International Core Journal of Engineering
ISSN: 2414-1895
Volume 7 Issue 10, 2021
DOI: 10.6919/ICJE.202110_7(10).0040

Short-term Power Load Forecasting based on Big Data

Jian Di[1], Boyu Qi[1]

[1]School of Control & Computer Engineering, North China Electric Power University, Baoding Heibei 071003, China.

Abstract

The short-term power load forecasting method had been researched based on the big data. And combined the local weighted linear regression and cloud computing platform, the parallel local weighted linear regression model was established. In order to eliminate the bad data, bad data classification model was built based on the maximum entropy algorithm to ensure the effectiveness of the historical data. The experimental data come from a smart industry park of Gansu province. Experimental results show that the proposed parallel local weighted linear regression model for short-term power load forecasting is feasible; and the average root mean square error is 3.01% and fully suitable for the requirements of load forecasting, moreover, it can greatly reduce compute time of load forecasting, and improve the prediction accuracy.

Keywords

Big Data; Cloud Computing; Load Forecasting; Local Weighted Linear Regression.

1. Introduction

With the emergence of massive data of intelligent electricity consumption, it is necessary to find a new method to meet the requirements of big data analysis of massive electricity consumption. The existing prediction algorithms cannot meet the requirements of prediction speed and accuracy. The traditional local weighted linear regression prediction has the advantages of fast training speed and small prediction error rate when used for small data prediction. However, when the data volume is very large, because the algorithm needs to find the nearest neighbor for each test point, the computation is very large, and the single computing time will reach several hours or days. Therefore, it is very important to solve the prediction problem based on massive data.

Based on the massive data of intelligent industrial parks, this paper combines the local weighted linear regression forecasting algorithm and the cloud computing Mapreduce model to carry out short-term power load forecasting methods. This method first divides the massive data into multiple data sub-blocks, then analyzes and processes the data of each sub-block at the same time through the cloud platform, and finally merges the results. This processing process reduces the time processing overhead of the massive data. At the same time, this article also processed the enumerated data and added it to the distance calculation to improve the accuracy of the prediction.

2. Organization of the Text

2.1 Partially weighted linear regression model based on cloud computing

Locally weighted linear regression model based on partial data fitting of polynomial regression curve, the rule and trend of observation data in the local display. To determine the nearest data point around the prediction point, the common method to determine the local data point is the KNN algorithm [16-17], whose main idea is to calculate the distance between the prediction point and all data points in the feature space, and find the set of K points closest to the prediction point.

Let any one instance be described by X = {s1, s2, …, sn}, and the distance between the two instances X1 and X2 can be obtained by equation (1):

ISSN: 2414-1895 DOI: 10.6919/ICJE.202110_7(10).0040

$$d(X_1, X_2) = \sqrt{\sum_{i=1}^{n}(s_i - s_j)^2} \quad (1)$$

Establish regression formula:

$$f(x) = \omega_0 + \omega_1 a_1(x) + \omega_2 a_2(x) + \cdots + \omega_n a_n(x) \quad (2)$$

In the formula, w_i represents the weight calculated according to the distance formula (1), and its calculation formula is:

$$\omega_i = \frac{1}{d(x_q, x_i)^2} \quad (3)$$

Where, x_q is the prediction point;x_i is the point near x_q; The reciprocal of the distance between the two is the weight.

In Formula (2), ω_0 is the regression constant term, $\omega_1, \omega_2, \ldots, \omega_n$ are the regression coefficient, and f(x) is the regressor predictive value.α_i (x) represents the ith attribute value of instance x. When fitting the linear function of the above form to a given training set, the gradient descent method is usually adopted to find the coefficients $\omega_1, \omega_2, \ldots, \omega_n$ that minimize the error, namely:

$$E(x) = \frac{1}{2}\sum_{x=i}\left(f(x) - f(i)\right)^2 \quad (4)$$

By satisfying the error criterion and local approximation, the gradient descent training rule can be obtained:

$$\Delta_{\omega_j} = \mu \sum_{x=i} K\left(d\left(x_q, x\right)\right)\left(f(x) - f(i)\right)a_j(x) \quad (5)$$

Where, μ is the learning rate.

It can be seen from the description in Section 1 that the traditional local weighted linear regression algorithm has serious defects, that is, when the number of regression data to be increased, the computation amount generated by determining the set of nearest neighbor data points from massive data is very large. In this paper, combined with cloud computing technology, LWLR algorithm and MapReduce model framework are combined to realize concurrent power load prediction.

MapReduce is a parallel programming model and computing framework for processing massive amounts of data. It uses a "divide and conquer" approach. Therefore, the parallel locally weighted linear regression model in this paper includes three stages: Map stage, merge stage and Reduce stage, and the data in each stage will be exchanged in the way of < key, value >.

2.1.1 Map phase.

Firstly, the input data set is divided into several data subsets, which are represented by. Key is the relative offset of the current data, and value is resolved into the coordinate values of each dimension of the current data. The distance between the test point and the nearest k center points of the data subset is calculated based on the local minimum distance algorithm, and the intermediate results of the calculation will be put into the intermediate library.

2.1.2 Merge phase.

The task in this phase is to merge the processed data at the local level.Reordering the intermediate keys to the set produces a new tuple where the same keys are grouped together.

2.1.3 Reduce phase.

The reduce function first analyzes the number of samples and the accumulated coordinate value of each dimension of the corresponding node, calculates the k points closest to the prediction point in each data subset, and calculates the weighted value of each attribute based on the Gaussian mixture model. The result will be updated to the distributed file system and the next iteration will be carried out until the algorithm converges.

Due to the existence of human factors or some special reasons, usually the abnormal data collected by sampling will affect the accuracy and reliability of the prediction results.In this paper, the sample historical data is preprocessed and the bad data classification model is established based on the maximum entropy algorithm.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202110_7(10).0040

The principle of maximum entropy was proposed by E.T. Jaynes in 1950. Its main idea is that when predicting unknown hypotheses with limited knowledge, the probability distribution with the maximum entropy should be selected in line with known assumptions. That is, the most reasonable inference about the unknown distribution is the most uncertain or the most random inference that conforms to the known knowledge under the premise of known partial knowledge.

In the maximum entropy model, information is expressed in the form of features, where the features are binary features fi(x,y). If fi is useful for the model, a constraint model that can generate training sample p(x,y) model is constructed.

The data in this paper are time series data, so the normalization process is carried out first, and then it is fed into the maximum entropy model for iteration. All elements are normalized according to formula (6).

$$r_i = \frac{x_i - min(x_i)}{max(x_i) - min(x_i)} \tag{6}$$

In LWLR algorithm, the number of maps is first solved. The data source and its data structure, parallelism, increment field, exception handling mode and other parameters are read, and the data set is divided and adjusted according to the maximum value of the increment field to determine the number of maps. Secondly, using KNN, K points closest to the prediction point are selected for the data blocks processed by each Map.Finally, the distance between K points of each Map and the prediction points was compared to select the smallest K points, and the weight was calculated based on the Gaussian mixture model, and the parameters were determined to complete the task of model building.

2.2 Load forecasting experiment and result analysis

Since electric load forecasting is an estimation of future electric load based on historical data, there is a gap between the predicted value and the actual value, resulting in electric load forecasting errors. There are many reasons for the error, which can be summarized as: 1) the simplification of the mathematical model and the ignorance of various factors; 2) the historical data is not complete; 3) the error caused by improper parameter selection.

The evaluation indexes adopted in this paper are as follows:

Let y(i) and p(i) represent the actual load and predicted value at time i, respectively, then we have:

Absolute error:

$$E = y(i) - p(i) \tag{7}$$

The relative error:

$$e = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y(i)-p(i)}{y(i)}\right| \times 100\% \tag{8}$$

The data source of this paper is the load data and weather data collected by a power grid enterprise in Gansu Province. The training data range is the electricity consumption data from November 24, 2011 to November 30, 2011. The sampling interval of each device is 15min, as shown in Table 1. The influence of load influencing factors such as temperature, humidity, working day, holiday and season on load fluctuation of power users is considered, and the correlation strength with load is calculated to provide a basis for establishing a more accurate load forecasting model.

Table 1. Training data

The date	hours	minutes	The highest temperature	Minimum temperature	electricity
24	00	15	7	-3	52679.1842
24	00	30	7	-3	58495.6730
24	00	45	7	-3	57386.6410
…	…	…	…	…	…
30	23	30	3	-5	84704.7841
30	23	45	3	-5	72975.8408

 Volume 7 Issue 10, 2021
DOI: 10.6919/ICJE.202110_7(10).0040

When small data samples, the prediction time difference between the two is not big, on the contrary, the traditional linear regression method the time required to slightly better than parallel local linear weighting algorithm, the reason: parallel local linear weighted algorithm will data under small sample set is divided into several sub sample set, the communication between different data subset increased cost instead of impact prediction speed; However, as the sample set increases, the iteration time required by the prediction algorithm is significantly different, and the time required by the parallel weighted linear regression algorithm is much less than that of the traditional method.

In this paper, the comparison results between the predicted load value and the actual load value obtained based on the parallel local weighted linear regression algorithm are shown in Table 2.

Table 2. Comparison results

The serial number	Predictive value	The actual value	error
1	50976.5818	53100.6061	−4.0
2	54386.0209	53105.9723	2.4
3	56034.5931	54000.8643	3.8
4	52065.2806	53400.2878	-2.5
…	…	…	…
96	73517.7189	71100.3085	3.4

It proves that the partial weighted linear regression method based on cloud computing is feasible. The system software has been running normally in a smart park, and it has played a very important role in managing the power load of the park for the power company.

3. Literature References

Power load forecasting important in ensuring power system planning and reliability is of great significance. With the rapid development of China's economy today, it has become an important and arduous task to solve the problem of power load forecasting. High-quality load forecasting requires accurate mathematical models. With the continuous progress of modern technology and the deepening of intelligent electricity consumption [1], load forecasting theory and technology have been greatly developed, and theoretical research has been gradually deepened [2-3]. Over the years, theories and methods of power load forecasting have been constantly emerging. Neural network [4-7], time series [8-9], Bayes [10], fuzzy theory [11], wavelet analysis [12], regression analysis [13-14], support vector machine [15] and other technologies have provided powerful tools for power load forecasting. However, the existing methods still have limitations. Neural network method: One is unable to avoid in the training process of learning deficiency or over-fitting phenomenon; Second, the convergence rate is slow and easy to fall into local minimum. Time series method: it requires high accuracy of historical data, and is not sensitive to weather factors in short-term power load forecasting, so it is difficult to solve the problem of inaccurate short-term load forecasting caused by meteorological factors. Regression analysis method is to quantitatively describe the quantitative relationship between observed variables in the statistical mean meaning, which often has a limit on the amount of data.

4. Conclusion

Aiming at the serious defect of traditional local weighted linear regression algorithm, this paper studies the short-term forecasting problem of power load with massive data. After removing bad data model by maximum entropy and carrying out data preprocessing, a parallel local weighted linear regression algorithm was proposed by combining Mapreduce, which has parallel programming model and computational framework, with the local weighted linear regression algorithm, which solved the problem of massive data computation and greatly shortened the prediction time. At the same time, the prediction accuracy is guaranteed to meet the requirements of load forecasting. Next, the problem of factor correlation in load forecasting is solved by combining multiple models.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202110_7(10).0040

References

[1] The white paper for the development of Chinese electric power big data[M]. Beijing: China Electric Power Press, 2013 (in Chinese).

[2] Zhong Qing, Sun Wen, Y u Nanhua, et al. Load and power forecasting in active distribution network planning [J]. Proceedings of the CSEE, 2014, 34(19):3050-3055(in Chinese)

[3] Xiao Bai, Zhou Chao, Mu Gang. Review and prospect of the spatial load forecasting methods [J]. Proceedings of the CSEE, 2013, 33(25):78-92(in Chinese).

[4] Liang Z S. The short term load forecast of power system based on adaptive neural network[J].Journal of Northeas China Institute of Electric Power Engineering, 1994,14(1):27-35.

[5] Xie Kaigui, Li Chunyan, Zhou Jiaqi. Research of the combination forecasting model for load based on artificial neural network [J]. Proceedings of the CSEE, 2002, 22(7): 85-89(in Chinese).

[6] Liu Ling, Y an Dengjun, Gong Dengcai, et al. New method for short term load forecasting based on particle swarm optimization and fuzzy neural network [J]. Proceedings of the CSU-EPSA, 2006, 18(3): 47-50(in Chinese).

[7] Fu Zhongyun. Application of PSO-BP algorithm in electric power system short-term load forecast [J]. Journal of Chongqing Institute of Technology (Natural Science Edition), 2007, 21(10):93-96(in Chinese).

[8] Hagan M T, Behr S M. The time series approach to short-term load forecasting [J]. IEEE Transactions on Power System, 1987, 2(3):25-30.

[9] Zhao Hongwei,Ren Zhen,Huang Wenying.Short-term load forecasting considering weekly period based on PAR[J]. Proceedings of the CSEE,1997,17(3):211-213(in Chinese).

[10] Tao Wenbin, Zhang Lizi, Pan Hong, et al. Spatial electric load forecasting based on double-level Bayesian classification[J].Proceedings of the CSEE,2007,27(7):13-17(in Chinese).

[11] Bakirtzis A G, Theocharis J B. Short term load forecasting using fuzzy neural networks[J].IEEE Transactions on Power System,1995,10(3):1518-1524.

[12] Y ao Lixiao, Liu Xueqin. A wavelet analysis based combined model for monthly forecasting[J]. Powe System Technology, 2007, 31(19):65-68(in Chinese).

[13] Lei Shaolan, Sun Caixin, Zhou Quan, , et al. The research of local linear model of short-term electrical load on multivariate time series[J]. Proceedings of the CSEE, 2006,26(2):25-29(in Chinese).

[14] Zhang Fusheng,Wang Hong,Han Ti,et al.Short-term load forecasting based on partial least-squares regression [J]. Power System Technology,2003,27(3):36-40(in Chinese).

[15] Niu Dongxiao, Gu Zhihong, Xing Mian, et al. Study on forecasting approach to short-term load of SVM based on data mining [J]. Proceedings of the CSEE,2006,26(18):6-12(in Chinese).

[16] Thomas M C, Peter E H. Nearest neighbor pattern classification [J]. IEEE Transaction Theory, 1967, 13(1):21-27.

[17] Bremner D, Demaine E, Erickson J, et al. Output-sensitive algorithms for computing nearest-neighbor decision boundaries [J]. Discrete and Computational Geometry, 2005, 33(4):593-604.