Electrical Engineering (2022) 104:2045–2059
https://doi.org/10.1007/s00202-021-01457-5

ORIGINAL PAPER

Data-driven random forest forecasting method of monthly electricity consumption

Xinfu Pang[1] · Changfeng Luan[1] · Li Liu[1] · Wei Liu[1] · Yuancheng Zhu[2]

Received: 10 July 2021 / Accepted: 15 November 2021 / Published online: 12 January 2022

Abstract
Accurate forecast of monthly electricity consumption has guiding significance for the economic dispatch of the power system, and it is also a prerequisite for the power company to formulate a reasonable sales plan. The traditional forecasting method of monthly electricity consumption performs poorly in processing the sequence of monthly electricity consumption with a dual trend, and it cannot consider multiple influencing factors at the same time and cannot screen the influencing factors of monthly electricity consumption. This paper proposes a random forest prediction method of monthly electricity consumption based on the maximum mutual information coefficient. First, the maximum mutual information coefficient between monthly electricity consumption and its influencing factors is calculated; second, high-relevance factors are filtered out based on the maximum mutual information coefficient value; third, the data of high-relevance factors are combined, and random forest is used to predict monthly electricity consumption; finally, the program of the abovementioned method is compiled in Python language with the electricity consumption data of the whole society in Shenyang, Liaoning Province as an actual calculation example, and the method is compared with the method that does not use correlation factor identification. Simulation results show that the proposed method has high prediction accuracy and can provide a basis for making reasonable grid operation plans and making power decisions correctly.

Keywords Forecast of monthly electricity consumption · Mutual information · Identification of related factors · Random forest · CART decision tree

1 Introduction

1.1 Literature review

Prediction of monthly electricity consumption is important in the power system [1]. Accurate forecast of monthly electricity consumption is the premise for power departments to allocate power resources and power companies to make reasonable sales plans [2]. According to the study of the relationship between different socio-economic factors and electricity consumption, the monthly electricity consumption forecast can help power enterprises better understand the service demand of all walks of life and provide data support for the future development of power grid and the formulation of power demand-side response policies [3].

Traditional forecasting methods of monthly electricity consumption often use time series method [4, 5], exponential smoothing method [6], Arima [7], gray model method [8], and regression analysis method [9]. These methods require high stability of the original data, and the curve of monthly electricity consumption is a typical nonstationary and nonlinear time series. Moreover, the traditional forecasting methods have poor performance in dealing with the double trend monthly electricity consumption series with growth and volatility. In reference [4], a prediction method of monthly electricity consumption based on STL (Seasonal and Trend decomposition using Loess) model is proposed. The STL model is used to decompose the time series of electricity consumption, and then, the decomposed components are predicted. However, the method ignores other factors, such as economy. In recent years, quantile regression method has been widely used in prediction of monthly electricity consumption abroad; this method has the advan-

✉ Xinfu Pang
pangxf@sie.edu.cn

1 Key Laboratory of Energy Saving and Controlling in Power System of Liaoning Province, Shenyang Institute of Engineering, Shenyang 110136, China

2 State Grid Yingkou Electric Power Supply Company, Yingkou 115200, China

tage of being insensitive to outliers [10]. In reference [11], a constrained quantile regression averaging (CQRA) method is proposed; this method creates an improved overall prediction from multiple individual probability predictions. The parameter estimation problem of CQRA is described as a linear programming problem, and its goal is to minimize the loss of marbles. In reference [12], a new joint forecasting system is established; this system improves the online probability forecasting of single load, and the refining process is based on multiple quantile regression. However, the main disadvantage of quantile regression is that the complexity of calculation in the process of solving leads to long prediction time [13]. In reference [14], a load residual forecasting method based on quantile regression is proposed; this method significantly improves the accuracy of load forecasting. However, the framework only considers the conditional distribution of load errors, and it ignores the relationship between multi-point load errors.

Data-driven forecasting methods of monthly electricity consumption mainly include artificial neural network [15–18] and support vector machine [19–21]. These methods can consider nonlinearity and have self-learning ability. In reference [22], a medium-term load forecasting method based on singular spectrum analysis and neural network is proposed. Singular spectrum analysis method is introduced to filter and decompose monthly power consumption series to obtain each sub series. Then, neural network model is used to predict each sub series. Finally, the predicted power consumption is reconstructed. However, this method only uses the power consumption series to forecast, and it ignores many influencing factors of monthly power consumption. In reference [23], a conditional probability density forecasting method of residential load based on deep hybrid network is proposed, and an end-to-end probabilistic residential load forecasting composite model composed of convolution neural network and gating recursive unit is designed. However, the convergence speed of neural network algorithm is slow and easy to fall into the problem of local minimization, which leads to the failure of network training. In reference [24], a new online integrated learning method is proposed; this method combines batch learning with online learning for load forecasting to ensure the adaptability of online application. In reference [25], subspace clustering is used to analyze the power consumption-related factors of different types of users. Finally, random forest algorithm is used for prediction, and good prediction results are obtained.

1.2 Motivation

The traditional forecasting method of monthly electricity consumption cannot fully consider various affecting factors of the monthly electricity consumption and is sensitive to outliers and noise. The data-driven electricity consumption method also has some problems, such as easy over fitting and slow convergence speed. Random forest algorithm is applicable to all kinds of data sets and has the advantages of preventing over fitting, being insensitive to outliers and noise [10], and having many input variables and fast convergence speed.

Table 1 Relationship between correlation coefficient and degree

Absolute value of correlation coefficient	Degree of correlation
$0.00 \le \|r\| < 0.10$	Irrelevant
$0.10 \le \|r\| < 0.19$	Very low correlation
$0.20 \le \|r\| < 0.39$	Low correlation
$0.40 \le \|r\| < 0.69$	Moderate Correlation
$0.70 \le \|r\| < 0.89$	Highly correlated
$0.90 \le \|r\| < 1.00$	Very high correlation

On the basis of the abovementioned discussion, this paper proposes a random forest prediction method based on the maximum mutual information coefficient. The maximum mutual information coefficient is used to identify the correlation between the monthly electricity consumption and its influencing factors, screen out the strong correlation factors, simplify the input variables of prediction, and use the random forest algorithm for prediction. Finally, the monthly electricity consumption and socio-economic variable data of Shenyang City in Liaoning Province from 2005 to 2014 are taken as the training set, and the monthly electricity consumption data of Shenyang City in 2015 are used as the verification set. The effectiveness of the studied prediction method of monthly electricity consumption is proven by an example.

2 Problem description of forecast of monthly electricity consumption

In the forecast of monthly electricity consumption, we need to consider the historical data of monthly electricity consumption, GDP, the total fixed assets investment (electricity, heat, gas, water production and supply, transportation, storage, and postal industry), the total number of hotel accommodation, the total registered residence, tap water sales, natural gas sales, and the production of industrial steel above designated size. The monthly electricity consumption is predicted on the basis of the factors related to electricity consumption, such as crude oil output of industries, above designated size, added value of industries above designated size, sales volume of wholesale and retail trade enterprises above designated size, and total value of import and export. However, more factors considered do not mean better effect of prediction of monthly electricity consumption. Thus, the strong correlation factors of monthly electricity consumption should be

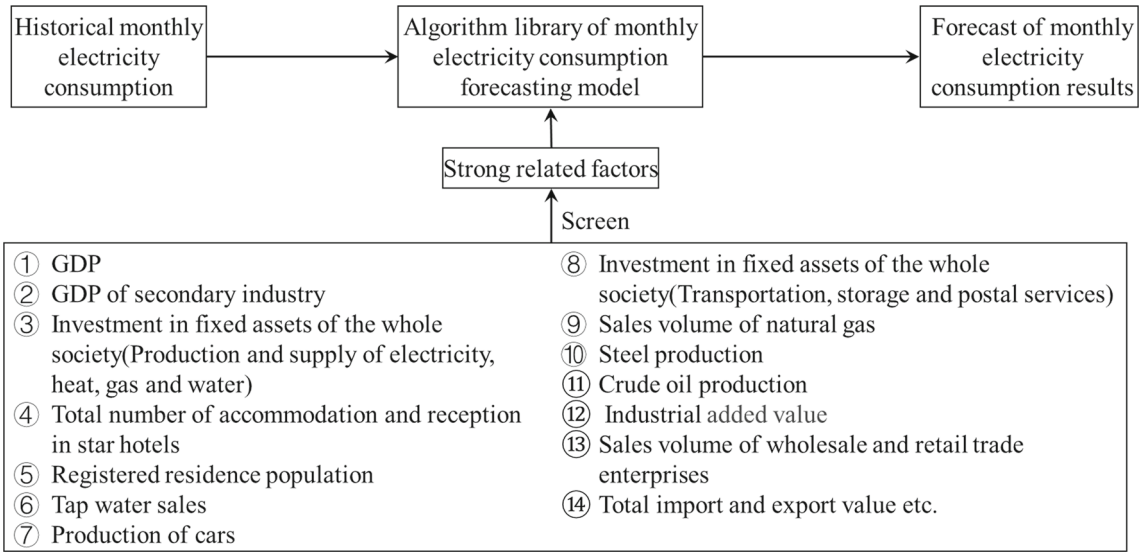

Fig. 1 Problem description of forecast of monthly electricity consumption

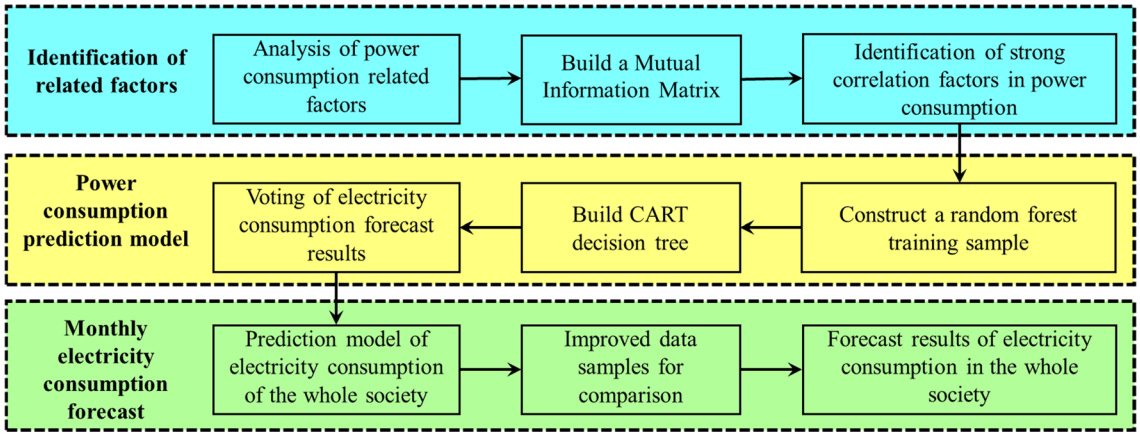

Fig. 2 Strategy chart of forecasting method of monthly electricity consumption

identified. At the same time, outliers and noise will be present considering that the forecast of monthly electricity consumption is the operation of the data set, and the data set contains many variables. Thus, an appropriate algorithm for the forecast of monthly electricity consumption should be selected. The problem description of prediction of monthly electricity consumption is shown in Fig. 1.

3 Strategy structure of stochastic forest forecasting method of monthly electricity consumption considering mutual information

The main contents of the prediction method of monthly electricity consumption based on the maximum mutual information coefficient are as follows. The maximum mutual information coefficient is used to analyze the correlation between the monthly electricity consumption and the potential correlation factors, and the strong correlation factors are screed out. The training sample set is constructed on the basis of the data of monthly electricity consumption and its strong correlation factors. After the parameters of the decision tree are optimized, the random forest algorithm is used to predict the monthly electricity consumption of the whole society. The specific implementation strategy of the method is shown in Fig. 2.

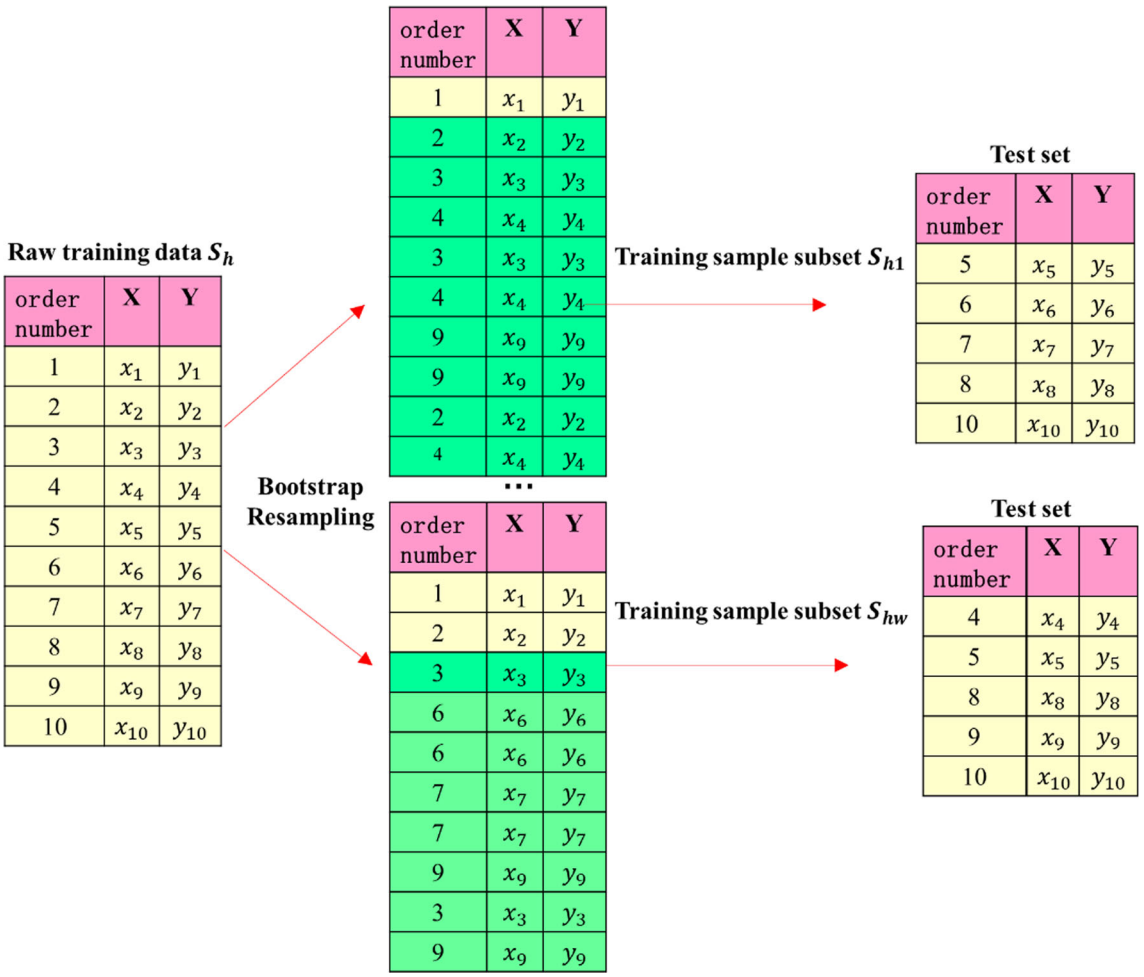

Fig. 3 Random selection of training sample subset

3.1 Identification of related factors

Monthly electricity consumption has different correlation with many factors. The maximum mutual information coefficient is used to analyze and sort the relevant influencing factors of monthly electricity consumption, and the factors that have a greater effect on the monthly electricity consumption, and a stronger correlation are screened out [26]. The factors that have a lower correlation to the monthly electricity consumption are eliminated, the input of monthly electricity consumption prediction and the complexity of modeling are reduced, and the prediction accuracy is improved.

The maximum information coefficient is based on information and mutual information theories [27], and it can better measure the linear and nonlinear relationship between variables by dividing the data interval with grid [28]. The maximum mutual information coefficient is a standard to determine the correlation between two variables.

X and Y are the monthly electricity consumption and related factors of the whole society in data set D, where $X = \{x_i, \quad i = 1, 2, \ldots\}$, $Y = \{y_j, \quad j = 1, 2, \ldots\}$. The mutual information between X and Y is defined as.

$$\mathrm{MI}(X;Y) = \sum_{y\in Y}\sum_{x\in X} p(x,y)\log\frac{P(x,y)}{P(x)P(y)}, \tag{1}$$

where $p(x, y)$ is the joint probability density of X and Y; $p(x)$ and $p(y)$ are the edge probability density of X and Y, respectively.

All values of monthly electricity consumption X and related factor Y in data set D are divided into two grids a and b, respectively, and such grid division is called $a \times b$, which is recorded as $R = (a, b)$. Many kinds of grid partition methods are available for the same $a \times b$, and data set D has different distributions under different partition methods. If the maximum value of $\mathrm{MI}(X; Y)$ in different partition methods is taken as the mutual information value of partition R, then the maximum mutual information can be defined as

$$\mathrm{MI}^{\max}_{D|R}(X;Y) = \max_{R=(a,b)} \mathrm{MI}_{D|R}(X;Y), \tag{2}$$

where $D|R$ is the partition of data set D under grid R.

The detailed steps of maximum mutual information coefficient are given in Algorithm 1.

Algorithm 1: Calculation method of maximum mutual information coefficient

Input: Sample set $D\left(X_i, Y_j\right)$

Output: Maximum mutual information coefficient matrix

initialization: Number of samples n ; $c=15$; $alpha=0.6$

for all $\left(X_i, Y_j\right) \in D$ **do**

The monthly electricity consumption X_i of the whole society is divided into equal parts a , and the related factor Y_j is divided into equal parts b

if $a>ci$, let $a'=ci$ **then** Divide the data space with $a' \times b$ grid, mark $R=a' \times b$

else

Divide the data space with $a \times b$ grid

if $R<n^{0.6}$; **Break** (go to **step 12**)

else

go back to **step 3**

end

end

Calculate the mutual information value of the data falling in the (x, y) grid with different grid division methods

$$MI(X;Y)=\sum_{y \in Y} \sum_{x \in X} p(x,y) \log \frac{P(x,y)}{P(x)P(y)}$$

Select the grid generation method with the maximum mutual information value of the data in the (x, y) grid

Calculate the mutual information coefficient of data set D in grid R

$$MI_{D|R}(X;Y)=\frac{MI_{D|R}(X;Y)}{\log \min(a,b)}$$

Calculate the maximum mutual information coefficient and get the maximum mutual information coefficient matrix

$$MIC(D)=\max_{ab<B(n)}\left\{MI_{D|R}(X,Y)\right\}$$

end

$$\mathrm{MI}_{D|R}(X;Y)=\frac{\mathrm{MI}_{D|R}(X;Y)}{\log \min(a,b)} \tag{3}$$

The maximum information coefficients of X and Y are defined as.

$$\mathrm{MIC}(D)=\max_{ab<B(n)}\left\{\mathrm{MI}_{D|R}(X,Y)\right\}, \tag{4}$$

where $ab < B(n)$ is the upper bound of mesh generation; $n=\max(i, j)$, $B(n)=n^{0.6}$.

The correlation criteria of maximum mutual information coefficient analysis are shown in Table 1.

The sample set D is composed of monthly electricity consumption X_i and related factors Y_j.

3.2 Power consumption prediction modeling based on random forest

3.2.1 Random selection of training sample subset

The random selection of training sample subset is realized by Bootstrap method [29]. Bootstrap method forms different data sets by repeatedly extracting samples from the original data set and putting them back instead of repeatedly dividing the original data into separate data sets. Each Bootstrap data set is based on extraction and then put back, which is the same size as the original data set. Specifically, if the size of the original data set is N, and N samples are put back from it, then the size of the formed Bootstrap data set is N. An

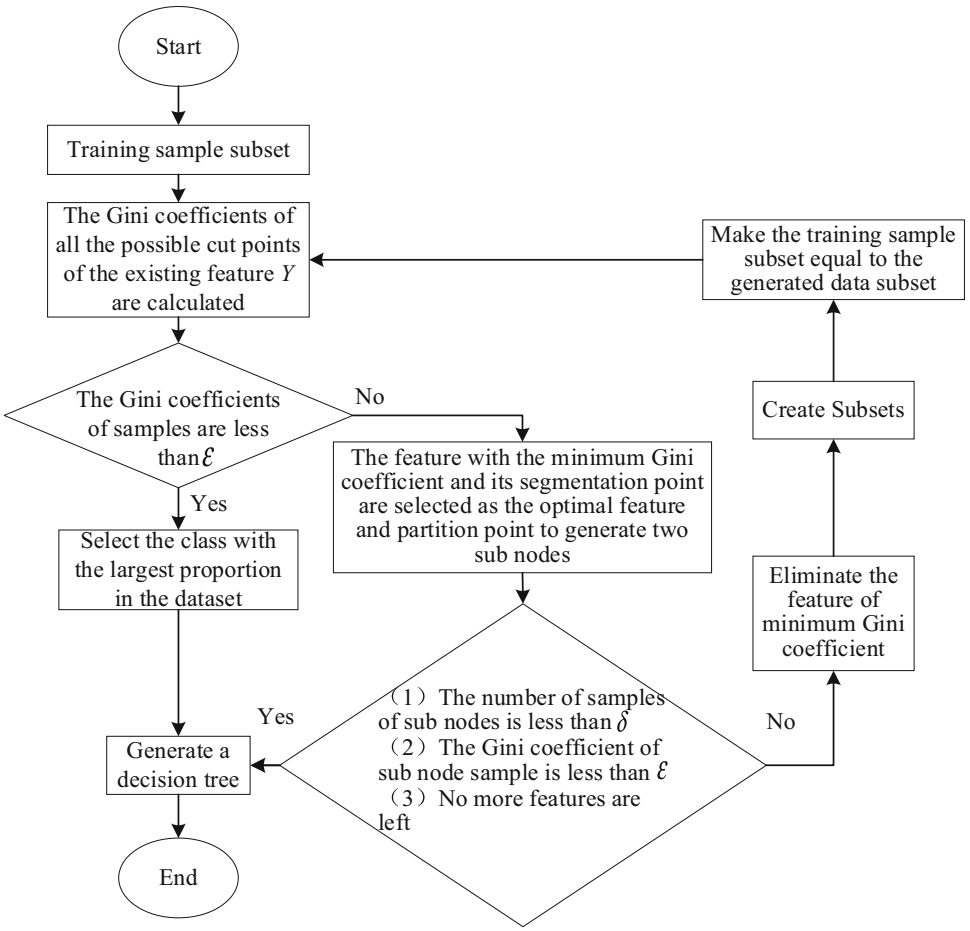

Fig. 4 Flowchart of decision tree construction

observation may appear many times in the bootstrap sample, or it may not appear at all.

With the original training sample set S_h as the input, S_h is composed of the power consumption of the whole society and its potential related factors, including GDP, investment in fixed assets of the whole society, number of star hotels and accommodation, population, sales of tap water, sales of natural gas, steel production, industrial added value, total import and export value, and automobile production. The resampling of S_h is conducted, and its working process is shown in Fig. 3.

Using Bootstrap sampling method, we randomly select w training sample subsets $S_{h1}, S_{h2}, \ldots, S_{hw}$ (each subset contains the abovementioned two types of data) from a to construct w classification and regression tree (CART). The test set is used to estimate the error of CART decision tree. By averaging the error estimates of w decision trees, the generalized error estimates of random forest can be obtained, and the accuracy of prediction model of power consumption can be quantitatively measured.

3.2.2 Construction of CART decision tree

Random forest is a combination of multiple decision trees. By voting the prediction results of each decision tree, the decision tree with the most votes is regarded as the final random forest prediction result.

CART algorithm constructs binary decision tree [30]. The CART algorithm selects the features by Gini coefficient when constructing decision tree. The principle is as follows:

If the sample data are divided into K classes and with the probability that the data belong to k class is p_k, then the Gini coefficient of the sample data is defined as.

$$\mathrm{Gini}(p) = \sum_{k=1}^{K} p_k(1 - p_k). \tag{5}$$

For the training sample set D, the number is $|D|$. With the assumption of K categories and the number of the k category is $|C_k|$, the Gini coefficient is as follows:

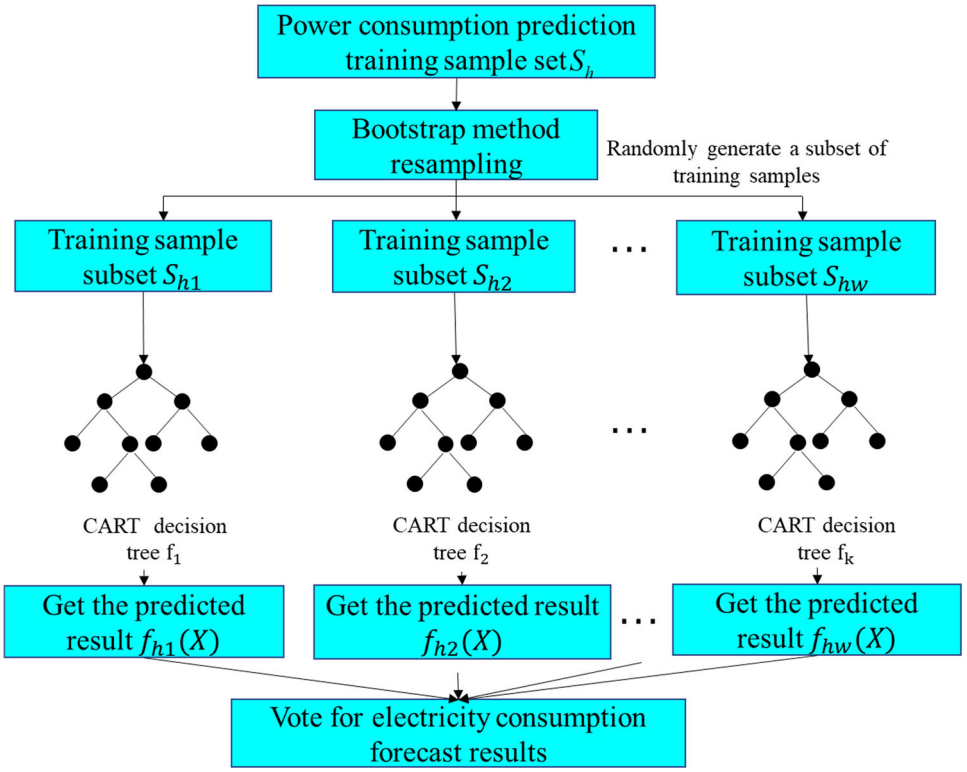

Fig. 5 Forecasting modeling of monthly electricity consumption based on random forest

$$\mathrm{Gini}(D) = \sum_{k=1}^{K} \frac{|C_k|}{|D|}\left(1 - \frac{|C_k|}{|D|}\right). \tag{6}$$

For the training sample set D, the number is $|D|$. According to the value of feature A_m, it can be divided into two parts D_1 and D_2. The number of D_1 and D_2 is $|D_1|$ and $|D_2|$, respectively. Under the condition of feature A_m, the Gini coefficient of D is as follows:

$$\mathrm{Gini}(D, A_m) = \frac{|D_1|}{|D|}\mathrm{Gini}(D_1) + \frac{|D_2|}{|D|}\mathrm{Gini}(D_2). \tag{7}$$

Gini coefficient $\mathrm{Gini}(D)$ represents the uncertainty of training sample set D under the condition of feature A_m, and the Gini coefficient of D represents the uncertainty of D under the condition of feature A_m. Therefore, Gini coefficient can represent the ability of feature A_m to classify data set D.

With the training sample subset S_{hj} as an example, S_{hj} is composed of feature set Y (Y is composed of GDP, investment in fixed assets of the whole society, number of star-rated hotels and accommodation, population, sales of tap water, sales of natural gas, steel production, industrial added value, total import and export value, and automobile production). For data set X of monthly electricity consumption of the whole society, the threshold value of node sample number is δ, and the threshold value of Gini coefficient is ε. CART binary decision tree is the output, and the construction process of decision tree is shown in Fig. 4.

First, the Gini coefficients of all the possible segmentation points of the feature set Y in the training sample subset S_{hj} are calculated. Then, whether the Gini coefficients of the samples are less than the given threshold ε is determined. If they are all less than the given threshold ε, then a single node tree is generated, whose category is the class with the largest number of samples in S_{hj}; otherwise, the feature with the smallest Gini coefficient and the corresponding segmentation point α are selected as the eigenvalue and segmentation standard of the root node. S_{hj} is divided into two subsets S_{hj1} and S_{hj2}, and S_{hj1} and S_{hj2} are allocated to the two sub nodes, respectively. Next, whether the number of samples of the sub nodes is less than the given threshold δ is judged. Whether the Gini coefficients of sub node samples are less than ε is also determined. If it is true, then the child node is a leaf node. If the two child nodes are leaf nodes, then the decision tree is generated; otherwise, for the non-leaf node, S_{hj} is equal to the corresponding data set of the child node, and the feature with the smallest Gini coefficient is removed. The Gini coefficients of all possible segmentation points of S_{hj} are recalculated.

ART algorithm is used to generate a decision tree for each subset of training samples based on the principle of minimum Gini coefficient, and w decision trees are symbiotic to form a "forest." Half of the strong correlation factors are randomly selected to participate in the node splitting process of decision tree for ensuring the randomness of decision tree

construction and avoiding over fitting problem. In addition, the number of decision trees in the whole random forest should be adjusted according to the prediction results.

The detailed steps of generating CART decision tree are given in Algorithm 2.

Algorithm 2: Decision tree generation based on CART algorithm

Input: Training set $S_{hj}=(X,Y_i)$; Node sample number threshold δ; Gini coefficient threshold ε

Output: A CART decision tree

1: **initialization:** $i=(1,\ldots,I)$; $k=1$
2: **if** all samples in S_{hj} belong to the same category C **then** generate a single node decision tree
3: **end**
4: **for** all $(X,Y_i)\in S_{hj}$ **do**
5: **Calculate** the Gini coefficient of all possible cut points of the current sample

$$Gini(p)=\sum_{k=1}^{K}p_k(1-p_k)$$

6: **if** $Gini(p)<\varepsilon$ **then** generate a single node tree, whose category is the class with the largest number of samples in S_{hj}
7: **else**
8: **Select** the feature with the smallest Gini coefficient and its division point as the optimal feature and optimal division point to generate two subsets $S_{hj,1}$ and $S_{hj,2}$, and assign them to the two sub nodes
9: **end**
10: **if** $Gini(S_{hj,1})<\varepsilon; Gini(S_{hj,2})<\varepsilon;$ $|S_{hj,1}|<\delta; |S_{hj,2}|<\delta$ and no more features are left **then** generate a tree with two child nodes as leaf nodes
11: **else**
12: For non-leaf nodes, let S_{hj} be equal to the data set corresponding to the child node, and remove the feature with the smallest Gini coefficient
13: **Continue:** (go to **step 4**)
14: **end**
15: **end**

3.3 Voting of forecast results of electricity consumption

The final output of the prediction model based on random forest algorithm is generated by voting:

$$F_h(X)=\arg\max_{Y}\sum_{i=1}^{w}I(f_{hi}(X)=Y), \tag{8}$$

where F_h is the prediction model of monthly electricity consumption; f_{hi} is a single decision tree prediction model; I (□) is an indicative function.

For the same data set, when w CART decision trees are constructed, w prediction results will be obtained. At this time, we need to vote and select the prediction result with the highest number of votes. The sample data are tested for simulation, the data X of correlation factors related to power consumption Y are taken as the input, the prediction result series $\{f_{h1}(X), f_{h2}(X), \ldots, f_{hw}(X)\}$ of each decision tree model are obtained, and voting is conducted to derive the final prediction result of power consumption. The process is shown in Fig. 5.

4 Example analysis

4.1 Data source

The monthly electricity consumption of the whole society, agriculture, forestry, animal husbandry and fishery, industry, finance, real estate, business and residential services, urban and rural residents' life, and the potential related factors of the abovementioned monthly electricity consumption are all from Huibo database, with the time span from January 2005 to December 2015. The aforementioned industries

Table 2 Data source

Month	Monthly electricity consumption of Shenyang (10^4kw h)	Monthly electricity consumption of agriculture, forestry, animal husbandry, and fishery in Shenyang (10^4kw h)	Accumulated value of GDP in Shenyang (10^{-2} billion RMB)	Shenyang registered residence population (person)	Water sales in Shenyang (10^4 t)	Sales volume of natural gas in Shenyang (10^4m^3)	Import and export value of Shenyang (10^4 dollar)	…	Shenyang automobile output (Vehicles)
2015/12	298,822.00	3317.00	2426.83	7,304,051.00	3685.40	5795.20	122,331.00	…	131,022.00
2015/11	283,009.00	3647.00	2426.83	7,302,132.00	4091.60	6410.30	109,438.00	…	109,751.00
2015/10	240,884.00	6474.00	2426.83	7,299,970.00	3457.80	4498.40	102,193.00	…	72,407.00
2015/9	235,462.00	10,108.00	1734.00	7,298,205.00	3233.00	4325.00	101,537.00	…	71,603.00
2015/8	264,925.00	10,363.00	1734.00	7,297,945.00	3571.80	4480.60	91,997.00	…	69,499.00
2015/7	266,009.00	8788.00	1734.00	7,300,029.00	3635.80	4388.10	102,433.00	…	69,820.00
2015/6	238,483.00	7724.00	1114.05	7,302,114.00	3922.90	4499.20	193,261.00	…	85,379.00
2015/5	236,996.00	4466.00	1114.05	7,304,082.00	3757.90	4834.70	146,381.00	…	84,932.00
…	…	…	…	…	…	…	…	…	…
2005/4	113,676.00	2171.00	287.85	6,956,000.00	2286.00	952.00	38,316.00	…	12,236.00
2005/3	136,472.00	1436.00	126.52	6,950,000.00	2367.00	1133.00	53,135.00	…	12,821.00
2005/2	121,315.00	923.00	126.52	6,942,000.00	2086.00	1192.00	31,001.00	…	6835.00
2005/1	147,616.00	1160.00	126.52	6,939,000.00	2136.00	1470.00	27,301.00	…	14,732.00

cover 6 categories: transportation, storage, postal, commerce, accommodation, and catering. A total of 14 potential related factors are also considered: GDP of Shenyang, GDP of the secondary industry of Shenyang, investment in fixed assets of Shenyang (production and supply of electricity, heat, gas, and water), investment in fixed assets of Shenyang (transportation, gas, and water), the total number of accommodation and reception in Shenyang star hotels, the total population of Shenyang registered residence, the volume of tap water sold in Shenyang, the sales of natural gas in Shenyang, the output of industrial steel above designated size in Shenyang, the industrial crude oil production above designated Size in Shenyang, the industrial added value above the Shenyang scale, the sales volume of wholesale and retail trade enterprises above the quota of Shenyang City, Shenyang's total import and export value, and Shenyang's industrial automobile output above designated size, as shown in Table 2.

The GDP data of potential related factors are quarterly data, but this study needs to use monthly data. Thus, the quarterly GDP data are processed by the way of average distribution to each month.

4.2 Identification of power consumption-related factors

The monthly power consumption data are taken as the explanatory variable X, and the matrix $X = \{X_1, X_2, X_3, X_4, X_5\}$ is set. Among them, X_1, X_2, X_3, X_4, X_5 represent the monthly power consumption of agriculture, forestry, animal husbandry, and fishery in Shenyang, the monthly energy consumption of Shenyang's industry, the monthly power consumption of finance, real estate, business, and residential service industry, the monthly consumption of urban and rural residents in Shenyang, and the monthly power consumption of the whole society of Shenyang.

The monthly data of potential related factors are taken as conditional variable Y, and $Y = \{Y_1, Y_2, \ldots, Y_{14}\}$ is set. Among them, Y_1 is Shenyang GDP, Y_2 is the second industry GDP of Shenyang, Y_3 is the fixed assets investment (production and supply of electricity, heat, gas, and water) in Shenyang, Y_4 is the fixed assets investment (transportation, storage, and postal industry) of Shenyang, Y_5 is the total number of hotel accommodation and reception in Shenyang City, Y_6 is the total population of Shenyang household registration, Y_7 is the sales volume of tap water in Shenyang, Y_8 is the sales volume of natural gas in Shenyang, Y_9 is the industrial steel production above Shenyang scale, Y_{10} is the industrial crude oil production above Shenyang scale, Y_{11} is the industrial added value above Shenyang scale, Y_{12} is the sales volume of wholesale and retail trade enterprises above the quota of Shenyang, Y_{13} is the total import and export value of Shenyang, and Y_{14} is the output of industrial vehicles above Shenyang.

Table 3 Maximum mutual information coefficient results

Y	X				
	X_1	X_2	X_3	X_4	X_5
Y_1	0.41	0.60	0.44	0.42	0.69
Y_2	0.43	0.55	0.48	0.43	0.70
Y_3	0.43	0.39	0.46	0.30	0.45
Y_4	0.40	0.53	0.36	0.38	0.49
Y_5	0.38	0.45	0.33	0.31	0.51
Y_6	0.32	0.82	0.75	0.80	0.96
Y_7	0.30	0.77	0.58	0.64	0.87
Y_8	0.35	0.75	0.81	0.83	0.90
Y_9	0.28	0.65	0.49	0.63	0.76
Y_{10}	0.28	0.42	0.41	0.38	0.42
Y_{11}	0.29	0.72	0.55	0.61	0.82
Y_{12}	0.42	0.48	0.29	0.32	0.51
Y_{13}	0.31	0.73	0.52	0.54	0.83
Y_{14}	0.34	0.75	0.50	0.65	0.88

The maximum mutual information coefficient of explanatory and conditional variables is analyzed by Python, and the maximum mutual information coefficient is obtained. Thus, a correlation coefficient table, as shown in Table 3, is formed.

Python's Seaborn and Matplotlib packages are used to analyze the maximum mutual information coefficient data in Table 3 for more intuitively identifying the strong correlation factors of monthly electricity consumption. The data are displayed in the form of a heat map, as shown in Fig. 6.

Figure 6 shows that the maximum mutual information coefficients of X_1 and Y_1, Y_2, Y_3,Y_4, Y_5, Y_{12} are relatively large and have strong correlation. The maximum mutual information coefficients of industrial monthly electricity consumption X_2 and Y_6, Y_7, Y_9, Y_{11}, Y_{13}, Y_{14} in Shenyang are relatively large and have strong correlation. The maximum mutual information coefficients of X_3 and Y_6, Y_8, Y_7, Y_{11}, Y_{13}, Y_{14} are relatively large, and the correlation is strong. The maximum mutual information coefficient of X_4 and Y_6, Y_7, Y_8, Y_9, Y_{11}, Y_{13}, Y_{14} is larger, and the correlation is strong. The maximum monthly mutual information coefficients of X_5 and Y_6, Y_7, Y_8, Y_9, Y_{11}, Y_{13}, Y_{14}, Y_1 and Y_2 in Shenyang are relatively large, and the correlation is strong. The correlation is also affected by the total population of Shenyang registered residence Y_6, the volume of Shenyang tap water sales Y_7, the sales volume of Shenyang natural gas Y_8, the industrial steel output above the scale of Shenyang Y_9, the industrial added value above Shenyang scale Y_{11}, the total import and export value of Shenyang Y_{13}, and the output of industrial vehicles above designated scale Y_{14}.

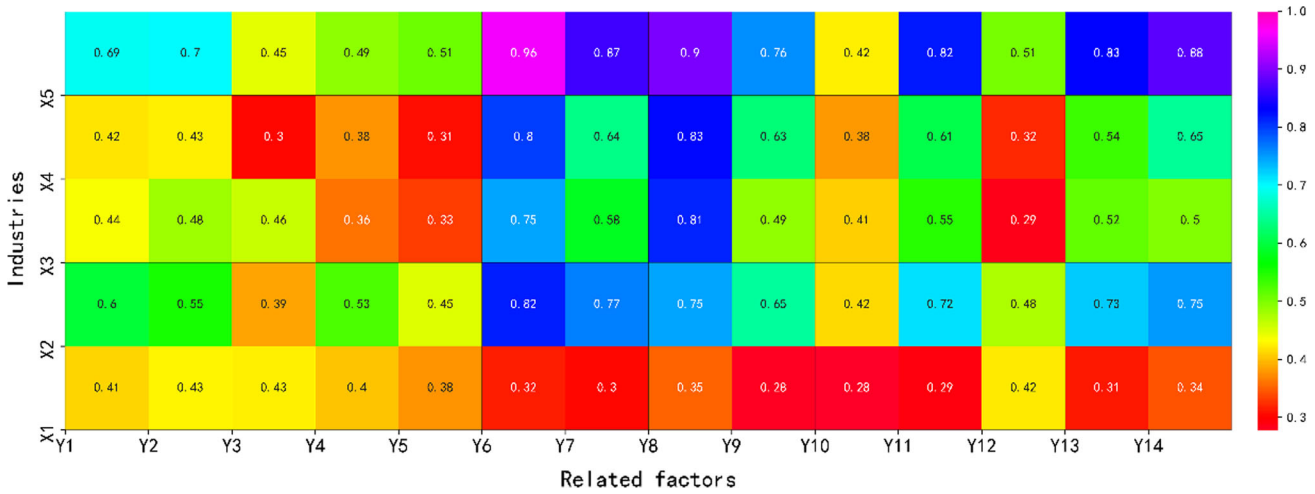

Fig. 6 Maximum mutual information coefficient results

4.3 Forecast of monthly electricity consumption

The forecast of monthly electricity consumption of Shenyang from January to December is conducted. With the monthly data of 9 strong correlation factors from 2005 to 2014 as the input and the monthly power consumption of Shenyang as the output, 12 original training sets from January to December are formed. Then, the random forest algorithm is used to predict the monthly power consumption. When using random forest algorithm to forecast electricity consumption, if we use monthly data to forecast directly, then the relationship between the data can be determined well, which greatly reduces the prediction accuracy. This study transforms the strong correlation factors and monthly electricity consumption data into monthly year-on-year growth rate as the input and output of the forecast to ensure the accuracy and stability of the forecast. The monthly year-on-year growth rate $R_{m,n}$ is calculated as follows:

$$R_{m,n} = \frac{d_{m,n} - d_{m-1,n}}{d_{m-1,n}} \times 100\%, \tag{9}$$

where $d_{m,n}$ represents the monthly data of the n month of the m year.

Python's train test split package is used to divide the training and test sets. Random forest classifier package is called, and bootstrap parameter is set to true.w training sample subsets are selected from the returned samples in the training set, and w decision trees are generated from these training sample subsets. The test set is used to estimate the error of random forest prediction model. When each decision tree is generated, half of the strong correlation factors are randomly selected as random characteristic variables to participate in the node splitting process.

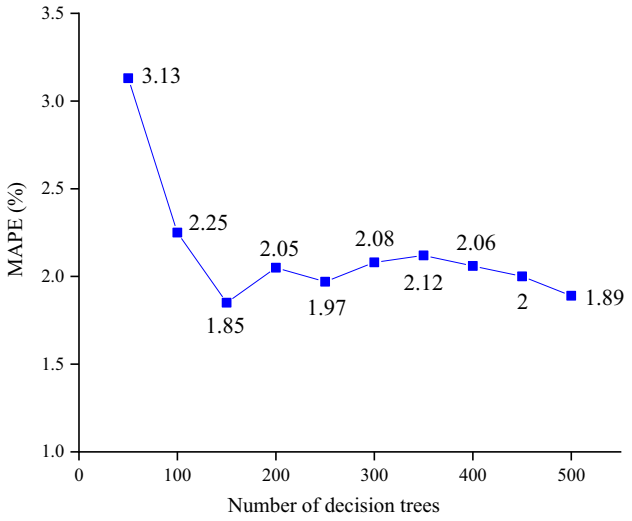

Fig. 7 Error analysis of different numbers of decision trees

Given different decision trees, the prediction accuracy of random forest algorithm will be different. The mean absolute percentage error (MAPE) is used to calculate the error value, and the formula is as follows:

$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{y_{f(t)} - y_{a(t)}}{y_{a(t)}}\right| \times 100\%, \tag{10}$$

where $y_{f(t)}$ is the predicted value, and $y_{a(t)}$ is the actual value.

With August as an example, Fig. 7 shows the MAPE between the predicted value and the actual value of the monthly electricity consumption forecast based on the random forest algorithm when taking different decision trees.

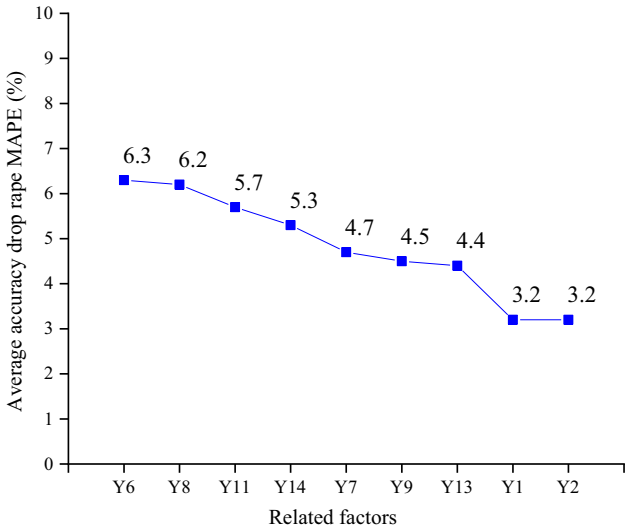

Fig. 8 Influence of different correlation factors on prediction accuracy

As shown in the figure, MAPE tends to a certain value with the increase in decision trees. However, when more decision trees are considered, the amount of calculation will increase rapidly and the prediction time will be longer. A total of 150 decision trees are selected to form a random forest to guarantee modeling speed and avoid prediction error.

The importance of each strong correlation factor to the prediction model differs. We adopt the way of average accuracy decline rate to intuitively determine the importance of each strong correlation factor to the prediction model. After a strong correlation factor is removed, the degree of prediction accuracy declines. The more decline in accuracy means the more important that this strong correlation factor is to the prediction model. The importance of the nine strong correlation factors is shown in Fig. 8. As shown in the figure, Y_6, Y_8, Y_{11}, and Y_{14} are of high importance, which is consistent with the analysis results in Fig. 6.

With the 73rd decision tree in the monthly electricity consumption forecast of random forest in August 2015 as an example, the working process of CART decision tree is analyzed. The 73rd CART decision tree is shown in Fig. 9.

In the figure, Y_1, Y_6, Y_7, Y_8, and Y_{13} are the strong correlation factors of Shenyang's social monthly electricity consumption in August 2015; gini is Gini coefficient, which is used for purity measurement. If all the training samples contained in a node are of the same category, then node is pure (gini $= 0$); samples represents how many training sample instances the current node is applied to; value represents the number of samples for each category in the current node; class is the classification result.

The specific prediction process of CART decision tree is as follows. Starting from the root node, whether Y_2 is less than or equal to 5.1 is determined; if yes, then the process moves to the left child node; otherwise, it moves to the right child node. In the specific power consumption prediction process, when $Y_2 = 0.33 < 5.1$, the process moves to the left sub node. Then, when $Y_2 = 0.33 < 4.215$, it moves to the left node. Finally, when $Y_7 = 4.6 > 3.35$, it moves to the right sub node because gini $= 0$ determines that the year-on-year growth rate of power consumption in August 2015 is the same as that in August 2014. The year-on-year growth rate of power consumption in August 2014 is taken as the forecast value of year-on-year growth rate of power consumption in August 2015, which is 3.78. It is converted into monthly power consumption of 2634.47 million kwh.

4.4 Analysis of prediction results

The random forest algorithm is used for prediction by using the data of strong correlation factors in 2015. On the basis of obtaining the monthly growth rate of electricity consumption, the monthly electricity consumption of the same period of the previous year is taken as the benchmark to obtain the monthly electricity consumption forecast value. At the same time, the monthly data considering all factors and the monthly growth rate data considering all factors are taken as the training samples, and the random forest algorithm is used for prediction and comparison with the proposed method. Forecast result 1 is that of the proposed method, and the training samples are the monthly growth rate data. Forecast result 2 uses the monthly year-on-year growth rate as the training sample and considers all factors. Forecast result 3 uses monthly data as training samples and considers all factors.

At the same time, because support vector machines are widely used in monthly electricity consumption forecasting, in order to further verify the effectiveness of the method proposed in this article, the support vector machine is used to compare with the method proposed in this article. The forecast result 4 is based on the monthly data considering all factors as the training sample, and the support vector machine is used for prediction. The actual monthly electricity consumption and forecast results of Shenyang in 2015 are shown in Table 4 and Fig. 10.

As shown in Fig. 10 and Table 4, the prediction accuracy of training samples using monthly year-on-year growth rate is higher than that of training samples using monthly data directly under the condition of not using mutual information to screen correlation factors. On the basis of using the monthly growth rate, the prediction accuracy of using the monthly growth rate data of strong correlation factors as the training sample is higher than that of using the monthly growth rate data of strong correlation factors as the training sample. Therefore, better results are obtained when more correlation factors are considered. Too many factors with low correlation will make the prediction result worse. When the

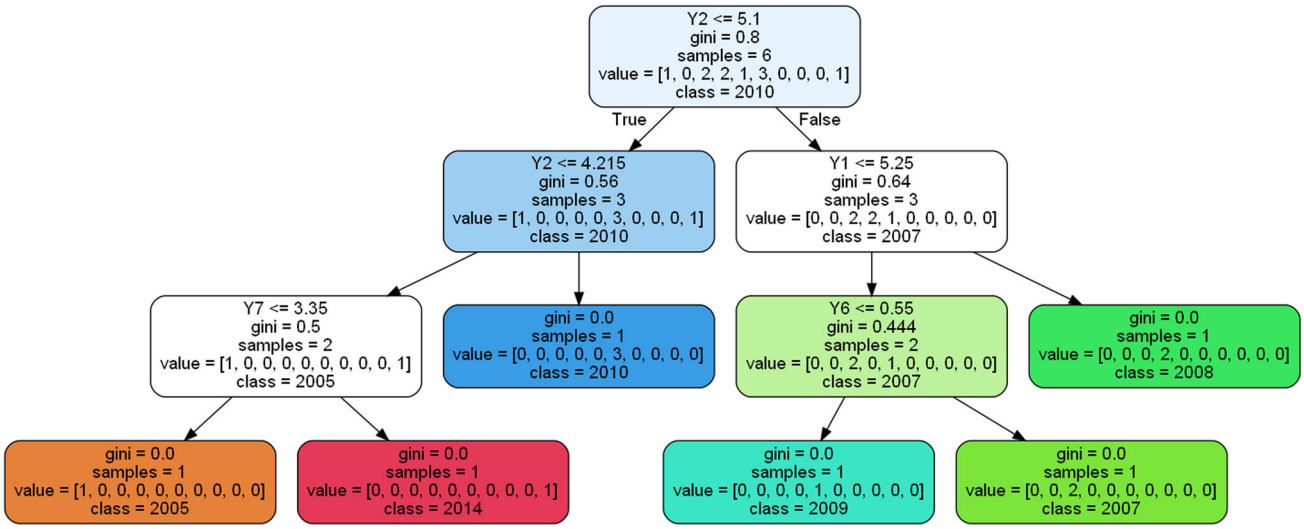

Fig. 9 CART decision tree

Table 4 Comparison of prediction result

Month	Actual value (10^4 kW h)	Forecast value (10^4 kW h)			
		Forecast results 1	Forecast results 2	Forecast results 3	Forecast results 4
1	291,777	294,179	294,527	278,649	256,573
2	225,499	228,880	228,880	224,041	211,243
3	258,137	264,281	267,074	258,643	241,613
4	232,390	229,050	229,050	232,115	219,709
5	236,996	245,140	245,140	232,031	219,945
6	238,483	240,859	240,859	235,030	217,422
7	266,009	265,975	273,931	242,063	240,573
8	264,925	263,447	263,447	253,851	240,003
9	235,462	237,586	237,586	213,471	212,331
10	240,884	239,168	239,168	230,029	223,022
11	283,009	276,543	276,543	258,735	249,380
12	298,822	287,618	287,922	278,979	264,577
MAPE		1.53%	1.88%	4.25%	8.82%

factors with high correlation are screened out by mutual information, the prediction error will be significantly reduced. Compared with the support vector machine algorithm, the method proposed in this paper improves the MAPE index by 7.29%. The experimental results show that the method proposed in this paper has a better prediction effect.

5 Conclusion and prospect

In this study, the maximum mutual information coefficient is introduced to identify the influencing factors of the monthly electricity consumption of the whole society in Shenyang, and the strong correlation factors of the monthly electricity consumption of the whole society are selected. The random forest algorithm is used to predict the monthly electricity consumption of the whole society with the strong correlation factors as the input. The predicted value of the monthly electricity consumption of the whole society is obtained. The effectiveness and correctness of the proposed method are verified by an example.

(1) The maximum mutual information coefficient of mutual information theory is used to quantitatively calculate the correlation between the influencing factors and the

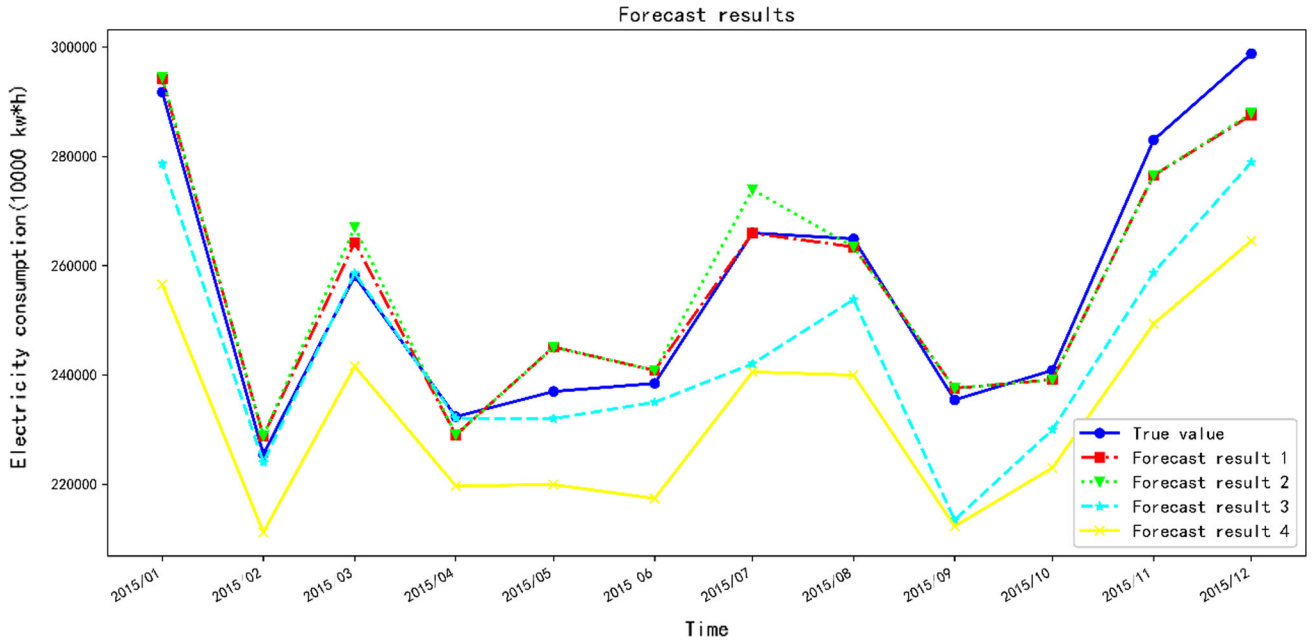

Fig. 10 Forecast results of monthly electricity consumption of Shenyang in 2015

monthly electricity consumption of the whole society. With the maximum mutual information coefficient, more effective correlation factors are screened out among many influencing factors.

(2) The random forest algorithm is used for prediction. The bootstrap resampling of random forest algorithm and the random selection of features enable the algorithm to avoid over fitting and make it suitable for all kinds of data sets. Combined with mutual information, the factors with low correlation to the whole society's monthly electricity consumption are eliminated. As a result, the prediction accuracy is higher.

(3) The monthly forecast of random forest is conducted using the strategy of monthly forecast with the historical data of the same month as the training set. As a result, the prediction accuracy is improved.

In addition to the factors mentioned in this article, monthly electricity consumption forecasts are also greatly affected by economic factors. The energy consumption of heating or cooling will be generated when the temperature is higher than the upper limit or lower than the lower limit of the comfortable temperature range. Thus, the weather data will be introduced in the next work to further improve the prediction accuracy.

Acknowledgements This work was partly supported by the National Natural Science Foundation of China (61773269), the Natural Science Foundation of Liaoning Province of China (2019-KF-03-08), the Program for Liaoning Excellent Talents in University (LR2019045), and the Program for Shenyang High Level Innovative Talents (RC190042).

Declaration

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Wu D, Wang B, Precup D, Boulet B (2020) 'Multiple kernel learning-based transfer regression for electric load forecasting.' IEEE Trans Smart Grid 11(2):1183–1192
2. Wang Y, Chen Q, Hong T, Kang C (2019) Review of smart meter data analytics: applications, methodologies, and challenges. IEEE Trans Smart Grid 10(3):3125–3148
3. Zhang S, Liu J, Zhao B et al (2013) Cloud computing-based analysis on residential electricity consumption behavior. Power Syst Technol 37(6):1542–1546 (**(in Chinese)**)
4. Liu L, Wang Y, Pang X et al (2020) A comprehensive forecasting method of monthly electricity sales based on STL model. Control Eng China 27(11):1930–1936 (**(in Chinese)**)
5. González JP, Muñoz San Roque AMS, Pérez EA (2018) Forecasting functional time series with a new Hilbertian ARMAX model: application to electricity price forecasting. IEEE Trans Power Syst 33(1):545–556
6. Medina Macaira P, Castro Sousa R, Cyrino Oliveira FL (2016) Forecasting Brazil's electricity consumption with Pegels exponential smoothing techniques. IEEE Latin Am Trans 14(3):1252–1258
7. Chen P, Pedersen T, Bak-Jensen B, Chen Z (2010) ARIMA-based time series model of stochastic wind power generation. IEEE Trans Power Syst 25(2):667–676
8. Zhang Y, Sun H, Guo Y (2019) Wind power prediction based on PSO-SVR and grey combination model. IEEE Access 7:136254–136267
9. Ceperic E, Ceperic V, Baric A (2013) A strategy for short-term load forecasting by support vector regression machines. IEEE Trans Power Syst 28(4):4356–4364

10. Aprillia H, Yang H-T, Huang C-M (2021) Statistical load forecasting using optimal quantile regression random forest and risk assessment index. IEEE Trans Smart Grid 12(2):1467–1480
11. Wang Y, Zhang N, Tan Y et al (2019) Combining probabilistic load forecasts. IEEE Trans Smart Grid 10(4):3664–3674
12. Bracale A, Caramia P, Falco P et al (2020) Multivariate quantile regression for short-term probabilistic load forecasting. IEEE Trans Power Syst 35(1):628–638
13. Xun G, Julian LC, Eduardo C et al (2019) Bottom-up load forecasting with Markov-based error reduction method for aggregated domestic electric water heaters. IEEE Trans Ind Appl 55(6):6401–6413
14. Wang Yi, Chen Q, Zhang N et al (2018) Conditional residual modeling for probabilistic load forecasting. IEEE Trans Power Syst 33(6):7327–7330
15. Taylor JW, Buizza R (2002) Neural network load forecasting with weather ensemble predictions. IEEE Trans Power Syst 17(3):626–632
16. Li B, Zhang J, He Y, Wang Y (2017) Short-term load-forecasting method based on wavelet decomposition with second-order gray neural network model combined with ADF Test. IEEE Access 5:16324–16331
17. Senjyu T, Takara H, Uezato K, Funabashi T (2002) One-hour-ahead load forecasting using neural network. IEEE Trans Power Syst 17(1):113–118
18. Ranaweera DK, Karady GG, Farmer RG (1996) Effect of probabilistic inputs on neural network-based electric load forecasting. IEEE Trans Neural Netw 7(6):1528–1532
19. Jiang H, Zhang Y, Muljadi E, Zhang JJ, Gao DW (2018) A short-term and high-resolution distribution system load forecasting approach using support vector regression with hybrid parameters optimization. IEEE Trans Smart Grid 9(4):3341–3350
20. Li G, Li Y, Roozitalab F (2020) Midterm load forecasting: a multistep approach based on phase space reconstruction and support vector machine. IEEE Syst J 14(4):4967–4977
21. Chen B-J, Chang M-W, Lin C-J (2004) Load forecasting using support vector machines: a study on EUNITE competition 2001. IEEE Trans Power Syst 19(4):1821–1830
22. Chen H, Liu W, Li Y (2020) Medium-term load forecast based on singular spectrum analysis and neural network. Power Syst Technol 44(4):1333–1347 ((**in Chinese**))
23. Afrasiabi M, Mohammadi M, Rastegar M et al (2020) Deep-based conditional probability density function forecasting of residential loads. IEEE Trans Smart Grid 11(4):3646–3657
24. Von Krannichfeldt L, Wang Y, Hug G (2021) Online ensemble learning for load forecasting. IEEE Trans Power Syst 36(1):545–548
25. Zhao T, Wang L, 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. Proceed CSEE 36(3):604–614 ((**in Chinese**))
26. Kiernan L, Kambhampati C, Mitchell RJ et al (1995) Automatic integrated system load forecasting using mutual information and neural networks. IFAC Proc Vol 28(26):503–508
27. Gu T, Guo J, Li Z, Mao S (2021) Detecting associations based on the multi-variable maximum information coefficient. IEEE Access 9:54912–54922
28. Zhen L, Karam LJ (2005) Mutual information-based analysis of JPEG2000 contexts. IEEE Trans Image Process 14(4):411–422
29. Xuan Y et al (2021) Multi-model fusion short-term load forecasting based on random forest feature selection and hybrid neural network. IEEE Access 9:69002–69009
30. Liu F, Dong T, Hou T, Liu Y (2021) A hybrid short-term load forecasting model based on improved fuzzy c-means clustering, random forest and deep neural networks. IEEE Access 9:59754–59765