Novel Back Propagation Algorithm for Reduction of Hidden Units and Acceleration of Convergence using Artificial Selection

MASAFUMI HAGIWARA

Dept. of Elec. Eng.
Facul. of Sci. and Tech.
Keio University,
3-14-1 Hiyoshi, Kohoku-ku, Yokohama
223 Japan

Abstract Novel back propagation algorithm with artificial selection is proposed. It is effective for both fast convergence and reduction of the number of hidden units. The main feature of the proposed algorithm is detection of the worst hidden unit. This is done by using the proposed "Badness factor" which indicates badness of each hidden unit. It is the sum of back propagated error component over all patterns for each hidden unit. For the fast convergence, all the weights connected to the detected worst unit are reset to small random values at suitable time. As for the reduction of the hidden units, detected bad units are erased precedently. Computer simulation results show the effectiveness of the proposed algorithm. For example, the numbers of hidden units in the EX-OR problems converge to 2 (theoretical number).

1. INTRODUCTION

Back propagation algorithm is the most popular learning paradigm, and has been used for various tasks such as recognition, generalization, control and classification [1-4]. However, there are two important problems associated with the back propagation learning of multi-layer network.

1) How to shorten the learning time.
2) How to reduce the number of hidden units.

As for the learning time problem, several methods have been proposed [5]-[9]. As for the hidden units, there is a paper proposing reduction of hidden units [10]. According to ref.[10], however, average number of hidden units in EX-OR problems is 2.8, whereas its theoretical number is 2. So, this method seems not to be effective. Therefore, it has been very difficult to reduce the number of hidden units effectively, and it has been quite general that the number is determined empirically [2].

In this paper, novel back propagation algorithm with artificial selection is proposed. It is effective for both fast convergence and reduction of hidden units. The main feature of the proposed algorithm is detection of the worst hidden unit. This is done by using the proposed "Badness factor" which indicates badness of each hidden unit. For the fast convergence, all the weights connected to the detected worst unit are reset to small random values at suitable time. As for the reduction of the hidden units, detected bad units are erased precedently. The proposed reduction method of hidden units will also offer great contribution to analyses of hidden units of layered neural networks. Computer simulation results show the effectiveness of the proposed algorithm. For example, the numbers of hidden units in the EX-OR problems converge to 2 (theoretical number).

In Sec.2, back propagation algorithm with artificial selection is explained. Computer simulation results are shown in Sec.3.

2. BACK PROPAGATION ALGORITHM WITH ARTIFICIAL SELECTION

2.1 Detection of the worst hidden unit

Detection of the worst hidden unit is done by using the "Badness factor". Now it is briefly explained using Fig.1.

At first, the total error input to the *i*-th unit of *(k-1)*th layer (hidden layer) for patter *p* is defined.

$$e_i^{k-1,p} = \sum_j w_{i\ j}^{k-1\ k}\ \delta_j^k \qquad (1)$$

where

$$\delta_j^k = (t_j - o_j^k)\, f'(net_j^k) \qquad \text{(Output layer)} \qquad (2)$$

$$\delta_j^{k-1} = \Big(\sum_l w_{j\ l}^{k-1\ k}\, \delta_l^k\Big)\, f'(net_j^{k-1}) \qquad \text{(Hidden layer)} \qquad (3)$$

$$net_j^k = \sum_i w_{i\ j}^{k-1\ k}\, o_i^{k-1} \qquad (4)$$

$$o_j^k = f(net_j^k) \qquad (5)$$

$w_{i\ j}^{k-1\ k}$ is the weight from *i*-th unit of *(k-1)*th layer (hidden layer) to *j*-th unit of *k*-th layer (output layer), t_j is the desired output, and o_j^k is the output.

Weight changes according to

$$\Delta w_{i\ j}^{k-1\ k} = \eta\, \delta_j^k\, o_i^{k-1}, \qquad (6)$$

where, η is the learning constant.

Here, "Badness factor" of *i*-th hidden unit is defined.

$$BAD_i^{k-1} = \sum_p (e_i^{k-1,p})^2$$

$$= \sum_p \Big(\sum_j w_{i\ j}^{k-1\ k}\, \delta_j^k\Big)^2 \qquad (7)$$

From Eq.(7) and Fig.1, the "Badness factor" indicates the degree of convergence. It is the sum of back propagated error component over all patterns for each hidden unit. All the weights connected to the hidden units which have large "Badness factor" should be changed largely, and vice versa. The hidden unit which has the largest "Badness factor" can be considered as the worst hidden unit which is the powerful unit to prevent the network from convergence.

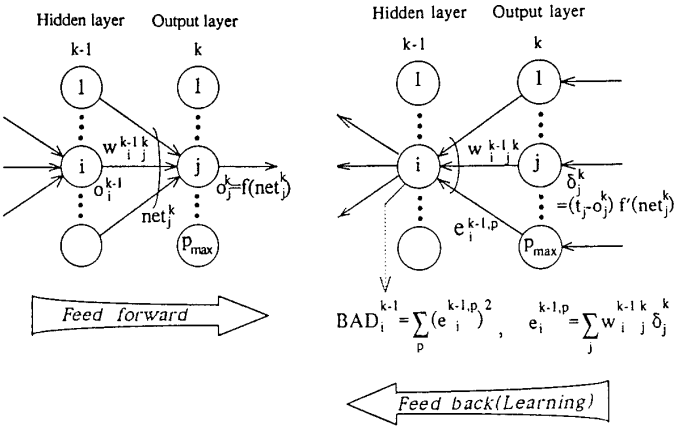

Fig.1 Network and notation.

2.2 Acceleration of convergence

Acceleration of convergence is based on the previous paper [8]. It is a reasonable concept for nature, "Artificial selection". The following is basic concept of the algorithm for acceleration of convergence.

ALGORITHM 1 (basic concept for acceleration of convergence)

1) Monitoring the total squared error at the output layer.
2) When the variation of the total squared error becomes small, the worst hidden unit is detected at suitable time.
3) All the weights connected to the detected worst hidden unit are reset to small random values (artificial selection).
4) Addition of a hidden unit is done when the number of reset reaches a certain number.

5) Repeat 1)~4).

It is often utilized to escape from local minimum that some disturbance is added to a network. The proposed method is more effective, because the proposed method which gives disturbance to the network is derived theoretically.

2.3 Reduction of the number of hidden units

After convergence of a network, reduction of the number of hidden units is possible by erasing the detected bad units.
The following is the basic concept of the algorithm for reduction of the number of hidden units.

ALGORITHM 2 (basic concept for reduction of the number of hidden units)

1) Confirm the convergence.
2) Continue the leaning for further convergence until some condition is satisfied.
3) Detect the worst hidden unit.
4) Erase the detected hidden unit. (All of the weights should be copied before erasing.)
5) Repeat 1)~4). (ALGORITHM 1 is used to accelerate convergence until the network is converged again.)

As for as the artificial selection proposed in this paper, heuristic method can be also considered.

3. SIMULATIONS

To demonstrate the effectiveness of the proposed algorithm, two examples are used ; one is relatively difficult problem (character recognition), another is an easy problem (EX-OR problem),

3.1 Relatively difficult problem

Here, as a relatively difficult problem, recognition of 20 characters shown in Fig.2 is done.

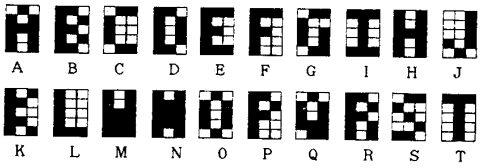

Fig.2 Used data of 20 characters.

The following simulation conditions are used.

1) Three-layer network is used.
2) Momentum α is zero in order to make clear the performance difference and reduce the influence by this parameter.
3) The number of input units is 15.
4) The number of output units is 20, namely 20 patterns which are shown in Fig.2 are used for learning and recognition.
5) The range of modified sigmoid function output is between -1 and 1.
6) True recognition is regarded as the state where every sign of each output layer unit is same to that of the corresponding desired output.
7) The maximum number of learning sets is 10000 (200000 times learning in total, because in this paper one learning set means one learning of each pattern).

Then detailed simulation algorithm is explained.

ALGORITHM 1 (for acceleration of convergence)

1) Monitor the total squared error at the output layer every 10 sets.
2) When the total squared error does not reduced more than 5% compared with the previous monitoring time, artificial selection (weights reset) is done. (Once the reduction of hidden unit is done, the artificial selection is done every 100 times of such occasions to decrease the frequency of resets.)
3) When the network does not converge after 30 times resets, one hidden unit is added.

ALGORITHM 2 (for reduction of the number of hidden units)

1) After convergence, more 50 sets learning are done. Then detected worst unit is erased.

Table 1 shows the convergent time and the minimum number of the hidden units. In the case where the conventional back propagation algorithm is used, only half trials can be converged. On the other hand for the proposed algorithm, all of the trials can be converged and that quickly. In addition, it can be seen that the hidden units can be greatly reduced.

Figs. 3 and 4 show the example of learning characteristics. As for the conventional back propagation algorithm (Fig.3), the learning speed is slow and it is obvious that the number of hidden units is constant. In the case of the proposed algorithm (Fig.4), the total error at the output layer is rapidly reduced and the network converges around 1700 learning sets. After that, reduction of hidden units is effectively continued until around 3500 learning sets. In this case, the final number of hidden units is 8.

		Learning constant η			
		0.1	0.2	0.3	0.4
Proposed	Learning times to convergence (sets)	900	1090	670	390
	Minimum number of hidden units	6	10	11	16
Conventional	Learning times to convergence (sets)	3186	1306	×	6334
	Minimum number of hidden units	20	20	20	20

(a) 20 initial hidden Units

		Learning constant η			
		0.1	0.2	0.3	0.4
Proposed	Learning times to convergence (sets)	1700	920	1140	470
	Minimum number of hidden units	8	11	11	22
Conventional	Learning times to convergence (sets)	×	2257	×	×
	Minimum number of hidden units	30	30	30	30

(b) 30 initial hidden Units

Table 1 Learning times to convergence and the minimum number of hidden units (character recognition problem).

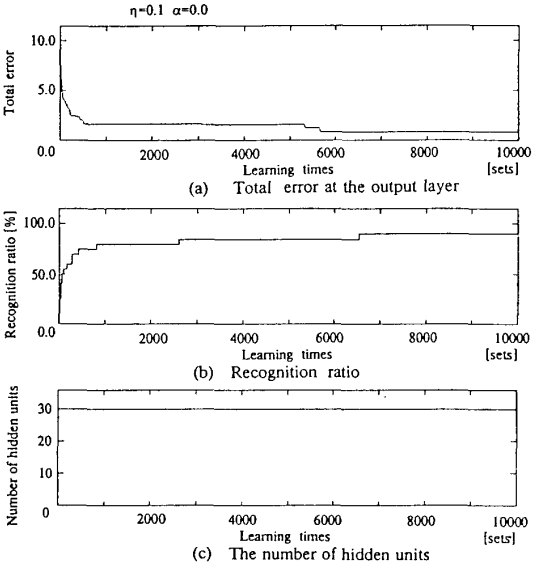

Fig.3 Learning characteristics for character recognition (conventional).

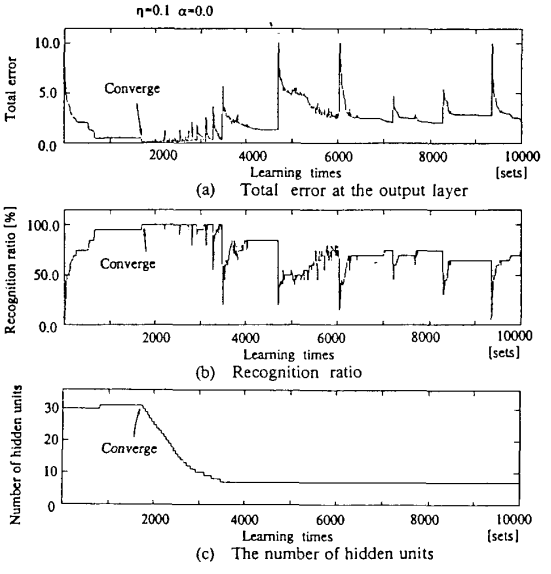

Fig.4 Learning characteristics for character recognition (proposed).

3.2 Easy problem (EX-OR problem)

Here, EX-OR problem as an easy one is done.

The simulation conditions different from 3.1 are following.

1) The number of input units is 2.
2) The number of output unit is 1.
3) The range of sigmoid function output is between 0 and 1.
4) True recognition is regarded when the output error is within -0.5 and 0.5.
5) One set learning means 4 times learning, because 4 patterns are learned.

Then detailed simulation algorithm is explained.

ALGORITHM 1 (for acceleration of convergence)

1) Monitor the total squared error at the output layer every 100 sets.
2) When the total squared error does not reduced more than 1% compared with the previous monitoring time, artificial selection (weights reset) is done. (Once the reduction of hidden unit is done, the artificial selection is done every 5 times of such occasions to decrease the frequency of resets.
3) When the network does not converge after 10 times resets, one hidden unit is added.

ALGORITHM 2 (for reduction of the number of hidden units)

1) After convergence, more 100 sets learning are done. Then detected worst unit is erased.

In general, the number learning times of neural network becomes large as the problem becomes difficult. And the frequency of disturbance to the network should be small when the problem is difficult. Therefore, parameters used in the algorithm 1 and 2 are not equal.

Number of initial hidden units		Trial number										Ave.
		#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	[sets]
1	①	2623	2170	1947	1663	1628	1658	1560	3147	2296	2274	2096. 6
	②	2623	2170	1947	1663	1628	1658	1560	3147	2296	2274	2096. 6
	③	×	×	×	×	×	×	×	×	×	×	×
2	①	3309	611	746	3495	487	745	1381	514	1117	1078	1348. 3
	②	3309	611	746	3495	487	745	1381	514	1117	1078	1348. 3
	③	×	821	746	×	487	×	×	519	×	×	9257. 3
3	①	7464	208	211	377	296	178	988	206	198	1596	1172. 2
	②	7564	308	311	477	396	278	1088	306	298	1696	1172. 2
	③	×	208	211	×	296	178	×	206	198	346	4664. 3
5	①	301	362	500	251	180	430	377	324	233	256	321. 4
	②	5564	1987	2076	747	674	1036	1854	1965	1542	1508	1895. 3
	③	301	362	500	251	180	384	377	324	233	256	321. 4
10	①	236	219	200	194	223	178	278	163	189	248	212. 8
	②	2873	1432	4061	2487	1701	4331	5106	2904	3991	3336	3222. 2
	③	236	219	200	194	223	178	278	163	189	248	212. 2

Note : ①(Upper line) : Learning times to convergence [sets] (proposed)
②(Middle line) : Learning times to 2 hidden units [sets] (proposed)
③(Lower line) : Learning times to convergence [sets] (conventional)

Table 2 Learning times to convergence and learning times to 2 hidden units (EX-OR problem).

Table 2 shows the convergent time and the minimum number of the hidden units where the initial number of hidden units are 1, 2, 3, 5, and 10. 10 times trials are done using different initial weights for each case. As for the conventional back propagation algorithm, only 40% trials can be converged when the initial number of hidden units is 2, and when that is 3, only 70% trials can be converged. As for the proposed algorithm, all of the trials can be converged and that quickly. In addition, all of the minimum numbers of hidden units are 2, which equals to the theoretical number.

Fig.5 shows the example of learning characteristics by the proposed algorithm for EX-OR problem, where the initial number of hidden units is 10. In this case, the network is converged for the first time at about 230 learning sets, and after that, reduction of hidden units is successfully done. At about 3000 learning sets, it reaches theoretical minimum number (2).

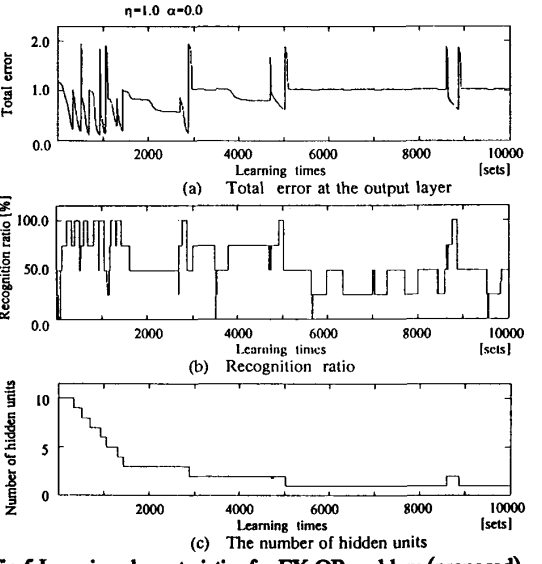

Fig.5 Learning characteristics for EX-OR problem (proposed).

4. CONCLUSIONS

Novel back propagation algorithm with artificial selection has beem proposed. It is effective for both fast convergence and reduction of the number hidden units.

The main feature of the proposed algorithm is detection of the worst hidden unit. This is done by using the proposed "Badness factor" which indicates badness of each hidden unit. For the fast convergence, all the weights connected to the detected worst unit are reset to small random values at suitable time. As for the reduction of the hidden units, detected bad units are erased precedently.

Computer simulation results show the effectiveness of the proposed algorithm. For example, the numbers of hidden units in the EX-OR problems converge to 2 (theoretical number).

The proposed algorithm can be used combined with other acceleration methods. And the proposed reduction method of hidden units will offer great contribution to analyses of hidden units of layered neural networks.

ACKNOWLEDGEMENT

The author is much grateful to Prof. Masao Nakagawa.

REFERENCES

[1] D.E.Rumelhart, J.L.McClelland and the PDP Research Group :"Parallel Distributed Processing", MIT Press, 1986.

[2] R.P.Gorman and T.J.Sejnowski :"Analysis of hidden units in a layered network trained to classify sonar targets", Neural Networks, vol.1, 1, pp.75-89, 1988.

[3] A.Waibel, T.Hanazawa, G.Hinton, K.Shikano, and K.J.Lang : "Phoneme recognition using time-delay neural networks", IEEE trans. Accoust., Speech, Signal Processing, vol-37, 3, pp.328-339, March 1989.

[4] D.Nguyen and B.Widrow :"The truck backer-upper: an example of self-learning in neural networks", Proc. of IJCNN'89, II p.357-363, June 1989.

[5] T.P.Vogl, J.K.Mangis, A.K.Rigler,W.T.Zink, and D.L.Alkon :"Accelerating the convergence of the Back-propagation method", Biological Cybernetics, 59, pp.257-263, 1988.

[6] T.Ash :"Increased rates of convergence through learning rate adaption", Neural Networks, Vol.1, 1988.

[7] H.Sawai, A.Waibel, P.Haffner, M.Miyatake, and K.Shikano :"Parallelism, hierarchy, scaling in time-delay neural networks for spotting Japanese phonemes/CV-syllables", Proc. of IJCNN'89, II p.81-88, June 1989.

[8] M.Hagiwara and M.Nakagawa:"Supervised learning with artificial selection", Proc. IJCNN'89, II p.611, June 1989.

[9] M.Hagiwara :"Accelerated back propagation using unlearning based on Hebb rule", Proc. IJCNN-90-WASH-DC, Jan. 1990.

[10] Y.Hirose, K.Yamashita, and S.Hijiya :"Back propagation method which varies the number of hidden units", Proc. Conf. of I.E.C.E. Japan, D-18, March 1989.

[11] Timur Ash :"Dynamic node creation in backpropagation networks", Proc. IJCNN'89, II p.623, June 1989.