Household Electricity Load Forecasting Based on Pearson Correlation Coefficient Clustering and Convolutional Neural Network

Minghao Xie, Chengwei Chai, Heng Guo and Minghao Wang

Software College of Northeastern University, Shenyang, 110000, CHN
xiet_email@163.com

Abstract. With the development and construction of country, the rapid growth of electricity consumption has caused the problem of undersupply of electricity. It has become a necessarily daily work to forecast the load of the electricity precisely. A new method, new clustering load forecasting method, is used to forecast residential electricity consumption. Distinguished from other methods with Euclidean distance as their evaluation index, however, the method in this paper is defined with the Pearson correlation coefficient. Furthermore, CNN is used in the experiment about residential load forecast. The result indicates that the new method offers more accurate forecasting data than the traditional methods.

1. Introduction

Electricity is closely related with our lives. Not only does forecast about electricity load enable us to recognize the coming load of power system effectively, but also react methodically. Currently, it is common that using clustering load forecast methods whose main idea is to cluster the residents first and forecast later in electrical system. Its rationality is that the diversity in different resident keeps considerable and the classification benefits the model to forecast. There are many clustering load forecasting methods whose main difference lies in the different clustering ways. For example, the K-means using the traditional Euclidean distance is popular. [1][2] The new K-means method basing on the Pearson correlation coefficient is put forward in the paper to cluster the residential electricity. The reason why we use the Pearson correlation coefficient will be introduced in coming paragraph.

Traditional Euclidean distance just indicates the distance similarity with the same coordinate but it can’t indicate the similarity in different time frame. The two curves are intuitively similar, but there is a time delay between them. As a consequence, Euclidean distance as evaluation index between curves is disabled, thus Pearson correlation coefficient is used whose definition is:

\[ r(X, Y) = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}[X] \text{Var}[Y]}} \]

X and Y are two different curves. Cov (X, Y) is the covariance between X and Y. Var[X] is the variance of X and Var[Y] is the variance of Y.

Correlation coefficient is a quantitative indicator to describe the statistical correlation between two groups of random variables. There are many methods to calculate the correlation coefficient, among which some are more classical like Pearson [3], Spearman [4] and Kendall [5]. These calculation methods have their applicable scope: Pearson correlation coefficient is applicable to binary Gaussian distribution. Spearman and Kendall correlation coefficients are suitable for nonlinear distribution. Although the calculation process is different, the above three correlation coefficients are similar, and
they can be abstracted as generalized correlation coefficient [6], that is, the correlation degree between two sets of data as a whole can be determined by comparing each pair of numbers in the array. Thus, the mismatch of lagging data is considered.

MAPE, MAE are used to evaluate the forecasting result. The mean absolute percentage error (MAPE), mean absolute error (MAE), and mean square error (MSE) of the predictor are regarded as indexes.

The MAPE is defined as follows:

\[
(2) \quad \text{MAPE} = \frac{1}{N} \sum_{n=1}^{N} \left| \frac{L_r - L_p}{L_r} \right| \times 100\%
\]

The MAE is defined as follows:

\[
(3) \quad \text{MAE} = \frac{1}{N} \sum_{n=1}^{N} \left| L_r - L_p \right| \times 100\%
\]

Where \(L_r\) is the real load, \(L_p\) is the forecasting load, and \(n\) is the number of \(L_p\) values.

The residential load data used in this paper are obtained from the Smart Metering Electricity Customer Behaviour Trials (CBTs) initiated by Commission for Energy Regulation (CER) in Ireland took place during 2009 and 2010 with over 5,000 Irish homes and businesses participating. It contains half-hour electricity consumption data. After cleaning the consumers with large number of invalid values, the data of 100 consumers form July 14,2009 to July 14,2010(1 year) are used for forecasting and testing.

We use the multiscale convolutional neural network [11]. MS-CNN is combined with a stack of blocks which has L convolution layers individually. The L convolutional layers are formed by the causal convolution and the empty convolution. Causal convolution ensure that any node only acquires information about the past. In the meanwhile, in order to solve the problems, too many convolution layers and complex training, caused by respective field that convolution needs, empty convolution is used to enlarge the receptive field and enhance the mobility of gradient.

2. Method

2.1. PSK-means algorithm
By means of distance optimization in PSK-means algorithm, this paper eliminates the defect of trend analysis caused by Euclidean distance and incorporates a new similarity analysis method: Pearson coefficient. This method provides processing steps for different value ranges of variables, so as to ignore the value ranges of different variables, and finally the obtained correlation measures the trend correlation. At the same time, the dimensionality of different variables is eliminated in the calculation process, which is equivalent to standardization, which also guarantees the accuracy of the results to some extent elation coefficient.

2.2. Pearson’s correlation coefficient
In formula (1), \(r (X, Y) \in [-1,1]\) is the degree of correlation. When \(r\) is -1 or 1, it means that the two samples are completely correlated. When \(r\) is 0, it means that the two samples are completely independent. In order to facilitate the measurement of sample correlation degree, the Pearson coefficient in formula (1) is appropriately deformed as follows:

\[
(4) \quad \rho(X, Y) = |r(X, Y)|
\]

Where \(\rho(X, Y)\) is in range of [0,1] [8].

2.3. Pearson’s correlation coefficient
The clustering process of PSK-Means algorithm for data is as follows: firstly, the data set is randomly divided into k classes, which contain several objects, and the center of each class cluster is calculated. Then for each remaining object, the most relevant cluster is assigned according to the correlation principle according to the correlation degree of the cluster mean. Finally, we continue to calculate the new mean of each cluster. Until the criterion function converges, otherwise the process will be repeated. In general, the squared error criterion is used, which is defined as follows [9]
In which, \( p \) represents the data object; \( c \) is the average of the cluster \( c_i \). The flow of the PSK-Means algorithm is as follows: Input: cluster number \( k \) and data set \( D = \{d_1, d_2, ..., d_n\} \).

Output: \( k \) class clusters, meeting the convergence of the square error criterion function.

1) Arbitrarily select \( k \) data objects from data set \( D \) as the initial cluster center;
2) calculate the Pearson correlation between other data objects and each initial cluster center and add them to the most relevant cluster;
3) update the cluster mean, that is, calculate the average value of the objects in each cluster;
4) calculate clustering criterion function;
5) until the value of the criterion function no longer changes, the algorithm ends [10]

At first 100 curves of household electricity are divided into \( n \) clusters \( n(\in 1, 2, ..., 8) \). By PSK-means method we offer, we combine multiple clustering clusters into a new set with richer information and features as the input of CNN. if \( k \) times clustering processes are gotten, the \( k \)th result will be \( C^k = \{C^k_1, C^k_2, ..., C^k_n\} \). The \( k \) clustering results are combined into a set as input \( C = \{C_1, C_2, ..., C^k\} \). As a result, we get 9 forecast results including self-forecasting result, clustering forecast results with 2-8 clusters and forecast result with all inputs.

3. Result

After a mass of experiments, the data is listed in following tables:

Tab 1. MAPE of the result between PSK-Means and k-means

| Cluster | P | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   |
|---------|---|-----|-----|-----|-----|-----|-----|-----|-----|
| PSK     | 10.119 | 10.399 | 10.489 | 9.915 | 10.886 | 10.720 | 10.249 | 10.22 | 10.792 |
| K       | 10.378 | 10.399 | 10.401 | 10.376 | 10.489 | 10.677 | 10.705 | 10.198 | 10.517 |

Tab 2. MAE of the result between PSK-Means and k-means

| Cluster | P   | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   |
|---------|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| PSK     | 3.981 | 4.080 | 4.110 | 3.917 | 4.272 | 4.183 | 4.034 | 4.044 | 4.258 |
| K       | 4.009 | 4.080 | 4.102 | 4.084 | 4.079 | 4.187 | 4.192 | 3.955 | 4.061 |

In the table, \( P \) means result of all inputs, numbers mean result with \( n \) clusters. PSK-means the method we offer. PSK–means and the k-means

It is obviously that if we put all clusters into CNN to forecast the data by using PSK-means we can get more concrete result which is the most convincing among 9 different clusters because it indicates the all characters of different cluster curves. By the way, the forecast result of k-means sometimes appears better than PSK-means.

4. Conclusion

This method directly improves the accuracy of prediction by fitting the user’s behaviour trend. The development of the method we can foresee is that weights can be given in different clusters to get better result according to the paper [1] or finding other clustering method that suits the character of household electricity curves.
Fig 1. K-means and PSK-means cluster load forecasting result

Acknowledgements
This project is supported by National Training Program of Innovation and Entrepreneurship for Undergraduates NO.201910145096. It is also supported by the Fundamental Research Funds for the Central Universities N182410001.

References
[1] Y. Wang, Q. Chen, M. Sun, C. Kang and Q. Xia, "An Ensemble Forecasting Method for the Aggregated Load With Subprofiles," in IEEE Transactions on Smart Grid, vol. 9, no. 4, pp. 3906-3908, July 2018, doi: 10.1109/TSG.2018.2807985:1-3.
[2] Huang, Nantian Wang, Wenting Wang, Sining Wang, Jun Cai, Guowei Zhang, Liang. (2020). Incorporating Load Fluctuation in Feature Importance Profile Clustering for Day-Ahead Aggregated Residential Load Forecasting. IEEE Access. PP. 1-1. 10.1109/ACCESS.2020.2971033.
[3] Pearson K. Note on regression and inheritance in the case of two parents[J]. Proceedings of the Royal Society of London, 1895, 58(1): 240-242.
[4] Spearman C. "Footrule" for measuring correlation [J]. British Journal of Psychology, 1906, 2(1): 89-108.
[5] Kendall M G. A new measure of rank correlation [J]. Biometrika, 1938, 30(1/2): 81-93.
[6] Daniels H E. The relation between measures of correlation in the universe of sample permutations [J]. Biometrika, 1944, 33(2): 129-135.
[7] Feng Zhang, Zhenhua Xie, Jiatao Cheng, et al. Combination weighting method based on vector Pearson correlation coefficient [J]. Firepower & command control, 2015, 40 (5): 83-86.
[8] Huanyi Yin, Zhicheng Wen, Zhengjian Ma. An improved KNN algorithm based on Pearson similarity and distance weight [J]. Computer knowledge and technology, 209, 15 (27): 208-210.
[9] Chao Feng. Research on k-means clustering algorithm [D]. Dalian: school of computer science and technology, dalian university of technology, 2007.
[10] Cunge Han, Changyong Liu. A modified K-Means algorithm [J]. Journal of institute of minjiang, 2019, 40 (05): 49-54+90.
[11] Zhuofu Deng, Zhiliang Zhu, Yanze Wang, Heng Guo, Chengwei Chai, Binbin Wang. 2020. "Unified Quantile Regression Deep Neural Network with Time-Cognition for Probabilistic Residential Load Forecasting," Complexity, Hindawi, vol. 2020, pages 1-18, January.