Short-term Power Load Forecasting Based on Balanced KNN

Xianlong Lv¹,Xingong Cheng¹,YanShuang²,TANG Yan-mei³

¹School of Electrical Engineering, University of Jinan, Jinan, China
²Zaozhuang power supply company of Shandong province, Zaozhuang, China
³China Electric Power Research Institute, Beijing, China

*Corresponding author e-mail: xianlonglv@163.com

Abstract. To improve the accuracy of load forecasting, a short-term load forecasting model based on balanced KNN algorithm is proposed; According to the load characteristics, the historical data of massive power load are divided into scenes by the K-means algorithm; In view of unbalanced load scenes, the balanced KNN algorithm is proposed to classify the scene accurately; The local weighted linear regression algorithm is used to fitting and predict the load; Adopting the Apache Hadoop programming framework of cloud computing, the proposed algorithm model is parallelized and improved to enhance its ability of dealing with massive and high-dimension data. The analysis of the household electricity consumption data for a residential district is done by 23-nodes cloud computing cluster, and experimental results show that the load forecasting accuracy and execution time by the proposed model are the better than those of traditional forecasting algorithm.

1. Introduction

The Energy Internet[1-4] is, in order to solve the fossil fuel depletion and environmental pollution problems, a new energy system for the characteristics of the deeply combined with a new energy technology and information technology in the background of the third industrial revolution. Compared with the smart grid, the energy Internet is more dependent on modern Internet technology[5].

The load forecasting[6] has always been a key link in the operation and control of power system, which affects the successful implementation of many analysis and decision-making of power system, such as economic dispatch, automatic power generation control, safety assessment, maintenance plan and operation of the electric power market. In terms of time, the load forecasting can be roughly divided into the medium and long-term load forecasting and short-term load forecasting. The latter predicts the load changes in the next few hours, one day up to several days, and mainly provides the data support[7] for the operation and control of the power system, which is the focus of this paper.

In the methodology point of view, short-term load forecasting methods can be broadly divided into time series forecasting method, intelligent forecasting method and combination forecasting method. The time series method constructs the forecasting model by using the association between successive time series elements, and predicts the future trend of load by extrapolation. The specific methods include regression analysis [8], load derivation method [9], exponential smoothing method [10], Kalman filter method [11] and so on. The intelligent prediction method is used to construct the learning structure which adapts to the highly nonlinear association relation. Through the supervised learning of input and output historical data, the nonlinear mapping relationship of input and output is obtained, and the future load changes are predicted. The specific methods include expert System
method [12], artificial neural network learning method [13], support vector machine learning method [14,15] and so on. The combination forecasting method focuses on the combination of two or more methods. It is difficult to find a kind of universal best forecasting method for different forecasting objects. The results of all kinds of prediction methods have certain credibility. In this way, the combined forecasting method improves the accuracy of load forecasting by integrating a variety of prediction models and methods [15]. The above method constitutes a solid theoretical basis for load forecasting, but its forecasting process does not take into account the application background of the massive data faced by the current power grid, and its applicability to the load forecasting problem in large data environment needs further study.

Cloud computing [16-17] is a set of servers located on the network to its computing, storage, data and other resources in the form of services provided to the requester to complete the information processing tasks and processes, cloud computing represents a large scale distributed computing model based on the Internet. Cloud computing first use the Internet to integrate a variety of wide-area heterogeneous computing resources to form an abstract, virtual, dynamically expand the computing resource pool, and then through the Internet to provide users with on-demand computing power, storage capacity, software platforms and applications. Software and other services. The paper[18] proposed the establishment of a computing platform of power system based on cloud computing, the realization of power system cloud computing platform is discussed in detail from the aspects of physical composition, system architecture and software technology; This paper[19] analyzes the sources and characteristics of large data in power generation, power transmission and power transmission, and analyzes in detail the advantages of large data processing technology in smart grid construction and large data processing. This paper[20] propose a platform based on local weighted linear regression and cloud computing, and established a parallel localized weighted linear regression model to reduce the load forecasting time and improve the prediction accuracy. The mean square error is 3.01%. The above-mentioned paper introduces the idea of parallel computing on the basis of traditional algorithms, and significantly improves the ability of the prediction algorithm to deal with large data of power load.

In the above background, this paper presents a parallel prediction model based on balanced KNN algorithm, which is applied to short-term power load forecasting. In this paper, the K-means clustering algorithm is used to segment the historical data of massive power load, so that the training based on the local weighted linear regression algorithm load forecasting model is more specific and the prediction result is more accurate. The balanced KNN algorithm is proposed to improve the traditional KNN algorithm, Load scene data to achieve equalization, improve the accuracy of the load scene identification. The algorithm is programmed in parallel using the Apark Hadoop framework of cloud computing to solve the problem of time-consuming training under mass data, which greatly shortens the model training and forecasting time based on massive data load forecasting.

2. Balanced KNN algorithm

2.1. Traditional KNN algorithm
KNN is a nonparametric classification algorithm that calculates the classification of the sample by finding the nearest category of the test sample. KNN classification steps can be expressed as: Given a sample data set T, if the sample to be tested has k nearest neighbor representations, and most of the k representative samples belong to a predetermined type, then the sample to be tested will also be attributed to this class.

KNN algorithm generally uses Euclidean distance to represent the distance between test samples and known class samples d:

\[ d(T_i, d_j) = \sqrt{\sum_{t=1}^{n} (\omega_{it} - \omega_{jt})^2} \]  

(1)

In the formula, \( M_t \) represents training samples, \( d_j \) represents test samples, \( n \) represents the number of samples, \( \omega_{it} \) and \( \omega_{jt} \) respectively represent the values of the tth feature in sample vectors \( T_i \) and \( d_j \).
The traditional KNN algorithm is simple and does not need to estimate the parameters. However, the computational complexity of the algorithm in the background of the massive set is amazing. The performance of the traditional KNN algorithm is susceptible to the training sample. When the sample set is unbalanced, the classification of the performance can easily be affected, or even become extremely poor.

2.2. Balanced KNN algorithm

The balanced KNN algorithm proposed in this paper greatly improves the accuracy of the original KNN algorithm in dealing with unbalanced data. The specific idea of the algorithm is shown in the following figure:

![Flow chart of balanced KNN algorithm](image)

The specific steps of the improved K- nearest neighbor method are as follows:

1. Input the training data set and the test data set, and standardize the sample data.

   \[ \hat{X} = \frac{X - \min(X)}{\max(X) - \min(X)} \]  

2. Determine representative-sample. Calculate the Euclidean distance between each sample data and the rest sample data in the class, and select the minimum value of the distances \( d_i \), select the minimum distance of the sample for this class as the representative-sample \( P_i \).

   \[ d_i = \min \sum_{j=1, j \neq i}^n d(x_i, x_j) \]  

   In the formula, \( x_i \) and \( x_\tilde{i} \) are respectively the \( \tilde{i} \)th and the \( i \)th samples in the class.

3. Calculate the threshold for each class. Calculate the Euclidean distance between each sample and the representative-sample in each class, and select the maximum value as the threshold of the class \( V_i \).

   \[ V_i = \max \sum_{j=1, j \neq i}^n d(x_i, x_j) \]  

   In the formula, \( x_i \) is the representative-sample of the \( i \)th class, and \( x_i \) is the \( \tilde{i} \)th sample in the \( i \)th class, \( V_i \) is the threshold of the \( i \)th class.

After determining the threshold for each class, calculate the distance between the test sample and the sample in each class, and exclude the distance and the class that is greater than the threshold, so as to improve the accuracy and speed of the classification.
(4) Calculate the sample mark and the class mark. The Euclidean distance of the sample and the sample in the class is calculated, and the distance between the sample and the sample in the sample is calculated and the ratio is calculated. That is the the sample mark \( R_{ci} \).

\[
R_{ci} = \frac{\sum_{i=1,j
ot= i}^n d(M_i - x_i)}{(n-1)d(M_i - x_i)}
\]  

(5) In the formula, \( M_i \) is the representative-sample, \( x_t \) and \( x_i \) are the \( t \)th and \( i \)th samples respectively.

(5) Calculate the distance including the sample mark. Calculate the Euclidean distance \( d(x_i, x_j) \) between each sample \( x_j \) in the test sample and the candidate class, and find the \( k \) nearest neighbor samples with the smallest distance.

\[
d(x_i, x_j) = R_{ci} \times \sqrt{\sum_{j=1}^l (\omega_{ij} - \omega_{cij})^2}
\]

(6) In the formula, \( R_{ci} \) is the sample mark of the \( i \)th sample of the \( c \)th class, \( \omega_{ij} \) and \( \omega_{cij} \) are the \( t \)th value the sample vectors \( x_i \) and \( x_j \) respectively.

(6) Calculate the weights of the selected class of the selected \( k \)-nearest neighbor samples; the weight that \( x_j \) belongs to \( c_t \) of the \( t \)th class:

\[
W_{jt} = \sum_{x_i \in X_t} v(x_i, x_j) \varphi(x_i, c_t)
\]

(7) In the formula, \( v(x_i, x_j) \) is the weight of the vote. In this paper \( d(x_i, x_j) \)

\[
\varphi(x_i, c_t) = \begin{cases} 1, & x_i \in c_t \\ 0, & x_i \not\in c_t \end{cases}
\]

(8) \( x_t \) is classified as the class with the largest weight;

3. **Design of the load forecasting algorithm based on Cloud Computing**

Combined with cloud computing technology, the combination forecasting method based on balanced KNN and the Hadoop model framework are combined to realize the short-term parallel prediction of the load. The system framework is shown below:

![Fig. 3.1 design of load forecasting system](image)

The combination forecasting method is divided into three steps:

(1) Divide load scenes. The K-means algorithm is used to cluster the massive data, and the clustering result is the load scene.

(2) Determine load scenes. The balanced KNN algorithm is used to analyze the unbalanced load scene in step 1 and make the scene decision.
(3) Fitting load scenes and forecasting load. The local weighted linear regression algorithm is used to fitting load scenes, and forecasting load.

3.1. Divide load scenes
In the load forecasting, the main reason for the analysis of the historical load curve is that the trend and direction of the load curve are inextricably linked with the factors such as climate and date type. If there is no targeted study of the user load under mass data, it will result in a very high resource loss for each user's electricity consumption. So to use a reasonable data mining technology, the law of electricity and similar to the load type is divided into the same scene. In this paper, the K-means clustering method is used to divide the load scene, and the time scale of the algorithm is improved by using the computing architecture of the Apache Hadoop.

![Figure 3.2 K-means algorithm clustering flow chart](image)

The concrete steps are as follows:
1. Read the power load history data in HDFS, and take the size of 64M for block processing. Select K cluster centers, and the K is the number of clustering;
2. Send the current cluster center data to all nodes, calculate the nearest cluster centers and corresponding distances of the data in the nodes, and add the class labels to the training data:
\[
c = \arg \min ||x_i - u_j||
\]

(9)
3. Calculate the sum of each class of each cluster center and the mean of each cluster center, then update the clustering center according to the mean:
\[
u_j = \frac{1}{|S_j|} \sum_{x_i \in S_j} x_i
\]

(10)
4. Determine whether the outcome is convergence to reach the convergence conditions. If not, repeat steps 2 and 3. Otherwise, end the iteration and output the result.

3.2. Determine Load scenes
In this paper, the K-mean algorithm is used to classify the nearest principle without considering the unbalance of the load scene data, which leads to inaccurate decision of the final load scene. In this paper, the balanced KNN algorithm is used to analyze the load scene and make the load scene decision.
Figure 3.3 computing structure diagram for the balanced KNN algorithm
The K nearest values of the distance between the load scene and the test data are calculated. The distance is the iterative distance between the test data and the entire load scene dataset. Since the whole calculation process is independent, it is possible to use the distributed computing model Apache Hadoop to realize the parallelization of the whole calculation process, thus shortening the execution time of the balanced KNN algorithm classification.

3.3. Load scene training and load forecasting
In the above steps, the load scene of the test data has been decided. Based on this, the fitting and prediction of the load have been carried out by using the local weighted linear regression algorithm.

The local weighted linear regression model fitting the polynomial regression curve based on the local data, observing the law and trend of the data appearing locally. KNN algorithm is used to determine the nearest data points around the prediction point, and the commonly used method to determine the local data points.

Figure 3.4 The local weighted linear regression algorithm flow chart
The specific implementation process is as follows:
(1) Read a sample data in HDFS, and take the size of 64M for block processing;
(2) Submit job to the cluster: Calculate all nearby point weight according to the "Gauss kernel". Set the weight based on Gauss function;
(3) Calculate the predicted value of samples and the above weights. Update the configuration file;
(4) Repeat steps 1~3 until that all the samples are computed.

4. Example analysis

4.1. Example data
The data source of this paper is the data collected from a district by metering equipment. The district covers an area of 4.2 thousand square meters, with a total population of about 5000 people. The types of data collected include electricity data, temperature, humidity, precipitation, wind speed, day type and season type. Load data were trained and predicted using 15 minutes intervals. The example
analysis data set uses the electricity data of March 2013 - February 2013 to predict the power load in April 2014.

4.2. Load forecasting error evaluation index
In this paper, a lot of reasons of prediction error prediction algorithm, summed up mainly include:
(1) the power of historical load data is not complete;
(2) the power load identification of the scene division is not enough;
The evaluation indicators adopted in this paper are as follows:

\[ e_i = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} |y(i) - \hat{y}(i)| \times 100\% \quad (11) \]

In the formula, \( e_i \) is the root mean square error, \( y(i) \) and \( \hat{y}(i) \) are the actual load value and the predicted load value respectively.

4.3. Calculation results of an example
According to the unbalanced analysis of the data after clustering, the decision of the load scene is carried out and the load value is determined after the equilibrium data is obtained.

(1) comparing the parallel algorithm with the traditional algorithm
In this paper, the load forecasting algorithm based on cloud computing is compared with the actual load value, as shown in table 1.

| number | 1  | 2  | 3  | 96 |
|--------|----|----|----|----|
| predicted value | 983.26 | 1035.15 | 1103.32 | 1306.28 |
| actual value | 968.63 | 1071.43 | 1141.30 | 1285.75 |
| Error rate/% | 1.5 | -3.4 | 3.3 | 1.6 |

As shown in the following figure, the predicted value curve is similar to the curve trend of the actual value, and the mean square error is 2.56%. The error of the prediction results conforms to the error standard of load forecasting. It is proved that the load forecasting algorithm based on cloud computing is feasible.

![Fig. 4.1](image)

Fig. 4.1 comparison chart of forecast results and mean square error of predicted result
(2) comparison of running time between parallel algorithm and traditional algorithm
It is shown from the results can be seen in the small sample data, predicting the time between the two are similar, on the contrary, the single is the time required by the algorithm is slightly better than that of parallel algorithm, parallel algorithm in reason: a small number of samples will be divided into several sub sample data sets between different subsets of the data communication cost. But the impact of the prediction speed; but with the increase of sample set, the iteration time prediction algorithm is required to have a significantly different, the time required for the parallel algorithm is much less than the single method.

Fig. 4.2 comparison of time between traditional algorithm and parallel algorithm

5. Conclusion
It shows that the predicted value curve is similar to the curve of the actual value, and the mean square error is 2.56% by calculating the electricity data of a residential district, and the time needed for the prediction is less than 6 minutes. The analysis results show that the prediction speed and accuracy are better than the traditional prediction algorithm, and the rationality and validity of the short-term load forecasting method are verified.

References
[1] HUANG Renle,PU Tianjiao,LIU Kewen,et al.Design of Hierarchy and Functions of Regional Energy Internet and Its Demonstration Applications[J].Automation of Electric Power Systems,2013,39(9):26-40.
[2] DONG Zhaoyang, ZHAO Junhua,WEN Fushuan, et al.From smart grid to energy internet:basic concept and research framework[J].Automation of Electric Power System,2014,38(15):1-11.
[3] YU Shenhang, SUN Ying,NIU Xiaona, et al.Energy internet system based on distributed renewable energy generation[J].Electric Power Automation Equipment,2010,30(5):104-108.
[4] PU Tianjiao, LIU kewen,CHEN Naishi, et al. Design of ADN Based Urban Energy Internet Architecture and Its Technological Issus[J].Proceedings of the CSEE,2015,35(14):3511-3521.
[5] SUN Hongbi, GUO Qinglai, et al.Energy Internet:Concept, Architecture and Frontier Outlook[J].Autamation of Electric Power Systems,2015,39(19):1-8.
[6] Kang Chongqing, Xia Qing, et al.REVIEW OF POWER SYSTEM LOAD FORECASTING AND ITS DEVELOPMENT.Autamation of Electric Power Systems,2004,28(17):1-11.
[7] LIAO Ni-huan, HU Zhi-hong,et al. Review of the short-term load forecasting methods of electric power system.Power System Protection and Control,2011,39(1):147-152.
[8] TANG Jun-jie,NIU Huan-na, et al.Periodic autoregressive short-term forecasting method based on the linear correlation analysis,2010,38(14):129-133.
[9] ZHANG Zhen-gao, YANG Zheng-ling, et al.Load Derivation in Short Term Forecasting Using Weather Factor[J].2006,18(5):79-83.
[10] NGO Vietcuong, WU Wenchuan,et al.Ultra-Short Term Load Forecasting Using Robust Holt-Winter in Distribution Network[J].2014,38(10):2810-2815.
[11] NAN Li,HAN Xueshan,et al.Method for Ultra-short Term Multi-node Load Forecasting[J].2007,31(21):30-34.
[12] Zhang Tao,ZHAO Dengfu,et al.Short-Term Load Forecasting Using Radial Basis Function Networks and Expert System [J].2001,35(4):331-334.
[13] LEI Shao-lan, SUN Cai-xin, et al. SHORT-TERM LOAD FORECASTING METHOD BASED ON RBF NEURAL NETWORK AND ANFIS SYSTEM [J], 2005, 25(22): 78-82.

[14] LI Xiao, WANG Xin, et al. Short-term wind load forecasting based on improved LSSVM and error forecasting correction [J]. Power System Protection and Control, 2015, 43(11): 63-69.

[15] ZENG Ming, LV Chunquan et al. Least Squares-support Vector Machine Load Forecasting Approach Optimized by Bacterial Colony Chemotaxis Method [J], 2011, 34(31): 93-99.

[16] ZENG Ming, LV Chunquan et al. Short-term Electric Load Forecasting with Combined Data Mining Algorithm [J], 2006, 30(14): 82-86.

[17] LUO Jun-zhou, JIN Jia-hui, et al. Cloud computing: architecture and key technologies [J]. Automation of Electric Power Systems, 2010, 34(15): 3-21.

[18] ZHAN Junhai, WEN Funshuan, XUE Yusheng, et al. Cloud Computing: Implementing an Essential Computing Platform for Future Power Systems [J]. Automation of Electric Power Systems, 2010, 34(15): 1-8.

[19] Song Yaqi, ZHOU Guoliang, ZHU Yongli. Present status and challenges of big data processing in smart grid [J]. Power System Technology, 2012, 37(1): 927-935.

[20] ZHANG Suxiang, ZHAO Bingzhen, et al. Short-term Power Load Forecasting Based on Big Data [J]. Proceeding of the CSEE, 2015, 35(1): 37-42.