ELSEVIER

Neurocomputing 48 (2002) 1009–1014

NEUROCOMPUTING

www.elsevier.com/locate/neucom

Speed-up of error backpropagation algorithm with class-selective relevance

In-Cheol Kim, Sung-Il Chien*

Department of Electronic and Electrical Engineering, Kyungpook National University, Buk-gu, Daegu 702-701, South Korea

Abstract

Selective attention learning is proposed to improve the speed of the error backpropagation algorithm for fast speaker adaptation. Class-selective relevance for measuring the importance of a hidden node in a multilayer Perceptron is employed to selectively update the weights of the network, thereby reducing the computational cost for learning.

Keywords: Class-selective relevance; Error backpropagation algorithm; Fast speaker adaptation

1. Introduction

A slow learning speed is one of the major drawbacks of the error backpropagation (EBP) algorithm generally employed in training a multilayer Perceptron (MLP). To accelerate the EBP algorithm, several modified methods, such as varying the learning rate during the learning process [6] or using different types of error function [3,5], have been suggested, which mainly focus on decreasing the number of iterations to improve the learning speed.

This paper proposes a selective attention method that reduces the computational cost required for MLP learning by selectively updating the weights of the network to speed up the learning process. The class-selective relevance is newly introduced to measure the importance of a hidden node in minimizing the mean square error (MSE) function for a given class. Those weights connected to the hidden nodes that are irrelevant to the considered class are fixed without updating during the learning process for the input patterns belonging to that class. The proposed method is particularly effective when

* Corresponding author. Tel.: +82-53-950-5545; fax: +82-53-950-5505.
E-mail address: sichien@ee.knu.ac.kr (S.-I. Chien).

0925-2312/02/$ - see front matter
PII: S0925-2312(02)00594-5

applied to a relearning task reconsidering already trained weights. Thus, its effectiveness is demonstrated using a representative example of a relearning task, speaker adaptation which is a training procedure for constructing a speaker-dependent speech system by adapting a speaker-independent system to a new speaker using a small amount of speaker-specific training data.

2. Class-selective relevance

The concept of relevance [4] has been introduced to measure the importance of a given hidden node for producing the appropriate output correctly. The relevance of the mth hidden node is defined by the incremental MSE computed without that node.

$$R_m = \sum_{p=1}^{P}\sum_{j=1}^{N}\left\{t_j^p - \varphi\left(\sum_{i=0}^{M} w_{ji}\rho_{mi}v_i^p\right)\right\}^2 - \sum_{p=1}^{P}\sum_{j=1}^{N}(t_j^p - o_j^p)^2. \tag{1}$$

Here, o_j^p and t_j^p denote the actual output and corresponding target value of the jth output node, respectively, and v_i^p denotes the output of the ith hidden node, upon presentation of input pattern p. The jth output node is connected to the ith hidden node via weight $w_{ij}\cdot\varphi(\cdot)$ is a sigmoid function, and ρ_{mi} is 0 for $m = i$ and one, otherwise. From Eq. (1), it is apparent that a hidden node with large relevance plays a very important role in learning, since removing that node results in a significant increase in the MSE. This idea has been successfully applied in several classification problems related to network pruning [2].

For selective attention learning, we propose to measure the relevance of each hidden node separately according to the class. Let Ω_k be a set of training patterns belonging to the kth class, ω_k. Then, the effect of the removal of the mth hidden node on increasing the MSE for class ω_k is measured as follows:

$$R_{mk} = \sum_{p\in\Omega_k}\sum_{j=1}^{N}\left\{t_j^p - \varphi\left(\sum_{i=0}^{M} w_{ji}\rho_{mi}v_i^p\right)\right\}^2 - \sum_{p\in\Omega_k}\sum_{j=1}^{N}(t_j^p - o_j^p)^2. \tag{2}$$

As an off-line step for speaker adaptation, this class-selective relevance, R_{mk} is calculated in advance using a baseline speech system. In an adaptation stage, the hidden nodes, depending on a class, are divided into relevant and irrelevant ones according to the threshold determined from the histogram of R_{mk} values. Then selective attention relearning is performed by updating only the incoming and outgoing weights of the relevant nodes.

Conceptually, this method has the same effect as weight pruning. However, the weights connected to the hidden nodes that are found to be irrelevant to a specific class are not actually pruned but rather frozen, and then become active again for input patterns belonging to other classes. Since the input patterns used for adaptation are characteristically similar with the initial training set whereby the relevance of each hidden node has been measured, it is expected that selective learning by the class-selective relevance can effectively remove redundant or unnecessary computation.

3. Baseline system build-up

A baseline speech system recognizing 20 Korean isolated-words was firstly constructed based on an MLP with one hidden layer. Two thousand speech data for 20 words were used for the initial training of the MLP. To extract the same dimensional feature vectors regardless of the signal length, the speech signal was partitioned into 19 frames with a length-dependent width; the longer the length of the speech signal, the wider the frame. Finally, 777-dimensional feature vectors consisting of magnitude coefficients derived from mel-scale filter banks [1] were extracted after applying a Hamming window with a half-frame shift.

The MLP consisted of 777 inputs, 35 hidden nodes, and 20 output nodes. The learning rate and momentum were assigned as 0.05 and 0.8, respectively. These learning parameters remained unchanged during the adaptation experiments to provide a fair comparison. For the initial MLP training before adaptation, the weights were initialized with random values drawn from a range $[-5 \times 10^{-3}, 5 \times 10^{-3}]$ and the learning termination criterion was determined as an MSE of 0.001. All experiments were carried out on a Pentium III-500 based PC with a Linux operating system. This MLP-based system is then to be adapted to a new speaker through the adaptation process.

4. Adaptation experiments

At the beginning of the adaptation task, the weights were initialized to the original weights obtained from the previous initial learning. Thereafter, the irrelevant hidden nodes, depending on the word class, were labelled based on their class-selective relevance computed in advance. In the experiment, a hidden node whose relevance value for a given class was lower than 0.01 was considered as irrelevant to that class. Fig. 1 shows the distribution of the irrelevant nodes on the hidden layer for 20 classes. On average, 20% of the nodes were determined as irrelevant, yet their distribution was strongly class dependent. The overall network size seems to be appropriate for our task because no nodes are commonly assigned as irrelevant across all classes. To evaluate our method, 20 simulations were performed with different input pattern presentations, and then the results were averaged. Furthermore, the adaptation simulation was repeated for five speakers to investigate the effectiveness of our method more exactly. As an adaptation database, 10 speech data for each word obtained from the person to be newly adapted were used. The results in Table 1 show that the selective attention method produced faster convergence than the standard EBP without lengthening the number of iterations, although the learning time somewhat varied depending on the speaker being adapted.

The next simulation was performed to show that our selective attention scheme could be successfully combined with other types of improved learning methods. We introduced Fahlman's learning method [3] that was proposed to shorten the learning iterations by solving the problem of premature saturation [5] inherent in EBP learning. Fig. 2 shows that Fahlman's method achieved significant time reduction of about 29% in an average sense when compared to the standard EBP. The learning time could

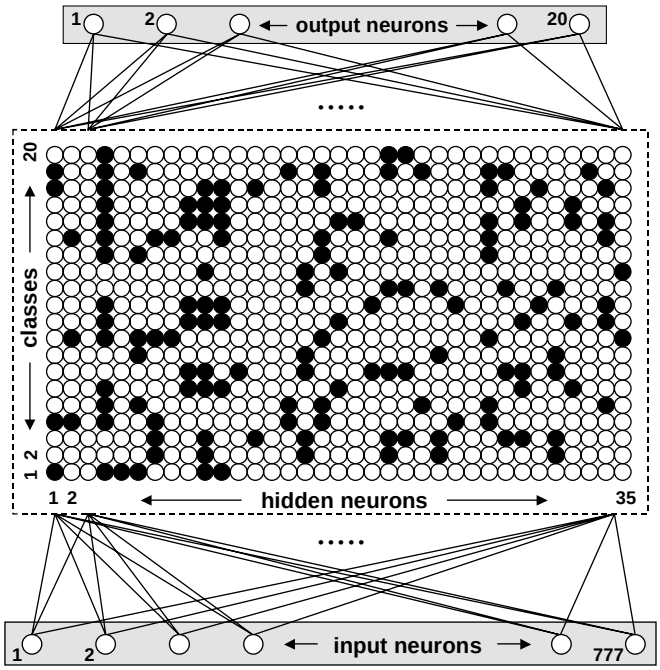


Fig. 1. Distribution of irrelevant hidden nodes (denoted by filled circle) depending on word class.

Table 1
Adaptation results using standard EBP and selective attention method

Speakers	Standard EBP		Selective attention	
	Iterations	Learning time (s)	Iterations	Learning time (s)
A	17.7	11.30	16.5	8.23
B	55.4	35.32	58.9	29.50
C	15.2	9.72	13.4	6.68
D	5.3	3.38	4.4	2.21
E	32.0	20.41	29.3	14.60

be further reduced with an average reduction ratio of 46% by combining it with the proposed selective attention.

5. Conclusions

A selective attention method based on class-selective relevance measuring the importance of a hidden node was proposed to accelerate the relearning speed of the EBP algorithm for fast speaker adaptation in an MLP. The weights of the unimportant

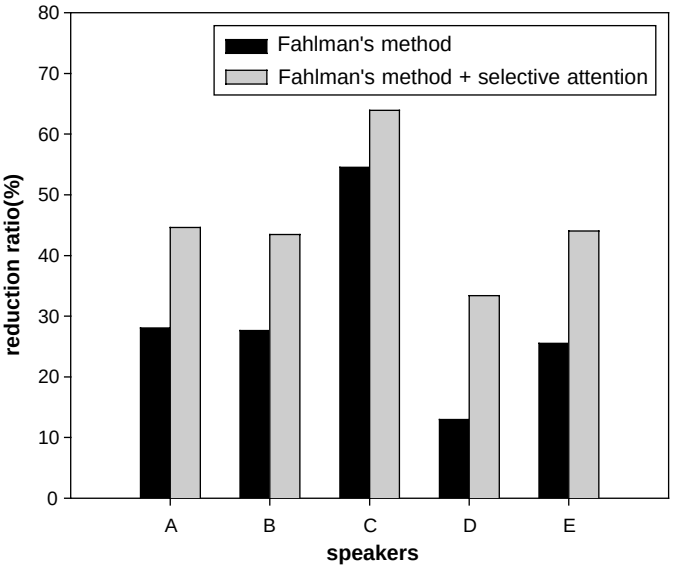


Fig. 2. Reduction ratios of relearning time compared to standard EBP for Fahlman's method and its selective attention version.

hidden nodes are frozen without updating to reduce the computational cost of relearning, thereby resulting in faster adaptation. From experimental results, we found that MLP learning could be considerably accelerated by the proposed attention technique and further improvement was also achieved when this method was combined with Fahlman's learning method.

Acknowledgements

This research was supported by the Brain Korea 21 Project and a grant no. (R01-1999-000233-0) from Korea Science and Engineering Foundation.

References

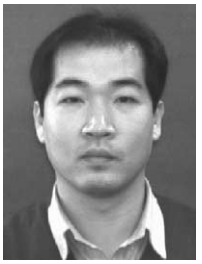
[1] S.B. Davis, P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, IEEE Trans. Acoust. Speech Signal Process. ASSP-28 (4) (1980) 357–366.

[2] S.S. Erdogan, G.S. Ng, K.H.C. Patrick, Measurement criteria for neural network pruning, in: Proceedings of the IEEE TENCON, Digital Signal Processing Applications, Vol. 1, November 1996, pp. 83–89.

[3] S.E. Fahlman, Faster-learning variations on back-propagation: an empirical study, in: Proceedings of the Connectionist Models Summer School, Carnegie Mellon University, 1988, pp. 38–51.

[4] M.C. Mozer, P. Smolensky, Using relevance to reduce network size automatically, Connect. Sci. 1 (1) (1989) 3–16.

[5] S.H. Oh, Improving the error backpropagation algorithm with a modified error function, IEEE Trans. Neural Networks 8 (3) (1997) 799–803.
[6] D.J. Park, B.E. Jun, J.H. Kim, Novel fast training algorithm for multilayer feedforward neural network, Electron. Lett. 28 (6) (1992) 543–545.



In-Cheol Kim received the B.S., M.S., and Ph.D. degrees in Electronic Engineering, all from the Kyungpook National University, Taegu, Korea in 1989, 1991, and 2001, respectively. From 1991 to 1996, he was a System Engineer in Computer Aided System Engineering Corp., Seoul, Korea. He is currently working as a postdoctoral researcher in BK'21 Information Technology Team, Kyungpook National University. His current research interests include pattern recognition, neural networks, artificial intelligence, and multi-modal human computer interaction.

Sung-Il Chien received the B.S. from Seoul National University, Seoul, Korea in 1977, the M.S. from the Korea Advanced Institute of Science and Technology, Seoul, Korea in 1981, and Ph.D. in Electrical and Computer Engineering from Carnegie Mellon University in 1988. Since 1981, he has been with the School of Electronic and Electrical Engineering, Kyungpook National University, Taegu, Korea, where he is currently a professor. His research interests are pattern recognition, neural networks, computer vision, and sensor fusion.