# Modified Error Function with Added Terms for the Backpropagation Algorithm

Weixing Bi[1], Xugang Wang[2], Ziliang Zong[3], and Zheng Tang[1]

[1] Faculty of Engineering, Toyama University, 930-8555 Toyama, Japan
biweixing613@hotmail.com, tang@iis.toyama-u.ac.jp
[2] Intelligence Engineering Laboratory, Institute of Software, The Chinese Academy of Science, Beijing 100080, China
wxg@iel.iscas.ac.cn
[3] Faculty of Computer Science, Shandong University, 250061 Jinan, China
pipizzl@hotmail.com

**Abstract.** We have noted that the local minima problem in the backpropagation algorithm is usually caused by update disharmony between weights connected to the hidden layer and the output layer. To solve this problem, we propose a modified error function with added terms. By adding one term to the conventional error function, the modified error function can harmonize the update of weights connected to the hidden layer and the output layer. Thus, it can avoid the local minima problem caused by such disharmony. Moreover, some new learning parameters introduced for the added term are easy to select. Simulations on the modified XOR problem have been performed to test the validity of the modified error function.

## 1 Introduction

We have noted that many local minima difficulties in the backpropagation learning for feedforward neural network are closely related to the neuron saturation in the hidden layer. Once such saturation occurs, neurons in the hidden layer will lose their sensitivity to input signals, and the propagation of information is blocked severely [1]. In some cases, the network can no longer learn [2]. The same phenomenon was also observed and discussed by Andreas Hadjiprocopis [1], Christian Goerick [2], Simon Haykin [3] and Wessels et al. [4].

In this paper, we explain this phenomenon as the weights update disharmony of between hidden and output Layers and propose a robust modified error function for the backpropagation algorithm in order to avoid the local minima problem that occurs due to neuron saturation in the hidden layer. Since a three-layered network is capable of forming arbitrarily close approximation to any continuous nonlinear mapping [5], our discussion will be limited to three-layered networks. To demonstrate the efficiency of the modified error function, we apply it to the backpropagation learning and conduct the simulation on the modified XOR problem. Furthermore, we compare the results with those of the backpropagation algorithm and simulated annealing method using the conventional error function.

F. Yin, J. Wang, and C. Guo (Eds.): ISNN 2004, LNCS 3173, pp. 338–343, 2004.

## 2 Weights Update Disharmony

The backpropagation algorithm is a gradient descent procedure used to minimize an objective function (error function) $E$. The most popularly used error function is the "*sum-of-squares*" that is given by

$$E = \frac{1}{2}\sum_{p=1}^{P}\sum_{j}^{J}(t_{pj} - o_{pj})^2, \tag{1}$$

where $P$ is the number of training patterns, $t_{pj}$ is the target value (desired output) of the $j$-th component of the outputs for the pattern $p$, $o_{pj}$ is the output of the $j$-th neuron of the actual output pattern produced by the presentation of input pattern $p$, and $J$ is the number of neurons in the output layer. To minimize the error function $E$, the backpropagation algorithm uses the following delta rule:

$$\Delta w_{ji} = -\eta \frac{\partial E}{\partial w_{ji}}, \tag{2}$$

where $w_{ji}$ is the weight connected between neurons $i$ and $j$ and $\eta$ is the learning rate. From the above equations, we can see that there is only one term related to the neuron outputs in the output layer. Weights are updated iteratively to make the neurons in the output layer approximate to their desired values. However, the conventional error function does not consider the neuron behavior in the hidden layer and what value should be produced in the hidden layer.

Usually, the activation function of a neuron is given by a sigmoid function:

$$f(x) = \frac{1}{1 + e^{-x}}. \tag{3}$$

There are two extreme areas that are called the saturated portion of the sigmoidal curve. If weights connected to the hidden layer and the output layer are updated so inharmoniously that all the hidden neurons' outputs are driven rapidly into the extreme areas before the output neurons start to approximate to the desired signals, no weights connected to the hidden layer will be modified even though the actual outputs in the output layer are far from the desired outputs. Therefore, the local minima problem occurs.

## 3 Modified Error Function

To overcome such a problem, and furthermore, to avoid the local minima caused by such disharmony, the neuron outputs in the output layer and those in the hidden layer should be considered together during the iterative update procedure. Motivated by this, we have proposed modified error function in [6]. The modified error function is given by:

$$\begin{aligned} E_{new} &= \frac{1}{2}\sum_{p=1}^{P}\sum_{j}^{J}(t_{pj} - o_{pj})^2 + \frac{1}{2}\sum_{p=1}^{P}(\sum_{j}^{J}(t_{pj} - o_{pj})^2)(\sum_{j}^{H}(y_{pj} - 0.5)^2) \\ &= E_A + E_B. \end{aligned} \tag{4}$$

where, $\sum_j^H (y_{pj} - 0.5)^2$ can be defined as the degree of saturation in the hidden layer for pattern $p$. $y_{pj}$ is the output of the $j$-th neuron in the hidden layer, and $H$ is the number of neurons in the hidden layer. Using the above error function as the objective function, the update rule of both the weight $w_{ji}$ and threshold $\theta_j$ can be given as:

$$\Delta w_{ji} = -\eta_A \frac{\partial E_A}{\partial w_{ji}} - \eta_B \frac{\partial E_B}{\partial w_{ji}}, \tag{5}$$

and

$$\Delta \theta_j = -\eta_A \frac{\partial E_A}{\partial \theta_j} - \eta_B \frac{\partial E_B}{\partial \theta_j}, \tag{6}$$

where $\eta_A$ and $\eta_B$ are the learning rates for $E_A$ and $E_B$, respectively. It has been proven that this modified error function could effectively help the network avoid some local minima problems [6]. However, the selection of learning rates of $E_A$ and $E_B$ often is a very difficult and is problem-dependent. That's, given a value of $\eta_A$, both an over-small and over-large value of $\eta_B$ will damage the learning. The setting of the learning rates $\eta_A$ and $\eta_B$ is a difficult and subtle task. Thus, we propose a novel modified error function that is defined as follows:

$$\begin{aligned} E_{new} &= \frac{1}{2}\sum_{p=1}^{P}\sum_{j}^{J}(t_{pj} - o_{pj})^2 + (\frac{1}{2}\sum_{p=1}^{P}\sum_{j}^{H}(y_{pj}(W) - 0.5)^2 \\ &+ \frac{1}{2}\sum_{p=1}^{P}\prod_{j}^{H}(y_{pj}(\theta_j) - 0.5)^2) \\ &= E_A + E_B. \end{aligned} \tag{7}$$

We can see that the new error function also consists of two terms—the first term is the conventional error function:

$$E_A = \frac{1}{2}\sum_{p=1}^{P}\sum_{j}^{J}(t_{pj} - o_{pj})^2, \tag{8}$$

and the second one is the added term concerning the hidden layer:

$$E_B = \frac{1}{2}\sum_{p=1}^{P}\sum_{j}^{H}(y_{pj}(W) - 0.5)^2 + \frac{1}{2}\sum_{p=1}^{P}\prod_{j}^{H}(y_{pj}(\theta_j) - 0.5)^2. \tag{9}$$

The added term also consists of two terms:

$$E_B = E_B(W) + E_B(\Theta), \tag{10}$$

where $E_B(W)$ is the added error function of the weights connected to the hidden layer, and $E_B(\Theta)$ is the added error function of the threshold parameters of hidden layer neurons, respectively. We redefine the added term $E_B$ and use different forms for the weights and thresholds since we found it is excellent trade-off for both the convergence speed and global optimization.

The derivatives of the added term $E_B$ corresponding the weights and thresholds are computed as deferent forms. Given a pattern $p$, for the weights connected to the hidden layer, $\frac{\partial E_B^p}{\partial w_{ji}}$ is easily obtained as follows:

$$\frac{\partial E_B^p}{\partial w_{ji}} = \frac{\partial E_B^p(W)}{\partial w_{ji}} = (y_{pj} - 0.5)\frac{\partial y_{pj}}{\partial w_{ji}} = (y_{pj} - 0.5)f'(\cdot)x_{pi}, \tag{11}$$

where $x_{pi}$ is the $i$-th input for pattern $p$ since we use the network with only one hidden layer. For the thresholds of the neurons in the hidden layer, $\frac{\partial E_B^p}{\partial \theta_j}$ can be computed as follows:

$$\begin{aligned}\frac{\partial E_B^p}{\partial \theta_{ji}} &= \frac{\partial E_B^p(\Theta)}{\partial \theta_j} = (y_{pj} - 0.5)\frac{\partial y_{pj}}{\partial \theta_j}(\prod_{h\neq j}(y_{ph} - 0.5)^2) \\ &= -(y_{pj} - 0.5)f'(\cdot)(\prod_{h\neq j}(y_{ph} - 0.5)^2). \end{aligned} \tag{12}$$

Since this added term is used to keep the degree of saturation of the hidden layer small when $E_A$ is large, the effect of term $E_B$ should be diminished and will eventually become zero while the output layer approximates to the desired signals. Therefore, for training pattern $p$, the learning rate $\eta_B$ at step $t+1$ is adapted according to the following rule:

$$\eta_B^p(t+1) = E_A^p(t)\eta_B(0), \tag{13}$$

where, $\eta_B(0)$ is the initial value of the learning rate $\eta_B$. It is set to the same value for all patterns. For the novel modified error function, the selection of learning rates $\eta_A$ and $\eta_B$ is much easier. Generally, if $\eta_B(0) < \eta_A$ is selected, the performance of the network is not affected too much with various $\eta_B$.

## 4 Simulation

In order to verify the effectiveness of the modified error function, we applied it to the backpropagation algorithm (denoted by “BP+Added-terms”). Then the modified XOR problem was used for simulation. For comparison, we also performed the backpropagation algorithm (denoted by “BP”) and a global search technology—the simulated annealing method [7] (denoted by “SA”) with the conventional error function. Here, weights and thresholds were initialized randomly from (-1.0, 1.0). Two aspects of the training algorithm performance, “success rate” and “training speed”, were assessed for algorithm. The upper limit epochs for the BP+Added-terms and BP were set to 10,000.

The modified XOR problem is different from the classical XOR problem because one more pattern is included (that is, inputs=(0.5,0.5), teacher signal=1.0) such that a unique global minimum exists. Furthermore, several local minima exist simultaneously in this problem [8]. We used the 2-2-1 neural network to solve this problem. To show how the BP+Added-terms can avoid the local minima,

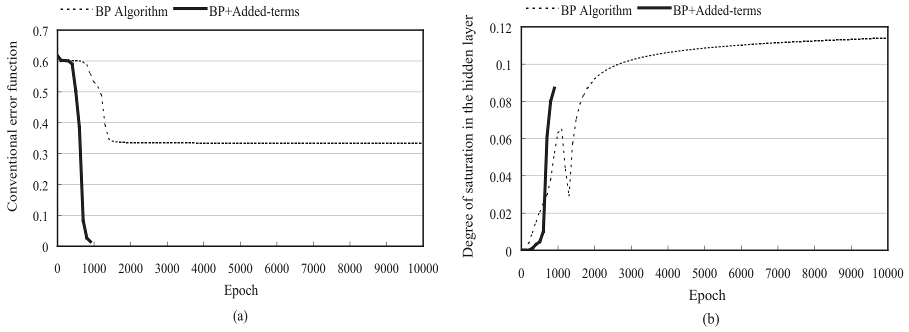

**Fig. 1.** Comparison of learning processes with local minima for the modified XOR problem between BP+Added-terms and BP: (a) conventional error function ($E_A$) vs. epochs and (b) degree of saturation in the hidden layer of all patterns vs. epochs

we compared a typical learning process of BP+Added-terms with that of BP in the case where there is the local minimum in the learning of BP. In Fig.1, the conventional error function ($E_A$) and the degree of saturation in the hidden layer of overall patterns are plotted as a function of epochs for both methods. We can see that the BP converged into a local minimum, while the degree of saturation in the hidden layer increased continually until it reached about 0.114. Meanwhile, the BP+Added-terms method avoided the local minimum and trained the network successfully about 900 epochs when the degree of saturation in the hidden layer was effectively neutralized by the modified error function. Table 1 shows the experimental results of the three methods based on 100 runs of this problem. For the BP, different learning rates $\eta = 0.3$, $\eta = 0.5$, and $\eta = 1.0$ were used. $\eta_B(0) = \eta_A = \eta$ was selected for the BP+Added-terms. The table shows that the backpropagation algorithm could obtain successful solutions for almost every run using the modified error function, while many failures in convergence to the global solution occurred both in the backpropagation algorithm and the simulated annealing method using the conventional error function. Although the average number of epochs of BP+Added-terms was a bit more than that of BP when $\eta = 0.3$ was used, it was almost the same as that of the BP when $\eta = 0.5$ was selected. Moreover, it was much less than those of the SA in all cases and BP in the case of $\eta = 1.0$. These results indicate that the proposed method could efficiently avoid the local minima for this problem.

## 5 Conclusions

In this paper, we proposed a modified error function with the added terms for the backpropagation algorithm to harmonize the update of weights connected to the hidden layer and those connected to the output layer. Therefore, local minima problems due to such disharmony could be avoided without much additional

**Table 1.** Experimental results for modified XOR problem

| Methods | | Success Rate | Average Number of Epoch |
|---|---|---|---|
| BP | $(\eta = 0.3)$ | 69% | 3582 |
| | $(\eta = 0.5)$ | 70% | 2181 |
| | $(\eta = 1.0)$ | 58% | 1106 |
| SA | $(\eta = 0.3)$ | 60% | 7164 |
| | $(\eta = 0.5)$ | 85% | 4811 |
| | $(\eta = 1.0)$ | 93% | 3593 |
| BP+Added-terms | $(\eta_B(0) = \eta_A = 0.3)$ | 99% | 4157 |
| | $(\eta_B(0) = \eta_A = 0.5)$ | 99% | 2257 |
| | $(\eta_B(0) = \eta_A = 1.0)$ | 97% | 904 |

computation and change in the network topology. And, the new learning parameters for the added term is not difficult to select. Finally, simulations performed on a benchmark problem demonstrated that the performance of the backpropagation algorithm was greatly improved by using the modified error function. More analysis on large problems is still required.

## References

1. Hadjiprocopis, A.: Feed Forward Neural Network Entities. Ph.D. Thesis. City University London UK. (2000)
2. Goerick, C., Seelen, W.V.: On Unlearnable Problems or A Model for Premature Saturation in Backpropagation Learning. In: Proceedings of the European Symposium on Artificial Neural Networks, Brugge Belgium April 24-26 (1996) 13–18
3. Haykin, S.: Neural Networks, A Comprehensive Foundation. MacMillan Publishing New York (1994)
4. Wessels, L.F.A., Barnard, E., van Rooyen, E.: The Physical Correlates of Local Minima. In: Proceedings of the International Neural Network Conference, Paris July (1990) 985
5. Funahashi, K.: On the Approximate Realization of Continuous Mapping by Neural Networks. Neural Networks, Vol. 2. (1989) 183–192
6. Wang, X.G., Tang, Z., Tamura, H., Ishii, M.: A Modified Error Function for Backpropagation Algorithm. Neurocomputing, Vol. 57. (2004) 477–484
7. Owen, C.B., Abunawass, A.M.: Application of Simulated Annealing to the Backpropagation Model Improves Convergence. In: Proceedings of the SPIE Conference on the Science of Artificial Neural Networks II, Vol. 1966. (1993) 269–276
8. Gori, M., Tesi, A.: On the Problem of Local Minima in Backpropagation. IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 14, No. 1. (1992) 76–86