COMPARISON OF ACTIVATION FUNCTIONS IN MULTILAYER NEURAL NETWORK FOR PATTERN CLASSIFICATION

Kazuyuki HARA† Kenji NAKAYAMA‡ †
†Graduate School of Nat. Sci. & Tech., Kanazawa Univ.
‡Faculty of Tech., Kanazawa Univ.

2–40–20 , Kodatuno , Kanazawa , 920 JAPAN
E–mal : nakayama@haspnn1.ec.t.kanazawa-u.ac.jp

ABSTRACT: This paper discusses properties of activation functions in multilayer neural network applied to pattern classification. A rule of thumb for selecting activation functions or their combination is proposed. The sigmoid, Gaussian and sinusoidal functions are selected due to their independent and fundamental space division properties. The sigmoid function is not effective for a single hidden unit. On the contrary, the other functions can provide good performance. When several hidden units are employed, the sigmoid function is useful. However, the convergence speed is still slower than the others. The Gaussian function is sensitive to the additive noise, while the others are rather insensitive. As a result, based on convergence rates, the minimum error and noise sensitivity, the sinusoidal function is most useful for both without and with additive noise. Property of each function is discussed based on the internal representation, that is the distributions of the hidden unit inputs and outputs. Although this selection depends on the input signals to be classified, the periodic function can be effectively applied to a wide range of application fields.

I INTRODUCTION

Advantage of multilayer neural networks (NNs) trained by the back-propagation (BP) algorithm is to extract common properties, features or rules, which can be used to classify data included in several groups [1]. Especially, when it is difficult to analyze the common features using conventional methods, the supervised learning, using combinations of the known input and output data, becomes very useful.

We studied the multi-frequency signal classification using multilayer neural network[2], [3]. Since the frequencies are assigned alternately to several groups, it is very difficult to distinguish the waveforms within a short period, and the limited number of samples by conventional methods.

The following advantages of the NN over conventional methods were confirmed. The neural network can classify the signals using a small number of samples and a short observation period with which Fourier transform can not classify. The number of calculation is sufficiently smaller than the convolution calculation, required in digital filters.

In the previous work, a sigmoid function was used. However, it is not always optimum. Therefore, properties of activation functions are investigated in this paper. For this purpose, some typical functions are taken into account. They include a sigmoid function, a radial basis function[2] and a periodic function. They will be compared with each other in classifying multi–frequency signals. Effects of noisy signals will be also discussed in the training and classification processes.

As a result, a rule of thumb for selecting the suitable functions and the combination of several kinds of functions will be provided.

II MULTI-FREQUENCY SIGNALS

Multi–frequency signals are defined by

$$\begin{aligned} x_{pm}(n) &= \sum_{r=1}^{R} A_{mr}\sin(\omega_{pr}nT+\phi_{mr}) \quad (1)\\ n &= 1 \sim N \ , \ \omega_{pr}=2\pi f_{pr} \end{aligned}$$

T is a sampling period. M samples of $x_{pm}(n), m = 1 \sim M$, are included in the group X_p as follows.

$$X_p = \{x_{pm}(n), m = 1 \sim M\}, p = 1 \sim P \quad (2)$$

In one group, the same frequencies are used.

$$F_p = [f_{p1}, f_{p2}, \ldots, f_{pR}]Hz, p = 1 \sim P \quad (3)$$

Amplitude A_{mr} and phase ϕ_{mr} are generated as random numbers, uniformly distributed in following ranges.

$$0 < A_{mr} \leq 1, \quad 0 \leq \phi_{mr} < 2\pi \quad (4)$$

0-7803-1901-X/94 $4.00 ©1994 IEEE

III MULTILAYER NEURAL NETWORK

3.1 Network Structure and Equations

A single-layer neural network is taken into account. N samples of the signal $x_{pm}(n)$ are applied to the input layer in parallel. The nth input unit receives x_{pm}(n). Connection weight from the nth input to the jth hidden unit is denoted w_{nj}. The input and output of the jth hidden unit are given by

$$net_j = \sum_{n=0}^{N-1} w_{nj} x_{pm}(n) + \theta_j \tag{5}$$

$$y_j = f_H(net_j) \tag{6}$$

Letting the connection weight from the jth hidden unit to the kth output unit be w_{jk}, the input and output of the kth output unit are given by

$$net_k = \sum_{j=0}^{J-1} w_{jk} y_j + \theta_k \tag{7}$$

$$y_k = f_O(net_k) \tag{8}$$

The activation function of the output layer is the sigmoid function.

The number of output units is equal to that of the signal groups P. The neural network is trained so that a single output unit responds to one of the signal groups.

3.2 Training and Classification

Signals are categorized into training and untraining sets, denoted X_{Tp} and X_{Up}, respectively. Their elements are expressed by $x_{Tpm}(n)$ and $x_{Upm}(n)$, respectively.

The neural network is trained by using x_{Tpm}(n), m = 1 ~ M_T, for the pth group. Here, M_T is the number of the training data. After the training is completed, the untrained signals x_{Upm}(n) are applied to the NN, and the output is calculated. For the input signal x_{Upm}(n), if the pth output y_p has the maximum value, then the signal is exactly classified. Otherwise, the network fails in classification.

IV SELECTION OF ACTIVATION FUNCTIONS

What kinds of activation functions should be selected is very important. At the same time, it is a very difficult problem. In this paper, the following typical functions are selected for the hidden layer.

When binary target can be considered, then the sigmoid function can be used in the output layer.

Sigmoid function:

$$y_j = f_{sig}(net_j) = \frac{1}{1 + e^{-(net_j)}} \tag{9}$$

Sinusoidal function:

$$y_j = f_{sin}(net_j) = \sin(\pi net_j) \tag{10}$$

Gaussian function:

$$y_j = f_{gau}(net_j) = e^{-net_j^2} \tag{11}$$

The input vectors are distributed in a N-dimensional space. Three functions divide the space as follows:

$$f_{sig}(net_j) \begin{cases} > \alpha_+, & net_j > T_{sig} \\ < \alpha_-, & net_j < T_{sig} \end{cases} \tag{12}$$

$$f_{sin}(net_j) \begin{cases} > \alpha_+, & |net_j - (2n\pi + \frac{\pi}{2})| < T_{sin} \\ < \alpha_-, & |net_j - (2n\pi + \frac{3}{2}\pi)| < T_{sin} \end{cases} \tag{13}$$

$$f_{gau}(net_j) \begin{cases} > \alpha_+, & |net_j| < T_{gau} \\ < \alpha_-, & |net_j| > T_{gau} \end{cases} \tag{14}$$

Here, n is integer.

These space division fundamental, and independent to each other. This is an idea behind selecting the above three functions.

Next step of selecting activation functions is how to combine them. It is also highly dependent on the distribution of the input signals, and is very hard to determine before hand. For this reason, both the homogeneous function and the composite functions are investigated.

V SIMULATION OF TRAINING AND CLASSIFICATION WITHOUT NOISE

5.1 Multi-frequency Signals

The number of frequency components is R = 3, and the signal groups is P = 2, respectively. The frequency components are located alternately between the groups as follows: F_1 = [1, 2, 3] Hz for Group 1 (#1) and F_2 = [1.5, 2.5, 3.5] Hz for Group 2 (#2). The sampling frequency is 10 Hz, that is T = 0.1 sec. The number of samples N is 10. Therefore, the observation interval is 1 sec.

5.2 Training and Classification

x_{Tpm}(n), m = 1 ~ 200 and x_{Upm}(n), m = 1 ~ 1800 are used. Simulation results are shown in Table 1. The training converged using three hidden units for all activation functions. In the case of the Gaussian and the sinusoidal function, the training almost converged with one hidden unit. Detailed discussion will be provided in Sec. 7.

Table 1:Classification rates by three functions[%]

Activation Function	Hidden Unit	Training		Untraining	
		#1	#2	#1	#2
Sigmoid	1	44.5	100	47.9	100
	3	100	100	97.4	100
Sinusoidal	1	86.0	99.0	79.8	99.0
	3	100	100	92.6	100
Gaussian	1	99.5	100	98.1	100
	3	100	100	99.1	99.9

VI SIMULATION OF TRAINING AND CLASSIFICATION WITH WHITE NOISE

6.1 White Noise

White noise, denoted *noise(n)*, is generated as random number, and is added to the signal $x_{pm}(n)$. Noisy signal $x'_{pm}(n)$ is given by

$$x'_{pm}(n) = x_{pm}(n) + noise(n) \qquad (15)$$

6.2 Training and Classification

The noisy multi-frequency signals are used for training. N is 10 and M is 200 for each group. After training, untraining signals with white noise are applied, and classification rates are evaluated. White noise is uniformly distributed in the range ±0.5. The results are shown in Table 2. Columns with (A) and (B) list the recognition rates using the training signals without and with white noise, respectively. The NN trained without noise is also used for comparison. From these results, it can be confirmed that training using noisy signals is useful to achieve robustness.

Table 2: Classification rates using training signals without (A) and with (B) white noise [%]

Activation Function	Hidden Unit	(A)		(B)	
		#1	#2	#1	#2
Sigmoid	1	47.0	52.9	92.8	28.5
	3	97.3	8.4	82.6	78.0
Sinusoidal	1	80.2	20.9	61.7	87.7
	3	65.9	36.2	79.9	82.7
Gaussian	1	98.2	4.8	71.7	65.9
	3	85.3	46.3	79.8	70.2

6.3 Convergence Rates

Figure 1 shows learning curves obtained using the three hidden units. The NN with the Gaussian function can converge faster than the other. However, the error does not well decreased. The NN with the sinusoidal function can also converge faster. At the same time, the error can be well decreased. A convergence rate using the sigmoid function is slow. However, the error can reach to the same level as in using the sinusoidal function.

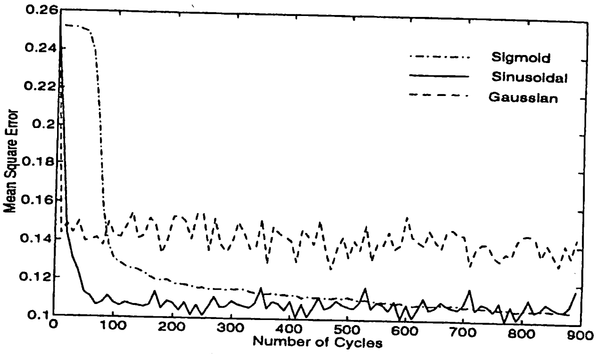

Figure 1: Learning curves

VII COMPARISON OF THREE ACTIVATION FUNCTIONS

7.1 Convergence Property Using Single Hidden Unit

The NNs trained without noise are further investigated by hidden unit output distribution. Figure 2 illustrates this distribution, using the sigmoid (a1), the sinusoidal (b2) and the Gaussian functions (c1).

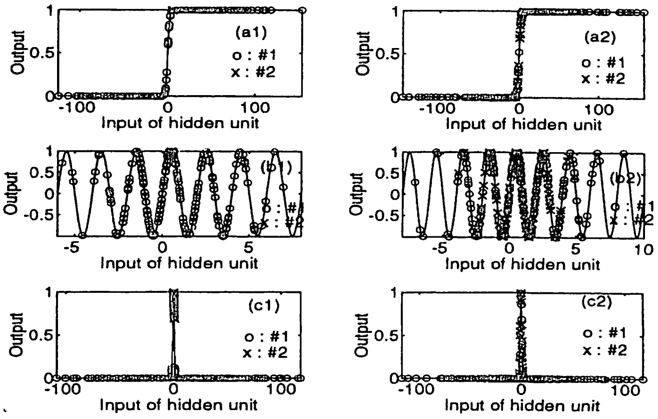

Figure 2: Hidden unit input and output distributions

In the case of using the Gaussian function, the #2 signals locate around the peak. Differential coefficients around the peak are large, then, the #2 data can be located in this area very fast. Most of the #1 data are distributed both sides. On the other hand, the distribution of the hidden unit inputs easily spread. Therefore, the Gaussian function is week for additive noise.

In the case of the sinusoidal function, the hidden unit inputs of the #2 locate near one of the peaks. The sinusoidal function have large differential coefficient except for the peak. Then the #2 data can locate around one of the peaks fast. The #1 data can locate in the region of $f_{sin}(net_j) < \alpha_{-}$.

In the case of the sigmoid function, the #1 and the #2 data have to locate the right or left side. Thus, the network have to adjust the weights , with which the input data are completely separated into the right or the left side. This requirement does not match the distribution of the input patterns.

From these results, the hidden unit inputs of the multi-frequency signals is concentrated on a narrow range for one group, and the other is distributed widely for the other group.

Figure 2 also shows the hidden unit inputs and output distributions, in which random noise is added. The networks are trained without noisy signals. In the case of the Gaussian, the #2 data distributed over the other region. Because a single peak is very narrow. Then these data easily move to the other group's region. Thus, the accuracy is decreased by noisy signals. The sinusoidal, the #2 data also widely distributed , but the sinusoidal function is periodic function. So, there are several extream values. Although the accuracy is decreased by the noise, it is higher than that of the Gaussian function.

7.2 Convergence Property Using Three Hidden Units

Homogeneous Activation Funtion:

Figures 3, 5 and 7 show distributions of the hidden unit inputs and outputs. The NN is trained by using the signals without noise. The sigmoid, the sinusoidal and the Gaussian functions are separately used. For each figure, (a), (b) and (c) correspond to one of the hidden unit. (a1), (b1) and (c1) are the response for the #1 data and (a2), (b2) and (c2) are for the #2 data. From these figures, there are two type of distributions, that is concentrated and dispersed distribution. One of two groups locate at near the peak of the functions and the other distribute widely.

In Fig. 3, it is very interesting that the #2 data locate at the middle of the slope. Since this region is not saturated to one or zero, accuracy will be easily changed by adding the noise. As shown in Table 2, the classification rates are 97.3 % for #1 and 8.4 % for #2. Thus, accuracy for #2 greatly reduced.

Figures 4, 6 and 8 shows distribution of the inputs of the two output units. In this figures, (a) and (b) correspond to the #1 and the #2 data, respectively. The region of overlap of the solid and the doted lines will cause miss classification. We can investigate from these figures, how the hidden units separate the signals into two groups. From the figures, the input space of the output units are well separated by the sigmoid and sinusoidal function. So, it can be concluded that three hidden units cooperate to make the distribution of the inputs to the output unit be in linear separable.

Composite Activation Functions:

Three functions can be combined in the same hidden layer. This combination is called, Composite Activation Function in this paper.

Table 3 shows classification rates using the signals without noise. In this table, the symbols A through F correspond to the combination of the functions. Training converged in all combinations.

The combination C, having three Gaussian functions achieve the best accuracy. The convergence rate is also the fastest among three function.

The combination D achieved better accuracy than A, B, E, and F. The composite activation functions from D to F, did not achieve better accuracy than the function C.

Table 3: Classification rates using signals without noise

	Combination of functions			Training		Untraining	
	Sig	Sin	Gauss	#1	#2	#1	#2
A	3	0	0	100	100	97.4	100
B	0	3	0	100	100	92.6	100
C	0	0	3	100	100	99.1	99.9
D	1	1	1	100	100	100	98.3
E	2	1	0	99.5	100	97.4	100
F	2	0	1	100	100	97.4	100

Table 4 shows classification rates of the network trained using noisy signals. Training did not converge in all cases in this table. The network using the homogeneous activation function A, B, C has the best accuracy. The composite activation function networks C, D, E, F did not work well than the homogeneous function network. The Gaussian function is very sensitive to the additive noise.

Then, it can be concluded that the combined activation functions have no advantage over the homogeneous one.

Table 4: Classification rates using signals with noise

	Combination of functions			Training		Untraining	
	Sig	Sin	Gauss	#1	#2	#1	#2
A	3	0	0	83.5	86.0	82.6	78.9
B	0	3	0	84.5	89.0	79.9	82.7
C	0	0	3	87.0	81.5	79.8	70.2
D	1	1	1	77.0	92.5	69.1	84.3
E	2	1	0	88.5	77.0	80.9	67.8
F	2	0	1	78.5	98.5	63.8	85.9

VIII CONCLUSIONS

Properties of the activation function for multi-frequency signal classification has been discussed using multilayer neural network supervised by BP algorithm. The Gaussian function can provide the highest performance for the input signals without noise. But it is sensitive to the additive noise. The sigmoid function is not useful for a single hidden unit. When several hidden unit is used, the sigmoid function is useful. This function is insensitive to the additive noise. Based on convergence rates, the minimum error and noise sensitivity, the sinusoidal function is the most useful for both with and without noise. Furthermore, the homogeneous activation function is much better than the composite type activation function in multi-frequency signal classification.

References

[1] D.E.Rumelhart and J.L.McCelland et al, "Parallel Distributed Processing", MIT Press, 1986.

[2] Philip D. Wasserman, "Advanced Methods in NEURAL COMPUTING", VAN NOSTRAND REINHOLD, pp.147-155, 1993.

[3] K.Hara and K.Nakayama, "Multi-frequency signal classification using multilayer neural network trained by backpropagation algorithm (in Japanese)", Tech., Rep. IEICE, NC92-75, pp.47-54.

[4] K.Hara and K.Nakayama, "High resolution of multi- frequencies using multilayer networks trained by back-propagation algorithm", Proc. WCNN'93, Portland Oregon, vol.IV, pp.675-678.

[5] K.Hara and K.Nakayama, "Classification of multi-frequency signals with random noise using multilayer neural networks", vol.I, pp.601-604.

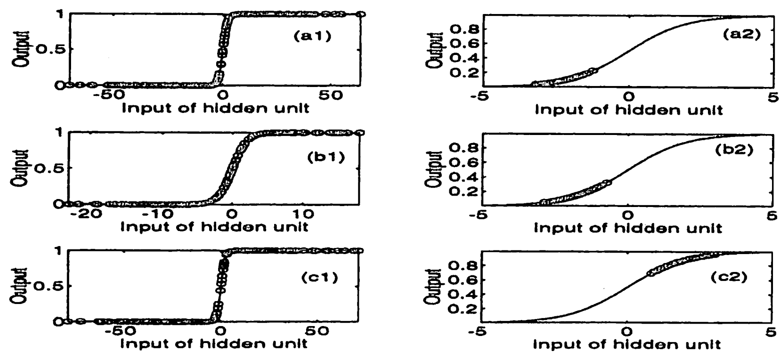

Figure 3: Distribution of sigmoid hidden unit outputs

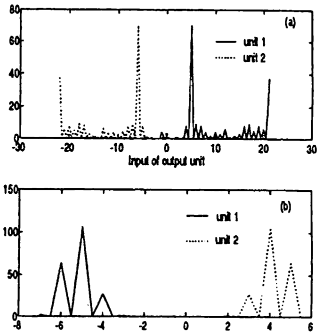

Figure 4: Distribution of output unit inputs

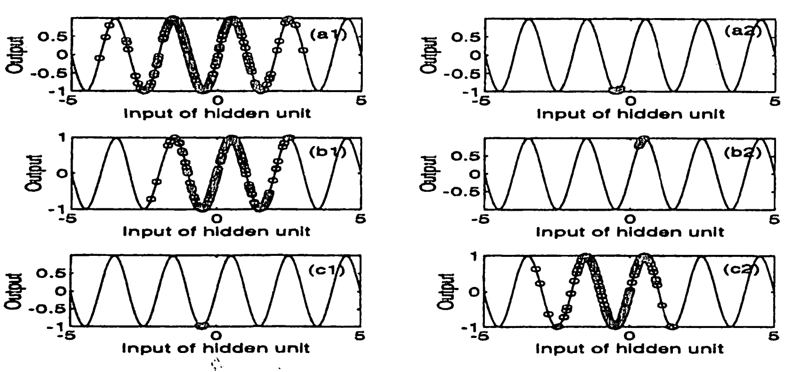

Figure 5: Distribution of sinusoidal hidden unit outputs

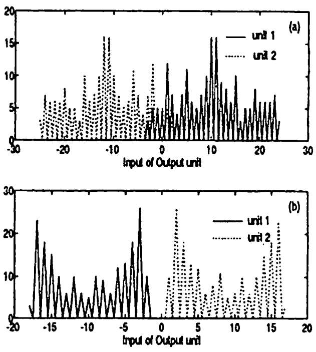

Figure 6: Distribution of output unit inputs

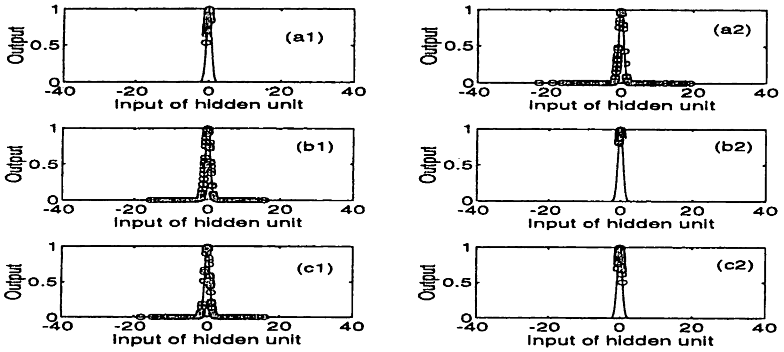

Figure 7: Distribution of Gaussian hidden unit outputs

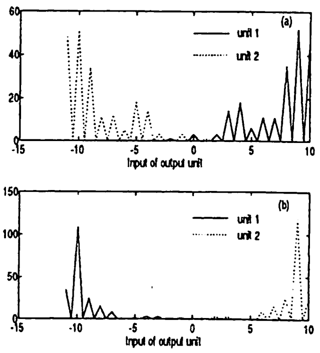

Figure 8: Distribution of output unit inputs