Combination of Text Mining and Corrective Neural Network in Short-term Load Forecasting

DongXiao Niu
School of Business Administration, North China Electric Power University, Beijing, China
Email: niudx@126.com

JianJun Wang
School of Business Administration, North China Electric Power University, Beijing, China
Email: wangjianjunhd@gmail.com

***Abstract*—Short-term load forecasting refers to short period load prediction of utility ranging from one hour to several days ahead. It is meaningful in planning and dispatching the load to meet the electricity system demand. The inaccuracy load forecasting can increase the electricity operating costs. In this paper, a novel method is presented and discussed which combines text mining and corrective neural network (TM-CNN) methods. Subsequently, a numeric example of daily maximum load forecasting is used to illustrate the performance of TM-CNN method, and the experiment results also reveal that TM-CNN method outperforms the autoregressive moving average(ARMA) and BP Artificial Neural Network(BPNN) approaches.**

***Index Terms*—load forecasting, text mining, artificial neural network, Autoregressive moving average (ARMA)**

I. INTRODUCTION

Short-term load forecasting is an important issue for electricity load planning and dispatching the loading of generating units in order to meet the electricity system demand. Accurate load forecasting is related to electricity company's economic, it is pointed out by Bunn and Farmer that a 1% increase in forecasting error implied a 10 million increase in operating costs[1]. Many studies have focused on how to improve the precision of the load forecasting, especially short-term load forecasting. Such as the famous Box-Jenkins' ARIMA method[2], which is the benchmark method in short-term load forecasting and the linear regression method[3] and so on. However, the short-term load forecasting is influenced by many factors, such as weather, holidays, special days, humidity and so on, the above methods are difficult to play an excellent role in short-term forecasting because they are difficult consider the non-linear factors.

In last two decades, the artificial intelligence methods are successfully applied in variety fields[4-6], and these methods are also employed to improve the performance of short-term load forecasting. Artificial Neural Network(ANN) is the most popular intelligence model in short-term forecasting, because this method has high capability of dealing with linear, near-linear or nonlinear relationships by any factor. Kun-Long Ho et al[7] applied a multilayer neural network to forecast an-hour ahead load demand of Taiwan power system. The results showed that the ANN method is very efficient and accurate. Chen Shin-Tzo et al[8] considered the weather factor influence in ANN, the analytical results showed that the models have good performance than other methods like ARMA. James W. Taylor[9] compared four different weather variables' performance in ANN to forecast from one to ten days ahead, the results showed that the average of the weather values is more accurate than traditional weather values. For more detail, Hippert[10] gave a very good review and evaluation to ANN in short-term load forecasting.

Recently, the neural network models are favored by more and more researchers, M. Ghiassi[11] proposed a dynamic artificial neural network for forecasting one month ahead load, and the experiment proofs that the dynamic ANN method outperformed the ARMA models. On the other hand, Support Vector Machines(SVM) method, proposed by Vapik[12], implements the structural risk minimization (SRM) principle rather than empirical risk minimization principle implemented by the traditional neural network models, and it is also used for load forecasting, Pai and Hong[13-14] employed SVM in Taiwan's load forecasting, the numerical results with actually Taiwan's yearly load data are superior to the ARIMA results.

According to above literatures, the neural network model performs better than other methods. However, it is complex to predict the short-term load, because the influencing factors depend on not only the temperature and other weather variables, but also social factors such as human social activities including work and entertainment, and the models are not competent for finding the implicit rules between the factors and load. So the expert system[15] and knowledge system[16] are used to get the rules assisted forecasting load accurately. The experiment result showed that it was helpful to gain more accurate forecasting values. However, Text mining is a new technology to extract the implicit, previously unknown, and potentially useful information from text[17]. It mainly contains the text classification, text clustering and text association analysis. It has been successfully used in WEB page classification. Therefore, in this paper, a novel method, combining text mining and corrective neural network(TM-CNN), is used in short-term load forecasting. The corrective neural network's is

doi:10.4304/jcp.4.12.1188-1194

composed by ARIMA method and corrected errors of ARIMA neural network. The text mining is used to find the implicit rules which can improve the load forecasting accuracy.

This paper is organized as follows: Section 2 presents the frame of TM-CNN method, and introduces the text structure and TM-CNN. Section 3 presents a numerical example and the comparison with other method. Finally, Section 4 presents the conclusion remark.

II. TM-CNN METHOD TO SHORT-TERM LOAD FORECASTING

The TM-CNN frame applied in short-term load forecasting is shown in Fig. 1. The whole TM-CNN method is divided into two phases and eight processing procedures. The two phases are CNN phase and TM phase, which are introduced as follows.

A. CNN Phase

CNN phase concludes a basic and two main processing procedures. First of all, the short-term load series should be prepared for forecasting. Secondly, the ARMA method is used for initial prediction. In ARMA model, the future value of a variable is assumed to be a linear function of several past observations and random errors, expressed as follows:

$$\phi(B)y_t = \theta(B)e_t \tag{1}$$

Where y_t is the actual value and e_t is the random error at time t ; B is the backward shift operator, i.e. $By_t = y_{t-1}; Be_t = e_{t-1}; B^2 y_t = y_{t-2}; B^2 e_t = e_{t-2}$ and so on, e_t is the random error, which is independently and identically distributed with a mean zero and a constant variance of σ^2,. and $\phi(B)$ and $\theta(B)$ can be calculated as follows:

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \phi_3 B^3 - \cdots - \phi_p B^p \tag{2}$$

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \theta_3 B^3 - \cdots - \theta_q B^q \tag{3}$$

Where p and q are integers; in addition, if the dth difference of $\{y_t\}$ is used to solve the non-stationary problems with ARMA method, then it is called an ARIMA (p,d,q) process. The value of p,d,q can be estimated by the difference time series characteristics of autocorrelation function (ACF) and a partial autocorrelation function (PACF).

Thirdly, CNN is employed to correct the errors of ARMA. After ARMA forecasting has been finished,it forms a forecast series, which is noted as $\{Y_t, t = 1, 2, \cdots\}$. Then the error series can be calculated by (4).

$$e_t = Y - Y_t \tag{4}$$

The e_t can be seen as the nonlinear component of the time series. Thus, a process of errors correction of time series is needed, and as mentioned above; the ANN is the preferred method for completing this mission. The training input vector is $e_t, e_{t-1}, \cdots, e_{t-n}$, and the training output is e_{t+1}. Then the forecast series E_t can be obtained by the CNN.

In this study, the BP training algorithm is employed in the CNN, which is the same as BPNN except the input and output variables. BP training algorithm is an iterative gradient descent algorithm designed to minimize the mean square error between the outputs and the desire values, it uses the gradient-descendent direction to train the weight between the layers. The process is made up of two directions though the layers, one is the forward, and the other is backward. The typical structure of the BPCNN is shown in Fig.2.

In CNN, the connection weights and node bias will be adjusted iteratively by a process of minimizing the forecasting errors by gradient descent algorithm after inputting the training set, the final computational equation of BPNN is[18]

$$e_t = b_0 + \sum_{j=1}^{q} w_j f(b_j + \sum_{i=1}^{p} w_{ij} e_{t-i}) + \varepsilon_t \tag{5}$$

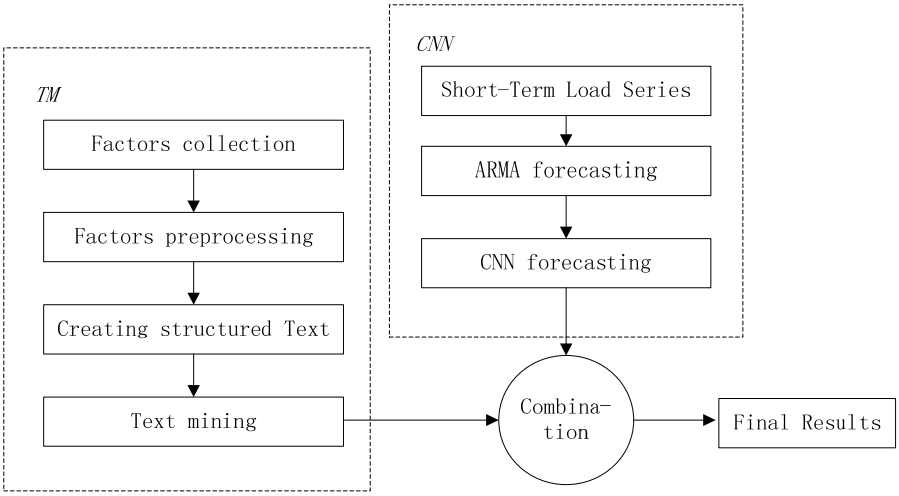

Figure 1. TM-SVM frame of short-term forecasting

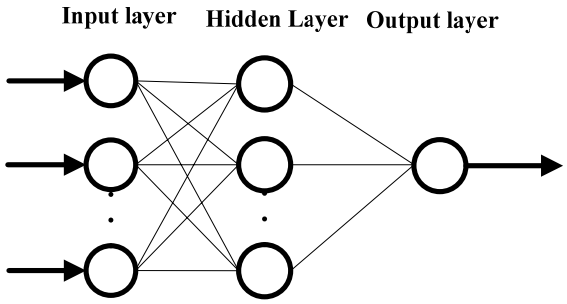

Figure 2. the BPCNN structure of load forecasting

Where b_j is a bias on the jth unit, w_{ij} is the connection weight between layers, $f()$ is the transfer function of the hidden layer, p is the number of input nodes and q is the number of hidden layer nodes.

When the CNN finished the training process, it can be used to forecast with new input values, such as test set, and then the forecast corrective values E_t can be obtained. Therefore, the CNN final results are calculated by $Y_t + E_t$, which is noted by C_t. Obviously, it must have some errors between C_t and target values, and some deviation may still very great. Thus, it is necessary to use text mining for improving the forecasting values.

B. TM Phase

TM phase extracts useful knowledge for load forecast. It contains factor collection, factor preprocessing, creating structured text and text mining, Factor collection collects the related variable data. Generally speaking, the maximum temperature, minimum temperature, humidity, human comfort index, day type, date, weather descriptive text and other related descriptive text will affect the short term demand. So the data of these variables should be collected at first, which contain both numeric, enumerate and pure texts in the data, it is necessary to preprocess the data of factors. For example, it contains continuous variable discretization, feature extraction from the descriptive text and so on. After factors preprocessing, a structured data set is generated. This data table is the typical data set in structured data set, and it is the base data structure of text mining. A typical data table should contain the columns as follows:

Date: Short-term load forecasting express obvious superimposed levels of seasonality, the seasonality trend of daily load forecasting always expresses the monthly and yearly trend. It needed the date variable to eliminate the seasonal trend to performance load forecasting accurately. It also helps finding the same day of last year and the same hours of yesterday. From data table perspective, the date is the best candidate of primary key in the variables.

Weekday: It is obvious that there are different load patterns between the weekdays and weekend, because there are different human activities patterns in it. Some researchers use this variable to divide the data set in order to train different neural networks[10]. The weekday is a binary variable, 0 expresses the weekday and 1 expresses the weekend.

Special Day: Special day means that some statutory holidays except weekend. In China, special days include New Year's day(Jan,1), Spring Festival Eve(the last day in Chinese Lunar Calendar), Spring Festival(from the 1st day to the 3rd day in Chinese Lunar Calendar), International Labor Day(May,1-May,3), National Day (Oct,1-Oct,3). In these days, the Chinese people will have their leisure time, so the load patterns in these days are different from the normal days, especially Spring Festival Eve and Spring Festival, the yearly minimum load always appears in these days. The special day is an enumeration type. It is pay attention to Spring Festival which is divided Spring Festival and During Spring Festival days.

Temperature: Temperature is the most important factors in using ANNs in short-term load forecasting. In summer, if temperature is high, air-conditions are used to chill in order to make people comfortable, the load also rises. The maximum and the minimum temperature often employ in ANNs for short-term load forecasting. Therefore, the maximum and minimum temperature variables are chosen in structured data, and they are continuous numeric variables.

Humidity: Humidity exerts a strong effect on the human sensation of thermal discomfort, as the temperature variable, it helps explaining the use of heating and cooling devices. Some researchers used humidity and temperature to classify the data or separate ANNs training[10]. It is numeric variable and the range is [0,1].

Human comfort index: The index is defined by the heat exchange between human body and air environment. Generally speaking, it is produced by temperature, humidity, barometric and wind speed factors. This index straightly explains the human feeling, as mentioned above, it will affect human activity and influent the use of devices. It is necessary to add these two factors so as to provide more information to classify or cluster data by text mining. The range of human comfort index or human feeling is from level 1 to level 11. The lowest degree express the coldest, the highest degree express the hottest, and when the degree is in 5-7, it means most of people feel comfortable. It is also an enumeration variable.

Weather descriptive text: This variable is some descriptive text, which includes some descriptive weather word, such as sunny, rain, cloudy, heavy rain and so on. It is also contains some special weather information. For example, Coastal areas may contain Typhoon information in summer or autumn, and arid areas may contain drought information. It is possible to have special load patterns in such weather. The role of this variable is to get more accurate load forecast values.

Other related descriptive text: As mentioned above, short-term load forecasting is a difficult task. In the end of the 2008, the global economy is hugely affected by the financial crisis in United States, many enterprises have to dismissal their employees and reduce their products to deal with the crisis, thus the electric demand declined. Besides, the change of political situation can also

influence the electricity demand. The energy conservation policy advocate people substituting their high-energy device to low-energy device, it also reduces electricity demand. Things like these will be recorded in the data table for text mining. It likes an abstract to make people remember what happened during a day, or a period. And with the development of information technology, these information can be easily got from html pages in the Web.

When the structured text data is prepared, the text mining process will be used to exact the implicit knowledge and rules. Classification is one of the most important tasks in text mining. The main goal of classification is to find the knowledge from the condition attributes to decision attributes. For load forecasting corrective, it mainly finds the if-then rules from the factors to the errors, and these rules can be integrated with the CNN's result. And the final result should be more accurate.

To summarize, the proposed TM-CNN approach consist three main steps. At the first step, an ARMA model should be constructed and the primary forecast results are computed from the ARMA model. At the second step, CNN model is used to model the error components in order to correct the primary results. At the final step, the effects of the other factors can be acquired by TM, and the integrated forecast results can be obtained. In order to verify the effectiveness of the proposed approach, a numerical example is given in next section.

III. A NUMERICAL EXAMPLE

This study employed the day maximum load series of Jiangmen's Power Company, which is in Jiangmen city, Guangdong province, China. The time series is from Jan, 1, 2005 to Dec, 31, 2007. Totally, 1095 data points are available, which is in Fig.3. The data is used for comparing with TM-CNN, ARMA, and BPNN models. To conduct the forecast performance on the same basis, it is necessary use the same test set, which is the last 31 points of the series from Dec, 1, 2007 to Dec, 31, 2007. Other data points are prepared for samples set of ARMA or training set of BPNN. The accuracy is measured by the mean absolute percentage error(MAPE), as given by (6).

$$e_{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{A(i)-F(i)}{A(i)}\right| \times 100\% \qquad (6)$$

Where n is the number of forecasting points; $A(i)$ is the actual load value at period i, and $F(i)$ is the forecasting load value at period i.

According to the mentioned TM-CNN method's steps, firstly, the sample set is used in ARMA's method and the parameters $p=1$ and $q=1$ is identified by the AIC criteria from the linear trend and periodicity eliminating time series.

In terms of forecasting result by ARMA(1,1), the error series e_t can be obtained by (4), and then CNN is used to correct the errors. By learning from repeated experiments, when the input factors are $e_t, e_{t-1}, e_{t-2}, e_{t-365}, e_{t-365*2}$, the output variables is e_{t+1}, and the hidden node number is eleven, the error forecasting is performance best. In which, the e_t means the error of the current period, e_{t-n} is the error of n periods before the current period. The hidden and output layers have the sigmoid activation functions; the training error level was set to 10^{-4}. In the end, the corrective error series E_t can be forecasted by CNN, and then the corrected result C_t is calculated by sum E_t and the ARMA results.

Naturally, it also has the errors between the corrected result and the target values. Then text mining is used to finally correct C_t. According to the collected data, the factors maximum humidity (max_humidity), minimum

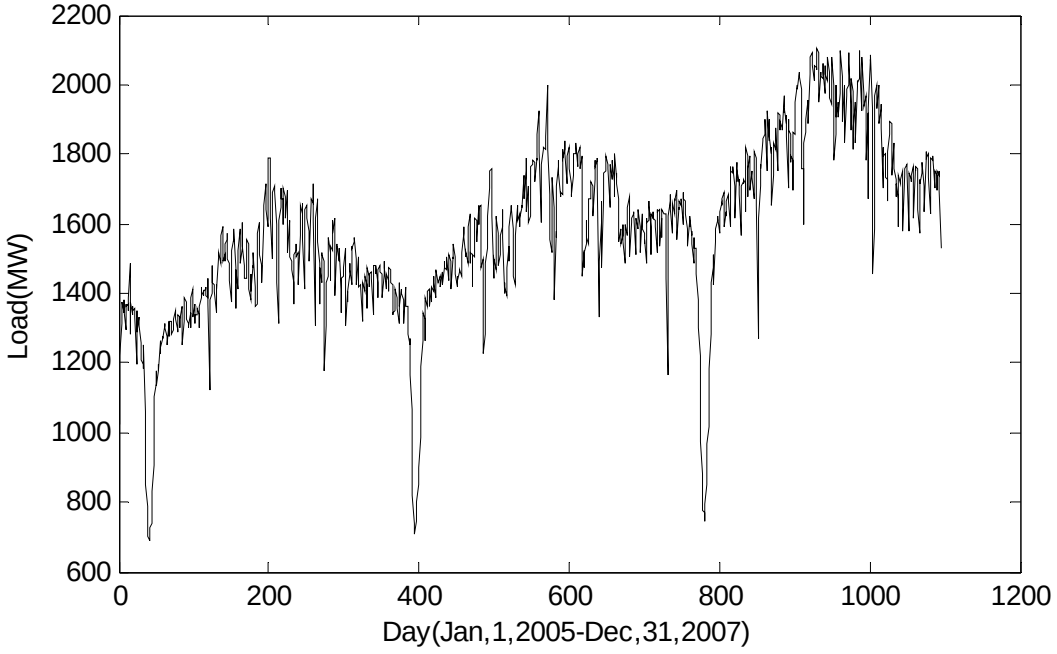

Figure 3. the time series of load from Jan,1,2005 to Dec,31,2007

```
month < 6.5
|  month < 4.5 : 0 (82/0) [38/0]
|  month >= 4.5 : 0 (44/0) [17/0]
month >= 6.5
|  max_humidity < 0.68
|  |  comfortable < 6.5 : 0.01 (13/0) [4/0]
|  |  comfortable >= 6.5 : 0 (2/0) [1/0]
|  max_humidity >= 0.68
|  |  wind < 2 : 0.01 (4/0) [0/0]
|  |  wind >= 2
|  |  |  max_humidity < 0.93 : 0 (66/0) [43/0]
|  |  |  max_humidity >= 0.93
|  |  |  |  max_vitoutla < 4 : 0 (2/0) [4/0]
|  |  |  |  max_vitoutla >= 4
|  |  |  |  |  comfortable < 8.5
|  |  |  |  |  |  month < 7.5 : 0 (3/0) [1/0]
|  |  |  |  |  |  month >= 7.5
|  |  |  |  |  |  |  min_humidity < 0.63 : 0 (2/0) [1/0]
|  |  |  |  |  |  |  min_humidity >= 0.63 : 0 (2/0) [1/0]
|  |  |  |  |  comfortable >= 8.5 : 0 (2/0) [2/0]

Size of the tree : 21
```

Figure 4. the corrective decision trees

humidity (min_humidity), maximum UV(max_vitoutla), human comfortable index(comfortable), wind speed(wind), month and weather descriptive text(weather) are chosen as the conditional factors, it is noted that the temperature factors is not included in the conditional attributes because when the factors contains, the rules is not satisfy for correcting the results, and the decision attribute is the relative error, which can be calculated as $(A(i)-F(i))/A(i)$, the symbol meaning is the same as (6).

After the attributes are determined, the implicit rules can be found by text mining. Weka is a famous software in data mining, it was developed at the University of Waikato in New Zealand, and the name stands for Waikato Environment for Knowledge Analysis. The software includes lots of the data mining algorithms; it also includes many of classification algorithms. In this study, the Weka's edition is 3.6, and the final decision tree is shown in Fig. 4. From the decision trees given by Weka, we can easily extract IF-THEN rules, which are listed in Table 1.

TABLE I. THE CORRECTIVE IF-THEN RULES

No	IF	THEN
1	month>=6.5 max_humidity<0.68 comfortable < 6.5	$A(i)=1.01\times F(i)$
2	month>=6.5 max_humidity>=0.68 wind <2	$A(i)=1.01\times F(i)$

Combining the rules and corrected result C_t, the final result is shown in Fig.5. To verify the significance of the accuracy improvement of the TM-CNN method, the BPNN forecast model is added to assess the significance of the same forecasting task, and the BPNN's structure is the same as CNN structure instead the input and output values of the load values, rather than ARMA errors. The test result is shown in Table 2.

It is clear that the TM-CNN forecasting methods reduce the errors more than the ARMA and BPNN methods. The MAPE is obvious lower than ARMA and BPNN method. Therefore, the TM-CNN results are more accurate than ARMA and BPNN. The main reason is the TM-CNN decline the error twice, the text mining and CNN processes makes effect result.

IV. CONCLUSIONS

Accurate short-term load forecasting is important for electricity industry. This study introduced a novel forecasting method, named TM-CNN, to investigate its feasibility in forecasting daily maximum electricity loads in Jiangmen city, Guangdong province, China, and the experimental results shows that the TM-CNN method outperformed the ARMA and BPNN models. The superior performance of TM-CNN method has two main

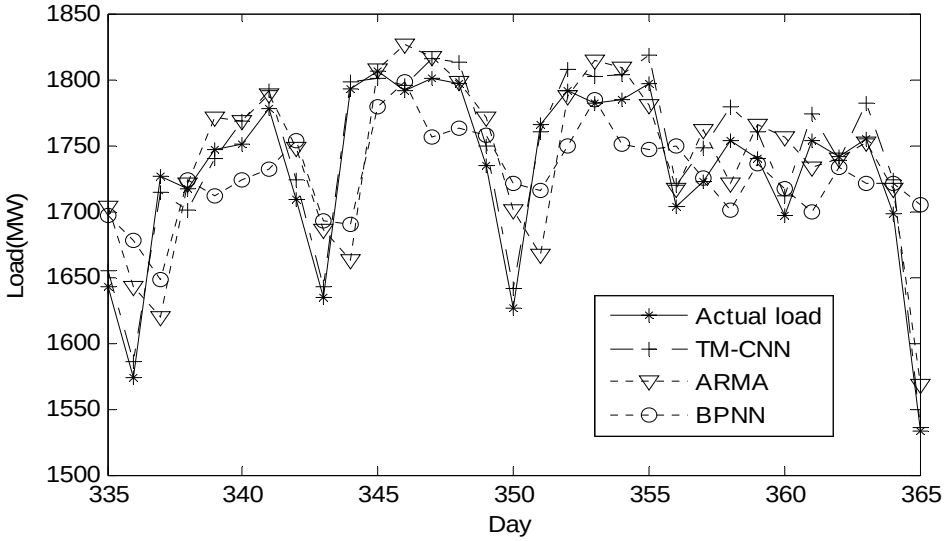

Figure 5. the result of three methods

TABLE II. THE RESULT OF THREE METHODS

Day	Actual load	TM-CNN	ARMA	BPNN	Day	Actual load	TM-CNN	ARMA	BPNN
2007-12-1	1643.23	1654.758	1702.994	1697.125	2007-12-17	1764.93	1760.63	1667.24	1715.563
2007-12-2	1573.58	1585.622	1642.311	1677.17	2007-12-18	1791.26	1807.504	1787.156	1748.745
2007-12-3	1726.41	1713.773	1619.763	1648.364	2007-12-19	1781.97	1801.841	1814.098	1784.778
2007-12-4	1716.95	1700.785	1721.459	1724.159	2007-12-20	1784.36	1803.654	1808.199	1750.336
2007-12-5	1747.28	1740.478	1771.512	1711.152	2007-12-21	1797.1	1818.397	1780.182	1746.403
2007-12-6	1750.04	1767.667	1767.931	1723.529	2007-12-22	1703.54	1719.207	1716.917	1749.619
2007-12-7	1777.37	1791.22	1788.825	1731.211	2007-12-23	1722.85	1747.318	1761.585	1725.098
2007-12-8	1709.12	1723.61	1747.306	1752.718	2007-12-24	1752.92	1778.828	1720.371	1700.249
2007-12-9	1633.8	1642.446	1685.149	1692.769	2007-12-25	1739.41	1759.728	1765.623	1736.367
2007-12-10	1791.91	1798.218	1662.327	1690.213	2007-12-26	1697.03	1711.974	1756.17	1717.081
2007-12-11	1805.43	1801.018	1807.15	1779.013	2007-12-27	1753.81	1773.226	1732.651	1698.961
2007-12-12	1791.06	1795.517	1826.087	1798.159	2007-12-28	1741.36	1738.851	1740.418	1733.65
2007-12-13	1800.87	1815.888	1816.454	1756.148	2007-12-29	1754.56	1781.543	1751.999	1721.512
2007-12-14	1796.37	1812.89	1797.762	1762.73	2007-12-30	1697.51	1723.132	1717.567	1721.184
2007-12-15	1734.78	1749.311	1770.484	1757.717	2007-12-31	1532.49	1536.306	1567.796	1704.138
2007-12-16	1626.62	1640.751	1701.035	1720.907		MAPE	0.0084	0.0208	0.0263

reasons as follows:

1) TM-CNN method combines the ARMA and ANN method, so it can deal with the linear and nonlinear influences of load forecasting, and the text mining can dig the influence rules between the other factors and the electricity patterns.

2) TM-CNN method reduces the error by CNN correction and text mining correction twice, and the processes can decline the deviation effectively.

ACKNOWLEDGMENT

This research was supported by the National Natural Science Foundation of China under Grant 70671039 and the New Century Excellent Talents in University under Grant (NCET-07-0281).

REFERENCES

[1] Bunn DW, Farmer ED, *Comparative models for electrical load forecast*, New York: John Wiley, 1985.

[2] S.Sp Pappas, L Ekonomou, D.Ch Karamousantas, G.E. Chatzarakis, S.K. Katsikas and P.LIatsis, "Electricity demand loads modeling using AutoRegressive Moving Average (ARMA) models", *Energy*, vol 33, pp. 1353-1360, September 2008.

[3] J.H. Park, Y.M Park, K.Y. Lee, "Composite modeling for adaptive short-term load forecasting", *IEEE Trans. Power Syst*, Vol 6, pp. 450-457, May 1991.

[4] D.G. Khairnar, S.N. Merchant, Desai, B. Uday, "Radar signal detection in non-Gaussian noise using RBF neural network", *Journal of Computers*, Vol 3, pp. 32-39, January 2008.

[5] Alan McCabe, Jarrod Trevathan and Wayne Read," Neural network-based handwritten signature verification", *Journal of Computers*, Vol 3, pp. 9-22, August 2008.

[6] Jiang, Zhao-Hui, Ishita, Taiki, "A neural network controller for trajectory control of industrial robot manipulators", *Journal of Computers*, Vol 3, pp. 1-8, August 2008.

[7] Ho Kun-Long, H.Y.-Y., Yang Chlsn-chuen, "short term load forecasting using a multilayer neural network with an adaptive learning algorithm", *Transactions on Power Systems*, Vol 7, pp. 141-149, February 1992.

[8] Chen, S.-T., Yu, D.C., Moghaddamjo, A.R," Weather sensitive short-term load forecasting using nonfully connected artificial neural network", *IEEE Transactions on Power Systems*, Vol 7, pp. 1098 – 1105, August 1992.

[9] James W. Taylor, R.B.,"Neural Network Load Forecasting With Weather Ensemble Predictions", *IEEE Transactions on Power Systems*, Vol 17, pp. 626-632, August 2002.

[10] Henrique Steinherz Hippert, C.E.P., and Reinaldo Castro Souza, "Neural Networks for Short-Term Load Forecasting: A Review and Evaluation", *IEEE Transactions on Power Systems*, Vol 16, pp. 44-55, February 2001.

[11] M. Ghiassi, D.K.Z., H. Saidane," Medium term system load forecasting with a dynamic artificial neural network model", *Electric Power Systems Research*, Vol 76, pp. 302-316, March 2006.

[12] V.Vapnik, S. Golowich, A.Smola, "Support vector machine for function approximation, regression estimation, and signal processing", *Adv. Neural Inf. Process. Syst.*, Vol 9, pp. 281-287, 1996.

[13] Ping-Feng Pai, W.-C.H.," Forecasting regional electricity load based on recurrent support vector machines with genetic algorithms", *Electric Power Systems Research*, Vol 74, pp. 417-425, June 2005.

[14] Ping-Feng Pai, W.-C.H. "Support vector machines with simulated annealing algorithms in electricity load forecasting", *Energy Conversion and Management*, Vol 46, pp. 2669-2688, October 2005

[15] S. Rahman, R. Bhatnagar, "An expert system based algorithm for short-term load forecast", *IEEE Transactions on Power Systems*, Vol 3,pp. 392-399, May 1988.

[16] Rahman S, Hazim O, "A generalized knowledge-based short-term loadforecasting technique", Vol 8, pp.508-514, May 1993.

[17] Ian H. Witten, Eibe Frank, Data Mining: Practical Machine Learning Tools and Techniques, San Francisco: Morgan Kaufmann, 2006.
[18] Wang Shouyang, YU Lean, K.K. Lai, "crude oil price forecasting with TEI@I Methodology", Journal of Systems Science and Complexity, Vol 18, pp. 145-166, April 2005.

Dongxiao Niu was born in China in 1962, and obtained his Ph.D. and Professor degree in North China Electric Power University in Beijing in China. His research interests are load forecasting theory and application, technical and economic management theory and application.

Jianjun Wang was born in China in 1981, and he is now major in North China Electric Power University in Beijing in China for his Ph.D. His research interests are load forecasting theory and application, data mining, and risk analysis in project management.