2008 International Symposium on Nonlinear Theory and its Applications
NOLTA'08, Budapest, Hungary, September 7-10, 2008

Weight Distribution Criterion of Neural Networks for modeling Chaotic Attractors

Guoqiu Zhang, Yi Zhao, and Hong Hu

Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, China
Email: zhao.yi@hitsz.edu.cn

Abstract In this paper, we describe a new weight distribution criterion to investigate generalization of large neural networks for modeling chaotic attractors. The technique of minimum description length is first employed to estimate the optimal neural network as the standard. Weight distribution criterion then reveals that large networks whose effective structures are consistent with this standard usually exhibit good performance while others whose effective structures are distinct from this standard perform poor. We illustrate our approach to the two kinds of time series: the Rössler system and normal human pulse data.

1. Introduction

Nonlinear time series analysis based on neural networks modeling has been intensively studied [1]. However, overfitting is a common and serious problem for neural networks, especially large networks modeling. Many techniques have proposed to avoid overfitting including early stopping [2], Bayesian learning [3], and minimum description length [4]. The former two techniques focus on the improvement of the known networks while the last one emphasizes the selection of the optimal model. Meanwhile, we also notice that some large networks make small prediction errors on the training set and also respond properly to novel inputs while others become over fitted. So the topic of this manuscript is to study this phenomenon, i.e. selection of large networks with adequate generalization.

We propose a novel method, weight distribution criterion, to estimate the effective structure of a large network, i.e. to analyze which neurons make significant contribution to the model and which neurons take little affection. We find that when applied to the same time series prediction, large networks whose effective neurons are consistent with the optimal neural networks estimated by minimum description length obtain almost the same good performance as the optimal one. The technique of minimum description length has been proved to be a good criterion to select optimal neural networks for diverse time series prediction [5]. So we adopt such optimal network as the standard to compare with effective structures of large networks described by weight distribution criterion.

In the next section we will discuss our method at length. In Section 3 we validate our method with the computational chaotic data and experimental data. Finally, we take the conclusion.

2. Methodology

Although large networks easily over fit they are still popular in complicated time series modeling as they have more potential to capture the underlying dynamic and researchers are inclined to employ larger networks for modeling. So the large neural network is the object of this research.

2.1. Weight Distribution Criterion (WDC)

As we know, using the back-propagation neural network to predict time series must adjust its weights and basis to obtain a smaller mean square error on the training data set. The network we used consists of three layers which are the input, hidden, and output layers. Here we aim to observe changes of the weights between input and hidden layers. Input vectors denoted by $\{P_1, P_2, \cdots P_R\}$ are fed to the neural network, and connected with the neurons in the hidden layer denoted by $\{N_1, N_2, \cdots N_S\}$. We thus obtain a weight matrix, W, whose row and column is the neuron index (S) and input vector (R). The elements of W are the weights of connection between the input and hidden layer.

If values of some elements in W are near zero in comparison with the others it means that connections weighted by such values are weak. In topological terms, one may neglect these links. If connections between one neuron and the entire input vector are weak (i.e. the neuron is separated from inputs), it is approximated to be neglected [6]. Note that we normalize the weight matrix so as to make a fair comparison between distinct experimental data.

When projecting weight distribution we can estimate the significant neurons in a network and the number of those significant neurons will be compared with the optimal network determined by the minimum description length. It is demonstrated that when its effective structure is consistent with estimation of method of minimum description length the large network can avoid overfitting and capture underlying dynamics. The weight distribution criterion working together with the minimum description length forms the comprehensive model selection criterion.

We introduce the technique of minimum description length method, as follows.

2.2. Minimum Description Length (MDL)

The minimum description length is to choose the optimal model which can accurately capture the dynamics of chaotic time series. The description length of a time series *D(k)* is consisted of two parts: the description length of the model prediction errors *E(k)* and the description length of a model of that time series *M(k)*.

$$D(k) = M(k) + E(k) \qquad (1)$$

The arithmetic in detail of this method can be found in [7].

As model size *k* increases, the cost to describe the model prediction error will decrease while the cost to describe model parameters will increase, and that there will be minimum of the sum *M(k)*. The MDL principle states that the minimum point of *M(k)* is the optimal model size (i.e. optimal number of neurons in a network). Since DL curves fluctuate dramatically and this probably affects the estimation of the true minimum point, nonlinear curve fitting is adopted to smooth the fluctuation and provide a more accurate estimation of the original one. The explanation of the nonlinear fitting idea can also be found in [4].

3. Application and Example

3.1. The Rössler System

The equations of the Rössler system are given by

$$\begin{cases} \dot{x}(t) = -y(t) - z(t) \\ \dot{y}(t) = x(t) + a * y(t) \\ \dot{z}(t) = b + z(t) * [x(t) - c] \end{cases} \qquad (2)$$

For a=0.398, b=2, and c=4, the system exhibits “single band” chaos. We integrated these equations with the step size, 0.5. By adding dynamic noise to *x*-component at each step we generate 2000 data points, of which 1588 points are used to train the networks and the rest is regarded as the testing data. From Figure 1(a), we observe that the optimal model is composed of five neurons.

We use the optimal network to perform the free-run prediction for the testing set. The prediction is then converted into three vectors, *x*(*t*), *x*(*t+3*), and *x*(*t+5*) ($t \in$ [1,395]) to construct its dynamics, as shown Figure 1(c) [8][9]. In Figure 1(b) the weight distribution exhibits that after training only the first, third, seventh, ninth and eleventh neurons have strong connection with the input vector while the weights between the other neurons and input vector are almost zero. That is, the effective neurons in this large network are five neurons which play an important role in the following calculation and final output. It indicates that the large network whose effective structure is consistent with the MDL-optimal network can avoid overfitting and also make the good prediction.

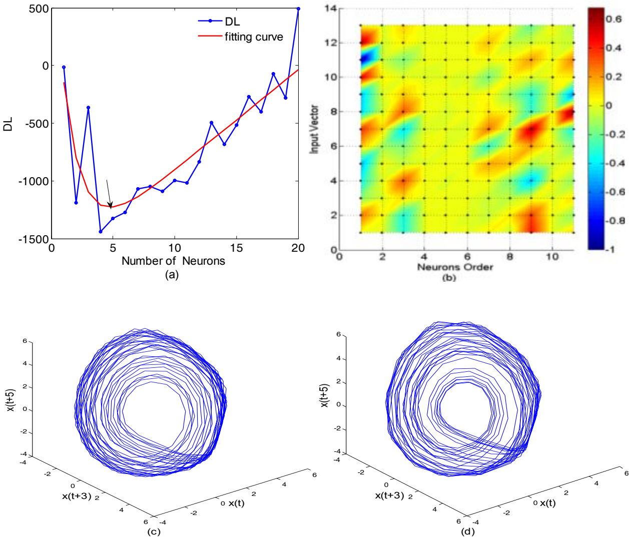

Figure 1(a) Description length curve from one to twenty neurons. The blue line is the original curve and the red line is the fitted curve. (b) The weight distribution Projection of a large network with 11 neurons. (c) The Rössler system predicted by the optimal one. (d) The Rössler system predicted by this large network.

On the contrary, if the effective neurons shown by the distribution are much larger than the MDL estimation (i.e. five neurons), the network is apt to over fit. As shown in Figure 2, for the same data set the trained network contains too many effective neurons and its prediction for the testing data become over fitted. So based on both results we consider to employ the weight distribution to test whether the trained model has the adequate generalization. To further validate the feasibility and utility of the proposed weight distribution criterion we apply it to estimate the performance of large networks for the practical data, the human pulse data.

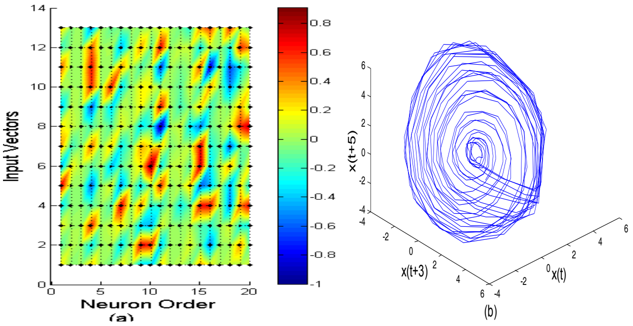

Figure 2(a) Weight distribution project of a trained network with 20 neurons. (b) The Rössler system predicted by this network

3.2. Pulse data

The typical human pulse data we use was collected from one healthy young man in the relaxed situation. We select 2550 data as the training data and the other consecutive 350 data as the testing data. According to the minimal point of the fitted description length curve in Figure 3(a), we obtain the optimal network with ten neurons. Again, the prediction of the optimal network exactly follows the original data.

Meanwhile, the large network with twenty-three neurons also makes the good prediction. Its weight distribution projection exhibits that its effective neurons is almost the same as the number estimated by MDL. That is, the effective structure of the large network is equivalent to the optimal network with ten neurons. So it indicates that the trained network whose effective structure is consistent with the standard structure generalize well.

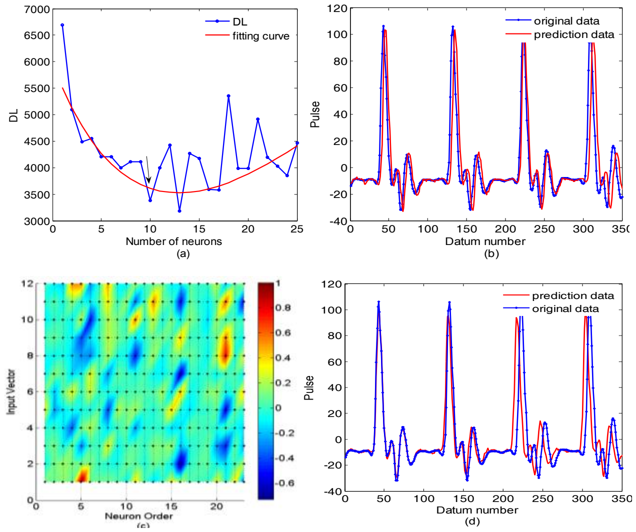

Figure 3(a) Description length curve from one to twenty-five neurons. The blue line is the original curve and the red line is the fitted curve. (b) The free-run prediction of the optimal network. The blue line is the original pulse data and the red one is the prediction. (c) The weight distribution projection picture of a trained large network with 23 neurons. (d) The free-run prediction of this large network. The notation of lines is the same as (b).

However, the weight distribution criterion reveals that the effective structures of other trained large networks are still quite larger than the standard one. Those networks are apt to over fit. An example for such case is illustrated in Figure 4. Obviously, when the effective neurons indicated by the weight distribution criterion are much more than the MDL estimation, the network completely fails to capture the dynamics of the pulse data.

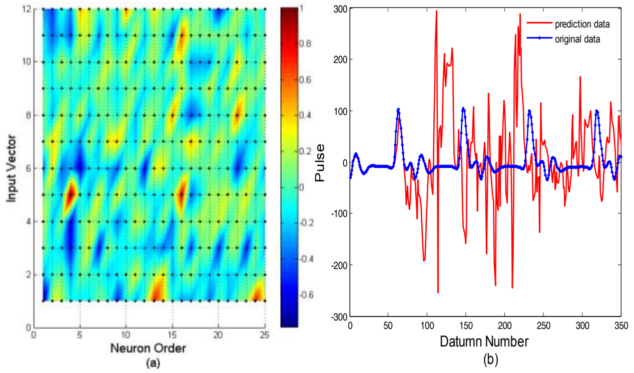

Figure 4 (a) Weight distribution projection of a trained network with 25 neurons. (b) The free-run prediction obtained by this network.

The reason is that the trained network is too large to avoid overfitting based on the observation of its weight distribution projection. We therefore conclude that the criterion of weight distribution clearly discriminates large networks with good generalization from the others.

4. Statistical Analysis of Weight Distribution Criterion

Here we calculate the basic statistics of the weight distribution, including mean, standard deviation, minimum and maximum. For example, say the mean, we calculate the average value of original elements in a column of the weight matrix before normalization, i.e. the average value of weights exerted to all the input vectors for one neuron.

Actually, if the weights between a neuron and its input vectors are weak, then both the mean and variance (or standard deviations), which measure the center and deviation of the data respectively, are near zero and small; if there exists strong connection between a neuron and some its input vectors, it also reflect on those statistics. We take example of the Rössler system, where another trained large network with fifteen neurons follows the dynamics well for the same data set as Section 3.1. Table 1 shows the statistical values of its weight matrix.

Table 1 the mean, standard deviation, minimum and maximum of weights for each neuron in a large network with 15 neurons.

Neurons	1	2	3	4	5	6	7	8
Mean	0.1103	2.7363	0.0183	0.0302	2.1626	-0.0196	-0.0634	-0.0012
Standard Deviation	2.6135	25.1281	0.5124	0.3432	24.3276	0.3654	0.3285	0.2171
Maximum	2.7238	27.8644	0.5307	0.3734	26.490	0.3458	0.2651	0.2159
Minimum	-2.5031	-22.3919	-0.4942	-0.313	-22.165	-0.3850	-0.3919	-0.2182

Neurons	9	10	11	12	13	14	15
Mean	0.0489	0.1507	-0.0727	-1.7075	0.059	0.1475	0.1003
Standard Deviation	0.4644	2.2951	1.014	17.503	0.592	0.7992	2.6761
Maximum	0.5133	2.4459	0.9412	15.795	0.651	0.9467	2.7763
Minimum	-0.4154	-2.1444	-1.0867	-19.2106	-0.533	-0.651	-2.5758

As presented in Table 1, the first, second, fifth, tenth, twelfth, and fifteenth neurons have strong connection with the input vector while values of the others stay in a narrow range around zero.

So the statistical analysis indicates that the effective structure of this network (i.e. six neurons) is close to the estimation determined by MDL. We also observe that this trained network captures the dynamics of the Rössler system well.

5. Conclusion

In this paper, we describe a novel idea, the weight distribution criterion, to estimate the performance of trained large network and explain the reason. The weight distribution projection is used to investigate the effective structure of those networks.

The technique of minimum description length is used to determine the optimal model for the given time series and such optimal model is regarded as the standard. The effective structure of large networks estimated by the weight distribution criterion is then compared with the standard. We conclude that large networks whose effective structures are consistent with the standard one can avoid overfitting while others whose effective structures are far away from the standard one are apt to over fit.

We illustrate the proposed approach with two kinds of time series prediction, the Rössler system and the human pulse data. Certainly, the application is not limited to the above data.

Finally, we give some statistical analysis on weight distribution criterion so as to quantitatively determine the effective structure of a large network. Combination of weight distribution criterion and MDL is demonstrated to be a useful method to determine the large networks with adequate generalization.

Acknowledgments

This work was supported by Natural Scientific Research Innovation Foundation in Harbin Institute of Technology under Project 2008.

References

[1] J. C. Principe, A. Rathie, J. M. Kuo, "Prediction of chaotic time series with neural networks and the issue of dynamic modeling," *International Journal of Bifurcation and Chaos.*, vol.4, pp. 989–996, 1992.

[2] A. Weigend, "On overfitting and the effective number of hidden units," in *Proc. 1993 Connectionist Models Summer School.*, pp. 335–342, 1994.

[3] D. J. C. MacKay, "Bayesian interpolation," *Neural Comput.,*vol. 4, no.3, pp. 415–447, 1992.

[4] Y. Zhao and M. Small, "Minimum Description Length Criterion for Modeling of Chaotic Attractors with Multiplayer Perceptron Networks," *IEEE Transaction on circuits and systems.,* vol.53, no.3, pp. 722–732, 2006.

[5] M. Small and C. K.Tse, "Minimum description length neural networks for time series prediction," *Physical Review E.,* vol.66, p.066701, 2002.

[6] Y. Zhao, G. Zhang, J. Sun, and M. Small, "Distribution Criterion for Large Multilayer Neural Networks with Application of Chaotic Attractors Modeling," *Circuits, Systems, and Signal Processing.,* 2008.(1strevision).

[7] J. Rissanen, "Modeling by the shortest data description," *Automatica.*, vol.14, pp. 465–471,1978.

[8] H. Kantz and T. Schreiber. *Nonlinear Time Series Analysis.*, New York: Cambridge Univ. Press, pp. 30–36, 2004.

[9] G. Sugihara and R. M. May, "Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series," *Nature.*, pp. 734–741, 1990.