Multi-variable Echo State Network Optimized by Bayesian Regulation for Daily Peak Load Forecasting

1Dongxiao Niu
1School of Economic and Management, North China Electric Power University, Beijing, China
Email: ncepujl@126.com

1Ling Ji, 2Mian Xing and 3Jianjun Wang
1,3School of Economic and Management, North China Electric Power University, Beijing, China
2School of Computer, Mathematical and Physical Sciences, North China Electric Power University, Beijing, China
Email: hdjiling@126.com

Abstract—In this paper, a multi-variable echo state network trained with Bayesian regulation has been developed for the short-time load forecasting. In this study, we focus on the generalization of a new recurrent network. Therefore, Bayesian regulation and Levenberg-Marquardt algorithm is adopted to modify the output weight. The model is verified by data from a local power company in south China and its performance is rather satisfactory. Besides, traditional methods are also used for the same task as comparison. The simulation results lead to the conclusion that the proposed scheme is feasible and has great robustness and satisfactory capacity of generalization.

Index Terms—Echo State Networks; Bayesian regulation; multi-reservoir; load forecast; daily peak load

I. INTRODUCTION

Echo state network is a novel recurrent neural network, and many studies has been reported on it [1-2]. Compared with the conventional recurrent neural networks, ESN represents great improvement on nonlinear system identification. First of all, the spectral radius of the internal connection weight matrix is predefined to ensure the stability of the recurrent neural network; besides, the weight of output is the only one need to be modified in the learning process, and it’s usually global optimum, thus it can avoid the universal local minimum phenomenon in traditional neural network. In addition, the training process of ESN is relatively simple. ESN was firstly applied to predict chaotic time series and obtained the best result on the benchmark task of Mackey–Glass series prediction [3]. As the meanwhile, ESN has a prominent prediction performance for other chaotic time series. Due to the excellent performance of ESN in nonlinear system identification, it’s suitable for the load forecasting model.

When it comes to the application of echo state network in multi-variable time series, there are still several problems need to study. In most study cases, the multi-variable time series are put into single reservoir, which means the dimension of different variables’ phase space have to be the same. In this way, the dynamical characteristics of complex system are unable to be described. To some extent, it limits the flexibility of the model and also leads to imprecise result. In addition, the key process in the echo state network construction is the determination of the weights to the readout. The issue of the estimation in multi-variable ESN is more complicated than that in standard ESN. In traditional study, linear regression and Least Squares (LS) have been widely used. However, it easily produce over-fitting phenomenon. Especially when the number of training samples is far less than that of dimension of reservoir, the phenomenon is particularly evident.

To deal with the problems mentioned above, the multi-reservoir echo state network optimized by Bayesian regularization is presented in this paper. The features of multi-variable can be extracted and sufficiently mapped into though multi reservoirs. By this way, the flexibility of the forecasting model can be improved. What’s more, as an alternative, Bayesian regularization is adopted to estimate the weight and parameters in the model. This can avoid the over-fitting phenomenon during linear regression.

The rest of this work is organized as follows: Section 2 introduces the principle of echo state network and gives a short review to some fundamental knowledge of generalization, especially Bayesian regularization algorithm. In section 3, the forecasting model based on a multi-variable ESN utilizing Bayesian regularization is established; and its application in short-time load forecasting is followed. Comparative results are reported and discussed in section 4. At last, some conclusions are drawn in section 5.

II. GENERALIZATION OF MULTI-VARIABLE ESN
A. The Architecture of ESN with Multi-reservoir

The nonlinear dynamic characteristics of ESN are produced by a massive reserve pool, which contains a number of randomly generated and sparse connection neurons. The reservoir contains the operation state of the system and has the memory function. With the external input, an “input-state-readout” drive system is produced, which is illustrated by Figure 1. Suppose there are L inputs, M outputs and N neurons in the reservoir pool. The weights of the connections from input layer to reservoir, within reservoir and from reservoir to readout are notes as \( W^i \), \( W \) and \( W^o \) respectively, their sizes are \( N \times L \), \( N \times N \) and \( N \times M \) in order. During working, the reservoir state and the readout are updated as follows[4-5].

\[
\begin{align*}
  x(k) &= \tanh(W \cdot x(k-1)+W^o \cdot u(k)+v(k-1)) \\
  y(k) &= W^o \cdot x(k)+b
\end{align*}
\]

(1)  (2)

Where, \( \tanh(\cdot) \) is the hyperbola tangent Sigmoid function, \( x(k) \) is the reservoir’s state variable at time \( k \) and its dimension is usually high, from 300~1000. \( u(k) \) represents the input vector at time \( k \). \( v(k-1) \) is the random noise vector, and \( y(k) \) is the output of the echo state network. \( W \) is a high dimensional matrix random generated with sparse connection (usually 1%-5%)[6-7]. Once the reservoir is set, the conjunction weights remain unchanged and the output weight \( w \) is the only one need to be modified. Therefore, the learning mechanics of ESN is quite different from the traditional neural network in training process. Generally, the output weight \( w \) is calculated by linear regression algorithm:

\[
w = (X^T X)^{-1} X^T y_d
\]

(3)

In the multivariable forecasting model, due to the different features of various variables, it is necessary to build corresponding reservoir to map each variable. The structure of echo state network with several reservoirs is hierarchical, and the core of each lay is the reservoir. The schematic of the multi-reservoir echo state network is shown by Figure 1.

\[
\begin{array}{c}
\text{Input} \\
\text{Multi-reservoir} \\
\text{Output}
\end{array}
\]

Figure 1. The schematic of multi-reservoir echo state network

Take the echo state network with two reservoirs as an example, its mathematical model can be expressed as:

\[
\begin{align*}
  x_1(k) &= \tanh(W_{11}x_1(k-1)+W^i_{11}u_1(k-1)) \\
  x_2(k) &= \tanh(W_{21}x_1(k-1)+W^i_{21}u_2(k-1)) \\
  y(k) &= w^1_1x_1(k-1)+w^1_2x_2(k-1)+\varepsilon
\end{align*}
\]

(4)

Where, \( W \) and \( W^i \) are the weights of the connections within Dynamic Reservoir (DR) and from input layer to DR respectively. \( x_1 \) and \( x_2 \) are denoted as the reservoir states. \( u_1 \) and \( u_2 \) are the input variables of each reservoirs; \( w_1 \) and \( w_2 \) is the output weight matrix for each reservoir.

With the given samples of input and output for the network training, the steps in building ESN are as following:

STEP 1: Select suitable value for the scale \( N \) of the reservoir, the sparsity of internal weights \( W \) and spectral radius for the reservoir, and Initialize the ESN randomly.

STEP 2: Choose the input and output samples for the network. The internal state is inspired by the input. Calculate and record the value of state variable of the reservoir at each time.

STEP 3: According to the regression relation between the state variable and output variable, we can calculate the output weight \( w_1 \) and \( w_2 \). In this paper, this process is under Bayesian regulation theory.

B. Generalization of the Model

Since the neural network is the nonparametric model, the only information is derived from the training sample. As a result, the training result is extremely unstable, and the caused over-fitting will bring down the generalization [8-10]. Voluminous literatures have studied the function approximating in network [11-12].

The generalization of the network is depend on the character of the training sample (like the size and the quality of the training set), the structure of network (the number of hidden nodes and hidden layer, the function characteristics of the hidden nodes), and the complexity of the problem. If there are few training samples, the learning can’t capture enough information to reflect the essential problem, and the reliability of the network is affected, much less its generalization. On the contrast, if the sample is too much, the network easily produces the over-fitting phenomenon because of some exceptional individual and the noise, as a result, we can’t find the real law of the issue. In practice, there are two main methods for improving the network’s generalization.

(1) Regularization

Given a set of training samples \( D = \{(x_1, y_1), \ldots, (x_n, y_n)\} \), the purpose of neural network learning is to find a function \( f \) to approximate these samples effectively. In traditional method, it’s usually realized by minimizing the objective function, that is:

\[
E_D = \frac{1}{2} \sum_{i=1}^{n} (y_i - f(x_i))^2
\]

(5)

This function reflects the distance between the expected output and the actual data. But, in fact, there are infinite ways to regain function from limited samples, which means the uncertainty remains. Whereas, regularization is a very effective process do deal with this problem. It is achieved by adding \( E_p \) to standard error.

\[
E_p = \sum_{i=1}^{n} w^2_i
\]

(6)
Where, \( w \) denotes the parameters of the neural network, \( m \) is the number of the parameters, thus the total error function is defined as following:

\[
F(W) = \beta E_p + \alpha E_w
\]

(7)

Where, \( \alpha \) and \( \beta \) are called regularization parameters and can control the distribution of other parameters. \( E_w \) represents the sum of squares of weights.

Mackay found that minimizing the loss function is equal to maximize the posterior probability of the parameters. If we minimize the loss function in the learning process, the redundant weight will decay to zero as the learning continued, and be wiped out eventually. Obviously, once the redundant weights or hidden nodes are pruned, the structure is simplified while the accuracy remains. And the generalization ability will be improved [13-14].

However, a problem of choosing optimal regularization parameters in the modified objective function is appearing. The regularization parameters have direct influence on the training effect. As a rule of thumb, a larger \( \alpha \) leads to low fitting degree; and greater \( \beta \) courses over-fitting. In many cases, these parameters are usually obtained by cross-validation techniques, which is time-consuming.

(2) Early stopping criterion

Stopping learning at a suitable time is also an important method to improve the generalization ability [14-15]. Once the extended error achieves the minimum (the optimal stopping point), we can stop the learning process. Minh et al. have studied the choice of stopping criterion for neural-network subpixel-classification application [15].

In this paper, there are several reservoirs in the network, thus the output weights are enormous inevitably. In order to avoid the over-fitting in traditional linear regression, Bayesian theory is adopted to modify the output weight to improve the generalization ability. In other word, try to wipe out redundant connections and neurons without deteriorating the accuracy of the network. As a result, reduce the complexity of the network and gain better generalization capability.

C. Introduction of Bayesian Regularization

Bayesian approach is come from the inductive method. For certain issues, people usually depend on the deductive method. When it comes to the uncertain things, some inductive and inference methods will be adopted. Bayesian approach describes things in probability theory. Its idea is to evaluate the model using prior knowledge and data, which is also the main difference between Bayesian and classic statistics.

A large number of statistic and relative experts have done abundance study on the application and popularize of Bayesian theory [16-19]. In recent years, Bayesian evidence framework, which is firstly proposed by MacKay, has been widely studied to solve problems like pattern recognition and parametric estimation. It also has been successfully employed to design the neural network. Bayesian regularization is one of the effective methods used for adjusting the self-adapting parameters in the training process of neural network.

At present, there are three methods for the calculation of posterior probability: it is firstly proposed by Mackay based on Gaussian approximation; then Neal has developed a method based on MCMC numeric integration method; besides, variational Bayes is newly studied. Among them, the first one is the most widely studied, thus it is adopted to calculate the posterior probability.

Suppose \( D = (x_i, t_i) \), \( i = 1, 2, \ldots, n \) is the training sample set for neural network. Where, \( n \) is the number of the training samples, \( w \) is the parameters of the network, \( m \) is the number of the parameters. Given the network framework \( H \), supervised neural network model is expressed as \( y = f (x, W, H) \). Without the sample data, the prior distribution is \( P(w|\alpha, \beta) \), once given sample data \( D \), we can get the posterior distribution \( P(w|D, \alpha, \beta, H) \), according to the Bayesian theory, it can be written as following:

\[
P(w|D, \alpha, \beta, H) = \frac{P(D|w, \alpha, \beta)P(w|\alpha, \beta)}{P(D|\alpha, \beta, H)}
\]

(8)

Where, \( P(D|w, \alpha, \beta) \) is a likelihood function, \( P(D|\alpha, \beta, H) \) is full probability, as a normalization factor. Thus, it can be written as follows:

\[
P(D|\alpha, \beta, H) = \int_{\alpha} P(D|\alpha, \beta, H)P(w|\alpha, \beta)dw
\]

(9)

Prior distribution \( P(w|\alpha, \beta) \) usually takes index distribution, such as Gaussian distribution, which is commonly used, and it is expressed as:

\[
P(w|\alpha, \beta) = \frac{1}{Z_\alpha} \exp\left(-\frac{\alpha E_w}{2}\right)
\]

(10)

If we assume the prior distribution is Gaussian distribution with zero mean and variance \( 1/\alpha \), and \( Z_\alpha \) represents the normalization factor:

\[
Z_\alpha = \int \exp\left(-\frac{\alpha E_w}{2}\right)dw = \left[\frac{2\pi}{\alpha}\right]^{m/2}
\]

(11)

Suppose the training sample independent and from the same distribution, then the likelihood probability of data \( D \) is given by:

\[
P(D|w, \alpha, \beta, H) = \prod_{i=1}^{n} P(t_i | x_i, w, H)
\]

(12)

At the same time, the output error of the neural network is:

\[
e = y(x_i | w, H) - t_i
\]

(13)

We suppose \( e \) subject to Gaussian distribution with zero mean and variance \( 1/\beta \) (hyperparameter \( \beta \) control the noise of the network’s output), then corresponding to the real error of network, the likelihood probability of data \( D \) is as following:

\[
P(D|w, \alpha, \beta, H) = \frac{1}{Z_\beta} \exp\left(-\frac{\beta E_e}{2}\right)
\]

(14)

Where, \( Z_\beta(\beta) \) is the normalization factor:

\[
Z_\beta(\beta) = \left[\frac{2\pi}{\beta}\right]^{m/2}
\]

(15)

Therefore, the posterior distribution of the parameters \( P(w|D, \alpha, \beta, H) \) of the network is expressed as:
\[
P(w|D, \alpha, \beta, H) = \frac{1}{Z_r(\alpha, \beta)} \exp(-\beta E_p - \alpha E_w)
\]

(16)

Where, \( Z_r(\alpha, \beta) = \int \exp(-\beta E_p - \alpha E_w) \) \( dw \), since the value of \( Z_r(\alpha, \beta) \) does not depend on \( w \), the maximum of posterior distribution can be obtained by minimizing \( F(W) \). Mackay used Gaussian distribution to approximate the posterior distribution of \( P(w|D, \alpha, \beta, H) \). Expand \( F(w) \) at \( w^* \) point where its \( F(w^*) \) value is the smallest, due to the gradient is 0, and it is somewhat similar to the following.

\[
F(w) = F(w^*) + (w - w^*)^T \nabla^2 F(w^*) (w - w^*)
\]

(17)

Where, Hessian matrix \( H(w^*) = \nabla^2 F(w^*) \) is symmetric positive semidefinite at least, which can be written as \( H = H_k^T H_k^T \), and order \( u = H_k^T (w - w^*) \). We can substitute (17) into (16), make both sides integral and obtain the following:

\[
1 = \frac{(2\pi)^{\frac{\alpha}{2}}}{Z_r(\alpha, \beta)} e^{-(w^*)^T [\text{det}(H(w^*))^{-\frac{1}{2}}]} \]

(18)

\[
Z_r(\alpha, \beta) = (2\pi)^{\frac{\alpha}{2}} e^{-(w^*)^T [\text{det}(H(w^*))^{-\frac{1}{2}}]}
\]

(19)

Combine (8) with (16), we can get

\[
P_l(\alpha, \beta, H) = \frac{Z_r(\alpha, \beta)}{Z_r(\alpha, \beta)}
\]

(20)

Put formula (19) into (20), obtain logarithm from both sides, then use the one-power condition of optimal value to find the optimal regularization parameter:

\[
\alpha^* = \frac{\gamma}{2E_p(w^*)}, \quad \beta^* = \frac{n - \gamma}{2E_w(w^*)}
\]

(21)

Where, \( \gamma = N - 2 \text{attr}(H) \) represents the number of effective parameters in the network, reflecting the actual scale of the network. Levenberg-Marquardt (LM) algorithm is improved Gauss-Newton. Literature (Man, 2004) indicates that as to Hison matrix \( H \), training network by Levenberg-Marquardt is easy to approach by Gauss-Newton method. Hence the training process of echo state network optimized by Bayesian regulation is detailed in the following:

(1) Initialize the value of \( W_{\text{init}}, W \) and \( W_{\text{back}} \);

(2) Calculate the reservoir state series \( x(k) \) though the samples and formula (1);

(3) Setting the initial value of \( \alpha \) and \( \beta \);

(4) Look for the minimum point \( w \) by training network with L-M algorithm;

(5) Calculate the number of effective parameters in the network, then update the value of \( \alpha \) and \( \beta \);

(6) The iterative process (4) to (5) may be continued until some criterion for convergence is met;

(7) According to the status series \( x(k) \) and output \( x(k) \), we can obtain output weights matrix \( W_{\text{out}} \) though formula (2) and regression method (3).

III. MODELLING AND SIMULATION

Meteorological factor is very important to the load forecasting. With the development of the modern science and technology, it is easier to get the meteorological data, and the accuracy of weather prediction has also increased. Therefore, it’s necessary and possible to consider the meteorological factors in the load forecasting model. In theory, more factors taken into account, the better the fitting precision is. However, superabundent variables will lead to too complex forecast model and influence the efficiency, which it’s not practical. Here, we only choose the daily average and the daily peak load as the input variable. The framework of the new forecasting model is illustrated in Figure 2.

![Figure 2: The new structure of forecasting model](image)

To verify the predictive ability of the proposed scheme, we perform numerical simulations. The data used in the simulation are the daily peak load and daily average temperature of a southern city in China ranging from July 1, 2008 to August 1, 2009, thus two variables time sequences are formed. The curves of original data have been plotted in Figure 3. The data from the first 264 days are used as learning samples, the rest 132 samples as test samples to evaluate the generalization and prediction precision of new model.

![Figure 3: The curves of daily peak load and daily average temperature](image)

It’s necessary to carry out the normalization before various input variables getting through different reservoirs separately, as formula (22). As a result, they are equally important status at the beginning. Figure 3 shows the normalization results of both variables.

\[
A = \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}}
\]

(22)

Where, \( x_{\text{max}} \) and \( x_{\text{min}} \) are the maximum and minimum of the variable respectively.

![Figure 4: The normalization result of daily peak load and daily average temperature](image)
Based on correlative dimension, we can compute the phase space parameters of daily peak load ($x_1$) and daily average temperature ($x_2$) respectively. Delay time for $x_1$ and $x_2$ are 4 and 1, embedding dimension are 2 and 6 respectively. The prediction system is built based on a multi-reservoir ESN with 100 internal units in each reservoir, the sparse matrix with 5% percentage of connectivity. The spectral radius $\rho$ is always considered as one of the most important parameters determining ESN performance. Since there is no standard or rules to determine it, in most study $\rho$ is obtained by trial and error. After several experiments, the spectral radius is set at 0.8. A hyperbolic tangent activation function is used for the reservoir neurons and a linear activation function for the output neurons. The training of the $W^{\text{out}}$ is referred to the previous discussion.

At the mean while, standard echo state network and traditional BP network are applied for the same task as contrast. The parameters of standard ESN model are set the same values as the ESN-Bayesian. Bayesian regulation is the only difference between them. The BP network has three layers with sigmoid transfer function in the hidden layer and linear transfer function in the output layer. The node number of the input layer, hidden layer and output layer are 10, 6 and 1 respectively. The system error is set as 0.001.

To evaluate the performance of different network, we computed the RMSE and MAPE:

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{n=1}^{N} (y_n - \hat{y}_n)^2}$$

$$\text{MAPE} = \frac{1}{N} \sum_{n=1}^{N} \left| \frac{y_n - \hat{y}_n}{y_n} \right|$$

IV. CONCLUSIONS

Figure 5-7 show the comparisons between the real records of hourly load and the predicted ones by the models of ESN-Bayesian, standard ESN, and BP. Observing the graphs, we can find that these methods are useful for the task in general. From Figure 5, it can be found that the bias at the peak and bottom is greater, which is also obvious in the results of ESN and BP. Especially, during June to July, due to the high temperature and variable weather, the peak load is the gap between peak and valley becomes greater. And ESN performs not well as ESN-Bayesian. As to the BP, the predicted curve is not as close fit to the actual curve as the whole. Hence, the optimized ESN achieves higher predictive accuracy.

The influence of Bayesian regulation can be observed by comparing the performance of ESN-Bayesian and standard ESN. Relative to the network with single reservoir, the delay time and embedding dimension of various input variables can be set differently, which can fully reflect the dynamic characteristics of them separately. Therefore, the multi-reservoir echo state network improves the flexibility and the performance of the network.

The prediction results obtained via BP network illustrate that the percentage absolute errors have a mean equal to 2.51% and a standard deviation equal to 0.577%. In terms of forecasting ability, ESN performs better than BP network, with a mean and standard deviation of absolute percentage errors equal to 2.36% and 0.593% respectively. While, it is clearly that the optimized ESN with Bayesian regulation is superior with smaller maximum error. Besides, there is a reduction in the standard deviation. An inspection of Figure 5 and Table 1 reveals that the model proposed here is helpful for daily peak load forecasting. It should be noted that with the scale of the reserve pool expanding, the dispersiveness and conditional number of the state matrix’s singular value increase. Hence, the model becomes more complicated, which will lead to over-fitting, as a result, its application for forecasting is not satisfying.
V. CONCLUSIONS

This paper presents a new model for daily peak load forecasting based on echo state network and a specially designed training algorithm. Multi-reservoir model can reflect the dynamic characteristics of different variables to obtain better prediction precision. Bayesian regularization makes a good balance between the fitting degree of training sample and the complexity of the model. We compare this model with other classical methods, standard ESN and traditional BP network. Simulation examples show that the forecasting errors by ESN-Bayesian described in the paper are far less than those by other typical models. This proves that the effort to improve the forecast precision in the paper is obviously fruitful. The novel model has strong robustness and generalization ability.

In fact, except the temperature and historical load data, the short term load forecasting is also effected by other factors, like wind speed, humidity, date type and so on. These factors should be taken into consider in further research. In addition, the regularization process is sensitive to the regularization coefficient, which influences the size of the network and the prediction result. More work should be done to optimum the regularization coefficients.

ACKNOWLEDGMENT

This work was supported in part by NSFC under Grant Nos. 71071052, as well as “the Fundamental Research Funds for the Central Universities” under Grant Nos. 12QX23.

REFERENCES


Dongxiao Niu, Ph.D., is a Professor at North China Electric Power University, Beijing, China. His area of expertise is power load forecasting and comprehensive evaluation.

Ling Ji was born in Zhejiang, China. She is pursuing her Ph.D. in Management Science and Engineering at the School of Economic and Management, North China Electric Power University, Beijing, China. Her current research interests include power load forecasting and project management.

Mian Xing, is a Professor at the College of Computer, Mathematical, and Physical Sciences, North China Electric Power University, Beijing, China. Her research interests include mathematical modeling and applied mathematics.

Jianjun Wang, Ph.D., is an instructor at the School of Economic and Management, North China Electric Power University, Beijing, China. His current research interests include power load forecasting and power economics.