. When I_{j,x_t} is equal to 1, $\left(x_t-m_j(t)\right)$ is equivalent to $\left(x_t-m_v\right)$, which is the vector of the winning point m_v pointing to the data point x_t, and denotes the rewarding direction of the winning point, i.e., the winning point is moving towards the direction of approaching the data point x_t. As shown in Fig. 1, the green direction indicates the reward direction of the winning point m_v and the blue direction indicates the penalty direction of the losing point m_f. The formula is:

$$\beta_j=\frac{\min\left(\left\|m_j-m_v\right\|,\left\|x_t-m_v\right\|\right)}{\left\|m_j-m_v\right\|} \tag{3}$$

β_j denotes the penalty strength of the failure point, which takes the value of $(0,1]$.

2.2.2 Subclass generation

In order to solve the problem of the number of predefined centroids, this paper proposes an adaptive competitive learning method which determines the number of centroids by adaptively adding new centroids.

In order to add a new centroid, one can choose to copy an existing centroid. The selection criteria are based on two factors. The first factor is the number of wins n_j of the centroid m_j and the second factor is the maximum density gap δ_j of the subclass C_j. In order to compute δ_j, the local density ρ_i of each data point x_i is first computed by means of a Gaussian kernel density function, Eq:

$$\rho_i=\sum_{i\in X,\,j\neq i} e^{-\left(\frac{d_{ij}}{d_c}\right)^2} \tag{4}$$

where d_{ij} is the Euclidean distance between data point x_i and data point x_j and d_c is the truncated distance threshold.

For the unbalanced dataset, a new formula is designed to compute the local density ρ_i^j of each data point x_i in subclass C_j as follows:

$$\rho_i^j = \sum_{i \in C_j / (i), d_{ij} \le d_c} e^{-\left(\frac{d_{ij}}{d_c}\right)^2} \tag{5}$$

It can be found that computing the local density of a point using only data points in a neighborhood of radius d_c greatly reduces the computational effort of the algorithm and the local density impact of the larger class on the smaller class.

The maximum density gap δ_j of subclass C_j is calculated as shown:

$$\delta_j = \max_{i \in C_j} \min_{i' \in C_j; \rho_i > \rho_i} \frac{\left\| x_i - x_{i'} \right\|}{\bar{d}_j} \tag{6}$$

where x_i and $x_{i'}$ are data points in subclass C_j and the local density $\rho_{i'}$ of $x_{i'}$ is greater than the local density ρ_i of x_i, $\left\| x_i - x_{i'} \right\|$ denotes the local distance between data points x_i and $x_{i'}$, $\bar{d}_j$ is the average distance between all data points in subclass C_j.

Considering the number of wins n_j and the maximum density gap δ_j of the subclasses, the existing centroids can be selected to be replicated by using Eq. (7) to obtain the location of the additional centroids. Eq:

$$j^* = \arg\min_{j=1,2,\cdots,K} n_j \delta_j \tag{7}$$

2.2.3 Merging of subclasses

After the subclass generation algorithm, the set of subclasses $C = \left\{ C_1, C_2, \cdots, C_{\hat{K}} \right\}$ is obtained, and in order to get the true cluster class, a subclass merging algorithm is then designed. The objective of this algorithm is to merge the subclasses belonging to the same cluster class.

Firstly, the definition of shared point is given, and the degree of overlap between subclasses is obtained based on the shared point, and the closeness of the cluster class is further calculated.

Definition 1 (Shared point): for any two subclasses A and B, a point is said to be a shared point of subclasses A and B if the data point x_i belongs to subclasses A (or B) and at least one of its inverse k nearest neighbors belongs in subclasses B (or A).

The overlap of subclasses A and B is shown in Equation (8):

$$OL_{(A,B)} = \left(\sum_{x_i \in A} SN_i + \sum_{x_j \in B} SN_j \right) \Big/ \min\left(|A|, |B| \right) \tag{8}$$

where $|A|$ and $|B|$ are the number of data points in subclasses A and B, respectively. The total number of shared points for subclasses A and B is $\left(\sum_{x_i \in A} SN_i + \sum_{x_j \in B} SN_j \right)$.

For subclasses A and B, the distance $d_{(A,B)}$ between the subclasses is:

$$d_{(A,B)} = \min_{x_i \in A, x_j \in B} \left\| x_i - x_j \right\| \tag{9}$$

Suppose there are $\hat{K}$ subclasses in the dataset and the set of subclasses is

$C=\left\{C_1,C_2,\cdots,C_{\hat{K}}\right\}$ with the initial subclass groupset $\mathcal{G}^{\hat{K}}=\left\{G_1,G_2,\ldots,G_{\hat{K}}\right\}$, each subclass group initially contains one subclass $G_i=\{C_i\}$. The compactness com_K of the set of subclass groups $\mathcal{G}^K$ is defined as the maximum value of the overlap $OL\left(G_i,G_j\right)$ between all subclass groups in $\mathcal{G}^K$:

$$com_{K-1}=\max_{1\le i\le j\le K}\ \max_{C_r\in G_i,C_r\in G_j} OL_{(C_r,C_r)} \tag{10}$$

If the overlap of subclass groups G_i and G_j is $\mathcal{G}^K$ the largest overlap among all subclass groups, subclass groups G_i and G_j are merged to produce a new subclass group $\left\{G_i\cup G_j\right\}$, and G_i and G_j are deleted from $\mathcal{G}^K$ to obtain $\mathcal{G}^{K-1}$. Thus, $\mathcal{G}^{K-1}$ contains $K-1$ subclass groups recorded as $\mathcal{G}^{K-1}=\left\{G_1,G_2,\ldots,G_{K-1}\right\}$. The set of subclass groups $\mathcal{G}^{K-1}$ is:

$$\mathcal{G}^{K-1}=\mathcal{G}^K\setminus\left\{G_i,G_j\right\}\cup\left\{G_i\cup G_j\right\} \tag{11}$$

where K decreases from $\hat{K}$ to 2 and the subclass merging algorithm stops when $K=1$. The higher the compactness com, the more compact the subclass sets are to each other and the less difficult the merging is. In addition to considering the compactness of the set of subclass groups, it is also necessary to define the divisibility sep of the set of subclass groups by the distance between subclasses. Eq:

$$sep_{K-1}=\min_{1\le i\le j\le K}\ \min_{C_r\in G_i,C_r\in G_j} d_{(C_r,C_r)} \tag{12}$$

The higher the separability sep, the more dispersed the subclass groups are from each other and the more difficult it is to merge.

2.3 Comparison Experiments of Improved Clustering Algorithms

Experimental environment: operating system is windows 11, processor is Intel(R) Core(TM) i7-10210U CPU, memory is 32GB, and running tool is Matlab R2021a.

In order to verify the effectiveness of the multi-center clustering algorithm based on adaptive competitive learning for unbalanced data, experiments will be conducted on synthetic dataset and UCI real dataset respectively. Considering to test the effect of dataset imbalance on the experimental results, four datasets, Guassian, Ids2, Banana and Lithuanian, are selected for the synthetic dataset. All of these datasets have unbalanced features and different shapes and sizes, and are commonly used as experimental datasets for testing unbalanced clustering algorithms. Among them, since Guassian and Ids2 are spherical datasets and Banana and Lithuanian are non-spherical datasets, the algorithm can also be tested for irregularly shaped datasets. Considering to test the applicability of the algorithm on datasets with different dimensions, different number of classes and different class sizes, the UCI real dataset selects two commonly used datasets, Ecoli and Vehicle, as the experimental dataset.

In order to verify that the multi-center clustering algorithm for unbalanced data based on adaptive competitive learning has some effectiveness, the proposed algorithm is experimentally compared with the popular clustering algorithms in recent years. They are DPC algorithm,

SNN-DPC algorithm, DPADN algorithm and SMCL algorithm. Specifically, DPC algorithm is the traditional density peak clustering algorithm.SNN-DPC algorithm and DPADN algorithm are the algorithms to improve DPC, and SMCL algorithm is the clustering algorithm for unbalanced datasets. Meanwhile, DPAND algorithm and SMCL algorithm divide the clustering process into two stages as well, which can reflect the advantages of the algorithms in this chapter among the multi-stage algorithms.

The MCCL algorithm and the comparison algorithm have been experimented on the selected datasets, and the experimental results of the accuracy of the comparison algorithm are shown in Fig. 2.

The ACC of the MCCL algorithm on all six datasets is above 0.97, and the MCCL algorithm has excellent accuracy.

In Guassian dataset, MCCL algorithm outperforms the optimal comparison SMCL algorithm by 0.074. In Ids2 dataset, DPADN algorithm performs poorly. In Banana dataset, SNN-DPC algorithm has only 0.721 accuracy, which is lower.In Lithuanian dataset, SMCL algorithm and MCCL algorithm have 0.982 and 0.989 accuracy respectively.In Ecoli dataset, MCCL algorithm outperforms DPAND algorithm by 0.372.

The accuracy of DPAND algorithm in Vehicle dataset is only 0.362. The accuracy of DPC algorithm, SNN-DPC algorithm, and SMCL algorithm in Vehicle dataset are 0.721, 0.702, and 0.752 respectively. And the accuracy of the MCCL algorithm proposed in this paper is 0.978, which is higher than the DPC algorithm, SNN-DPC algorithm, and SMCL algorithm by 0.257, 0.276, and 0.226, respectively.

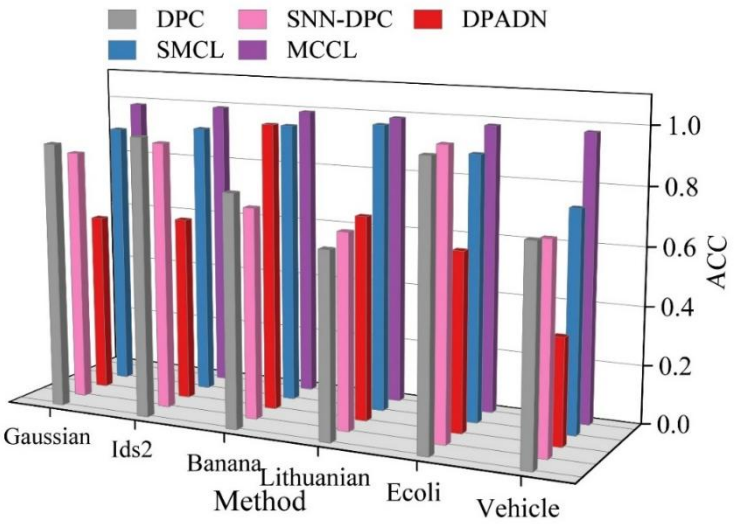

Figure 2: Comparison algorithm accuracy experiment results

3 Short-term load forecasting based on MCCL-BILSTM

3.1 Power Data Acquisition System

The composition of the smart grid is shown in Figure 3. Smart grid consists of power distribution network and communication network, the power distribution network is responsible for power generation and transmission work, and the communication network is responsible for monitoring the operation status of smart devices. On this basis, the AMI system comes into being and becomes a great asset of the communication network, which makes the communication network work efficiently, conveniently and orderly.The AMI system mainly

consists of three parts: the home area network (HAN), the neighboring area network (NAN), and the wide area network (WAN), of which the HAN's role is to connect the HAN and the AMI during the bidirectional communication, and the NAN controls the smart meters and smart collectors in two-way communication, which is usually capable of controlling thousands of smart meters as well as collectors, and the WAN is responsible for connecting the home area network and end-system connections.

As the cornerstone to support the continuous development of the power industry, the smart grid is developing vigorously at a new pace, and the AMI group is the core component, and its development largely determines the direction of the future of the smart grid, and in-depth research on it is imperative.AMI mainly consists of four parts: smart meter, customer gateway, AMI communication network, and AMI front-end. It has the following characteristics:

(1) Provide massive consumption report, which can view the trend of electricity consumption in time.

(2) The extensive use of software and communication facilities in AMI provides new loopholes for power thieves. Remote meter reading means that staff have fewer opportunities for on-site inspections and are less likely to detect unusual behavior.

(3) AMI is connected to the Internet and has almost all the characteristics of the Internet, but it cannot be installed with firewalls and antivirus software like the Internet, so it is vulnerable to attack and then tamper with smart meter data.

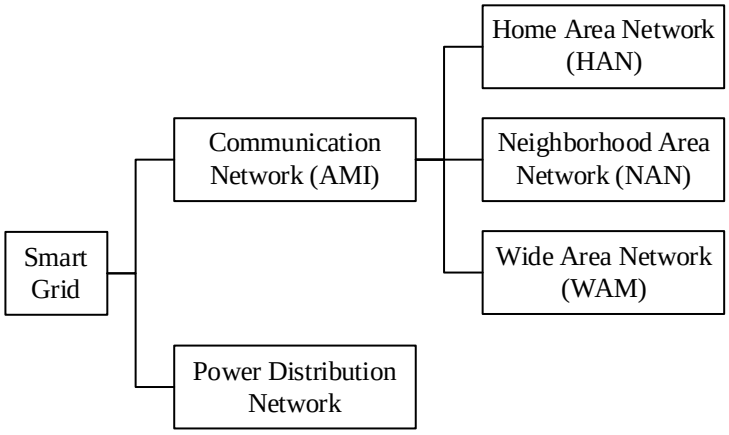

Figure 3: The composition of the smart grid

3.2 Basic steps of power load forecasting

Electricity load forecasting can be broadly categorized into the following seven steps:

Step 1: Define the needs and objectives of the forecast. The type of load forecasting needs to be clarified, choosing ultra-short-term, short-term, medium-term or long-term forecasting. A detailed power load forecasting plan is developed based on these.

Step 2: Collect reliable historical data. Accurate power load forecasting relies on real and reliable historical power load data, and inaccurate data will bring about poor forecasting accuracy, model underfitting or poor generalization ability.

Step 3: Analyze and process historical information. Processing historical load data is an essential and important part of achieving the goal of accurate load forecasting. Although professional means have been adopted as much as possible in the data collection stage, there are inevitable errors and missing data in electricity load data. Only professional pre-processing can ensure the integrity and reliability of the data to ensure that the later forecasting work to achieve good results.

Step 4: Determine the appropriate method and construct the model.

Step 5: Debug the parameters of the prediction model.

Step 6: Evaluate the model. Evaluate the model through the evaluation indexes to determine whether the model meets the forecasting requirements, if it does not meet the requirements, return to step 5 for iterative modification until it meets the forecasting requirements.

Step 7: Load forecasting. Use the constructed good load forecasting model to complete the forecasting work and output the forecasting results.

3.3 BILSTM model

Bidirectional Long Short-Term Memory Network (BiLSTM) is an improvement of LSTM network that not only learns forward information but also utilizes backward information effectively.

BiLSTM operation is as follows:

(1) $\vec{h}_t$ is the state of the hidden layer of the forward LSTM network at the moment t, which is calculated as shown in equation (13). It can be viewed as a single-layer LSTM network, and the state $\vec{h}_{t-1}$ at moment $t-1$ and the input at moment t are used to calculate the state $\vec{h}_t$ at moment t. That is:

$$\vec{h}_t = f\left(\vec{w} \cdot x_t + \vec{v} \cdot \vec{h}_{t-1} + \vec{b}\right) \tag{13}$$

where $\vec{h}_t$ denotes the forward LSTM hidden layer state at moment t. The $\vec{h}_{t-1}$ denotes the state of the LSTM hidden layer at moment $t-1$. The x_t denotes the input at moment t.

(2) $\overleftarrow{h}_t$ is the state of the hidden layer of the reverse LSTM network at moment t, and the formula is shown below:

$$\overleftarrow{h}_t = f\left(\overleftarrow{w} \cdot x_t + \overleftarrow{v} \cdot \overleftarrow{h}_{t-1} + \overleftarrow{b}\right) \tag{14}$$

where $\overleftarrow{h}_t$ denotes the state of the reverse LSTM hidden layer at moment t. The $\overleftarrow{h}_{t-1}$ denotes the state of the reverse LSTM hidden layer at moment $t-1$. x_t is the input at moment t.

(3) BiLSTM forward computed hidden vector is denoted by $\vec{h}_t$ and reverse computed hidden vector is denoted by $\overleftarrow{h}_t$, the final output of y_t is:

$$y_t = g\left(U\left[\vec{h}_t \ \overleftarrow{h}_t\right] + c\right) \tag{15}$$

The power load sequence is essentially a time-series data, whose input power load data before and after are interconnected, which is suitable for establishing a suitable short-term power load forecasting model by utilizing a bidirectional long- and short-term short-term memory network.

3.4 Load forecasting process based on data clustering

The short-term load forecasting model of BiLSTM neural network is constructed by clustering analysis based on meteorological factors on historical load data. The load forecasting process

is divided into two stages, which are the clustering analysis of load training sample data and load forecasting.

Phase 1: Multi-center clustering of load training data.

Construct a multi-center clustering method for imbalance data based on adaptive competitive learning, realize adaptive competitive learning clustering of multi-featured meteorological factor sample data sets, and construct load sample training data through index evaluation.

Phase II: BiLSTM neural network prediction.

Construct a short-term load forecasting model for load forecasting based on the training data, and optimize the neural network parameters by using the prediction error evaluation index in the process of neural network training.

4 Example forecasts of short-term electricity loads under uncertain scenarios

4.1 Similarity Day Clustering Model

The method of similar days is to form a daily feature vector based on a variety of factors affecting power loads, and use some similarity assessment method to select a number of similar days from historical days based on the degree of similarity of the daily feature vectors. In this paper, a similar day selection method based on PCA and K-means is proposed based on the theory described earlier. The method carries out the following steps for the selection of similar days: firstly, various factors affecting the load are selected. Then the principal component analysis method is adopted to form the similar day influence factors for the load influence factors and the corresponding N-1th day load data. Then, the principal components were used to reduce the dimensionality of the similar day influence factors, and the principal component contribution rate of more than 85% was selected as the input variable for cluster analysis. Finally, the analysis method of K-means is used to realize the selection of similar days and output the results of clustering according to the input principal components.

The flow of similar day selection using principal component analysis and K-means is shown in Figure 4.

The specific realization steps are as follows:

(1) According to the situation of the region and research information query with the region's electricity load data and related influence factors, construct similar day influence factor matrix $X_{(m*n)}$, where $X_{(m*n)}$ indicates that there are m influence factors in the n dimension.

(2) Normalize the matrix $X_{(m*n)}$ to obtain a new matrix $Y=(y_{ij})_{m*n}\left(j=1,2,\ldots,p\right)$.

(3) Create a matrix of correlation coefficients $R=(r_{ij})_{m*m}$ for the m factors that were normalized.

(4) Calculate the eigenvalues $\lambda_1>\lambda_2,\ldots,\lambda_n$ of the new matrix $R=(r_{ij})_{m*m}$ and its corresponding eigenvectors $\mu_1,\mu_2,\ldots,\mu_n$ and take the first r principal components $F_j\left(j=1,2,\ldots,r\right)$ whose contribution η_i is greater than 85%.

(5) Take the principal component $F_j\left(j=1,2,\ldots,r\right)$ as a variable, and use the K-means clustering method to select the similarity day for it, in which the selection of the K value in the K-means method utilizes the elbow method and the contour coefficient method to select.

(6) Output the classification results.

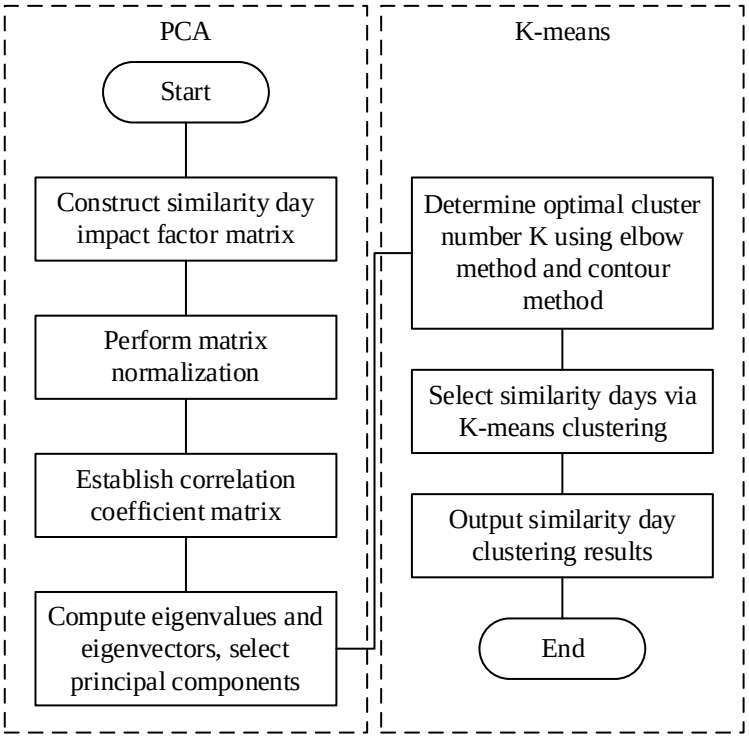

Figure 4: The process of the similar day selection of PCA and K-means

4.2 Short-term load forecasting based on clustering to screen similar days

4.2.1 Similarity day screening process

Under the principle of ensuring the temporal continuity of the original training data on the forecast target days without missing, 30 daily load profiles are randomly selected from the load data in the region from January 1, 2024 to December 31, 2024 as the forecast target, and the load data of the remaining 336 days are used as a clustering dataset, which participates in the clustering model and serves as the preparatory training data for the forecast model.

In addition to selecting the nine real influencing factors of the load in the region, the electric load data are added to form the daily load feature matrix to compose the original clustering dataset.

The composition of the daily load characteristic matrix is: the 24-hour load data of 336 load days in 2024 and the nine main influencing factors that have been selected. At this point, the dimension of the original clustered data matrix is 45. PCA principal component analysis needs to be taken to downscale the original clustered data and cluster the combined variables after PCA principal component analysis. After PCA principal component dimensionality reduction, the 10 composite variables with the largest eigenvalues are shown in Table 1.

The four variables with eigenvalues greater than 1 were selected as the composite variables of the original clustered data. At this point, the cumulative contribution rate of the 4 variables with selected eigenvalues greater than 1 has exceeded 94%, and it can be assumed that these 4 variables have contained most of the information of the original clustered data.

At this point, the 4 variables are combined to form the input matrix of the K-means clustering model, and the Euclidean distance is chosen as the quantitative criterion for clustering. Determine the maximum number of clusters, the number of clusters within the range of values will be input into the clustering preprocessing model respectively, and save the clustering results after completing the clustering.

Table 1: Part of the results of the PCA main component analysis

	Initial eigenvalue			Extracting the load of the load		
	Total	Percentage of variance	Cumulative %	Total	Percentage of variance	Cumulative %
PAC1	30.526	69.427	69.427	30.526	69.427	69.427
PAC2	7.889	15.526	84.953	7.889	15.526	84.953
PAC3	1.704	6.013	90.966	1.704	6.013	90.966
PAC4	1.156	3.264	94.23	1.256	3.264	94.23
PAC5	0.993	2.008	96.238			
PAC6	0.857	1.026	97.264			
PAC7	0.685	0.912	98.176			
PAC8	0.421	0.755	98.931			
PAC9	0.253	0.423	99.354			

The K-means algorithm was used in order to select typical days for each category for load characterization, and the center of clustering for each category in the K-means clustering results is shown in Figure 5.

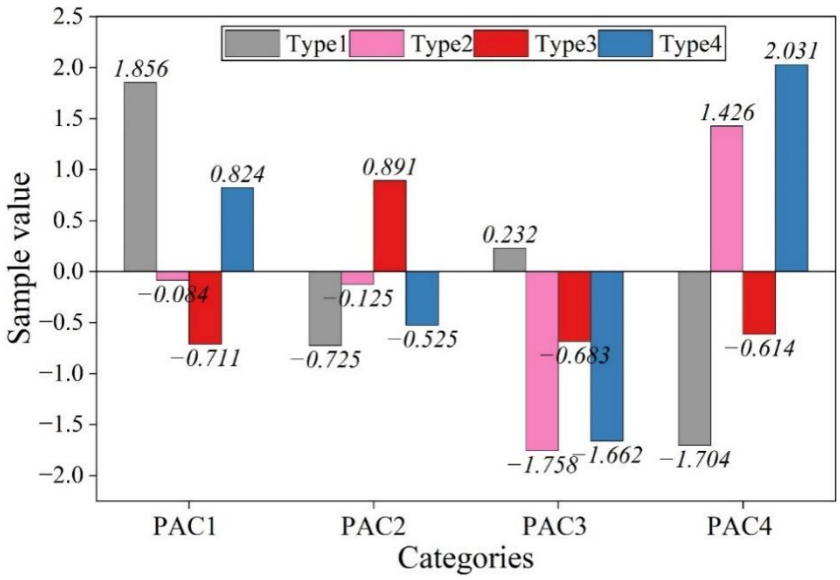

Figure 5: The k-means clustering results in various categories of clustering centers

4.2.2 Prediction results based on clustering screening for similar days

The clustering results classified the load data involved in training and their influencing factors into four categories, based on the MCCL-BILSTM prediction model built in this paper, the training data will be classified into the same category according to the clustering model for prediction.

Here, one day in the validation set of each category is selected as the prediction day, which is July 4, 2024 (the first category), May 5 (the second category), November 18 (the third category), and February 22 (the fourth category) for prediction, respectively.

In order to select the optimal load forecasting model, the performance of different forecasting models is compared. Four different base forecasting models, RNN, BP, LSTM and generalized regression neural network GRNN, are constructed for forecasting. And compare the prediction performance of different base models as well as verify the superiority of the

combined model MCCL-BILSTM based on clustering to screen similar days. The prediction results of the four categories of predicted days in the five prediction models are presented below.

The prediction results of MCCL-BILSTM model for category 1 are shown in Fig. 6. Comparing the actual load values, the prediction result curve of the LSTM model has similar ups and downs with the actual load curve, but the prediction result of the combined MCCL-BILSTM model is superior to the LSTM model.

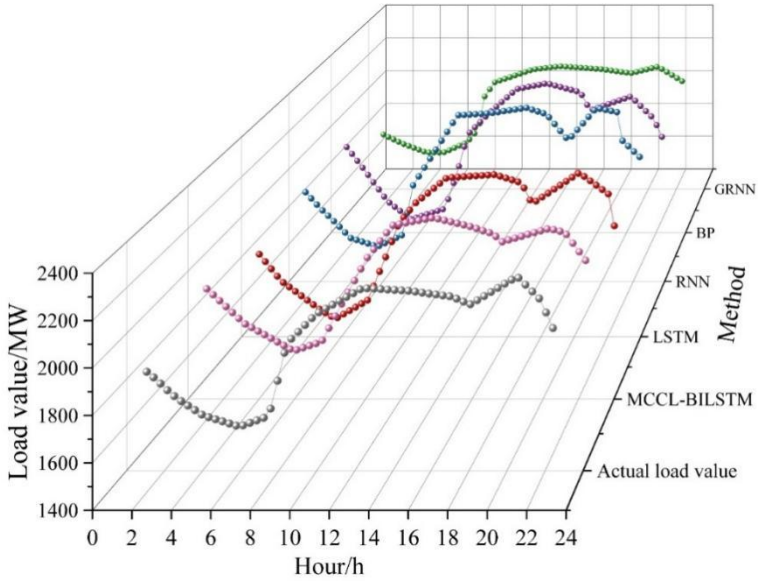

Figure 6: The MCCL-BILSTM model is the prediction of category 1

The prediction results of multiple models for category 2 are shown in Fig. 7. It can be seen that the fluctuation of the prediction result curves of LSTM model, RNN model, BP model, GRNN model and MCCL-BILSTM model in this paper are basically the same.

The prediction results of MCCL-BILSTM model at 12h and 24h are 1623.422 MW ~1668.6167 MW and 1441.0866 MW ~1531.7334 MW, respectively.The prediction results of MCCL-BILSTM model at 12h and 24h are more in line with the actual load values.

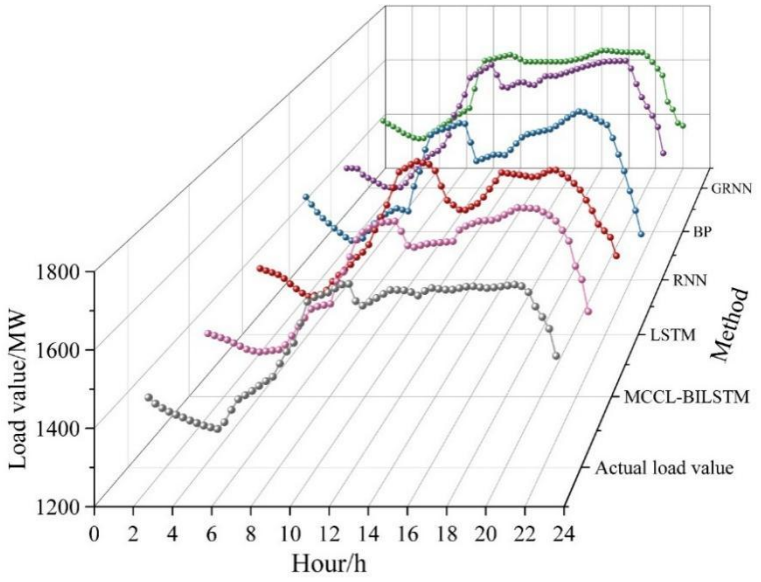

Figure 7: Multi-model prediction of category 2

The forecast results for category 3 are shown in Figure 8.

The predicted results of load values for category 3 by each model are kept within the range of 1210 MW~1650 MW.

The load prediction curves of LSTM model and RNN model at 22h~24h are lower than the actual load curves, and the predicted load values of LSTM model and RNN model at 22h are 1455.9790 MW~1517.3910 MW and 1373.3125 MW ~1448.5203 MW, respectively, and the model prediction accuracy is low. And the load prediction curve of MCCL-BILSTM model always fit the actual load value.

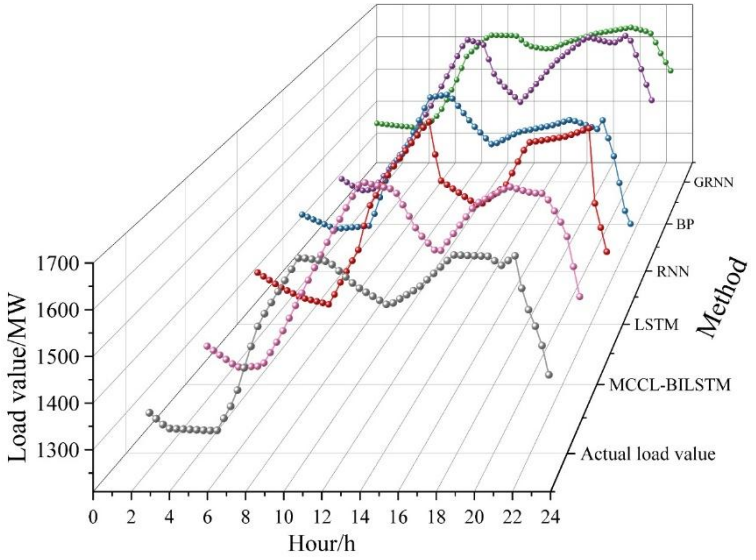

Figure 8: Category 3 prediction results

A comparison of the prediction results for category 4 is shown in Fig. 9.

In category 4, the actual load values for 12h are 1787.6544 MW ~ 1834.1367 MW, and the predicted values from the MCCL-BILSTM model are 1777.2324 MW ~ 1804.9865 MW. The predicted values of LSTM model are 1736.9591 MW ~1748.0220 MW. Compared to the LSTM model, the MCCL-BILSTM model based on clustering to screen similar days is more appropriate in predicting peaks and valleys to actual loads.

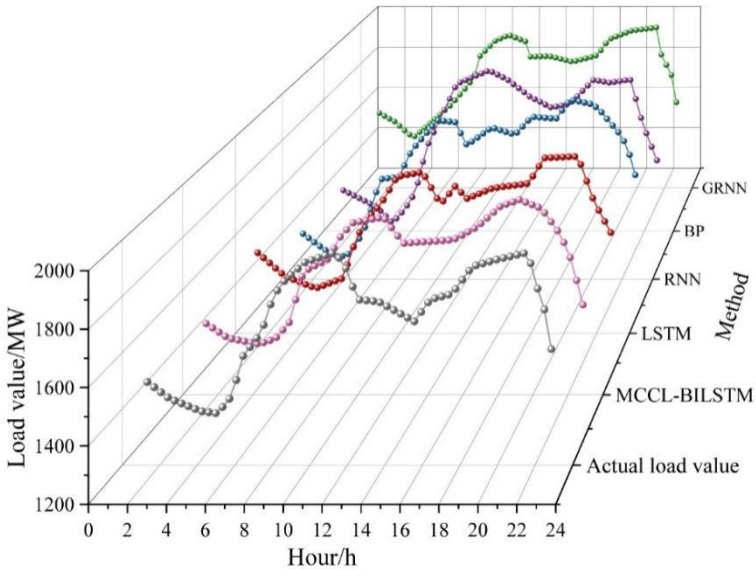

Figure 9: The prediction results of category 4 are compared

Combining the prediction results of each model for categories 1, 2, 3, and 4, it can be seen that among the four original prediction models without adopting the adaptive competitive learning-based multicenter clustering algorithm for imbalanced data to screen similar days as training data, their prediction results can reflect the general trend of the peaks and valleys change of the load curves, but there is a certain gap in prediction effect. The MCCL-BILSTM prediction model, which adopts clustering to screen similar days, shows more superior prediction results.

5 Conclusion

In this paper, adaptive competitive learning based multi-center clustering algorithm for unbalanced data is applied to process the electricity data, combined with BILSTM model to design the short-term electricity load forecasting process based on MCCL-BILSTM. Power load forecasting is performed for the similar days screened by clustering.

(1) The accuracy of the multi-center clustering algorithm based on adaptive competitive learning for imbalance data is above 97% on Guassian dataset, Ids2 dataset, Banana dataset, Lithuanian dataset, Ecoli dataset, and Vehicle dataset. The algorithm performs well on datasets with unbalanced features.

(2) PCA+K-means clustering filters out the similar days and divides them into four categories.The MCCL-BILSTM model performs power load forecasting for the four categories.The MCCL-BILSTM model consistently fits the actual load values in power load forecasting for the similar days in different categories. In category 4, the actual load values for 12h are 1787.6544 MW ~ 1834.1367 MW, and the predicted values of the MCCL-BILSTM model are 1777.2324 MW ~ 1804.9865 MW. it has better prediction effect than the original LSTM model.

Funding

This research was supported by the Study on power load characteristics in Sichuan under industrial upgrading and extreme weather influence (52199623000H).

About the Author

Lin Guo, male, Han, born in Chengdu, Sichuan, August 1973, bachelor's degree, senior engineer, research direction: power grid planning, plan management, etc.

Gang Wu, male, Han, born in Guangyuan, Sichuan, August 1995, master's degree, engineer, research direction: power market, load forecasting, etc.

Jieying Liu, female, Han, born in Chengdu, Sichuan, December 1990, master's degree, engineer, research direction: power system and automation, etc.

Yuxin Xiao, female, Han, born in Guiyang, Guizhou, April 2000, master's degree, assistant engineer, research direction: power system analysis, stability and control, etc.

Wei Wang, male, Han, born in Suizhou, Hubei, March 1984, Ph.D., senior engineer, research direction: power system, etc.

Ruiguang Ma, male, Han, born in Jining, Shandong, May 1987, Ph.D., senior engineer, research direction: energy policy, energy economy, power market, etc.

References

[1] Yu, Z., Liu, W., Chen, L., Eti, S., Dinçer, H., & Yüksel, S. (2019). The effects of electricity production on industrial development and sustainable economic growth: A VAR analysis for BRICS countries. Sustainability, 11(21), 5895.

[2] Bauer, N., Calvin, K., Emmerling, J., Fricko, O., Fujimori, S., Hilaire, J., ... & van Vuuren, D. P. (2017). Shared socio-economic pathways of the energy sector–quantifying the narratives. Global Environmental Change, 42, 316-330.

[3] Makarov, A. A., Mitrova, T. A., Veselov, F. V., Galkina, A. A., & Kulagin, V. A. (2017). Perspectives of the electric power industry amid the transforming global power generation markets. Thermal Engineering, 64(10), 703-714.

[4] Ugboke, P., Ebimaro, J., Olanrewaju, P., & Orukpe, P. (2024). The Impact of Digitization Towards Technological Revolution in the Power Industry. Journal of Energy Technology and Environment, 6(2), 193-200.

[5] Liang, M., Liu, L., Liang, W., Mi, W., Ye, K., & Gao, J. (2024). Intelligentization helps the green and energy-saving transformation of power industry-evidence from substation engineering in China. Scientific Reports, 14(1), 8698.

[6] Andrae, A. S. (2019). Prediction studies of electricity use of global computing in 2030. International Journal of Science and Engineering Investigations, 8(86), 27-33.

[7] Shen, C., Zhu, W., Tang, X., Du, W., Wang, Z., Xu, S., & Yao, K. (2024). Risk assessment and resilience enhancement strategies for urban power supply-demand imbalance affected by extreme weather: A case study of Beijing. International Journal of Disaster Risk Reduction, 106, 104471.

[8] Xu, Y., Yin, M., Dong, Z. Y., Zhang, R., Hill, D. J., & Zhang, Y. (2017). Robust dispatch of high wind power-penetrated power systems against transient instability. IEEE Transactions on Power Systems, 33(1), 174-186.

[9] Wang, F., Xu, H., Xu, T., Li, K., Shafie-Khah, M., & Catalão, J. P. (2017). The values of market-based demand response on improving power system reliability under extreme circumstances. Applied energy, 193, 220-231.

[10] Aquila, G., Morais, L. B. S., de Faria, V. A. D., Lima, J. W. M., Lima, L. M. M., & de Queiroz, A. R. (2023). An overview of short-term load forecasting for electricity systems operational planning: Machine learning methods and the Brazilian experience. Energies, 16(21), 7444.

[11] Li, C., Li, Z., Zhu, H., Tian, Z., & Feng, W. (2022). Study on operation strategy and load forecasting for distributed energy system based on Chinese supply-side power grid reform. Energy and Built Environment, 3(1), 113-127.

[12] Mounir, N., Ouadi, H., & Jrhilifa, I. (2023). Short-term electric load forecasting using an EMD-BI-LSTM approach for smart grid energy management system. Energy and Buildings, 288, 113022.

[13] Carvallo, J. P., Larsen, P. H., Sanstad, A. H., & Goldman, C. A. (2018). Long term load forecasting accuracy in electric utility integrated resource planning. Energy Policy, 119, 410-422.

[14] Habbak, H., Mahmoud, M., Metwally, K., Fouda, M. M., & Ibrahem, M. I. (2023). Load forecasting techniques and their applications in smart grids. Energies, 16(3), 1480.

[15] Wen, L., Zhou, K., Yang, S., & Lu, X. (2019). Optimal load dispatch of community microgrid with deep learning based solar power and load forecasting. Energy, 171, 1053-1065.

[16] Acaroğlu, H., & García Márquez, F. P. (2021). Comprehensive review on electricity market price and load forecasting based on wind energy. Energies, 14(22), 7473.

[17] Nti, I. K., Teimeh, M., Nyarko-Boateng, O., & Adekoya, A. F. (2020). Electricity load forecasting: a systematic review. Journal of Electrical Systems and Information Technology, 7(1), 13.

[18] Xu, L., Wang, S., & Tang, R. (2019). Probabilistic load forecasting for buildings considering weather forecasting uncertainty and uncertain peak load. Applied energy, 237, 180-195.

[19] Dab, K., Nagarsheth, S. H., Amara, F., Henao, N., Agbossou, K., Dubé, Y., & Sansregret, S. (2024). Uncertainty Quantification in Load Forecasting for Smart Grids Using Non-parametric Statistics. IEEE Access, 12, 138000-138017.

[20] Kumar, K. P., & Saravanan, B. (2017). Recent techniques to model uncertainties in power generation from renewable energy sources and loads in microgrids–A review. Renewable and Sustainable Energy Reviews, 71, 348-358.

[21] Li, Z., & Zhang, Z. (2021). Day-ahead and intra-day optimal scheduling of integrated energy system considering uncertainty of source & load power forecasting. Energies, 14(9), 2539.

[22] Luo, J., Hong, T., & Fang, S. C. (2018). Robust regression models for load forecasting. IEEE Transactions on Smart Grid, 10(5), 5397-5404.

[23] Wang, Z., Wen, Q., Zhang, C., Sun, L., & Wang, Y. (2024). DiffLoad: Uncertainty quantification in electrical load forecasting with the diffusion model. IEEE Transactions on Power Systems.

[24] Sheng, Y., & Hu, M. (2022). A Multi-Stage Planning Method for Distribution Networks Based on ARIMA with Error Gradient Sampling for Source–Load Prediction. Sensors, 22(21), 8403.

[25] Mi, J., Fan, L., Duan, X., & Qiu, Y. (2018). Short-term power load forecasting method based on improved exponential smoothing grey model. Mathematical Problems in Engineering, 2018(1), 3894723.

[26] Bessani, M., Massignan, J. A., Santos, T. M., London Jr, J. B., & Maciel, C. D. (2020). Multiple households very short-term load forecasting using bayesian networks. Electric

Power Systems Research, 189, 106733.

[27] Dong, X., Deng, S., & Wang, D. (2022). A short-term power load forecasting method based on k-means and SVM. Journal of Ambient Intelligence and Humanized Computing, 13(11), 5253-5267.

[28] Aseeri, A. O. (2023). Effective RNN-based forecasting methodology design for improving short-term power load forecasts: Application to large-scale power-grid time series. Journal of Computational Science, 68, 101984.

[29] Li, K., Huang, W., Hu, G., & Li, J. (2023). Ultra-short term power load forecasting based on CEEMDAN-SE and LSTM neural network. Energy and Buildings, 279, 112666.

[30] Imani, M. (2021). Electrical load-temperature CNN for residential load forecasting. Energy, 227, 120480.

[31] Liao, W., Ge, L., Bak-Jensen, B., Pillai, J. R., & Yang, Z. (2022). Scenario prediction for power loads using a pixel convolutional neural network and an optimization strategy. Energy Reports, 8, 6659-6671.

[32] Tang, C., Zhang, Y., Wu, F., & Tang, Z. (2024). An improved cnn-bilstm model for power load prediction in uncertain power systems. Energies, 17(10), 2312.

[33] Li, T., Wang, Y., & Zhang, N. (2019). Combining probability density forecasts for power electrical loads. Ieee transactions on smart grid, 11(2), 1679-1690.

[34] Wang, Y., Zhang, N., Tan, Y., Hong, T., Kirschen, D. S., & Kang, C. (2018). Combining probabilistic load forecasts. IEEE Transactions on Smart Grid, 10(4), 3664-3674.

[35] Jiang, C., Zheng, J., & Han, X. (2018). Probability-interval hybrid uncertainty analysis for structures with both aleatory and epistemic uncertainties: a review. Structural and Multidisciplinary Optimization, 57(6), 2485-2502.

[36] Yang, Y., Li, W., Gulliver, T. A., & Li, S. (2019). Bayesian deep learning-based probabilistic load forecasting in smart grids. IEEE Transactions on Industrial Informatics, 16(7), 4703-4713.

[37] Xu, L., Hu, M., & Fan, C. (2022). Probabilistic electrical load forecasting for buildings using Bayesian deep neural networks. Journal of Building Engineering, 46, 103853.

[38] Rodríguez, F., Bazmohammadi, N., Guerrero, J. M., & Galarza, A. (2021). A very short-term probabilistic prediction interval forecaster for reducing load uncertainty level in smart grids. Applied Sciences, 11(6), 2538.

[39] Zuniga-Garcia, M. A., Santamaría-Bonfil, G., Arroyo-Figueroa, G., & Batres, R. (2019). Prediction interval adjustment for load-forecasting using machine learning. Applied Sciences, 9(24), 5269.

[40] Li, B., Mo, Y., Gao, F., & Bai, X. (2023). Short-term probabilistic load forecasting method based on uncertainty estimation and deep learning model considering meteorological factors. Electric Power Systems Research, 225, 109804.

[41] Lv, X., Cheng, X., & Tang, Y. M. (2018, March). Short-term power load forecasting based on balanced KNN. In IOP Conference series: materials science and engineering (Vol. 322, No. 7, p. 072058). IOP Publishing.

[42] Zhang, C., Li, J., Zhao, Y., Li, T., Chen, Q., Zhang, X., & Qiu, W. (2021). Problem of data imbalance in building energy load prediction: Concept, influence, and solution. Applied Energy, 297, 117139.

[43] Zhu, D., Sun, Y., Cui, J., Hu, Z., Huang, J., & Zhang, J. (2023, November). Novel Power Load Pattern Recognition Algorithm Based on Unbalanced Data Clustering. In 2023 3rd International Conference on New Energy and Power Engineering (ICNEPE) (pp. 670-673). IEEE.

[44] Munguía Mondragón, J. C., Rendón Lara, E., Alejo Eleuterio, R., Granda Gutirrez, E. E., & Del Razo López, F. (2023). Density-based clustering to deal with highly imbalanced data in multi-class problems. Mathematics, 11(18), 4008.