Computer Science and Information Systems 20(2):657–678 https://doi.org/10.2298/CSIS220807013W

Predicting Smart Cities' Electricity Demands Using K-Means Clustering Algorithm in Smart Grid

Shurui Wang[1], Aifeng Song,[2, *]and Yufeng Qian[3]

[1] School of Electric and Electronic Engineering, Huazhong University of Science and Technology, Wuhan, 430000, China
[2] School of Management and Economics, North China University of Water Resources and Electric Power, Zhengzhou, China
saf0217@126.com
[3] School of Science, Hubei University of Technology, Wuhan, 430000, China

Abstract. This work aims to perform the unified management of various departments engaged in smart city construction by big data, establish a synthetic data collection and sharing system, and provide fast and convenient big data services for smart applications in various fields. A new electricity demand prediction model based on back propagation neural network (BPNN) is proposed for China's electricity industry according to the smart city's big data characteristics. This model integrates meteorological, geographic, demographic, corporate, and economic information to form a big intelligent database. Moreover, the K-means clustering algorithm mines and analyzes the data to optimize the power consumers' information. The BPNN model is used to extract features for prediction. Users with weak daily correlation obtained by the K-means clustering algorithm only input the historical load of adjacent moments into the BPNN model for prediction. Finally, the electricity market is evaluated by exploring the data correlation in-depth to verify the proposed model's effectiveness. The results indicate that the K-mean algorithm can significantly improve the segmentation accuracy of power consumers, with a maximum accuracy of 85.25% and average accuracy of 83.72%. The electricity consumption of different regions is separated, and the electricity consumption is classified. The electricity demand prediction model can enhance prediction accuracy, with an average error rate of 3.27%. The model's training significantly speeds up by adding the momentum factor, and the average error rate is 2.13%. Therefore, the electricity demand prediction model achieves high accuracy and training efficiency. The findings can provide a theoretical and practical foundation for electricity demand prediction, personalized marketing, and the development planning of the power industry.

Keywords: Smart city, smart grid, electricity prediction model, K-means clustering algorithm, back propagation neural network.

1. Introduction

With the rapid development of economic globalization, urbanization is accelerating. The problem of energy shortage is becoming even more obvious, such as water resource

[*] Correspondence author

lack, lack of non-renewable fossil resources, including coal and oil, and a serious waste of electricity. These increasingly prominent energy crises and environmental problems drive people to connect a considerable amount of distributed energy and energy storage devices. Meanwhile, green development and green economy have become the theme of economic development [1, 2]. The electric system is a critical element of urbanization, and intelligent electricity consumption is an inevitable factor in the construction of smart cities. The industrial society also puts forward higher reliability and quality requirements for power supply [3]. Consequently, scholars in related fields focus on applying intelligent prediction algorithms to intelligent electricity demand prediction.

The State Grid Corporation of China vigorously advocates the construction of a smart grid, and it leads the integration of various information technologies. It pursues the safety performance of the power system, the stability and distribution of power supplies, and the efficient utilization of energy, to reduce energy consumption and alleviate environmental pollution simultaneously in the power supply and distribution process [4]. The smart grid aims to optimize the production, distribution, and consumption of electric energy by obtaining more electricity consumption information from customers. Souri and Hosseini affirmed the business processing performance of big data technology in smart cities in their research [5]. The development of the Smart Grid makes accurate power load/demand prediction possible and increasingly urgent, helping power enterprises' scientific load dispatching and power production. Ultimately, the Smart Grid aims to achieve renewable energy in the power system. Introducing consumer decisions to reduce energy use will significantly reduce power load usage, which may reduce consumption in the coming years. Understanding the types of load prediction is helpful for the power system to manage the demand response plan effectively. Based on this, the load demand change can be predicted to calculate the required power generation and improve energy efficiency.

The electricity demand prediction for the smart grid is crucial for the power industry and critical to operation, planning, and control. Electricity demand prediction is a vital foundation for ensuring the safe operation of the power system, achieving scientific management and unified dispatching of the power grid, and formulating a reasonable development plan [6]. Electric energy consumption is affected by multiple factors. However, traditional electricity demand prediction methods take several factors as variables. It only considers the internal data of the power system, such as maximum load and regional average electricity consumption, ignoring numerous external factors [7]. Therefore, it is significantly crucial for the development of the power industry to investigate the electricity demand of the smart grid. Many high-resolution user load data can describe the user's power consumption habits and predict the user's power consumption. Therefore, mining user load data is of great value for grid system scheduling optimization, fine operation and management, and serving market users. According to the temporal and nonlinear characteristics of power load data, short-term load forecasting models can be divided into two categories. One is the time series approach, which usually regards the power load as a collection of time series. The prediction model is constructed according to the historical power load data and related influencing factors. In this way, the future load value can be predicted. The real-time load prediction method based on the user clustering strategy first takes a load of users at the hour as the feature. Then, it uses a clustering algorithm to classify users of different power consumption modes and a regression algorithm to forecast the load of classified users. By superimposing the predicted load of various users, the total real-time predicted power load is obtained, and then the generation is predicted.

In summary, according to the new characteristics of the current development of smart cities and the new needs of the smart grid, it is of great practical value to predict the demand for smart electricity in the smart city. This work uses the K-means algorithm to cluster the historical load curves of users' electricity consumption. BPNN is used to construct different feature extraction network structures for different types of users, and the prediction model is trained. The contributions of this work are as follows. (1) A customer segmentation model is constructed based on traditional customer segmentation methods by analyzing the customer differences through the demographic, legal person, economic and geographic information provided by the Smart City Basic Database. As corporate reconstruction continues in China, the power industry is also experiencing such a change. The electric power companies with a fixed source of customers and stable economic benefits in China are previously owned by the government, and the power generation, power supply, and power distribution of the companies are under unified management by the national government. Nowadays, Chinese power companies have also been promoted to the market, which will face more competition. Thus, the power industry in China must comply with market rules. Meanwhile, private enterprises' practical power demands must be fully considered to strengthen the consciousness of serving customers and help them increase profits. (2) The relationship between the total electricity consumption and regional population, Gross Domestic Product (GDP), per capita GDP, the number of industrial enterprises above a specific size, and total imports and exports are analyzed by data from the Smart City Basic Database. An electricity demand prediction model is established by combining the internal data from power companies. Finally, its performance is analyzed through simulation to provide a reference for green energy use in subsequent smart cities.

2. Recent Related Work

There have been many studies on electricity demand prediction in the past few years. Classic and traditional electricity prediction methods have failed to adapt to the changing electricity market. Some new methods have been applied to electricity demand prediction. Mirjat et al. (2018) predicted that Pakistan's average electricity growth rate would be 8.35% in subsequent years using deep learning and electricity data from 2015 to 2017; this data was19 times the existing base [8]. Khalifa et al. (2019) used electricity consumption data to model the Qatar electricity market. They found that more energy consumption would be generated around 2030 and proposed to improve electricity efficiency by reducing electricity consumption [9]. Kim et al. adopted the Long Short-Term Memory (LSTM) algorithm of deep learning to predict electricity consumption and applied the one-hot coding method to the input and output values of electricity demand. This model had higher prediction accuracy than other general algorithms [10]. Then, the observed daily load curves were represented by a set of periodic smooth-spline basis functions. The basis function coefficients were obtained following the evolution of a linear Gaussian state space model. Nam et al. (2020) developed an energy prediction model by renewable energy technologies and implemented it in South

Korea's energy policy. They utilized the deep learning algorithm to predict fluctuations in electricity demand and power generation. Through experiments, they proved that this model achieved the lowest economic and environmental costs, generated stable electricity to meet demand, and realized the policy of 100% renewable energy [11]. The general LSTM hybrid model only optimizes the prediction front end. The optimization treatment of the residual at the back end of LSTM prediction is ignored, and the optimization treatment of the residual is missing. However, BPNN, as a multi-layer mapping network with forwarding information transmission and back error propagation, can achieve the purpose of error correction.

The above research shows that different models have been devised for electricity demand prediction and have achieved better model performance than traditional models in electricity demand prediction. However, some models only consider the demand of the electricity market without the influencing factors of electricity demand from other aspects during prediction. Therefore, it is necessary to determine the critical influencing factors using appropriate methods and massive data to establish a prediction model for electricity demand.

An electricity demand prediction model is proposed according to the actual situation by investigating literature about the smart city and smart grid. The model considers the new characteristics of current smart cities and the new needs of smart grids. A segmentation model is created by classifying electricity consumers and analyzing data from the smart city base database to study demographic, economic, and geographic differences. In addition, a new electricity quantity prediction algorithm is innovatively proposed, achieving high accuracy based on data correlation. The results can provide a theoretical and practical basis for smart grid construction. It provides new ideas for smart city construction.

3. Methods

3.1. Methods of electricity demand prediction

The three most commonly used electricity prediction methods are the classic prediction method, the traditional prediction model, and the intelligent prediction model. (1) Classic prediction methods include the trend extrapolation method, classified electricity demand prediction method, and load density method. Although these prediction methods are widely used, most analyze the relationship between some simple variables without deep data analysis. Thus, they cannot provide precise prediction results [12]. (2) Traditional prediction methods contain the regression analysis method, time series method, and random time series method. The regression analysis method establishes the relationship between the dependent variables and known load data. Then, it predicts the electricity system's load through mathematical analysis. By comparison, the time-series methods cover the exponential smoothing and the Census-H Decomposition methods. The random time series methods include the state space method, the Box-Jenkins method, and the Markov method [13]. According to the given data, the relationship between the independent and dependent variables is determined, and the regression equations and various parameters are confirmed. Based on the given equation, the dependent variable is obtained from the existing independent variables, and finally, the electricity prediction data are obtained. (3) When there are large random factors in historical electricity demand, there may be errors in the prediction result caused by bad data in the time series. In recent years, the electricity market has become increasingly complicated. Classic prediction methods cannot adapt to the electricity market's nonlinear, multi-variable, time-varying, and random characteristics. Hence, some new prediction methods are used in electricity demand prediction. The laws are extracted to establish a knowledge base for reasoning and judgment based on real experience [14]. Table 1 illustrates the comparison results of the advantages and disadvantages of different prediction methods.

Table 1. Comparison of different electricity demand prediction methods

Recognition methods	Advantage	Disadvantage
Classic prediction method	The broadest range of applications and the most extended use time	Lack of scientific theory, and low prediction accuracy
Traditional predictive model	Substantial data analysis capabilities and high model accuracy	The algorithm runs for a long time and requires high system configuration
Intelligent prediction model	High prediction accuracy	Time-consuming database construction

3.2. K-Means Clustering (KMC) algorithm

(1) Algorithm utilization: at present, the construction of a smart grid must comply with market laws, and electric power enterprises must rely on its competitiveness for subsequent survival and development. It is crucial for enterprises to fully understand all customers, improve customer experience, and enhance customer loyalty. For customer segmentation, the traditional segmentation method uses a single indicator and cannot effectively divide customers. With the development of smart cities and the advancement of big data technology, a large amount of data can be obtained, while data mining technology can be used to extract the required indicators to segment power customers [15]. Currently, among various data mining algorithms, the K-Mean clustering algorithm has attracted the attention of many scholars due to its simple implementation and high efficiency.

The KMC algorithm uses the distance between two targets as an evaluation indicator to measure the similarity. When the distance between two objects is small, the similarity between the data is relatively high. This algorithm usually consists of relatively close objects. The final goal is to obtain a data group with a compact distance and a high degree of separation [16].

(2) Algorithm principle: the initial dataset is set to (x1, x2 … xn), and each data unit is a p-dimensional vector (the p-dimensional vector is composed of p eigenvalues). The KMC algorithm aims to divide the original dataset into K categories G= {G1, G2, …,

Gk} with a given number of categories k (k=n). Each iteration of the KMC algorithm must check whether the classification of each data unit is correct. The data must be adjusted if it is classified into the wrong category. The adjusted data is clustered with k points in the space as the center, and each cluster center's value is updated until the cluster center is a constant. The stable state of the cluster center indicates that the clustering criterion function has converged, and the best clustering result is obtained [17]. Figure 1 illustrates the implementation scheme of the KMC algorithm.

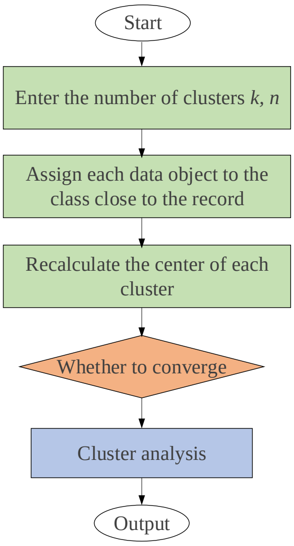
Start
Enter the number of clusters *k*, *n*
Assign each data object to the class close to the record
Recalculate the center of each cluster
Whether to converge
Cluster analysis
Output

Fig. 1. The implementation scheme of the KMC algorithm

First, K data objects are arbitrarily selected from N data objects as initial clustering centers. The remaining objects are assigned to the cluster that is most similar to them (represented by the cluster center) according to their similarity (distance) to these cluster centers. Then, the cluster center of each new cluster (the mean of all objects in the cluster) is calculated, and the process is repeated until the criterion function begins to converge.

For consumer segmentation based on the KMC algorithm, the clustering result depends on the random selection of the initial cluster center. In practice, it has the advantages of simple description and easy implementation. It is scalable and efficient for processing large data sets. The specific implementation steps of the power consumer segmentation model are as follows. (1) K objects are randomly selected from n pieces of sample data as the initial cluster center. (2) The distance from each sample to each cluster center is calculated. The sample is assigned to the nearest cluster center category. (3) After all samples are allocated, it recalculates the centers of k clusters. (4) Compared with the k cluster centers obtained from the previous calculation, if the cluster center changes, turn to step (2); otherwise, turn to step (5). (5) When the center does not change, the algorithm flow stops and outputs the clustering results. Before any operation, the user load data in the prediction area and is preprocessed to identify and correct bad data. According to the different correlations of its time series, the user load series of different categories are converted into tensors and then input into BPNN for abstract feature extraction. Furthermore, according to the differences in total and industry load characteristics, the k-means algorithm is used to aggregate load curves to obtain industry-typical load curves, and the user clustering curves are further extracted to form group loads. Finally, the group load is predicted by selecting appropriate monthly and annual forecasting methods.

3.3. Back propagation neural network (BPNN) algorithm

The Artificial Neural Network (ANN) is composed of numerous interconnected neurons and has a strong nonlinear mapping ability [18]. The BPNN is a multi-layer feedforward network trained according to the backpropagation algorithm [19]. The topological structure of the BPNN includes an input layer, a hidden layer, and an output layer, as shown in Figure 2.

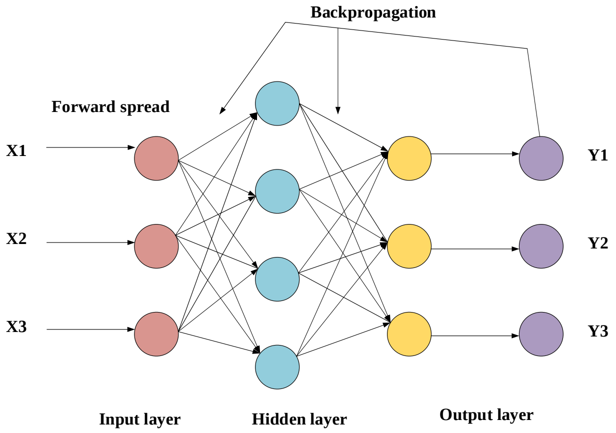

Fig. 2. Structure of BPNN

The BPNN is usually composed of multiple layers and multiple neurons, generally including an input layer, a hidden layer, and an output layer [20]. The specific flow of the BPNN prediction model is shown in Figure 3. The input vector can be expressed as:

$$x = \left[x_1, x_2, x_3 ... x_i, ... x_m \right], i = 1, 2, ... m \tag{1}$$

The output vector can be written as:

$$y = \left[y_1, y_2, y_3 ... y_{\mathrm{k}}, ... y_n \right], k = 1, 2, ... n \tag{2}$$

Equation (3) describes the neuron input of the hidden layer.

$$h^{(l)} = \begin{bmatrix} h^{(l)}{}_1, h^{(l)}{}_2, h^{(l)}{}_3 \\ \dots h^{(l)}{}_j, \dots h^{(l)}{}_{sl} \end{bmatrix}, j = 1, 2, \dots .sl \quad \textbf{(3)}$$

In Equation (3), *sl* denotes the number of neurons in the *l*-th layer. Assumed that $w^{(l)}{}_{ij}$ represents the connection weight associated with the *j*-th neuron in the *l*-th layer, $b^{(l)}{}_i$ refers to the threshold of the *i*-th neuron in the *l*-th layer, and $net^{(l)}{}_i$ stands for the input of the *i*-th neuron in the *l*-th layer. Then:

$$h^{(l)}{}_i = f(net^{(l)}{}_i) \quad \textbf{(4)}$$

$$net^{(l)}{}_i = \sum_{j=1}^{sl-1} w^{(l)}{}_{ij} h^{(l\text{-}1)}{}_j + b^{(l)}{}_i \quad \textbf{(5)}$$

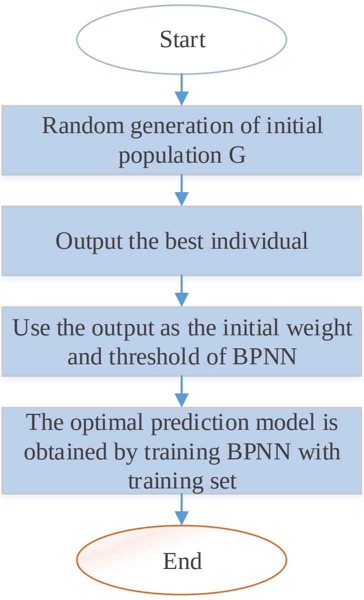
Start
Random generation of initial population G
Output the best individual
Use the output as the initial weight and threshold of BPNN
The optimal prediction model is obtained by training BPNN with training set
End

Fig. 3. Specific flow of BPNN prediction model

3.4. Power consumer segmentation model

(1) Overall framework: based on the above theory, the Smart City Basic Database is utilized to establish a functional structure model for power consumer segmentation (Figure 4). First, a data warehouse is established. Then, relevant customer segmentation data is extracted for data analysis. Moreover, data is cleaned and conversed. The association analysis method is adopted for data mining, and finally, the mining results are analyzed. The Smart City Basic Database is the foundation of the entire model. Pre-processing of data is the guarantee for real and effective mining results. The effectiveness of customer segmentation depends largely on selecting customer consideration standards and establishing measurements. The adopted mining method based on actual needs is the key to the entire model.

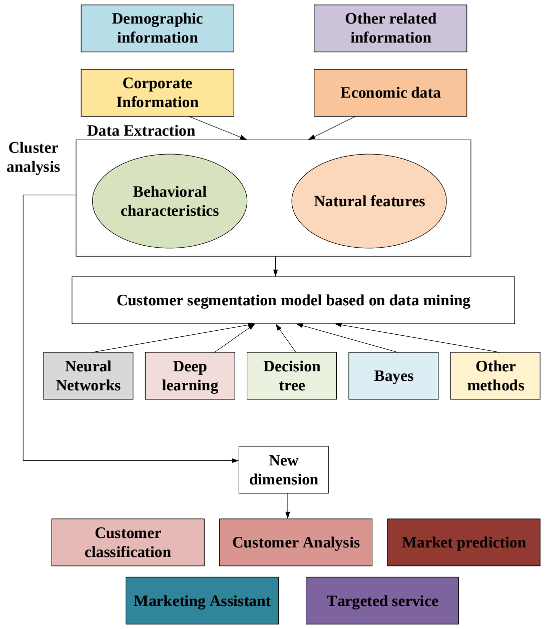

Fig. 4. Functional structure of power consumer segmentation

(2) Customer segmentation: Currently, electricity data analysis research and technology use traditional data extraction methods and statistics. The Smart City Basic Database's data is utilized to analyze electricity data to optimize the accuracy and usability of the algorithm. Besides, the big data method is utilized for prediction to maximize the type of experimental data. The previously impossible prediction task can be completed by exploring the data association relationship at present, which can ensure high precision simultaneously. The population electricity value information directly reflects the individual electricity value of customers. It indirectly demonstrates the gathering area of high-potential customers through personal information and social insurance information. In load (electricity demand) prediction, users are subdivided and predicted separately. That is, the major categories of power consumption characteristics are understood. Then, a regression algorithm is used to model and predict each cluster's load and add up each cluster's prediction results to form the final urban load forecast. Additionally, the prediction results can be compared with the actual historical data to generate an evaluation. The evaluation will be fed back to the prediction model. The accuracy of the prediction model can be improved by adjusting the modeling parameters accordingly.

(3) Data processing: first, the Dongfangtong TI-ETL tool and data desensitization technology are utilized to transform sensitive or confidential information to protect private data. Some missing data, including names, gender, and address of customers, is extracted from other information. Various social insurance databases are integrated, and the collected information is used to roughly restore the demographic information and provide the basis for power consumer segmentation.

(4) Algorithm realization: it is divided into population electricity value information, enterprise commercial value distribution, and macroeconomic information. Figure 5 reveals the population electricity value information. The power consumers are segmented. The resident data is arranged in a table from high to low in the order of individual units according to residents' social insurance information, social security information, corporate information, and the potential and influence of electricity use. Each administrative district is taken as a unit. As for the data of the corporate legal person, the legal person's registered capital is used as the analysis target, and the social insurance information is determined according to the insurance amount. In addition, the social security information depends on the subsidy amount, and the housing provident fund is based on the monthly payment amount [21].

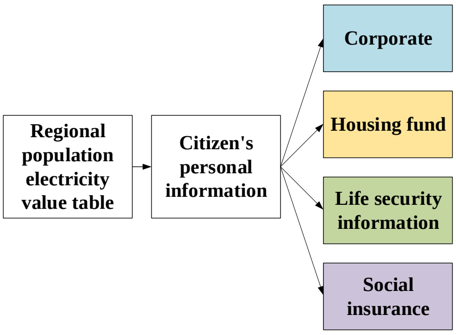
Regional population electricity value table
Citizen's personal information
Corporate
Housing fund
Life security information
Social insurance

Fig.5. Structure of population electricity value realization

The enterprise commercial value distribution is presented in Figure 6. Obviously, the organization code is used as the search basis to match the source data from each commission, office, and bureau. The evaluation and ranking are based on four dimensions: business category, registered capital, annual turnover, and the number of employees. Each administrative district is taken as a unit. For corporate legal person's data, KMC analysis is performed based on turnover, registered capital, and the number of employees. Among various business categories, the conversion weight of turnover for the construction industry is 10%, that for the manufacturing industry is 100%, that for the wholesale and retail industry is 30%, and that for the service industry is 15% [22].

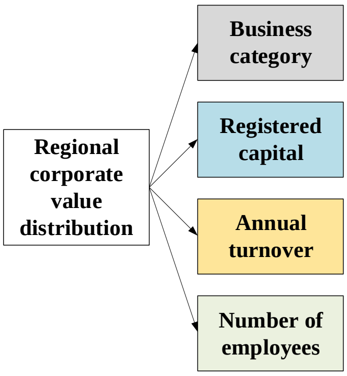

Fig. 6. Structure of enterprise commercial value realization

In terms of macroeconomic information, the macro value evaluation aims to evaluate the administrative districts' electricity consumption potential to segment power consumers. The data from some districts are accurate to the administrative streets. Previous studies have selected several significant data categories, such as regional GDP, per capita disposable income, per capita GDP, total asset investment, and trade data [23].

In a city, five districts, A, B, C, D, and E, are chosen for empirical research. In District A, four companies are chosen from the industrial area A11-A14. In District B, four companies are chosen from the office building areas B11-B14. In District C, four neighborhoods are selected from the residential areas C11-C14. District D and District E are comparison controls.

3.5. Model of electricity demand prediction

(1) Influencing indicators: electric energy consumption is affected by multiple factors. However, traditional prediction methods only consider a few factors as variables. These methods only consider the internal data of electricity companies and fail to fully consider the impact of changes in other factors on prediction, ultimately providing limited prediction accuracy [24]. The external information of electricity companies provided by the Smart City Basic Database is fully utilized. Given the impact of changes in various relevant objects on the prediction value, the regional annual electricity consumption is predicted considering multiple influencing factors.

In addition to the electricity companies' factors, the total electricity consumption is also affected by many factors such as population, corporate trends, economic conditions, energy policies, and electricity price adjustments. In particular, the development of the social economy needs to consume a large amount of electric energy, and there is a correlation between electricity consumption and economic indicators [25]. There are many statistical dimensions of financial data. The district's social product, per capita GDP, price index, total import and export, and other economic indicators of the electricity companies are chosen to establish an electricity consumption prediction model [26]. The specific names of the data indicators are as follows: (1) macroeconomic data; (2) policies and other external data; (3) regional electricity consumption data in past years; (4) power consumer segmentation data.

(2) Normalization processing: load prediction is still very challenging because it depends on external factors such as weather and exogenous variables. The model's characteristics (including the prediction period or the variables used) are reflected in the structure of the neural network. The empirical selection method is used to determine the model input variables, and then the power load prediction model is established. ANN usually normalizes the data before training. The training effects of different transfer functions are different to avoid neuron oversaturation [27]. The input data value must be within [0,1], which is the characteristic requirement of the transfer activation function. Therefore, the original data of the network must be processed. The original data are normalized, and the equation is as follows:

$$x_{i0} = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \tag{6}$$

In Equation (6), x_{i0} denotes the *i*-th feature parameter after normalization, x_i represents the original *i*-th feature parameter, $x_{\min}$ signifies the minimum value of the *i*-th feature parameter, and $x_{\max}$ stands for the maximum value of the *i*-th feature parameter.

(3) Hyperparameter setting: after preliminary experiments, a three-layer network model structure with a hidden layer is determined. The number of neurons is 18, and the logsig transformation function is used. The number of neurons in the second layer is the same as the number of output variable vectors, and the output layer uses a pure linear transformation function. The input feature parameters are 65; that is, the number of input layer nodes is 65, and the number of output layer nodes is 5. Generally, increasing the number of nodes in the hidden layer can reduce the network's training error rather than increasing the number of hidden layers. The BPNN algorithm can be set as a three-layer structure to map the n-dimensional input layer to the *m*-dimensional output layer. Therefore, the number of hidden layers in the network is determined as 1. When applying a neural network for electricity prediction, the reference equations for selecting the number of hidden layer neurons are as follows:

$$h = 2m + 1 \tag{7}$$

$$h = \sqrt{n+m} + \alpha * \tag{8}$$

In Equations (7) and (8), *h* represents the number of nodes in the hidden layer, *m* denotes the number of nodes in the input layer, and *n* refers to the number of nodes in the output layer. After a comprehensive comparison of experiments, the number of nodes in the hidden layer is selected as 18. The S-tangent tansing is selected as the activation function for hidden layer neurons, and the activation function for output layer neurons is the S-type logarithmic logsig function.

(4) Data source: the streaming data in the power grid is collected from smart meters, PMUs, and various sensors. The data is extensive in scale, diverse in structure, and fast in speed. The electric power company has installed many smart meters to accurately obtain customers' electricity consumption data of different electric equipment. The meters will send real-time electricity consumption information to the grid every 5 minutes. Consequently, streaming data collection requires fast collection speed, high reliability, real-time monitoring of data changes, and simple data processing. Therefore, the collection system is a distributed, reliable, and highly available system of massive log aggregation, which can monitor and receive data from the client and send it out. The log file is transferred to other nodes without loss when a node fails, ensuring data integrity. The collected data is divided into a training set and a test set by the ratio of 8:2, and the percentage of each data type in the two datasets is consistent.

3.6. System improvement and verification

(1) System improvement: the influence of changing trends on the error surface. The BPNN may fall to a local minimum, which the additional momentum can prevent. The adjustment equations for weight and threshold with additional momentum factor are:

$$\Delta w_{ij}(k+1) = (1-mc)\eta\sigma_i p_j + mc\Delta w_{ij}(k) \tag{9}$$

$$\Delta b_{ij}(k+1) = (1-mc)\eta\sigma_i p_j + mc\Delta b_{ij}(k) \tag{10}$$

In Equations (9) and (10), *w* denotes the weight vector, *k* refers to the number of trainings, and *mc* represents the momentum factor. In addition, η stands for the learning rate, Δb_{ij} signifies the gradient of the error function, and σ_i and p_j are the correlation coefficients.

Power load prediction is susceptible to many elements. These influencing factors are used as independent variables to explain the changes in dependent variables. When there is a linear relationship between multiple dependent and independent variables, the regression problem becomes a multiple linear regression. The prediction model is as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \varepsilon \tag{11}$$

In Eq. (11), β_0 is a constant term. $\beta_1, \beta_2, ..., \beta_n$ represents a regression coefficient. $x_1, x_2, ..., x_n$ refers to the n influencing factors related to the load. Y stands for the predicted value, and ε indicates a random disturbance term.

(2) Clustering algorithm evaluation: here, the KMC algorithm uses the square error and criterion function to evaluate the clustering performance. *X* represents the given dataset, and each data unit is a *p*-dimensional feature vector. It is set to *K* categories. The algorithm randomly selects *k* data as the starting cluster center to analyze the distance from each data unit to the cluster and divides the data into the array sink where the corresponding cluster center is located. It is supposed that X contains k data groups X_1, X_2 … X_k, the amount of data units in each data group is m_1, m_2, ..., m_k; the cluster centers of each data group are n_1, n_2, ..., n_k. The used square error equation [28] is:

$$E = \sum \sum \| p - m_i \| \tag{12}$$

(3) Training of BPNN: during the training, there are 18 neurons in the hidden layer of the BPNN network with the logsig transform function. The number of neurons in the second layer is the same as the number of output variable vectors, and the output layer adopts the pure linear transform function. The selection of expected training errors is related to the number of neurons and the training time. The minimum error boundary E of the general neural network ranges from 0.001 to 0.1. When the error value of the cyclic calculation is smaller than E, the calculation is stopped, and the calculation results are output. Otherwise, the calculation continues until the error meets the design requirements. Here, E is set to 0.02. In the selection of the learning rate, the excessively large learning rate will affect network stability. In contrast, the extremely small learning rate will prolong the learning time of the network, and convergence will deteriorate. Therefore, the appropriate learning rate should be selected according to the specific network model structure. Here, the learning rate is set to 0.05, and the network will be trained 10,000 times. The preprocessed data matrix is imported into MATLAB and normalized. The Newff function is employed to establish a BPNN model.

The data for the experiment comes from the London household electricity Usage Record (UK-DALE) provided by the UK Energy Centre. The dataset is collected using a smart meter. The original power demand data are household load data collected with 6s as the acquisition frequency. Then, the model forecasts 1-hour household electricity demand. Therefore, the data unit is converted first, and the load data is converted to the electricity consumption within 6 seconds, and then the data is merged. 90% of the combined data are selected as the training set, and 5% are used as the validation set. 10% is used as the test set. After the calculation, denormalization processing is performed on the data. The MATLAB toolbox calculates the relative error percentage and the predicted electricity demand value.

4. Experimental Results

4.1. Results of power consumer segmentation

Different electricity consumption areas can be divided through the above customer segmentation model, as shown in Table 2 and Figure 7. The four companies in District A are in the high electricity consumption area in their administrative district. The medium electricity consumption area is the office area of District B. The area with low electricity consumption is the residential electricity area of District C. This is consistent with the actual result. Then, all areas are divided into residential electricity and commercial electricity areas. According to the model, the number of all residential customers and commercial customers in the city can be subdivided into four levels from P1-P4, from more to less electricity consumption. The number of residents with the highest electricity consumption reached 25,678. The number of commercial customers with the highest electricity consumption reached 298. Such results also confirm the effectiveness of the segmentation model proposed here.

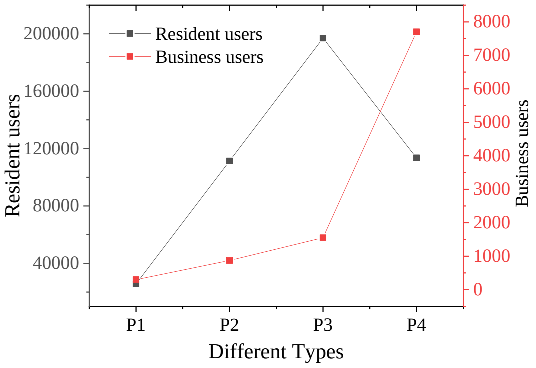

Fig. 7. Performance analysis of the model based on BPNN

Table 2. Regional electricity potential distribution

Area category	High electricity consumption area	Middle electricity area	Low electricity area
Bend over	A11	B11	C11
	A12	B12	C12
Squat down	A13	B13	C13
	A14	B14	C14

Each feature clustering's center value is analyzed for clustering user groups. According to the difference in performance features of user categories, the power user value is grouped to formulate the most reasonable value response strategy. The strategy can optimize the power sales side. The prototype system of power user payment data is designed and developed to visualize the clustering results of power users. The system can characterize the users of power enterprises.

4.2. Performance analysis of the electricity demand prediction model

Figure 8 shows the results of electricity demand prediction calculated by the BPNN model. Figure 9 shows the model's performance analysis results based on BPNN. The electricity demand prediction results of different areas are close to the actual value, and the largest relative error is 5.129%. The minimum value is 2.294%, and the mean relative error (MRE) is 3.2671%. Hence, the BPNN has a good prediction effect. As shown in Figure 8, the model's training error is analyzed, and the results suggest that the model tends to be stable when it is trained for 5,000 s. In contrast, the training speed of the algorithm of related research is excessively long, failing to meet the actual demand.

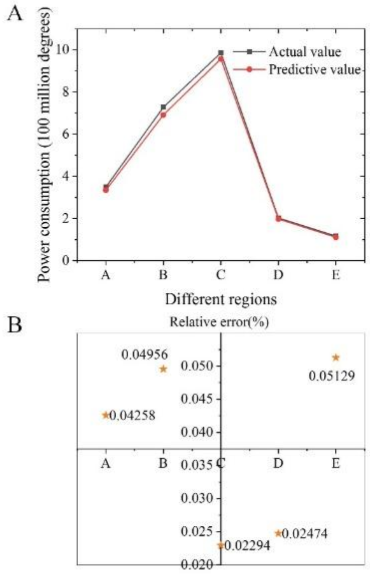

Fig. 8. Electricity demand prediction results based on BPNN

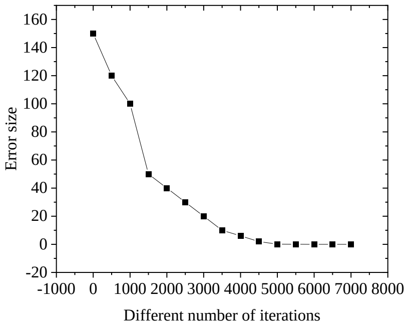

Fig. 9. Performance analysis of the model based on BPNN

4.3. Performance analysis results of the improved algorithm

After optimization, according to the results shown in Figures 10 and 11, the improved BPNN algorithm has a smaller MRE, with an MRE of 2.13%, and a faster training speed compared with the unimproved BPNN. Besides, the improved algorithm reached a stable state in the 2000s, and its training accuracy is higher, which is more advantageous in predicting electricity demand. Hence, the improved algorithm meets the basic requirements of the electricity company for electricity prediction and has particular practical application value.

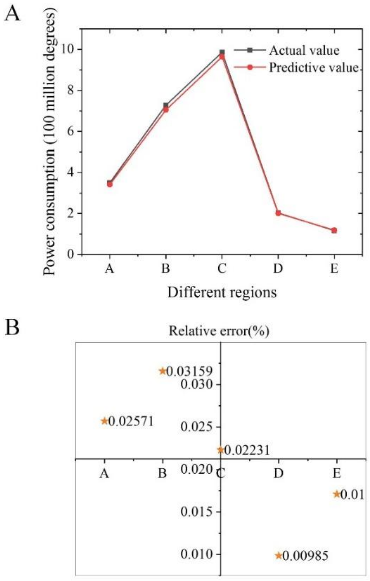

Fig. 10. One kernel a electricity demand prediction results based on improved BPNN

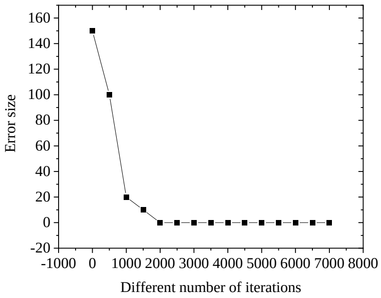

Fig. 11. Model performance analysis based on improved BPNN

4.4. Performance comparison of different models

Table 3 summarizes the comparison results of the model reported here and the latest models. The accuracy of the BPNN reported in the literature [29] is relatively high and remains at 79.63%. Compared with methods in other literature, the proposed model has the best performance. The accuracy rate of the model is 85.25%, and the average value is also the largest, 83.72%. Hence, the proposed model is significantly better than the latest models in terms of algorithm performance. Accuracy (%)

Table 3. Performance comparison results of different models

Models	Accuracy (%)	Recall (%)	F1-value	Mean value
Literature [29]	0.7963	0.6625	0.8253	0.761366667
Literature [30]	0.8142	0.6768	0.8649	0.7853
Literature [31]	0.7125	0.6972	0.8165	0.742066667
Literature [32]	0.8073	0.7323	0.8246	0.788066667
The proposed model	0.8525	0.7848	0.8745	0.837266667

5. Discussion and Analysis

The smart city system's comprehensive application of big data can provide trend judgments and information sharing. Doing so promotes the innovative management and interconnection of cities, the healthy competition of various industries, and the sustainable development of society. Besides, with the increasingly prominent energy crisis and environmental problems, many distributed energy sources and energy storage devices are connected, and industrial society has put higher requirements for power reliability and quality. The load forecast is of great significance to the dispatch of the power system, the fundamental basis of formulating power generation and transmission plans, and an influential aspect of the modern development of the power market. Improving the accuracy of power system load forecasting can effectively improve the power sector's economic benefits and promote the power grid's safe and economical operation. Here, the demand forecasting of the power system is studied by consulting many relevant domestic and foreign documents.

The characteristics and significance of power system demand forecasting are considered under the big data from smart cities. Traditional power data analysis methods rely primarily on limited sample data. New algorithms are proposed and improved based on proprietary data in this industry and field to achieve higher prediction accuracy and faster processing speed. The big data method for forecasting requires as many types of data as possible. The previously impossible forecast task is completed by exploring the data association relationship, ensuring high accuracy. During power information forecasting, the focus of work is shifted from researching complex algorithms to preparing big data acquisition. On this basis, the existing methods of electricity demand prediction are compared, and their characteristics are analyzed, which is also reported in the relevant literature [33]. Here, the data mining analysis is performed on the constructed data set, and the KMC algorithm is employed to establish a power customer segmentation model. After setting the specific hyperparameters of the model reported here, the initial data on the smart city is processed deeply. Finally, the validity of the customer segmentation model is verified, and the errors are quantitatively analyzed. The MRE is 3.2671 %, indicating that the model system is stable. Through cleaning, conversion, and establishment of new data warehouses, its intrinsic data characteristics are pointed out, and the intrinsic value of data is tapped [34]. Through the experiment, the model's accuracy reaches 85.25%, with the biggest mean of 83.72%. The MRE remains below 5%. Apparently, unlike traditional methods, the proposed customer segmentation model apparently depends on the smart cities' big data. It can describe customer behavior more comprehensively and accurately. Furthermore, it provides a reference for power companies to formulate appropriate marketing strategies and other in-depth research.

6. Conclusion

A smart grid customer segmentation model is constructed based on the KMC algorithm for electricity demand prediction by establishing a database and mining related data through big data analysis. The intrinsic value of data is deeply explored through data cleaning, data conversion, and establishing new data warehouses. Compared with the traditional customer segmentation method, the proposed model can describe customer behavior more comprehensively and accurately, relying on the Smart City Basic Database. In addition, the results of power consumer segmentation are incorporated into the input samples. The BPNN method is utilized for electricity demand prediction on the electricity system to establish a BPNN model. Finally, the input data is normalized. According to the simulation result, the accuracy rate of the constructed model algorithm reaches 85.25%, and the error value is also below 5%, which verifies the model's validity. This model can provide a reference for future green use of power energy in smart cities. The subdivision of the power consumers' user groups can help understand the demand of the power side and then predict the power generation. Such a division is of great value for the planned power generation of the Smart Grid and for improving the user service experience and satisfaction.

Although a suitable power demand forecasting model has been established in this study, there are some deficiencies. Firstly, there is no mature multi-index forecasting model for the relationship between time, social factors, and electricity demand. In addition, the constructed power demand forecasting model is still in the research stage based on actual cases, which lacks general applicability. Secondly, it is necessary to model the impact of annual weather conditions on power load. In power load prediction, electricity consumption has been accurately analyzed according to weather changes. However, the weather changes are macroscopic within a year, and it is difficult to measure them accurately in local areas. Therefore, in follow-up studies, it is worth striving to generate accurate digital reports on economic activities, people's lives, trade, and transport in specific regions. In addition, with the deepening development of the power market, the electricity demand prediction will show different characteristics.

Building suitable and precise prediction models for different business needs is also essential.

References

1. Rout, S. S., *et al.*, "Smart water solution for monitoring of water usage based on weather condition," International journal, Vol. 8, No. 9, 1-9. (2020)
2. Alhazmi, M., *et al.*, "Optimal integration of interconnected water and electricity networks," energy, 2021,4: 6. Vol. 4, 6. (2021)
3. Zhang K., *et al.*, "Security and privacy in smart city applications: Challenges and solutions," IEEE Communications Magazine, Vol. 55, No. 1, 122-129 (2017)
4. Anthopoulos, L., "Smart utopia VS smart reality: Learning by experience from 10 smart city cases," Cities, Vol. 63, No. 2, 128-148. (2017)
5. Souri, A. and Hosseini, R. "A state-of-the-art survey of malware detection approaches using data mining techniques," Human-centric Computing and Information Sciences, Vol. 8, No. 1, 1-22. (2018)
6. Mengelkamp, E., *et al.*, "A blockchain-based smart grid: towards sustainable local energy markets," Computer Science-Research and Development, Vol. 33, No. 1-2, 207-214. (2018)
7. Ahmad, T. and Chen, H. "Potential of three variant machine-learning models for forecasting district level medium-term and long-term energy demand in smart grid environment," Energy, Vol. 160, 1008-1020. (2018)
8. Mirjat, N. H., *et al.*, "Long-term electricity demand forecast and supply side scenarios for Pakistan (2015–2050): A LEAP model application for policy analysis," Energy, Vol. 165, 512-526. (2018)
9. Khalifa, A., Caporin, M. and Di Fonzo, T. "Scenario-based forecast for the electricity demand in Qatar and the role of energy efficiency improvements," Energy Policy. Vol. 17, No. 2, 155-164. (2019)
10. Kim, K. H., *et al.*, "Deep Learning Based Short-Term Electric Load Forecasting Models using One-Hot Encoding," Journal of IKEEE, Vol. 23, No. 3, 852-857. (2019)
11. Nam, K., *et al.*, "A deep learning-based forecasting model for renewable energy scenarios to guide sustainable energy policy: A case study of Korea," Renewable and Sustainable Energy Reviews, Vol. 122, 109725. (2020)
12. Kim, M., *et al.*, "A hybrid neural network model for power demand forecasting," Energies, Vol. 12, No. 5, 931. (2019)
13. Ghalehkhondabi, I., *et al.*, "An overview of energy demand forecasting methods published in 2005–2015," Energy Systems, Vol. 8, No. 2, 411-447. (2017)
14. Boroojeni, K. G., *et al.*, "A novel multi-time-scale modeling for electric power demand forecasting: From short-term to the medium-term horizon," Electric Power Systems Research, Vol. 142, 58-73. (2017)
15. Azaza, M. and Wallin, F. "Smart meter data clustering using consumption indicators: responsibility factor and consumption variability," Energy Procedia, Vol. 142, 2236-2242. (2017)
16. Bai, L. *et al.*, "Fast density clustering strategies based on the k-means algorithm," Pattern Recognition, Vol. 71, 375-386. (2017)
17. Zhang, G., Zhang, C., and Zhang, H. "Improved K-means algorithm based on density Canopy," Knowledge-based systems, Vol. 145, 289-297. (2018)
18. Whittington, J. C and Bogacz, R. "Theories of error back-propagation in the brain," Trends in cognitive sciences. 2019;23(3):235-250. Vol. 23, No. 3, 235-250. (2019)
19. Neftci, E. O. *et al.*, "Event-driven random back-propagation: Enabling neuromorphic deep learning machines," Frontiers in neuroscience, Vol. 11, 324-331. (2017)

20. Scellier, B. and Bengio, Y. "Equilibrium propagation: Bridging the gap between energy-based models and backpropagation," Frontiers in computational neuroscience. 2017;11:24-31. Vol. 11, 24-31. (2017)
21. Kim, H. W. and Jeong, Y. S. "Secure authentication-management human-centric scheme for trusting personal resource information on mobile cloud computing with blockchain," Human-centric Computing and Information Sciences, Vol. 8, No. 1, 11-23. (2018)
22. Li, F. *et al.*, "A clustering network-based approach to service composition in cloud manufacturing," International Journal of Computer Integrated Manufacturing, Vol. 30, No. 12, 1331-1342. (2017)
23. Della Peruta, M. "Adoption of mobile money and financial inclusion: a macroeconomic approach through cluster analysis," Economics of Innovation and New Technology, Vol. 27, No. 2, 154-173. (2018)
24. Olsson, T., Barcellos, L. F. and Alfredsson, L. "Interactions between genetic, lifestyle and environmental risk factors for multiple sclerosis," Nature Reviews Neurology, Vol. 13, No. 1, 25-33. (2017)
25. Gupta, V. and Pal, S., "An overview of different types of load forecasting methods and the factors affecting the load forecasting," International Journal for Research in Applied Science & Engineering Technology (IJRASET), Vol. 5, No. 4, 729-733. (2017)
26. Amara, F. *et al.*, "Household electricity demand forecasting using adaptive conditional density estimation," Energy and Buildings, Vol. 156, 271-280. (2017)
27. Pořízka, P. *et al.*, "Impact of laser-induced breakdown spectroscopy data normalization on multivariate classification accuracy," Journal of Analytical Atomic Spectrometry, Vol. 32, No. 2, 277-288. (2017)
28. Satapathy, S. C. *et al.*, "Multi-level image thresholding using Otsu and chaotic bat algorithm," Neural Computing and Applications. Vol. 29, No. 12, 1285-1307. (2018)
29. Wang, D. *et al*., "Multi-step ahead electricity price forecasting using a hybrid model based on two-layer decomposition technique and BP neural network optimized by firefly algorithm," Applied Energy, Vol. 190, 390-407. (2017)
30. Casteleiro Roca, J. L. *et al.,* "Short-term energy demand forecast in hotels using hybrid intelligent modeling," Sensors, Vol. 19, No. 11, 2485-2496. (2019)
31. Dinesh, C. *et al.*, "Residential power forecasting using load identification and graph spectral clustering," IEEE Transactions on Circuits and Systems II: Express Briefs, Vol. 66, No. 11, 1900-1904. (2019)
32. Johansson, C. *et al.,* "Operational demand forecasting in district heating systems using ensembles of online machine learning algorithms," Energy Procedia, Vol. 116, 208-216. (2017)
33. Al Ogaili, A. S. *et al*., "Review on scheduling, clustering, and forecasting strategies for controlling electric vehicle charging: challenges and recommendations." Ieee Access, Vol. 7, 128353-128371. (2019)
34. Kazemzadeh, M. R. *et al*., "A hybrid data mining driven algorithm for long term electric peak load and energy demand forecasting," Energy, Vol. 204, 117948-117956. (2020)

Shurui Wang was born in Tangshan, Hebei, P.R. China, in 1999. He is studying for an undergraduate degree at Huazhong University of Science and Technology, P.R. China. His research interest include electric power system.

Song Aifeng, Ph.D., lecturer, North China University of Water Resources and Electric Power, research direction: decision-making theory and method, energy efficiency, sustainable supply chain.

Yufeng Qian was born in Wuhan, Hubei, China, in 1986. He received the B.S. degree in information and computing science from Central China Normal University, Wuhan, China, the M.S. degree and the Ph.D. degree in computational mathematics from Wuhan University, Wuhan, China, in 2008, 2010 and 2013, respectively. Currently, he works as a lecturer in the School of Science, Hubei University of Technology, Wuhan, China. His research interests include applied mathematics, complex systems and complex networks.

Received: August 07, 2022; Accepted: December 02, 2022.