A WAVELET NEURAL NETWORK FOR FUNCTION APPROXIMATION AND NETWORK OPTIMIZATION

KUNIKAZU KOBAYASHI AND TOYOSHI TORIOKA
Department of Computer Science and Systems Engineering
Faculty of Engineering, Yamaguchi University
2557 Tokiwadai, Ube-shi, Yamaguchi 755, Japan
E-mail: k@nn.csse.yamaguchi-u.ac.jp

ABSTRACT:
A new mapping network combined wavelet and neural networks is proposed. The algorithm consists of two process: the self-construction of networks and the minimization of errors. In the first process, the network structure is determined by using wavelet analysis. In the second process, the approximation errors are minimized. The merits of the proposed network are as follows: network optimization, partial retrieval of the approximated function, fast convergence and escaping local minima. The computer simulations confirmed these merits

INTRODUCTION

Recently, it has been shown that neural networks (NNs) can realize any mappings (e.g., Hecht-Nielsen, 1987). These are important to theoretically explore the potential of NNs not practically.

Backpropagation (BP) networks are now the most popular mapping network (Rumelhart, Hinton and Williams, 1985). It is, however, well known that BP networks have few problems such as trapping into local minima and slow convergence. In addition, the network structures are determined by trial and error.

Recently, many researchers proposed various network optimization schemes in order to solve such problems. There are two approaches: recruiting (adding) method (e.g., Azimi-Sadjadi, Sheedvash and Trujillo, 1993) and pruning (deleting) method (e.g., Hagiwara, 1991).

It is expected that wavelets will be a new powerful tool for signal analysis (Chui, 1992). Wavelets can approximately realize the time-frequency analysis using a mother wavelet. The mother wavelet has a square window in the time-frequency space. The size of the window can be almost freely variable by two parameters. Thus, wavelets can identify the localization of unknown signals at any level.

In this paper, wavelets and NNs are combined and a *self-recruiting wavelet neural network* (SERWANN) is proposed. The combination of wavelets and NNs have been studied but the number of hidden units was determined before learning (Zhang and Benveniste, 1992; Pati and Krishnaprasad, 1993).

SERWANN has four merits: self-construction of networks, partial retrieval of approximated function, fast convergence and escaping local minima. Incorporating the idea of wavelets, the output function is localized in both the time and frequency domains. Therefore, each hidden unit has a square window in the time-frequency plane. Thus, SERWANN can capture function approximating problems as two tasks: optimizing the net-

work structure and minimizing errors. In this connection, the proposed algorithm comprises two processes: construction of networks and minimization of errors. In the first process, the network gradually recruits hidden units to effectively and sufficiently cover the time-frequency region occupied by a given target. Simultaneously, the network parameters are updated to preserve the network topology and take advantage of the later process. In the second process, the parameters of the initialized network is updated using δ rule (Rumelhart, Hinton and Williams, 1985) in order to minimize the errors of approximation. This rule is only applied to the hidden units where the selected point falls into their windows. Therefore, the learning cost can be reduced. In addition, the localization of the output function results in the partial retrieval of approximated function.

WAVELETS

In this paper, we will use the underlying Hilbert space $L^2(R)$, whose inner product and norm are defined as follows.

$$< f, g > \stackrel{\triangle}{=} \int_R f(x)\,\overline{g(x)}dx$$
$$\|f\| \stackrel{\triangle}{=} < f, f >^{1/2}$$

where $\overline{\psi}$ denotes the dual of ψ.

Wavelet transform

The integral wavelet transform of function $f \in L^2(R)$, $T(a,b)$, is defined as:

$$T(a,b) \stackrel{\triangle}{=} < f, \psi^{(a,b)} > \qquad (1)$$

where $\psi^{(a,b)}(x) = \frac{1}{\sqrt{a}}\,\psi(\frac{x-b}{a})$.

This means the correlation between function f and $\psi^{(a,b)}$, which is obtained from a basic function by dilation by a and translation of b.

In wavelet analysis, the basic function ψ is called mother wavelet but we call it *fitting wavelet* (FW) because the objective is function approximation. This FW must satisfy the following admissibility condition.

$$c_\psi = \int_R \frac{|\Psi(\lambda)|^2}{\lambda}\,d\lambda < \infty \qquad (2)$$

where Ψ is Fourier transform of ψ.

Inversion formula

The NNs must reconstruct the target function from partial information of $T(a,b)$. In this paper, the dyadic partition $a = 2^j$, $b = kb_0 2^j$ $(j,k \in Z)$ is used (b_0 is a sampling rate). This inversion formula is defined as:

$$f(x) \stackrel{\triangle}{=} \sum_{j,k\in Z} < f, \psi_{j,k} > \psi^{j,k}(x) \qquad (3)$$

where $\psi_{j,k}(x) = 2^{-j/2}\psi(2^{-j}x - kb_0)$ and $\{\psi^{j,k}\}$ means the dual basis of $\{\psi_{j,k}\}$.

The FW should satisfy the following stability condition because f must be reconstructed from partial information of $T(a,b)$.

$$A\|f\|^2 \le \sum_{j,k\in Z} | < f, \psi_{j,k} > |^2 \le B\|f\|^2 \qquad (4)$$

where $0 \le A \le B < \infty$ $(A, B \in R)$.

Window of a wavelet

In wavelets, one of the most important concepts is window of a wavelet. This refers to a rectangular region in the time-frequency plane defined for a FW. That is, the FW can *see* such region and not other region. This property results in the identification of localization.

For FW ψ, the support in the time domain, $\mathrm{supp}_\epsilon(\psi)$, is defined as:

$$\mathrm{supp}_\epsilon(\psi) \stackrel{\triangle}{=} [x^{FW}_{\min}, x^{FW}_{\max}] \qquad (5)$$

where $x^{FW}_{\min}$ and $x^{FW}_{\max}$ satisfies the following inequality.

$$1 - \epsilon < \frac{\int_{x^{FW}_{\min}}^{x^{FW}_{\max}} |\psi(x)|^2\,dx}{\int_R |\psi(x)|^2\,dx} \qquad (6)$$

It is implied that the energy of ψ in $[x^{FW}_{\min}, x^{FW}_{\max}]$ is at least concentrated in the time domain at rate $1-\epsilon$. Similarly, for Fourier transform of FW, $\Psi(\lambda)$, the support in the frequency domain is defined as:

$$\mathrm{supp}_\epsilon(\Psi) \stackrel{\triangle}{=} [\lambda^{FW}_{\min}, \lambda^{FW}_{\max}] \qquad (7)$$

where

$$1-\epsilon < \frac{\int_{\lambda_{\min}^{FW}}^{\lambda_{\max}^{FW}} |\Psi(\lambda)|^2 \, d\lambda}{\int_R |\Psi(\lambda)|^2 \, d\lambda} \tag{8}$$

Therefore, ψ has a time-frequency window, $[x_{\min}^{FW}, x_{\max}^{FW}] \times [\lambda_{\min}^{FW}, \lambda_{\max}^{FW}]$.

In general, $\psi^{(a,b)}(x)$ has a window, $[b + ax_{\min}^{FW}, b + ax_{\max}^{FW}] \times [\lambda_{\min}^{FW}/a, \lambda_{\max}^{FW}/a]$. The size of the window is constant for any translation or dilation.

NETWORK EXPRESSION

The inversion formula cannot be expressed by finite NNs. Actually, however, most targets are restricted in both the time and frequency domains. Thus, the inversion formula can be approximately realized using NNs with finite hidden units.

Consider a three-layered network ($1 \times N \times 1$), whose input and output units are linear elements and the output function of the hidden units satisfies both admissibility and stability conditions, i.e. eq.(2) and eq.(4). It is assumed that the network sufficiently approximates the target. Intuitively, this means that the time-frequency region is effectively covered by their N windows. The estimate of the network, $\hat{f}$, is represented by:

$$\hat{f}(x) = \sum_{i=1}^{N} c_i \, \psi(a_i x - b_i) \tag{9}$$

APPROXIMATION AND OPTIMIZATION

Function approximating problem

Given sparse examples, NNs will internally construct a desired mapping through learning. SERWANN captures function approximating problems as two tasks:

- Effectively and sufficiently cover the time-frequency region of a given target by the windows of FWs.
- Approximate training data as precisely as possible and interpolate test data as plausible as possible.

The former implies optimizing the network structure and the later is realizing the best approximation and interpolation.

Construction of networks

The number of hidden units is determined and simultaneously the parameters of each hidden unit are updated by Kohonen's rule (Kohonen, 1989).

Consider that a target f is localized in $[x_{\min}, x_{\max}]$ and is sampled at sampling rate λ_s. The sampled set, T, is:

$$T = \{(x_\alpha, f_\alpha) \mid x_\alpha \in X,\ f_\alpha \in F,\ \alpha = 1 \sim M\} \tag{10}$$

where

$$X = \{x_1 = x_{\min}, \cdots, x_\alpha, \cdots, x_M = x_{\max}\}$$
$$F = \{f(x_1), \cdots, f(x_\alpha), \cdots, f(x_M)\}$$
$$M = \lambda_s (x_{\max} - x_{\min})$$

1. Estimate the band width of set T by DFT.

   $$\mathrm{supp}_\epsilon(T) = [\lambda_{\min}, \lambda_{\max}]$$

   The DFT requires discrete frequency. Thus, taking the inequality $a > 0$ into consideration,

   $$\lambda_{\min},\ \lambda_{\max} \in \Lambda = \{\lambda_1, \cdots, \lambda_\alpha, \cdots, \lambda_{M/2}\}$$

   where $\lambda_{M/2} \le \lambda_s/2$.

2. Using the following equations, transform $x \in X$ and $\lambda \in \Lambda$ to dilation and translation, respectively.

   $$a = \frac{\lambda_{\min}^{FW} + \lambda_{\max}^{FW}}{2\lambda}$$
   $$b = x - \frac{(x_{\min}^{FW} + x_{\max}^{FW})(\lambda_{\min}^{FW} + \lambda_{\max}^{FW})}{4\lambda}$$

   where the above equations can be easily derived using the fact that the center of the window exists in the time-frequency region occupied by the target.

3. Calculate wavelet spectrum of the target for such dilation and translation parameters.

4. Create training set S_{CL}.

$$S_{CL} = \{(x_\alpha, \lambda_\alpha) \mid x_\alpha \in X,\ \lambda_\alpha \in \Lambda,\ \alpha = 1 \sim M\} \quad (11)$$

where λ_α refers to the frequency at the maximum value of wavelet spectrums with respect to x_α.

5. Determine the number of initial hidden units, N^0, and initialize their parameters, a_i, b_i and c_i ($i = 1 \sim N^0$).

6. Pick up a training point $(x_\alpha, \lambda_\alpha)$ from S_{CL} at probability $p_\alpha = |T'(a_\alpha, b_\alpha)|$, where T' refers to normalized T.

7. Determine the nearest neighbor c of the selected point.

$$c = \arg \min_i \|(x_\alpha, \lambda_\alpha) - (b_i, a_i)\|$$

8. Determine its neighbor set N_c.

$$N_c = \{i \mid |i - c| \le l\}$$

where l is a positive integer.

9. Update the parameters of unit i if and only if it belongs to N_c and its window contains the selected point.

$$\begin{aligned} a_i^{t+1} &= a_i^t + \alpha_{CL}(\lambda_\alpha - a_i^t) \\ b_i^{t+1} &= b_i^t + \alpha_{CL}(x_\alpha - b_i^t) \\ &\text{for } i \in N_c \text{ and } (x_\alpha, \lambda_\alpha) \in W_i \end{aligned} \quad (12)$$

where α_{CL} is a learning constant and W_i is the window of unit i.

Otherwise, a new unit is created at rate ρ and its parameters are initialized as follows:

$$\begin{aligned} a^0_{N^{t+1}} &= \lambda_\alpha + \delta \\ b^0_{N^{t+1}} &= x_\alpha + \delta \\ c^0_{N^{t+1}} &= T(a^0_{N^{t+1}}, b^0_{N^{t+1}}) + \delta \end{aligned}$$

where $N^{t+1} = N^t + 1$ and δ denotes small fluctuations. The fluctuations will reduce the effects of *ghost*, which is involved in $T(a, b)$.

10. Repeat step 6 to 9 until the windows of FWs fully cover the time-frequency region of the target and the network settles down to a stable state.

Minimization of errors

The network parameters are updated using δ rule. if and only if the selected point falls into the windows of hidden units. We call it a localized backpropagation (LBP) algorithm.

1. Create training set S_{LBP}.

$$S_{LBP} = \{(x_\alpha, \lambda_\alpha; f_\alpha) \mid x_\alpha \in X,\ \lambda_\alpha \in \Lambda,\ f_\alpha \in F,\ \alpha = 1 \sim M\} \quad (13)$$

2. Select a training point from S_{LBP} at random.

3. Calculate the estimate of the network by eq.(9).

4. Calculate the squared error.

$$E_\alpha = \frac{1}{2}|f(x_\alpha) - \hat{f}(x_\alpha)|^2 \quad (14)$$

5. Update the parameters if and only if the selected point falls into the windows of their units.

$$\begin{aligned} a_i^{t+1} &= a_i^t - \alpha_{LBP}\,\delta_\alpha \frac{c_i^t (x_\alpha - b_i^t)}{(a_i^t)^2}\psi' \\ b_i^{t+1} &= b_i^t - \alpha_{LBP}\,\delta_\alpha \frac{c_i^t}{a_i^t}\psi' \\ c_i^{t+1} &= c_i^t + \alpha_{LBP}\,\delta_\alpha \psi \\ &\text{for } (x_\alpha, \lambda_\alpha) \in W_i \end{aligned} \quad (15)$$

where $\delta_\alpha = f(x_\alpha) - \hat{f}(x_\alpha)$ and α_{LBP} represents a learning coefficient.

6. Repeat step 2 to 5 until the errors are minimized.

EXPERIMENTS

The mapping networks are evaluated by two factors: the capabilities of approximation and interpolation. The former is how precise the network approximates training data. The later is how plausible the network interpolates test data.

We used $\psi(x) = \frac{x}{x_0}\exp\left(\frac{x^2}{2x_0^2}\right)$ as a FW, which satisfies both the admissibility and stability conditions. The x_0 was set to 0.07. We prepared the function $f(x) = \sin(3x)\cos(5(x - 0.5))$ defined in $[-1, 1]$

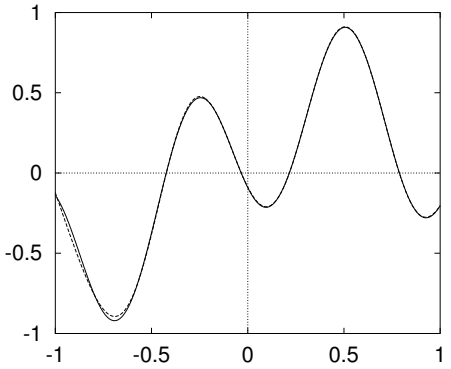

Figure1: Result by SERWANN.

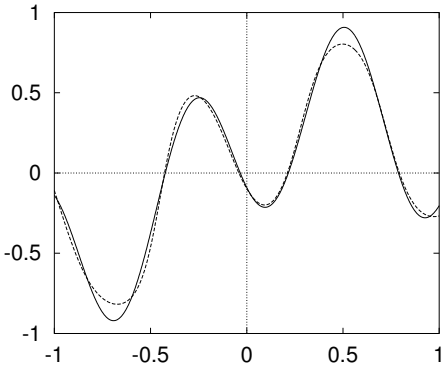

Figure2: Result by the BP network.

and sampled it at sampling rate 8[Hz] for training.

The supports of FW and the target are:

$$\mathrm{supp}_{0.05}(\psi)=[-0.1, 0.1]$$
$$\mathrm{supp}_{0.05}(\Psi)=[1.0, 7.0]$$
$$\mathrm{supp}_{0.05}(f)=[-1.0, 1.0]$$
$$\mathrm{supp}_{0.05}(F)=[0.5, 2.0]$$

The ranges of dilation and translation are $a \in [2.0, 8.0]$ and $b \in [-1.0, 1.0]$, respectively.

We experimented under the following conditions. The learning steps in the construction of networks and the minimization of errors were 5000 and 50000, respectively. The number of initial hidden units was zero, i.e. $N^0 = 0$. In this paper, the neighbors were not considered, i.e. $l = 0$. Furthermore, the learning constant and learning coefficient were both set to 0.03, i.e. $\alpha_{CL} = \alpha_{LBP} = 0.03$. The fluctuations in recruiting hidden units were determined from the interval $[-0.1, 0.1]$ at random.

SERWANN converged to 5 to 9 hidden units, 6.6 on average in 100 trials. The best result by SERWANN with 7 hidden units is illustrated in Fig.1. In this figure, the solid line denotes the target and the broken one is the approximated function.

For the comparison, the result by a BP network with 7 hidden units is shown in Fig.2. This is the best approximation in 100 trials. The values of learning and inertial coefficients were 0.05 and 0.5, respectively and the gradient of sigmoid function was 0.2. The initial values of weights and thresholds were determined from the interval $[-1.0, 1.0]$ at random. Furthermore, on-line algorithm were adapted as the updating method because we focus on the convergence speed.

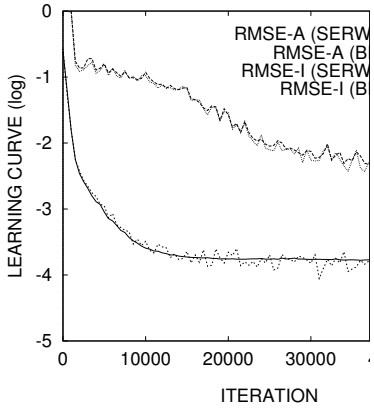

Figure3: Learning curve.

The learning curves of both networks are illustrated in Fig.3. Table 1 shows the figure of merit of each network. In this table, RMSE-A and RMSE-I represent RMSE for training data and test data, respectively.

SERWANN has better result than the BP network. In Fig.3, SERWANN shows fast convergence. On the total learning time, the computational cost of SERWANN was 70 percent of that of the BP network. This implies the fast con-

Table1: RESULTS.

Network	RMSE-A	RMSE-I
SERWANN	1.70×10^{-4}	2.02×10^{-4}
BP net	2.45×10^{-3}	2.26×10^{-3}

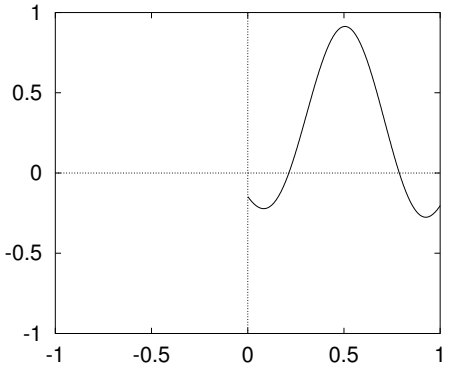

Figure4: Partial retrieval by SERWANN in $[0, 1]$.

vergence of the LBP algorithm. The learning steps for a convergence criterion RMSE-A $< 10^{-2}$ in SERWANN and the BP network were about 2000 and 27000 steps, respectively. The rate of convergence was 95 and 67 percent in SERWANN and the BP network, respectively.

The partial retrieval by SERWANN is shown in Fig.4, where the approximated function is reconstructed in the interval $[0, 1]$. This required 5 out of 7 hidden units. This is effective when the size of target is large.

CONCLUSIONS

This paper have proposed SERWANN, which was incorporated wavelets and NNs. SERWANN handled function approximating problems as two tasks: optimizing the network structure and minimizing errors.

The algorithms consisted of two processes. First, hidden units were gradually recruited to cover the time-frequency region of the target by the windows of FWs. Next, the network parameters were updated by the LBP algorithm to minimize the errors.

The capabilities of SERWANN by computer simulations through the comparison with the BP network. In addition, it was shown that SERWANN realize fast convergence and high convergence rate. Furthermore, it was shown that SERWANN can partially retrieve the approximated function using some hidden units.

REFERENCES

Azimi-Sadjadi,M.R., Sheedvash,S., Trujillo, F.O., (1993). Recursive dynamic node creation in multilayer neural networks, *IEEE Trans. on Neural Networks*, **4**, 207–220.

Chui,C.K., (1992). *An Introduction to Wavelets*, Academic Press.

Hagiwara,M., (1990). Novel back propagation algorithm for reduction of hidden units and acceleration of convergence using artificial selection, *Proc. ICNN*, 625–630.

Hecht-Nielsen,R., (1987). Kolmogorov's mapping neural network existence theorem, *Proc. ICNN*, 11–13.

Kohonen,T., (1989). *Self-Organization and Associative Memory*, Springer-Verlag.

Pati,Y.C., Krishnaprasad,P.S., (1993). Analysis and synthesis of feedforward neural networks using discrete affine wavelet transformations, *IEEE Trans. on Neural Networks*, **4**, 73–85.

Rumelhart,D.E., Hinton,G.E., Williams,R.J., (1985). Learning internal representations by error propagation, *PDP*, 318–362. MIT Press.

Zhang,Q., Benveniste,A., (1992). Wavelet networks, *IEEE Trans. on Neural Networks*, **3**, 889–898.