Master's Thesis 要旨

Analysis of relationship between structure and performance in echo state networks

Dept. of Mathematical Informatics 48-176230 Hei Ze

Supervisor Assoc. Prof. Gouhei Tanaka

1 Introduction

Echo state networks (ESNs) are powerful computing systems in predictions of time series [1]. However, the mechanism of the computational ability and the memory capacity of the ESN has not been fully clarified. In this thesis, we investigate how the structures of the ESN, including the activation function and the connection weights, influence its performance. In order to reveal the mechanism of the ESN in predicting time series, we compute the Jacobian matrix of the ESN system with respect to the past input signals, and analyze the relationship between the Jacobian matrix and the structures of the ESN. We find that it is the eigenvalue distribution of the recurrent connection matrix that determines the dynamical behavior of the ESN. Based on this result we propose a new method to design the recurrent connection matrix which shows better performance than that in the standard ESN.

2 Echo State Network

The dynamics of the ESN is described by:

$$\begin{aligned} \mathbf{x}^{[t]} &= \mathbf{C}\mathbf{r}^{[t-1]} + \mathbf{W}^{inp}\mathbf{u}^{[t]} + \mathbf{W}^{feed}\mathbf{z}^{[t-1]}, \\ \mathbf{r}^{[t]} &= f(\mathbf{x}^{[t]}), \\ \mathbf{z}^{[t]} &= \mathbf{W}^{out}\mathbf{r}^{[t]}, \end{aligned} \tag{1}$$

where $\mathbf{x}^{[t]}$ is the internal state at time t, and $\mathbf{u}^{[t]}$ and $\mathbf{z}^{[t]}$ are the input and output signals at time t, respectively. $\mathbf{C}$ represents the recurrent connection weights (inside the reservoir). The matrices $\mathbf{W}^{inp}$, $\mathbf{W}^{out}$ and $\mathbf{W}^{feed}$ represent the input, readout and feedback connections weights of the ESN, respectively, and f denotes the activation function which is element-wise in this formula. We only train the readout connection weights $\mathbf{W}^{out}$ of the ESN. The other connection weights are initialized randomly (Gaussian distribution is commonly used) and remained unchanged. Therefore the initialization method for other connections is of great importance. In this study, we investigate how the structure, such as the activation function f and the recurrent connection weights $\mathbf{C}$, influences the computational performance of ESNs.

3 Relationship between Structure and Performance in ESNs

3.1 Relationship between Feedback Connections and Long Term Memory

In the analysis of the role of feedback connections, we used the three-bit-flip-flop (TBFF) task as an example. TBFF task has three input channels and output channels. The input signals are given by sequences of pulses with irregular intervals. The peak value of all pulses is 1 or -1. This task requires the output to be 1 or -1, and to have the same sign as that of the most recent input pulse of the same channel.

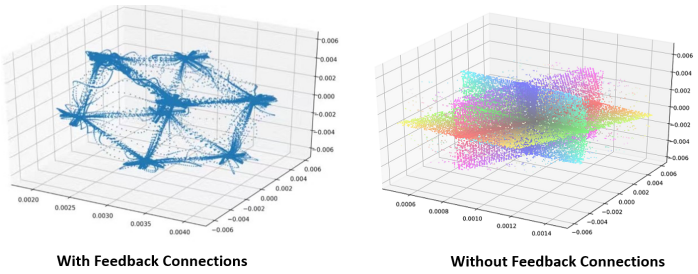

Fig. 1. PCs of ESNs with feedback connections[2] and those without feedback connections

To investigate the role of the feedback connections in TBFF tasks, we trained two ESNs with and without the feedback connections on the TBFF task. And compared the first three principle components (PCs) of the reservoir states of these two ESNs. The result is shown in Fig. 1. With the feedback connections, the ESN can perform the TBFF task. There are eight attractors in the dynamics of the ESN. When the output state was changed by the input pulse, the state of the reservoir jumped from one attractor to another, which is the origin of the long term memory. Without feedback connections, there is only one attractor. When there is a pulse of the input, the state of the reservoir will be pushed away, and soon fall into the attractor as the input pulse vanishes as shown in Fig. 1. The color denotes the channel of the input, and the values of three channels compose the RGB space of the color. From this result, we can draw the conclusion that feedback connections are important in the formation of the long term memory of the ESN.

3.2 Relationship between Recurrent Connections and Short Term Memory

The echo state property means that the influence of the past input signals gradually vanishes. Mathematically this means that the derivative of the state $\mathbf{x}^{[t]}$ with respect to the past input signal $\mathbf{u}^{[t-m]}$ converges to 0 as m increases. This can be regarded as the origin of the short term memory of the ESN. The value of the derivative is naturally included in the Jacobian matrix of the ESN. With the chain rule, we can compute the Jacobian matrix as follows:

$$\frac{\partial z^{[t]}}{\partial u^{[t-m]}} = \left(\mathbf{W}^{out}\right)^T \mathbf{F}^{[t]}\mathbf{C}\mathbf{F}^{[t-1]}\mathbf{C}\cdots\mathbf{F}^{[t-m]}\mathbf{C}\mathbf{W}^{inp}, \tag{2}$$

where $\mathbf{F}^{[t]}$ is an $n \times n$ diagonal matrix with the i-th diagonal term $f'(x_i^{[t]})$. We find that $\mathbf{F}^{[t]}$ can be regarded as constant signals with random perturbations if the scale of the input is small. Therefore we can make the approximation $\mathbf{F}^{[t]} \approx \mathbf{F}_a$, and for a hyperbolic tangent or sigmoid activation function, $\mathbf{F}_a$ approximately equals to the identity matrix multiplied by a constant. With the eigen-decomposition of $\mathbf{F}_a\mathbf{C}$, we can obtain the relationship between the eigenvalues of $\mathbf{F}_a\mathbf{C}$ and the short term memory. Especially for the k time delay task, which requires the ESN to output the time series value k time steps before the current time, we can compute the optimal readout weights as follows:

$$\mathbf{W}_{optimal} = \frac{1}{M\left(\lambda\lambda^T\right)^k}\begin{bmatrix}1 & \lambda_1^k & \lambda_2^k & \cdots & \lambda_n^k\end{bmatrix}^T, \tag{3}$$

where $\lambda = [\lambda_1, \lambda_2, \cdots, \lambda_n]$ is a vector of eigenvalues of $\mathbf{F}_a\mathbf{C}$, and n denotes the reservoir size and M is a variable which depends on the max value of λ, which satisfies $(max(\lambda))^M \neq 0$ and $(max(\lambda))^{M+1} = 0$. The matrix $\mathbf{W}_{optimal}$ is the optimal value of $\mathbf{W}$ which is a linear transform of the readout weights $\mathbf{W}^{out}$. When k is large, which means the required memory for this task is large, we can expect that the term $\lambda\lambda^T$ is neither rapidly expanding nor contracting.

3.3 The Effect of Activation Function on Computational Ability of ESN

We used the NARMA task in the analysis of the role of feedback connections. For a random time series u as the input signal, the target z is generated from u with the next recursive formula:

$$z_t = \alpha z_{t-1} + \beta z_{t-1}\sum_{i=1}^{k} z_{t-i} + \gamma u_{t-k}u_{t-1} + \delta, \tag{4}$$

where α, β, γ and δ are constant values, and k indicates the required memory of this task. Eq. (4) is extremely complicated and therefore the NARMA tasks can be used to evaluate the computational ability of the ESN. From Eq. (4) we computed the condition for the Jacobian matrix, and found that for tasks whose target time series are generated from a polynomial of degree n, the n-th derivative of the activation function is required to be non-zero. This can be proved by the next experiments, where we take the different order Taylor expansions of the hyperbolic tangent function as activation functions, and compare the performance of ESNs with these activation functions. The result is shown in Fig. 2 which shows that the computational ability of the ESN with the linear activation function is much worse than those of nonlinear activation functions.

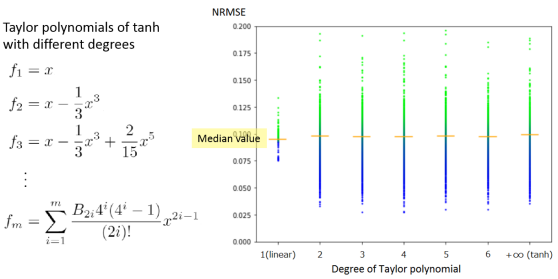

Fig. 2. Performance of ESNs with different order Taylor expansions of the activation function

4 Design Strategy for Echo State Networks

From the result obtained in section 3.2, we see that the mean value of the non-zero eigenvalues of $\mathbf{F}_a\mathbf{C}$ is important in the formation of the short term memory. Since $\mathbf{F}_a$ approximately equals to the identity matrix, we just consider the recurrent matrix $\mathbf{C}$. In order to adjust the eigenvalue distribution $\mathbf{C}$, we first generate a diagonal block matrix which is composed of multiple 2×2 matrices with controllable pairs of eigenvalues. By denoting the radius of each pair by r_i, we can define: $\mathbf{r}^{[m]} = [r_1^m, r_2^m, \cdots, r_n^m]$, where n is the reservoir size. Then we can adjust the radius distribution of the eigenvalues of the recurrent matrix by choosing different m which is shown in Fig. 3 (left). We can also compare the performance of ESNs with this matrix of different m on the time delay task (Fig. 3, right).

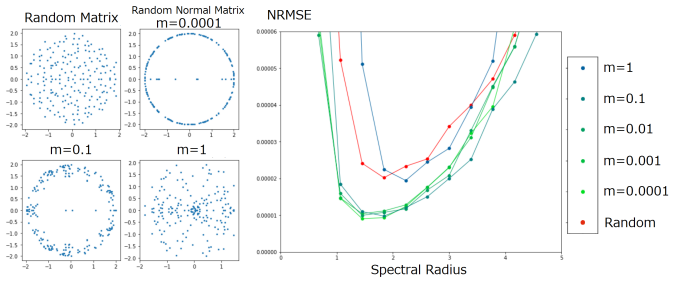

Fig. 3. Eigenvalue distributions of matrices $\mathbf{W}_m$ generated with $\mathbf{r}^{[m]}$ (left). And the performance of ESNs with the recurrent matrix as $\mathbf{W}_m$ on the time delay task (right).

From the result we found that when the eigenvalues are distributed near the edge of the circle (green lines, with $m = 0.0001, 0.001, 0.01, 0.1$), the performance is excellent. When $m = 1$, the eigenvalue distribution (bottom right corner in the left figure) is similar to that of the random matrix (upper left corner in the left figure), and their performance (red line and blue line) are both worse than the others.

Bibliography

[1] H. Jaeger, "The "echo state" approach to analysing and training recurrent neural networks-with an erratum note.," *German National Research Center for Information Technology GMD Technical Report*, vol. 148, no. 34, p. 13, 2001.

[2] D. Sussillo and O. Barak, "Opening the black box: low-dimensional dynamics in high-dimensional recurrent neural networks.," *Neural computation*, vol. 25, no. 3, pp. 626–649, 2013.