International Journal of Database Theory and Application
Vol.7, No.2 (2014), pp.13-24
http://dx.doi.org/10.14257/ijdta.2014.7.2.02

Research and Application of Data Mining and NARX Neural Networks in Load Forecasting

Li Xiaofeng[1] and Yang Chunshan[2]

[1]*Department of Informatic Science, Heilongjiang International University, Harbin 150025, China*
[2]*Department of Computer Science and Technology, Cheng-Dong College of Northeast Agricultural University, Harbin 150025, China*
[1]*mberse@126.com*, [2]*ycszy1999@126.com*

Abstract

The relationship between med-long term load forecasting and socio-economic indicators is very difficult to describe with an accurate mathematical model. The paper introduce data mining technology into the association analysis of China's electricity consumption growth, select many socio-economic indicators since 2000, constitute the relevant factors database, complement of a few missing data, and dig out a number of indicators closely related to the electricity consumption with cluster analysis, and the data of distortion indicators is corrected, thus, build a more scientific load forecasting model. Validate and test the correlation of electricity consumption and selected indicators by dynamic neural network time sequence tool. The results show that the prediction model has good convergence, and the effect is satisfactory.

Keywords: *load forecasting, Cluster analysis, Association analysis, Complement of data, NARX network*

1. Introduction

As the basis of planning, design, scheduling power system, importance of load forecasting has been recognized for a long time. In accordance with different prediction period, load forecasting can be divided into annual forecast, monthly forecast, day forecast, hour forecast. Except for per unit consumption, elastic coefficient, trend analysis and other traditional method, there are some other annual load forecasting methods, *e.g.*, regression analysis, gray prediction, neural network and so on [1-2].

The med-long term load forecasting required the introduction of socio-economic indicators, because almost all of socio-economic indicators have large or small impact on power Load, workload and amount of calculation are difficult for us to accept if all socio-economic indicators are introduced. Many researchers all over the world generally introduce a number of socio-economic indicators by experience, gross domestic product (GDP), total value of each industry, population and others are often selected [3-4]. However, there is not a precise theoretical foundation in that way.

With the development and application of data mining technology, its analysis way and computing algorithm have been effectively applied on prediction, classification, clustering analysis, correlation analysis and other areas of power systems, and improve the ability of researchers to deal with historical data [5-6]. In this paper, data mining technology is applied to mid-long term load forecasting, and select some socio-economic indicators as candidate from all socio-economic indicators from 2000 to 2010, then dig out a number of indicators

ISSN: 2005-4270 IJDTA

closely related to the electricity consumption by calculating the degree of association between the electricity consumption and those selected socio-economic indicators; Finally, verify the selected indicators based on NARX neural network, obtain general growth rule of the electricity consumption [7].

2. Relevant Work

2.1. The Complement of Missing Data

Being easy to missing the historical data used in med-long term load forecasting, the predictive performance of the model will be decreased if we abandon the socio-economic indicators that containing missing data [8]. Therefore, reasonable measures should be taken to the complement missing data. If the first or the end of the data is missing, trend in the proportion can be used to data complement, series $\{X\}$ is defined as:

$$\{X\} = [\phi(1), X(2), X(3)\cdots\phi(n)] \quad (1)$$

Here, $\phi(1)$ and $\phi(n)$ are missing data, and complement them as follows:

$$\phi(1) = [X(2)]^2 / X(3) \quad (2)$$

$$\phi(n) = [X(n\text{-}1)]^2 / X(n\text{-}2) \quad (3)$$

If the intermediate data is missing, Non-ortho mean generation method can be used, series $\{X\}$ is defined as:

$$\{X\} = [X(1), X(2), \cdots X(k-1), \phi(k), X(K+1), \cdots X(n)] \quad (4)$$

Here, $\phi(k)$ is missing data, and complement it as follows:

$$\phi(k) = 0.5X(k-1) + 0.5X(k+1) \quad (5)$$

2.2. Data Normalization Methods

The candidate indicators selected from the Yearbook have different dimensionless, and the magnitude of different indicators also differ greatly. So, before learning and training as well as the load forecasting based on neural network, input and output data must be normalized, so that all indicators have comparability to each other. The two most common methods of data normalization are:

2.2.1. The Extremum Method: The candidate indicators will be transformed into dimensionless number by the maximum and the minimum values of the indicators, common formula is defined as:

$$x_i' = \frac{x_i - \min x_i}{x_i} \quad (6)$$

$$x_i' = \frac{x_i - \min x_i}{\max x_i - \min x_i} \quad (7)$$

2.2.2. The Standard Deviation Method: This method is calculated by the following equations:

$$x'_i = \frac{x_i - \bar{x}}{s} \quad (8)$$

Where, $s = \sqrt{\frac{1}{n}\sum (x_i - \bar{x})^2}$, $\bar{x}$ is the average value of the indicators, x'_i is standardized value, and its results would exceed the range from 0 to 1.

2.3. Cluster Analysis

Data mining is a process of finding new meaningful relationships, patterns and trends from the large amounts of data by using pattern recognition, statistical and mathematical techniques. Data mining is an, quite simply, information technology to dig out some potential value from massive data. It is able to achieve data classification, clustering, correlation analysis and forecasting.

The cluster analysis can find the association rule between these dataset, then divide these dataset into several categories. Usually the cluster analysis is divided into the Q-type and the R-type. The former can measure the degree of similarity between samples. When describe the nature of different aspects of each sample by p indicators, there produce a p-dimensional vector. Then n samples are seen as n points in the p-dimensional vector space, consequently the degree of similarity between each samples can be measure by the distance between two points in the p-dimensional space. Let us suppose that d_{ij} is the distance from X_i and X_j. There are several common calculation formula of distance.

2.3.1. The Minkowski Distance: The formula of Minkowski distance is defined as the following equation:

$$d_{ij}(q) = (\sum_{k=1}^{p} | X_{ik} - X_{jk} |^q)^{1/q} \quad (9)$$

Here, q is a variable parameters.

When q=1, the equation is formula of Manhattan distance.

When q=2, the equation is formula of Euclidean distance.

When q=∞, the equation is formula of Chebyshev distance.

There are two problems when above all algorithms encounter multivariate data analysis. First, there not take the overall variation to the influence of distance into account; second, the Minkowski Distance is affected by the variable dimension, and which is unfavorable for the multivariate data analysis.

2.3.2. Standardization Euclidean Distance: As an improved scheme for the shortcomings of the Euclidean distance, standardization Euclidean distance would standardize each component to equal in mean value and variance, and then standardize variable mathematical expectation to 0, and variance to 1.The standardization process is described by formula:

$$X^*_i = \frac{X_i - m}{s} \quad (10)$$

Where X^*_i is standardized value, m is the mean value of component, s is the variance of component. We can obtain formula of standardization Euclidean distance between X_i vector and X_j vector after derivation.

$$d_{ij} = \sqrt{\sum_{k=1}^{n} \left(\frac{X_{ik} - X_{jk}}{s_k}\right)^2} \quad (11)$$

2.4. Correction in Distortion of Data

First, there is a preliminary analysis for chosen socio-economic indicators, and draw the dynamic line chart, then observe changes in the trajectory from the line chart, and analysis of the reason to bring abnormal values and turning point, and determine whether they are distorted data. If historical data sequence $\{X\}$ is defined as:

$$\{X\} = [X(1), X(2), \cdots X(n)] \quad (12)$$

Then, suppose $X_0 = [\sum_{i=1}^{n} X(i)] / n$.

If

$$X(i) > X_0 \times 1.2 \quad (13)$$

Or

$$X(i) < X_0 \times 0.8 \quad (14)$$

Then $X(i)$ can be seen as distorted data, regard it as missing data, and complement it use equation (2) ~ (5).

2.5. Dynamic Neural Networks

Neural networks can be classified into dynamic and static categories, and dynamic neural networks can also be classified into feedback and no feedback categories. In no feedback dynamic neural networks, the output of the network depends not only on the current input to the network, but also on the previous inputs to the network. In feedback dynamic neural networks, the output of the network depends not only on the current input to the network, but also on the previous inputs, outputs, or states of the network. One principal application of dynamic neural networks is in time series prediction [9].

Time series prediction studies the trends of process time series of the predicted target. Neural networks time series tools in MATLAB provide three categories to solve three different kinds of nonlinear time series problems [10-11].

- Nonlinear autoregressive with external input（NARX）
- Nonlinear autoregressive(NAR)
- nonlinear input-output

NARX used in the paper is a feedback dynamic neural network, and it also can be seen as the BP neural networks that depends on the previous inputs and outputs to the network [12]. The structure of NARX network used in the paper as shown in Figure 1.

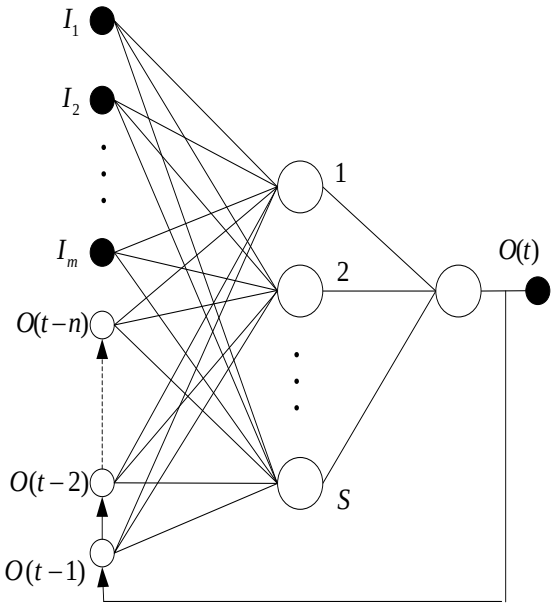

Figure 1. The Structure of NARX Network

As seen from Figure 1, the inputs of NARX network consist of two parts: the external input and the previous outputs to the network. We assume that $I(t)$ and $O(t)$ respectively are external input and output of network at time t, n_o is delay time of output feedback, n_i is delay time of external input, thus the output of NARX network as follows:

$$O(t) = f(O(t-1), O(t-2), \ldots, O(t-n_o), I(t-1), I(t-2), \ldots, I(t-n_i)) \quad (15)$$

Where the output of the network depends not only on $I(t-i)\ i = 1 \cdots n_i$, current and previous input to the network, but also on $O(t-i)\, i = 1 \cdots n_o$, previous outputs to the network.

3. Experimental Analysis and Results

3.1. Data Collection and Complement

The data set in the paper is selected from China statistical yearbook and power yearbook, and the core object is electricity consumption. It is advisable to select 5-10 years of data set in annual load forecasting. Therefore, the data set I selected span from 2000 to 2010. I select 22 indicators as candidate from the Statistical Yearbook, which can characterize development of economic from eight aspects. But consumption by industry in CHINA Statistical Yearbook 2011 is only statistical to 2009, so there need to complement its data for 2010 by formula (3).

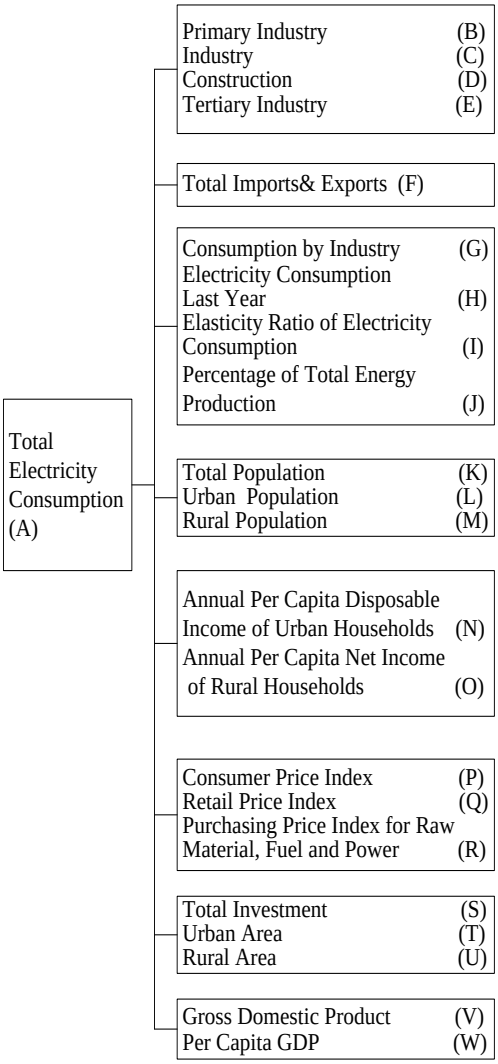

Figure 2. The Relevant Factors Database

3.2. Data Normalization

The Extremum method and the standard deviation method have their own characteristics; the Extremum method is not very stringent on the number of indicators and the requirement in distribution, and standardized value between 0 and 1. However, the standard deviation method is usually used when data of indicators show a normal distribution, and standardized value will exceed the range from 0 to 1, which affects further data processing. I consider that actual value, maximum and minimum impact on standardized value, then select the extremum method here, use formula (7) for data processing.

3.3. Cluster Analysis

According to standardized value and equation (9) ~ (11), I calculate the distance between classes, which indicate the distance between electricity consumption and other candidate indicators. If the distance between classes is smaller, the indicator is more closely related to

electricity consumption, and better reflects the general law of the electricity consumption growth. Computed results are shown in Table 1.

Table 1. The Distance between Electricity Consumption and other Candidate Indicators

if Indicators is	*Then* Euclidean distance is	*Then* Chebychev distance is	*Then* Manhattan distance is
B	0.2769	0.1823	0.6922
C	0.2786	0.1126	0.8138
D	0.5083	0.2109	1.5188
E	0.5676	0.2859	1.6031
V	0.3097	0.1363	0.9079
W	0.2748	0.1253	0.7951
F	0.3018	0.1404	0.8332
S	0.7411	0.2816	2.2615
T	0.7669	0.2889	2.3429
U	0.5722	0.2341	1.7253
N	0.1393	0.0759	0.3629
O	0.0995	0.0601	0.2481
P	1.3502	0.6241	3.8834
Q	1.3203	0.6062	3.7974
R	1.2940	0.6195	3.6814
J	0.9385	0.5226	2.4022
G	0.0321	0.0221	0.0791
I	1.1939	0.5504	3.5853
K	1.3676	0.6212	3.9540
L	0.8545	0.3681	2.4942
M	1.3518	0.6760	3.7209
H	0.1025	0.0720	0.1950

As can be seen from Table 1, elasticity radio of electricity consumption, total population, rural population and price index have larger distance between classes, which illustrate that there is less relevance between change of these indicators and growth of electricity consumption. But industry, consumption by industry, electricity consumption last year and GDP have smaller distance between classes, which illustrate that there is more relevance between change of these indicators and growth of electricity consumption. Also can be seen, three types of distance were basically consistent.

According to computed results in Table 1, there select minimum nine indicators, respectively primary industry, industry, consumption by industry, total imports& exports, electricity consumption last year, annual per capita disposable income of urban households, annual per capita net income of rural households, GDP and per capita GDP.

3.4. Analysis and Correction of Data

For further analysis based on the relationship between electricity consumption and nine indicators, the dynamic line chart is drawn, it as shown in Figure 3.

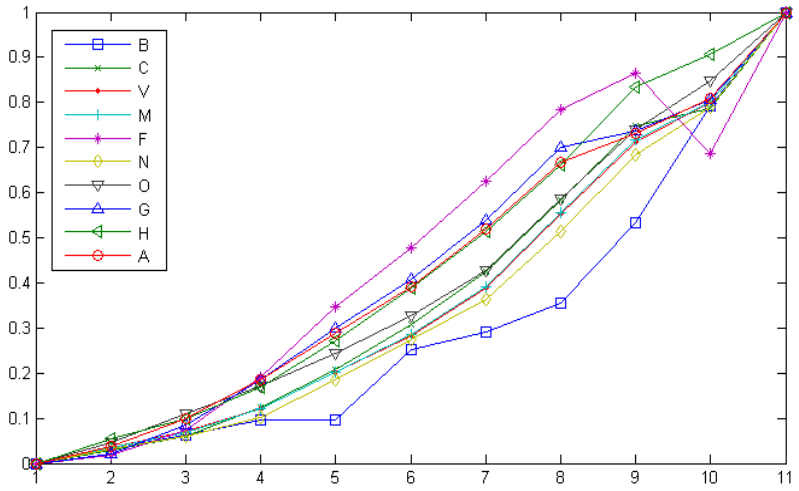

Figure 3. Before Correction

There is a greater volatility on F indicators (total imports& exports), and require further processing. Using formula (13) and (14), it can be calculated that there is distorted data at 10 point of F indicators. And then regard it as missing data, and complement it using the formula (5).

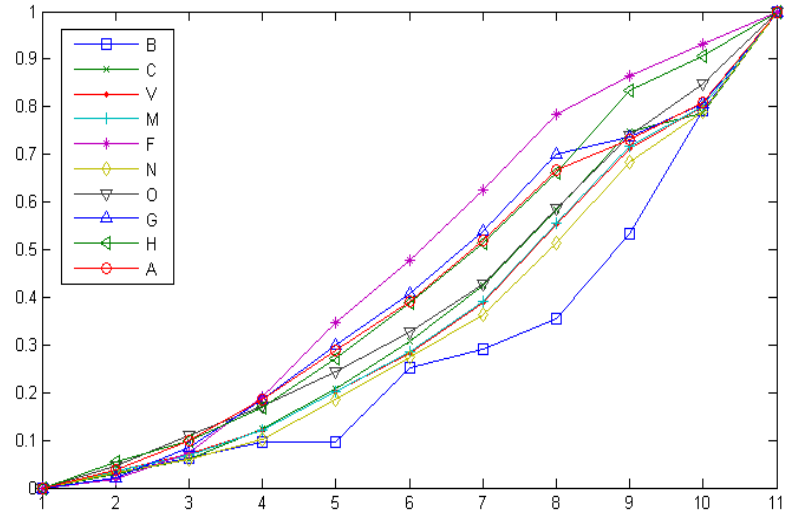

Figure 4. After Correction

The abscissa represents time Figure 3 and Figure 4, Time range is for 2000 to 2010, the ordinate represents relationship between the selected indicators and the whole social electricity. Dynamic line chart is drawn before complement as shown in Figure 3. Dynamic line chart is drawn after complement as shown in Figure 4.

3.5. The NARX Recursive Network Prediction

A NARX neural network is used to predict here. First, determine the number of neurons in hidden layer, training functions and sum squared error (SSE) by optimal training, its results as shown in Table 2.

Table 2. The Results of Optimal Training

***if* Training function is**	***And* Number of neurons in hidden layer is**	***Then* SSE is**
Trainbr	12	7.7408e-009
Trainrp	11	7.9882e-004
Trainlm	12	6.4846e-004
Traingdm	12	0.0071
Traingdx	13	0.0054

Through optimal training, determine that the number of neurons in hidden layer is 12, and training function is Trainbr. The error of neural network to achieve satisfactory results as shown in Figure 5.

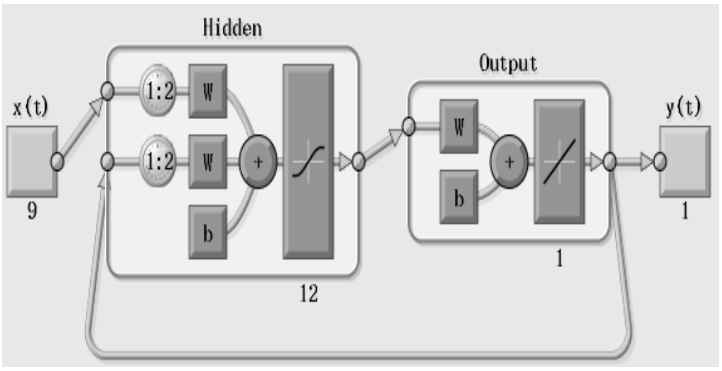

Figure 5. NARX Neural Networks

There are three parameters: error autocorrelation, input-error cross-correlation and time series response. For a perfect prediction model, there should only be one nonzero value of the autocorrelation function, and it should occur at zero lag; this would mean that the prediction errors were completely uncorrelated with each other; Except for the one at zero lag, fall approximately within the 95% confidence limits around zero, so the model seems to be adequate. As shown in Figure 6.

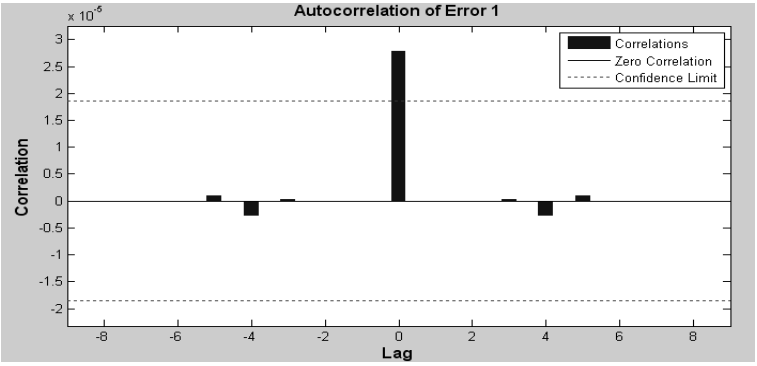

Figure 6. The Error Autocorrelation

The input-error cross-correlation illustrates how the errors are correlated with the input sequence; for a perfect prediction model, all of the correlations should be zero. And all of the correlations fall within the confidence bounds around zero. As shown in Figure 7.

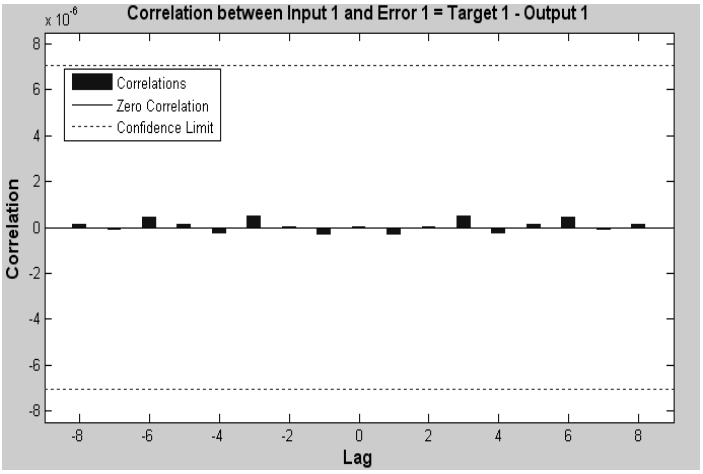

Figure 7. The Input-Error Cross-Correlation

Time series response displays the inputs, targets and errors versus time, and can also indicates which time points were selected for training, testing and validation. As shown in Figure 8.

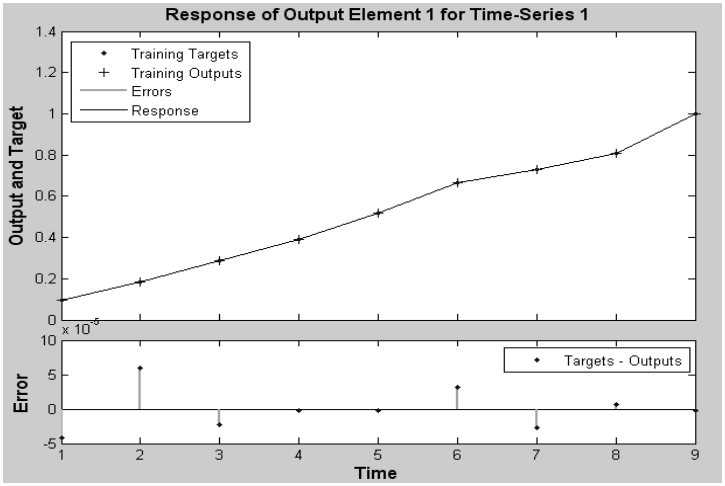

Figure 8. Time Series Response

4. Conclusion

The paper applies cluster analysis algorithms and ideas in data mining to correlation analysis of electricity consumption growth, find a number of indicators in historical data from 2000 to 2011 Statistical Yearbook, which closely related to China's electricity consumption growth. Expand the application of data mining, it provides a theoretical basis for the selection of input data in the long-term load forecasting model. The methods and ideas can also be used to analyze other similar problems in the power system.

References

[1] C.-Q. Kang, Q. Xia and B.-M. Zhang, "Review of Power System Load Forecasting and its Development", Automation of Electric Power Systems, vol. 28, no. 17, pp. 1-11, **(2004)**.
[2] Y.-gui Cheng, M. Li and M.-Y. Lin, "Forecasting and Analysis on long-term/mid-term Electric Load of city by GA-BP Neural Networks", Journal of Computer Applications, vol. 30, no. 1, **(2010)**, pp. 224-226.
[3] S. Zhang, R.-you Zhang and D.-W.ei Wang, "Medium/Long-Term Load Forecasting Based on DPCA-BP Neural Network", Journal of Northeastern University(Natural Science), vol. 31, no. 4, **(2010)**, pp. 483-485.
[4] J. Liu, H.-ma Yang and B.-X. Chen, "Application of Neural Network in Electrical Load Forecasting", Process Automation Instrumentation, vol. 33, no. 9, **(2012)**, pp. 21-24.
[5] E. M. Carreno, R. M. Rocha and A. Padiha-Feltrin, "A cellular automation approach to spatial electric load forecasting", IEEE Trans on Power Systems, vol. 26, no. 2, **(2011)**, pp. 532-540.
[6] A. Shiu and P. L. Lam, "Electricity consumption and economic growth in China", Energy Policy, vol. 32, no. 1, **(2004)**, pp. 47-54.
[7] Y. Fu, L. Zhu and J.-l. Cao, "A New Method to Obtain Load Density According to the Theory of Fuzzy Approach Degree", Automation of Electric Power Systems, vol. 31, no. 19, **(2007)**, pp. 46-49.
[8] L.-F. Mao, J.-G. Yao and Y.-S. Jin, "Abnormal Data Identification and Missing Data Filling in Medium-and Long-Term Load Forecasting", Power System Technology, vol. 34, no. 7, pp. 148-153, **(2010)**.
[9] J. Xiao, J. Zhang and T. Zhu, "Analysis of Urban Power Based on Association Rules", Automation of Electric Power Systems, vol. 31, no. 17, **(2007)**, pp. 103-107.
[10] Z.-Y. Li, Z.-G. Cheng and Z. Xu, "Identification of Dominant Factors in China's Power Consumption Growth", Automation of Electric Power Systems, vol. 34, no. 23, **(2010)**, pp. 30-35.
[11] Y.-H. Li and J.-H. Lei, "The Application and Research of Electric Power Load Forecasting Technology Based on the Time Series Model", Science Technology and Engineering, vol. 11, no. 4, **(2011)**, pp. 860-864.
[12] L. Cai, S.-Y. Ma and H.-T. Cai, "Prediction of SYM-H index by NARX neural network from IMF and solar wind data", Sci China Ser E-Tech Sci, vol. 40, no. 1, **(2010)**, pp. 77-84.

Author

Li Xiaofeng, He is an advanced member of China computer federation and he is the associate professor at Heilongjiang International University. His research interest includes Data mining, Text mining, intelligent algorithm so on.

