Linear Regression Electricity Prediction Method Based on Clustering of Electric Characteristics

Zhou Xiangyu, Miao Qingqing, Lin Tao, Zhang Lei, Zhou Jianquan
State Grid Jining Power Supply Company
No.28, Huoju Road, Gaoxin District, Jining City, Shandong Province, China
411754924@qq.com; sdjnmqq@126.com;lintao258@163.com;Yxbsck@163.com; yxbsck2009 @163.com

ABSTRACT
With the development of intelligent power technology, the prediction of users' electricity utilizes by relevant power enterprises tends to be meticulous gradually, and the accumulated data of users are various and large. On this basis of the users' data, Subspace clustering is carried out according to the evaluation index of user's electricity consumption characteristics, and various kinds of user's electricity consumption patterns are obtained. According to the different electricity consumption patterns, the users are divided into groups, and the related factors are judged by mutual information matrix for different groups of users, and then the electricity consumption is predicted by multiple linear regression algorithm. This method proposed in this paper can cluster automatically according to the electricity characteristic index, and can effectively identify the power related factors of different user groups. The simulation results show that the proposed method has better prediction effect than the comparative algorithm.

CCS Concepts
• Information systems → Information systems applications → Decision support systems → Online analytical processing.

Keywords
Electricity characteristic evaluation index; mutual information; multiple linear regression algorithm; electricity forecasting;

1. INTRODUCTION
The prediction of users' electricity consumption is of great significance to relevant power companies and government departments. The forecast of electricity consumption can not only help power companies to understand and serve users better, but also make corresponding plans for the development of the power grid, and also help the government to formulate relevant policies. With the passage of time and the continuous development of the economy, it is foreseeable that China's dependence on electricity will become higher and higher as well.

Users' electricity consumption behavior is different. Even for users in the same industry, the difference will become more and more obvious over time. Pattern recognition based on industry characteristics cannot mine users' information very well [1]. The electricity consumption characteristics of users are not only related to the relevant factors of their own industry, but also related to other social and economic factors. The change trend of power consumption characteristics of users in different regions and of different industries are similar, and the user's power consumption characteristics are diversified, which poses a challenge to the related electricity consumption prediction methods [2]. With the development of science and technology, especially the continuous advancement of intelligent technology, variety of smart power grid technologies have emerged in endlessly, and the construction of power grids has also been greatly improved. The existing power big data is sufficient to provide data support for the quantification of users' electrical consumption characteristics and the related predictions. Establishing a targeted prediction model can not only improve the accuracy of electricity consumption prediction, but also help enterprises to recognize users and their group effects [3].

This article is organized as follows: The first part mainly introduces the current research status; the second part introduces the research progress and the related work; the third part introduces the principle of the algorithm; the fourth part gives the experimental results and analysis; the fifth part summarizes and forecasts the future development direction.

2. RELATED WORK
At present, there are three main methods for electricity prediction: classical prediction method, traditional prediction method and emerging prediction method. The classical prediction method mainly evaluates the electricity value on the basis of the experience of relevant experts or the relationship between simple variables, and the prediction effect is not satisfactory. The common methods used in classical prediction are as follows: expert prediction method, average growth rate method, comprehensive electricity level method, exponential smoothing method, single consumption method, load density method, etc. The traditional prediction method mainly includes three types: time series method, regression analysis method and trend extrapolation method [4]. The time series method [5] is to model and predict through historical data, and to predict the development trend mainly through the trend extrapolation method, so as to obtain the method reflecting the development process and the regularity of data changes. The time series method is relatively mature in development, simple in modeling, and fast in prediction. The disadvantage of the method is that the change law of the

ICBDT2019, August 28–30, 2019, Jinan, China

ACM ISBN 978-1-4503-7192-6/19/08…$15.00
https://doi.org/10.1145/3358528.3358564

electricity is insufficiently fitted, and the prediction of the series of electricity with uniform changes is better. The principle of regression analysis method [6] is to establish a model based on relevant historical electricity data and factors, so as to predict future electricity. It's advantage is simple and easy to operate, while the disadvantage is that the prediction accuracy is susceptible to the inconsistency of relevant factors data. Emerging prediction methods mainly include neural network prediction method, support vector machine prediction method, chaos theory prediction method and gray prediction method, and researches in this area are also increasing. The advantage of neural network prediction method [7] is that the regression process of electricity and related factors can be realized without constructing a certain function. The method is easy to implement and has a good prediction effect. The disadvantage is that a large number of samples are needed as a support. Currently, BP neural network is the most representative one. According to relevant literature, it can be seen that this algorithm has two main advantages. First, under the condition that the hidden layer and the hidden node are sufficient, the BP neural network can approach any nonlinear mapping relationship. Second, it has good generalization ability. Support vector machine prediction method [8] takes into account both the model and the learning ability. The advantage is that it can achieve high efficiency from training to prediction, avoiding induction and deduction, and can automatically eliminate redundant samples, which has good robustness. The disadvantage is that it is only suitable for modeling prediction of small-scale samples and difficult to solve the classification problems. The chaos theory prediction method is based on the theory of spatial reconstruction. The advantage is that the inner "law" can be found from any messy data, and all states are not iterated within a certain range. The disadvantage is that the calculation time is long and the optimal solution may not be obtained. The grey prediction method is modeled and predicted by inconspicuous law or disorderly electricity. The advantage is that it does not require too many samples, and the method is simple and accurate. The disadvantage is that the prediction of the power sequence with weak regularity is not good.

Before the widely application of the power energy data acquire system, the data acquisition on the demand side was less, so the power forecasting was mainly based on the power energy datao of transmission lines. The power energy data of transmission lines are regular, and the number of samples used for prediction learning is less, which can not reflect the characteristics of users'electricity consumption behavior. Therefore, most of the prediction methods used are based on a small amount of sample data. Based on the work of predecessors, this paper proposes a new method of electricity prediction, which has two main contributions. Firstly, AP clustering algorithm is adopted on the basis of multi-dimensional indexes to realize the automatic clustering of user groups. Secondly, on the basis of clustering, multiple linear regression models are established for different user groups to predict users' electricity consumption. Compared with the traditional method, the method proposed in this paper makes more effective use of the massive data from power energy data acquisition, improves the number of samples, supports multiple iterations, which can greatly improve the accuracy and real-time prediction.

3. ALGORITHM PRINCIPLE

The content of the algorithm m includes: establishing multi-dimensional evaluation indicators through user data, and then analyzing the user's electricity consumption characteristics. Different users have different electricity consumption characteristics and clustering in different dimensions. AP clustering algorithm is adopted, which realizes the clustering of users in different subspaces, and use mutual information theory to analyze the factors associated with users in order to obtain the related factors of different user groups. On this basis, the power consumption data of various users and related factors are used to construct the corresponding sample dataset, multivariate linear regression analysis is used to analyze the correlation, so as to establish a corresponding electricity consumption prediction model to realize the users' electricity consumption prediction. The algorithm implementation process is shown in the figure below.

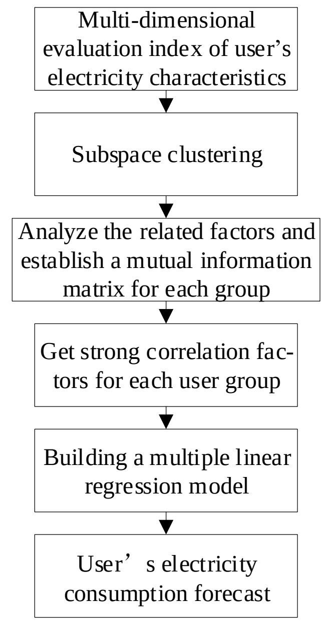


Figure 1. Flow chart of power prediction method.

According to the power consumption big data of users, the dataset V_D of power consumption characteristics of users and the dataset Y_D of associated factors (Y_D contains r associated factors, which are Y_1, Y_2,…, Y_r), the power consumption data of the i -th user in V_D can be expressed as $V_i(\alpha;\ \beta;\ \gamma)$, and clustering is performed on the three dimensions of α、β、γ by clustering algorithm to obtain n user groups ($G_k(k{=}1,2,\ldots,n)$), respectively calculating the mutual information between each user group and the associated factors. On this basis, the strong correlation factors of each user group are obtained. Finally, the prediction model is established based on the user's electricity consumption data and the strong correlation factor data. When the strong correlation factor data of the user to be tested is input, the value of related electricity consumption prediction is obtained.

3.1 User's Electricity Characteristics Analysis

In the aspect of selecting the evaluation index of electricity consumption characteristics, some relevant clustering feature vectors are selected from the time characteristics, such as annual electricity consumption data, quarterly electricity consumption data, and daily electricity consumption data [9]. Literature [10] selects evaluation indicators from the perspective of load, such as daily load curve, maximum load rate, and average user load rate. In addition, indicators with time series characteristics are selected from the perspective of peak and valley characteristics. Based on the selected indicators in, this paper selects the evaluation indicators including sequential and non-sequential for different user groups, and adds the daily load curve evaluation indicators in the time series evaluation indicators:

$$V_i = \{\alpha_{t1}, \alpha_{t2}, \cdots, \alpha_{tu}; \beta_{t1}, \beta_{t2}, \cdots, \beta_{tv}; \gamma_1, \gamma_2, \cdots, \gamma_w\} \in V_D \quad (1)$$

In the above formula i=1,2, $\cdots$,m. α_{t1}, α_{t2},$\cdots$,α_{tu} and β_{t1}, β_{t2},$\cdots$, β_{tv} are vectors with time-series features, respectively representing user annual electricity consumption data and monthly electricity consumption data; γ_1, γ_2,$\cdots$, γ_w are non- sequential feature vectors, and the specific contents include load density γ_1, daily average load rate γ_2, seasonal imbalance coefficient γ_3, maximum load utilization hours γ_4 and other load indicators.

Clustering the evaluation indicators of different dimensions of the user, so as to judge the user's power usage mode better. The literature uses the fuzzy C-means clustering method for clustering. The disadvantage is that the clustering data needs to be set in advance. Therefore, the clustering effect is affected by human factors. In the process of user clustering, with the increase of VC dimension (V_C={ V_α , V_β , V_γ }, V_α represents annual electricity consumption sequence, V_β represents monthly electricity consumption sequence, V_γ represents load characteristic data), it may Lead to "dimensional disaster"[11]. V_C contains various types of data, such as sequential data and non-sequential data. Dimensionality reduction can reduce the data attributes in V_C, but the interpretability is poor, and important data may be lost, and the processing of high-dimensional data has limitations. Based on the above, this paper uses subspace clustering method to cluster on different subspaces of V_D, divide the data set V_D into 3 different subspaces L_1~L_3, split the V_C, and perform AP clustering on the three dimensions of V_α , V_β , V_γ. r, s and t clusters can be obtained by clustering, and the membership degree of the sample data points for each cluster of the subspace can be expressed as

$$U = \begin{Bmatrix} u_{a,1}, \cdots, u_{a,j}, \cdots, u_{a,r} \\ u_{\beta,1}, \cdots, u_{\beta,j}, \cdots, u_{\beta,s} \\ u_{\gamma,1}, \cdots, u_{\gamma,j}, \cdots, u_{\gamma,t} \end{Bmatrix} \quad (2)$$

In the above formula, $u_{\alpha,i,}$, $u_{\beta,j,}$ $u_{\gamma,k}$$\in$[0,1], $\sum_{i=1}^{r} u_{a,i} = 1$, $\sum_{j=1}^{s} u_{\beta,i} = 1$ and $\sum_{k=1}^{t} u_{\gamma,i} = 1$ should be satisfied.

A cluster was extracted from the clusters obtained by clustering in each subspace for fusion, so as to determine a user power consumption mode. In the resulting cluster, the number of user power modes is n=r×s×t, according to the electricity mode, users are grouped into n groups G_i(i=1,$\cdots$,n). The membership degree of each sample point to different G_i can constitute a membership matrix:

$$U_{\max} = \begin{bmatrix} u_{\alpha,\max} \\ u_{\beta.\max} \\ u_{\gamma,\max} \end{bmatrix} = \begin{bmatrix} \max(u_{\alpha,1}, \cdots, u_{\alpha,i}, \cdots, u_{\alpha,r}) \\ \max(u_{\beta,1}, \cdots, u_{\beta,i}, \cdots, u_{\beta,r}) \\ \max(u_{\gamma,1}, \cdots, u_{\gamma,i}, \cdots, u_{\gamma,r}) \end{bmatrix} \quad (3)$$

Users can be divided into different user groups by formula (3).

3.2 Electricity Consumption Correlation Factors Analysis

Users' electricity consumption is related to a variety of social and economic factors, and different user groups have different responses to the same social and economic factor. The mutual information theory can analyze the correlation between users' electricity consumption characteristics and various social and economic factors, and find the factors most relevant to each user group. In this paper, the correlation between user electricity data X and correlation factor Y is obtained by calculating the mutual information between them. The correlation formula is shown as follows:

$$I(X,Y) = -\sum_{i=1}^{N_i} [\frac{M_i}{M} \log \frac{M_i}{M}] - \{-\sum_{u=1}^{N_j} P(y_u) \cdot [\sum_{v=1}^{N_i} \frac{M_{uv}}{M} \log \frac{M_{uv}}{M}]\} \quad (4)$$

In the above formula, M represents the sum of the numbers of all values of X and Y; N_i represents the number of intervals of X; M_i represents the number of values of X falling in the i -th interval; N_j represents the number of intervals of Y, and $P(y_u)$ represents The probability that Y falls in the u-th interval; M_{uv} is the number of values that X happens to fall into the v interval when Y falls into the u interval. The mutual information between the power consumption of each user in the user group and the potential correlation factor is expressed by formula (5).

$$I_k = \begin{bmatrix} I(X_1,Y_1) & \dots & I(X_1,Y_j) & \dots & I(X_1,Y_l) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ I(X_i,Y_1) & \dots & I(X_i,Y_j) & \dots & I(X_i,Y_l) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ I(X_p,Y_1) & \dots & I(X_p,Y_j) & \dots & I(X_p,Y_l) \end{bmatrix} \quad (5)$$

The $\{X_1,X_2,\cdots,X_p\}$ represents the data set constituted by data sequence of electricity consumption of p users, $\{Y_1,Y_2,\cdots,Y_l\}$ represents the dataset composed of potential association factors. The average mutual information between the correlation factors Y_j and X_1, X_2,$\cdots$, X_p is calculated by the formula (6).

$$\overline{I(X,Y_j)} = \frac{1}{p} \sum_{i=1}^{p} I(X_i,Y_j), \quad j = 1, \cdots, l \quad (6)$$

The average mutual information between users of different groups and related factors can be obtained through calculation, which shows the strength of the correlation. The larger the value is the stronger the correlation. Select the related factors represented by the value greater than 0 and sort them to obtain the sorted list of related factors, and select the top factors. These factors are used as strong correlation factors to construct the training sample set S_k(k=1,$\cdots$,n), together with the user's electricity data, combined with different user groups G_k(k=1,$\cdots$,n) for the prediction of power consumption.

3.3 Electricity Prediction Model

On the basis of Section 3.2, the strong correlation factors corresponding to each user group can be obtained, and the regression model is established by using multiple linear regression algorithm to predict the power consumption of each user group. Since the time complexity of the equations of the multiple linear regression algorithm, O(n^3), is relatively high, draw lessons from the random forest algorithm, for different user groups G_k, multiple training samples are randomly obtained from the original training sample set S_k, and each sample is modeled, and then each model is tested using the test set to obtain a power prediction result. The sampling method is Bootstrap, which randomly acquires w training samples (S_1, S_2,⋯, S_{kw}) from S_k for constructing w multiple linear regression models. The training sample subset extraction process uses a sampling method with a return. About 37% of the samples do not appear in the collected sample set. These are out-of-bag (OOB) data, and the out-of-bag data is used as a test set.

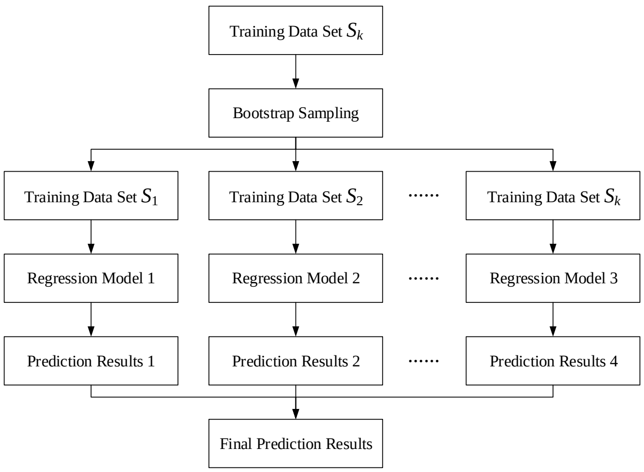


Figure 2. Electricity prediction modeling process

The original training data set S_k consists of two types of data, they are user power consumption data and M related factors data corresponding with it. The input of the model is the M related factor data, and the output is the time sequence data of the user's power consumption. w multiple linear regression models were established, and the test set was used for simulation test. The correlation factor data X_k related to the electricity consumption data Y_k was used as input, and the prediction result was calculated by formula (7):

$$H_k(X_k) = \frac{1}{n}\sum_{i=1}^{w} f_{ki}(X_k)\ k = 1, 2, \cdots, n \tag{7}$$

In the above formula, H_k is a power prediction model corresponding to G_k, and f_{ki} is a single linear regression prediction model. By linearly combining f_k, a power consumption prediction model for all users can be obtained.

4. CASE ANALYSIS

The analysis of this example uses the electricity consumption data of users in a certain region of Shandong Province. The data includes the electricity consumption information of 7,832 users from January 2014 to January 2017. The user's power consumption, daily load curve and other data are used to calculate the load characteristic index data such as the maximum load utilization hours γ_1, load density γ_2, peak-to-valley power ratio γ_3, seasonal imbalance coefficient γ_4, and daily average load rate γ_5. And combined with the user power consumption time sequence data, the user power characteristic data set V_D is constructed.

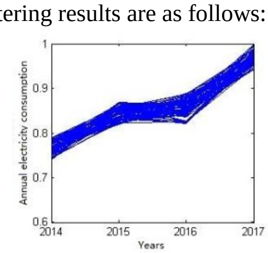
Users are mainly divided into two types: residents and non-resident users. Non-resident users cover industries such as industry, catering, transportation, etc. The research in reference [3] selects the corresponding related factors. For non-manufacturing users, 40 factors including the industry's total output value, fixed asset investment, industry sentiment index, main product ex-factory price index, and product inventory are selected as potential related factors. For the manufacturing users, 78 factors including product output, raw material price index, and product inventory are selected as related factors. In terms of region, 20 factors such as overall GDP, GDP of primary, secondary and tertiary industries, and fixed asset investment are considered. A total of 138 factors are used as user-related factors, and the potential correlation factor dataset YD is constructed through these factor data, and the user electricity data and YD are normalized. See Table 1 for details.

Table 1. Normalization of electricity consumption data and related factor data

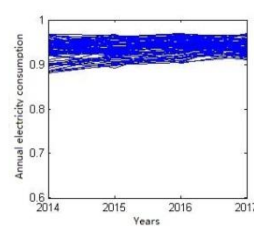
Users	2014.1	2014.2	2014.3	⋯	2017.1
X_1	0.632	0.578	0.726	⋯	0.864
X_2	0.862	0.821	0.929	⋯	0.882
X_3	0.684	0.532	0.625	⋯	0.824
⋮	⋮	⋮	⋮	⋮	⋮
X_{7832}	0.725	0.751	0.821	⋯	0.654
Potential correlation factor data	2014.1	2014.2	2014.3	⋯	2017.1
Y_1	0.651	0.540	0.751	⋯	0.924
Y_2	0.724	0.795	0.432	⋯	0.832
⋮	⋮	⋮	⋮	⋮	⋮
Y_{138}	0.842	0.256	0.536	⋯	0.642

4.1 Electricity Consumption Characteristics Clustering

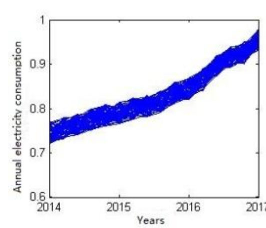
The user's electricity consumption dataset V_D includes three categories: user's annual power consumption time series data, average value of the monthly power consumption of the electricity users in each month, and load characteristic index. AP clustering is conducted on the basis of the three types of data, and the clustering results are as follows:



(a) Volatility

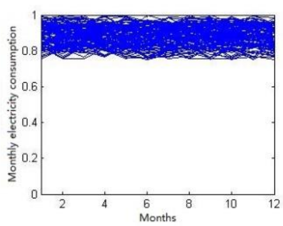


(b) Stable

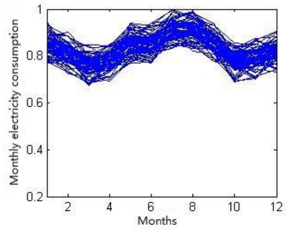


(c) Growth

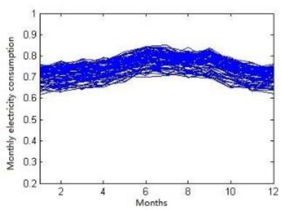
Figure 3. User's annual electricity consumption cluster



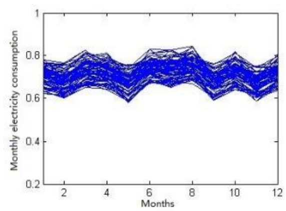
(a)Stable



(b) Bimodal



(c) Unimodal



(d) Multimodal

Figure 4. User's monthly electricity consumption clustering

Table2 Clustering results of load characteristic indexes

Load characterristic index	Index Segmentation Interval		
	Class 1	Class 2	Class 3
γ_1	[57.51, 311.62]	[12.95, 175.98]	[60.25, 221.32]
γ_2	[0.60,0.83]	[0.58, 0.69]	[0.71, 0.93]
γ_3	[0.60, 0.71]	[0.72, 0.96]	[0.62, 0.83]
γ_4	[2014,5908]	[493, 2894]	[4659, 6348]

Figure 3 shows the clustering results of the user's annual electricity consumption, which are divided into three categories: stable, volatility, and growth. The stable type includes three types of users: slow growth, slow decline, and stable. Volatility includes customers who are more affected by the business environment, and rapid growth includes users with good development momentum.

Figure 4 shows the clustering results of users' monthly electricity consumption, mainly including four types: stable, unimodal, bimodal, and multimodal. The stable type mainly refers the users with strong continuity of electricity consumption, such as the manufacturing industry. The unimodal type refers the customers whose peak electricity consumption occurs in May to October in a year. The bimodal type refers the customers who are heavily affected by seasonal factors. And the multimodal type refers to customers who are affected by multiple factors such as holidays, production cycles, etc. Table 2 shows the clustering results of different load characteristics indicators. Class 1 contains a small seasonal imbalance coefficient, indicating that such users are sensitive to seasonal factors, mainly including some light industrial users. The average load rate of class 2 is small, indicating that the users mainly refer to low-energy users such as public utilities. The load utilization hours and load density in class 3 are relatively large, mainly referring to the high-energy users such as heavy industry.

4.2 Electricity Consumption Prediction

Through clustering on the subspace, the user's power consumption mode can be divided into $3\times4\times3=36$ kinds, and all users are divided into 36 groups. On this basis, through the average mutual information of the related factors of each group of users, and selecting the top 15 related factors, the monthly data of the 15 related factor data is taken as input, and the monthly electricity consumption is used as the output to establish a sample set. The Bootstrap method is used to select m training sample subsets from the data set, and construct them into multiple linear regression models respectively. The remaining OOB data is used as a test set to test the error of the prediction model. Figure 5 shows the prediction error of different number of regression models under different electricity consumption modes, measured by MAPE value.

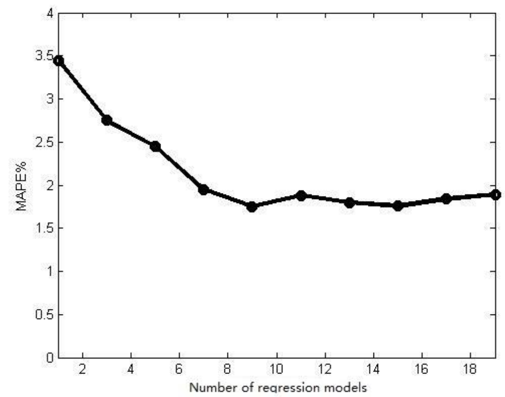


Figure 5. Prediction error of different regression model numbers

The predicted monthly electricity consumption can be obtained through the prediction of the regression model. In order to compare the prediction effect of the electricity prediction algorithm proposed in this paper, the same training data is compared with support vector machine prediction model (SVM) and random forest algorithm (RF) in literature [3]. The results are shown in Table 3. The average absolute error value of the predicted value and the true value of the user population in a random sample is shown. It can be clearly seen that the absolute percentage error and average absolute error of the algorithm proposed in this paper are both better than the algorithm compared, indicating that the algorithm proposed in this paper has higher precision. In order to further verify the effectiveness of the proposed algorithm, 27 user groups are predicted and modeled. The effect comparison is shown in Figure 6. The overall effect of the proposed algorithm is better than that of the multi-contrast algorithm.

Table 3 Comparison of prediction results

Groups	SVM(%)	RF(%)	Algorithm in this paper(%)
1	2.38	2.26	2.17
2	3.59	2.49	2.32
3	5.52	2.97	2.68
4	4.32	1.82	1.16
5	2.77	2.32	1.63
6	4.16	3.67	3.59
MAPE(%)	3.79	2.58	2.26

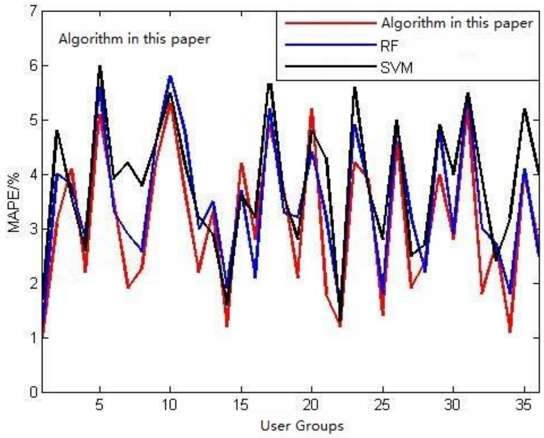


Figure 6. Prediction error of different user groups

5. SUMMARY AND OUTLOOK

In this paper, the subspace clustering method based on automatic user electricity characteristics is adopted to subspace clustering user's electricity characteristics, and mutual information theory is used to mine related factors among different user groups. On the basis of obtaining the correlation factors, the multivariate linear regression algorithm is used for prediction. The algorithm realizes the whole process data driving of electricity prediction, and expands the analysis method of users' electricity consumption behavior. With the continuous development of big data in the power grid, considering the time complexity of the AP algorithm is relatively high, the next work is to start with the optimization of the efficiency of the algorithm to adapt to larger data, and further combine advanced technologies such as machine learning and neural networks to improve the predictive effect of users' electricity consumption.

6. REFERENCES

[1] Zhao Y., Li L., and Liu J. Y., et al. 2010. Combinational recognition model for demand side load profile in shanghai power grid. *Power System Technology*. 34, 1(Jan. 2010), 145-151.

[2] Zhao T., Wang L. T.and Zhang Y., et al. 2016. Relation factor identification of electricity consumption behavior of users and electricity demand forecasting based on mutual information and random forests. *Proceedings of the CSEE*. 36,3(Feb. 2016),604-614

[3] Zhao T., Zhang Y., and Zhang D. X. 2014. Application technology of big data in smart distribution grid and its prospect analysis. *Power System Technology*. 38, 12 (Dec. 2014), 3305-3312

[4] Rallapalli S. R. and Ghosh S.. 2012. Forecasting monthly peak demand of electricity in India-A critique. *Energy Policy*. 45,2 (Jun. 2012), 516-520.

[5] Ali O. P. and H. K. C.. 2013. ANN hybrid model versus ARIMA and ARIMAX models of runoff coefficient. *Journal of Hydrology*. 500, 11(Jun. 2013),21-36

[6] Xu B. X. and Li X. F.. 2017. Research on the medium and long-term urban electricity consumption prediction methods. *Journal of Hunan University of Technology*. 31, 2(Feb. 2017)78-83.

[7] Mashud R. and Irena K. 2016. Forecasting electricity load with advanced wavelet neural networks. *Neurocomputing*. 182 (Jun. 2016), 118-132.

[8] Fan G. F., Peng L. and Hong W. C., et al. 2016. Electric load forecasting by the SVR model with differential empirical mode decomposition and auto regression. *Neurocomputing*. 173 (Jun. 2016), 958-970.

[9] Yang H., Zhang L. and He Q., et al. 2010. Study of power load classification based on adaptive fuzzy C means. *Power System Protection and Control*. 38, 16 (Aug. 2010), 111-115.

[10] Xue C. R., Gu J. and Zhao J. P., et al. 2014. Electricity retail tariff mechanism based on customers' electrical characteristics and cost apportionment. *East China Electric Power.* 42, 1 (Jan. 2014), 168-173

[11] Niu K., Zhang S. B. and Chen J. L. 2008. Subspace clustering through attribute clustering. *Journal of Beijing University of Posts and Telecommunications.* 30, 3(Jan. 2008), 1-5.