Research on electricity consumption forecast based on mutual information and random forests algorithm

Jing Shi [1,3]**, Yunli Shi** [2]**, Jian Tan** [1]**, Lei Zhu** [1] **and Hu Li** [1]

[1] State Grid Jiangsu Electric Power Company Economic Research Institute, Jiangsu Nanjing, 210008, China
[2] Qingyuan Pumped Storage Power Generation Co., Ltd,Guangdong Qingyuan 511500

[3] shijing0119@126.com

Abstract. Traditional power forecasting models cannot efficiently take various factors into account, neither to identify the relation factors. In this paper, the mutual information in information theory and the artificial intelligence random forests algorithm are introduced into the medium and long-term electricity demand prediction. Mutual information can identify the high relation factors based on the value of average mutual information between a variety of variables and electricity demand, different industries may be highly associated with different variables. The random forests algorithm was used for building the different industries forecasting models according to the different correlation factors. The data of electricity consumption in Jiangsu Province is taken as a practical example, and the above methods are compared with the methods without regard to mutual information and the industries. The simulation results show that the above method is scientific, effective, and can provide higher prediction accuracy.

1. Introduction

With the development of smart grid, electricity consumption forecast has an greater impact on long-term power system planning. To adapt the requirements of the market economy and ensure the economy of power supply, it’s necessary to research medium and long-term electricity consumption forecast.

Now the common forecasting methods have traditional prediction method such as Time series method[1]and artificial intelligence methods such as artificial neural network(ANN)[2]、gray models(GM)[3]、support vector machine(SVM)[4]. In recent years, with the rise of machine learning, artificial intelligence methods are more applied to electricity consumption forecast. However, the commonly artificial intelligence methods have some shortcomings: The generalization ability of ANN is not strong, and the hidden units are difficult to determine[5]; The key parameters of SVM are selected by experience, and its prediction accuracy is unstable. The random forests is a combination of Breiman’s Bagging method and Tin Kam Ho's random subspace method [6-7]. Firstly, Random forests (RF) use the bootstrap resampling to extract multiple samples from the original samples. Then, combine these decision trees together to get the final result of the classification or prediction by voting. RF algorithm can effectively avoid the "over-fitting" phenomenon and it is applicable to the operation of various data sets[8].Besides, it has a good tolerance for abnormal values and noise[9], and it has the advantages of high prediction accuracy, fast convergence speed, controllable generalized error [10].

In order to reduce the intervention by interpersonal parameters set, and accurately rely on quantitative calculation to filter out the socioeconomic variables related to electricity consumption,this paper uses the mutual information (MI) [11]in information theory. The method can identify the correlation between the electricity consumption and the socio-economic variables of each industry, and use the random forest algorithm to measure the power consumption. Taking the monthly data of electricity consumption and socioeconomic variables in Jiangsu Province as an example, the correctness and validity of the method in this paper are proved by the simulation.

2. Identification of the relationship between Electricity consumption and economic variables

2.1. Mutual information

The mutual information *I*(*x*;*y*) can not only represent the relationship between two random variables, but also can reflect the strength of the relationship between them. *I*(*x*;*y*) represents the amount of information about x obtained after receiving the message y [12].

$$\begin{aligned} I(x;y) &= \log\frac{P(x,y)}{P(x)P(y)} \\ &= \log\frac{1}{P(x)} - \log\frac{1}{P(x\mid y)} \end{aligned} \tag{1}$$

The average mutual information *I*(*X*;*Y*) is the result of the statistical average of the mutual information *I*(*x*;*y*) in the two probability spaces. For the input variable *X* and the output variable *Y*, the average mutual information is defined as formula(2) .

$$I(X;Y) = \int_X \int_Y P(X,Y)\log\frac{P(X,Y)}{P(X)P(Y)} \tag{2}$$

In the equation, *P*(*X*) and *P*(*Y*) represent the probability distribution of variable *X* and variable *Y* respectively; *P*(*X*,*Y*) is the joint probability distribution of *X* and *Y*.

Mutual information may be positive or negative. If the mutual information is negative, it means that the uncertainty of x is increased after receiving y. While as a statistical average, the average mutual information is always positive or zero. When the average mutual information is zero , it means that x and y are completely independent and don’t have any associated properties.

2.2. Identification of Correlation Factors between Industry Electricity Consumption and Economic Variables

This paper proposes to determine the correlation factors with the strongest correlation with the electricity consumption of each industry by identifying the correlation between the electricity consumption of each industry and the socioeconomic variables. And then the social electricity consumption is forecasted by modeling the industrial electricity consumption separately.

Let us set the monthly data of electricity consumption as variable X. X1 is the first industrial electricity consumption; X2 and X3 are respectively the second and third industrial electricity consumption; X4 is the residential electricity consumption. Let us set the monthly data of socioeconomic variable as variable Y. Y1 is the cumulative increase of industrial added value (%); Y2 is the cumulative value of steel production (million tons); Y3 is the cumulative value of cement production (million tons); Y4 is the cumulative value of pig iron production (million tons); Y5 is the accumulated value of finished products (100 million yuan). Y6 is the accumulated value of the inventory (100 million yuan); Y7 is the accumulated value of crude oil production (million tons); Y8 is the real estate investment accumulated value (100 million yuan); Y9 is the completion of fixed assets investment value (100 million yuan).

The average mutual information values are sorted to form a list of correlation coefficients, which is Table 1. In the list, the three related factors with the average mutual information value are selected for each industry electricity consumption, and the three kinds of social economic variables are used to predict the electricity consumption of each industry.

Table 1. Average mutual information values.

Y \ X	X_1	X_2	X_3	X_4
Y_1	0.107	0.101	0.208	0.073
Y_2	0.415	0.280	0.129	0.180
Y_3	0.406	0.270	0.123	0.207
Y_4	0.412	0.283	0.125	0.182
Y_5	0.301	0.164	0.233	0.065
Y_6	0.261	0.151	0.242	0.076
Y_7	0.405	0.240	0.133	0.229
Y_8	0.426	0.264	0.126	0.169
Y_9	0.390	0.242	0.132	0.144

Three values are selected according to the order of information values. When using the random forest algorithm to forecast the electricity consumption of each industry, we choose three of them in descending order. It can be seen from Table 1 that the average mutual information between the electricity consumption of the first industry *X1* and factors of *Y2*, *Y4* and *Y8* is larger, that is, the correlation among them is strong; and so on.

3. Random forest algorithm

3.1. The basic principle of random forest algorithm[6,10]

Given a series of classification trees $h_1(\mathbf{x}), h_2(\mathbf{x}), ..., h_k(\mathbf{x})$, the training set is arbitrarily selected from the distributions of the random vectors Y and X .The edge function is defined as follows.

$$mg(\mathbf{X}, Y) = av_k I(h_k(\mathbf{X}) = Y) - \max_{j \neq Y} av_k I(h_k(\mathbf{X}) = j) \tag{3}$$

Where *I(•)* is an instruction function, *j* is any of the X categories, and *Y* is the correct classification. The edge function measures the degree of the average number of votes for the correct classification *Y* of the vector *X* over the number of votes in the other categories. Besides, the greater the edge function value, the higher the confidence of the classification.

Definition 1 The generalization error is defined as:

$$PE^* = P_{\mathrm{X},Y}(mg(\mathbf{X}, Y) < 0) \tag{4}$$

In a random forest, $h_k(\mathbf{X}) = h(\mathbf{X}, \Theta_k)$. For many decision trees, the following theorem is based on the law of large numbers and the structure of the tree.

Theorem 1 When the number of trees increases, it is possible to determine all the sequences $\Theta_1, \Theta_2, ...$ and *PE* * converges on:

$$P_{\mathrm{X},Y}(P_\Theta(h(\mathbf{X}, \Theta) = Y) - \max_{j \neq Y} P_\Theta(h(\mathbf{X}, \Theta) = j) < 0) \tag{5}$$

The theorem shows that when the number of trees increases, the generalization error tends to a certain boundary, that is, the random forests has a good ability to prevent ‘over-fitting’.

3.2. Random forests flow

According to the correlation factors identified by mutual information, we use the random forests algorithm to model and forecast, as figure 1 shows:

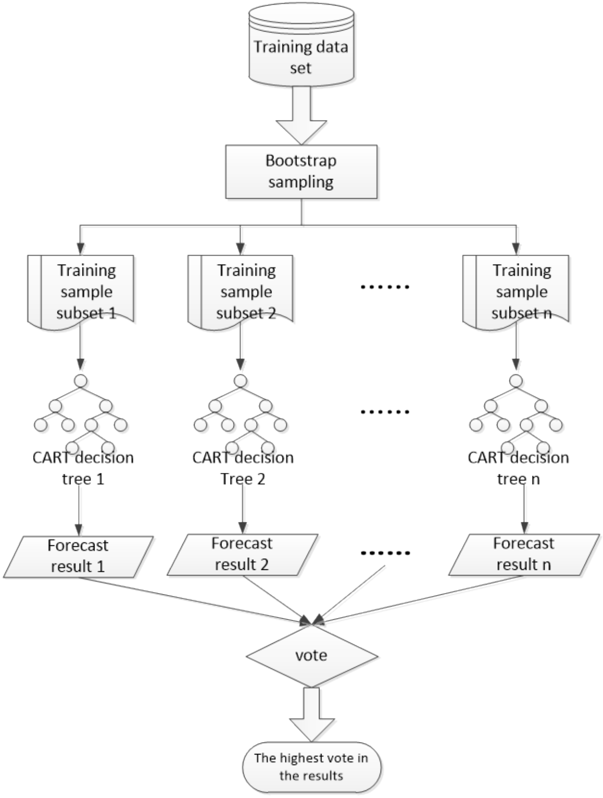

Figure 1. The flow chart for random forests.

4. Example analysis

4.1. Data source

This paper is according to related factors of socioeconomic variable in each industry determined by mutual information (Table 1),and training samples are selected from the monthly data from December 2011 to December 2014. Test samples are selected from the monthly data in Jiangsu Province from January 2015 to December 2015 ,and then they are modeled and predicted.

4.2. The analysis of Forecast result

4.2.1. The impact of the related factors on the forecast. Compare the forecast result from the following two points.

1. Do not consider the associated factors at all.

The error is shown in figure 2 below. The first prediction result represents the error value of the prediction result obtained without considering theassociated factor. The third prediction result represents the error value of the prediction result considering the mutual information.

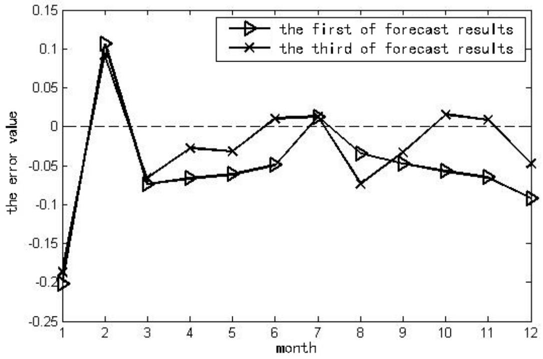

Figure 2. Forecast results comparison without considering interrelated factors.

When the correlation factors are taken into account, the prediction error value is obviously reduced.
2. Do not use mutual information, directly consider all the associated factors.

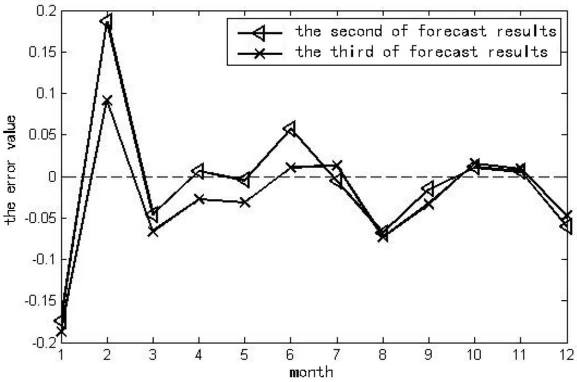

Figure 3. Forecast results comparison with considering all interrelated factors.

The error contrast is shown in figure 3. The second prediction result represents that we don’t use mutual information and consider all the nine socioeconomic variables directly to predict; The third prediction result represents the error value considering the prediction result of the mutual information.

When the factor with high correlation is selected by mutual information, the prediction error value is significantly reduced.

4.2.2. The impact of different industries on forecasting. The impact of different industries on the forecast is mainly in the following two conditions for comparison: 1. After identifying the relevant factors of electricity consumption in each industry, the whole society electricity consumption is obtained by adding respective industrial electricity consumption forecasts. 2. Establish the forecast model related to the whole society and socio-economic variables directly not considering the various industries.

The error is shown in figure 4 below. The forth prediction result represents the consideration of all the associated factors. It’s error is based on the the whole society calculation regardless of the sub-industry forecast. The third prediction result represents the error value considering the prediction result of the mutual information.

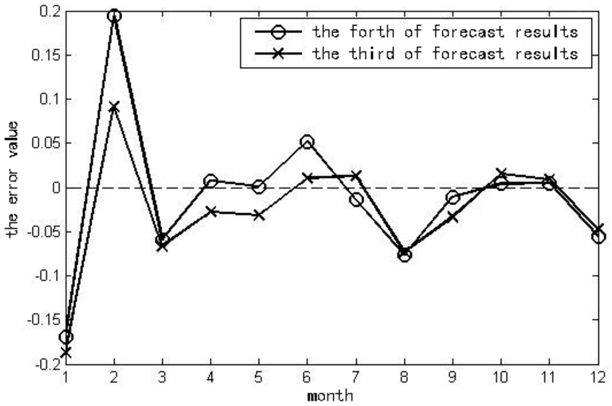

Figure 4. Industrial forecast results comparison.

It can be seen from figure 4, the error of value which is regardless of the industry is larger. The error of mutual information is significantly reduced by filtering out the high correlation factor.

The electricity consumption forecast results from January to December in 2015 is described in table 2.

Table 2. Forecast results for total electricity consumption.

Actual value	Forecast result 1	Forecast result 2	**Forecast result 3**	Forecast result 4
4586749	3660300	3786598	**3729825**	3812457
3072446	3399052	3646035	**3354530**	3670817
4385620	4061944	4182058	**4094582**	4124968
4114045	3838946	4138528	**3999795**	4145719
4204690	3946161	4183444	**4070726**	4210083
4173825	3965749	4411320	**4217592**	4389519
4631136	4689085	4605325	**4690388**	4567314
4812440	4646319	4483230	**4459713**	4442873
4182376	3983317	4117869	**4042118**	4136457
4112192	3876585	4155348	**4176626**	4128937
4221381	3944222	4244090	**4257693**	4243219
4650126	4221965	4366326	**4430131**	4392692

5. Conclusion

In this paper, the mutual information theory is introduced, and the related factors of each industry are identified before the establishment of electricity consumption forecasting model. Then, the advanced artificial intelligent random forest algorithm is adopted to forecast the electricity consumption of each industry, and finally we obtained the prediction value of the total social electricity consumption. Besides, the example shows that the method proposed in this paper can improve the prediction accuracy. In the future, we can consider simulation study and try to use artificial data with changeable parameters to further improve the forecast accuracy.

References

[1] Kang Chongqing, Xia Qing, Liu Mei. Power system load forecasting[M]. *China Electric Power Press*,2007.

[2] Xu Chen, Cao Li, Liang Xiaoxiao, et al. Research on electricity demand forecasting based on ABC-BP neural network[J].*Computer Measurement& Control*, 2014 **03** 912-914+922.

[3] Wang Yunping, Huang Dianxun, Xiong Haoqing, etal. Using relational analysis and multi-variable grey model forelectricity demand forecasting in smart grid environment[J].*Power System Protection and Control*,2012 **40(1)** 96-100.

[4] Liu Yunyun .Modeling and research on direct auxiliary power consumption of pumped storage power station based on the support vector machine[D].*South China University of Techbology*,2014.

[5] Li Yuancheng ,Fang Tingjian, Yu Erkeng, et, al. Study of support vector machines for short-term load forecasting[J]. *Proceedings of the CSEE*, 2003 **23(6)** 55-59.

[6] Breiman L. Random Forest[J].Machine Learning, 2001 **45** 5-32.

[7] Ho T K. The Random Subspace Method for Constructing Decision Forests[J].*IEEE Transactions on Pattern Analysis & Machine Intelligence*,1998 **20(8)** 832-844.

[8] Xu Baoxun. Research on optimization of random forest algorithm for high dimensional data[D]. *Harbin Institute of Technology*,2013.

[9] Cao Zhengfeng. Study on optimization of random forests algorithm[D].*Capital University of Economics and Business*,2014.

[10] Wu Xiaoyu, He Jinghan, Zhang Pei, etal. Power system short-term load forecasting based on improved random forest with Grey relation projection[J].*Automation of Electric Power System*,2015 **39(12)** 50-55.

[11] Yuan Yuan, Gu Hao, Huang Wei, etal. Application of mutual information to power system medium and long-term load forecasting[J].*East China Electric Power*,2009 **02** 236-239.

[12] Fu Zuyun. Information Theory: Basic Theory and Application (4th Edition) [M].*Electronic Industry Press*, 2015.