A New Learning Method Using Local and Global Information for Neural Networks

Lu, Baiquan
Department of Automation, Shanghai University

Murata, Junichi
Department of Electrical and Electronic SystemsEngineering, Kyushu University

Hirasawa, Kotaro
Graduate School of Information, Production and Systems, Wasada University

Gu, Hong
Nantong Cellulose Fibers Co., Ltd.

https://doi.org/10.15017/1516047

出版情報：九州大学大学院システム情報科学紀要. 9 (2), pp.55-60, 2004-09-24. 九州大学大学院システム情報科学研究院
バージョン：
権利関係：

九州大学大学院
システム情報科学紀要
第9巻 第2号 平成16年9月

Research Reports on Information Science and
Electrical Engineering of Kyushu University
Vol.9, No.2, September 2004

A New Learning Method Using Local and Global Information for Neural Networks

Baiquan LU *, Junichi MURATA **, Kotaro HIRASAWA *** and Hong GU†

(Received June 11, 2004)

Abstract: A new learning method is proposed, which can be free from local minima of error function by using prior information. Because prior information can describe some features of teach function, neural networks also must have the features after learning. For this, learning using the prior information must attain two targets: learning of the features of teach function and a good approximation accuracy. The proposed method is very promising for solving the generalization ability problem of neural networks and avoiding the convergence to local minima. A bound on learning rate is also given for stability of the proposed method. The simulation results indicate usefulness of the proposed method.

Keywords: Neural network learning, Global optimum, Local information, Global information

1. Introduction

Artificial neural networks(ANNs) have been used widely in different applications as very promising function approximation tools. However, there are two issues. One is fast convergence problem, and the other vital question arises from local minima of the error. For this latter problem, many attractive results have been reported recently[1)2)], such as Simulated Annealing algorithm, random search algorithm, tunneling algorithm, learning automata algorithm and special network structure method.

In the paper, we address the above problem from other aspect: how to make learning concentrate on the area around the global optimum as fast as possible using local information and global information. A steepest-descent learning algorithm stops when derivatives of error with respect to weights become zero. These give equations in terms of weights. A set of solutions to these equations must include the global optimum solution. If we use prior information to constrain this set of solutions, the set will shrink, and if we have sufficient prior information, the set will contain the global solutions only. Moreover, because global information can describe global features of teach function, neural networks also must have the characteristics after learning. So neural networks must do harder task of acquiring the features of teach function in addition to attaining good numerical approximation. Incorporating prior information is, however, very promising for escaping from local minima and improving generalization ability.

Based on the above, in the paper, a new learning method is proposed using various local and global information. Also, bounds on the learning rate are given that assure convergence of learning based on discussions on necessary conditions of convergent learning.

2. Learning Method

The error function E for a multi-layered neural network that is to approximate a teach function $f(X)$ is defined by

$$E(W(k)) = \sum_{i=1}^{T}\sum_{j=1}^{N}[f_j(X(i)) - g_j(X(i))]^2, \quad (1)$$

where $g_j(X)$ is the output of neural network, $W(k)$ is the vector of n weights $w_1, \cdots w_n$ at learning iteration k, N is the number of output units and T is the number of training samples. Note that in the sequel $E(W(k))$ may be also expressed as $E(k)$ or $E(W)$ for simplicity. Assume that we already have a network structure that can attain perfect approximation, i.e. there exists a weight vector W^* such that $E(W^*) = 0$.

Let us consider a weight updating rule

$$\Delta W(k+1) = \begin{cases} -\lambda \frac{E(k)}{C} \frac{\partial E(k)}{\partial W(k)} & \text{if } E \neq 0, \\ 0 & \text{otherwise,} \end{cases} \quad (2)$$

* Department of Automation, Shanghai University
** Department of Electrical and Electronic Systems Engineering
*** Graduate School of Information, Production and Systems, Wasada University
† Nantong Cellulose Fibers Co., Ltd.

where $C = \|\partial E(k)/\partial W(k)\|^2$. Then we have, by the mean-value theorem,

$$E(k+1) = E(k)(1-\lambda) + \frac{1}{2}\lambda^2 \cdot \frac{E^2(k)}{C^2}\left(\frac{\partial E(k)}{\partial W(k)}\right)^T H(\xi)\frac{\partial E(k)}{\partial W(k)}, \quad (3)$$

where $H(\xi)$ is Hessian matrix, $\xi=[\xi_1 \cdots \xi_n]$ and each ξ_i satisfies $\min\{W_i(k), W_i(k) - \lambda E(k)/C \cdot \partial E(k)/\partial W_i(k)\} < \xi_i < \max\{W_i(k), W_i(k) - \lambda E(k)/C \cdot \partial E(k)/\partial W_i(k)\}$. When W approaches a solution to $\partial E(k)/\partial W(k) = 0$, but $E(k) \neq 0$, then ΔW will be infinitely large. Only when W approaches a global solution W^*, ΔW will converge to zero. To examine this, we will give theorem 1 and theorem 2 based on the following lemma.

Lemma 1 : There exists a ball centered at the global solution W^*

$$B(\epsilon) = \{W : ||W - W^*|| < \epsilon, \epsilon > 0\}, \quad (4)$$

such that ΔW is bounded and continuous inside $B(\epsilon)$ (for proof see [2]).

Theorem 1 : Let us consider a ball

$$B1(\epsilon) = \{W : ||W - W^1|| < \epsilon, \epsilon > 0\},$$

centered at a local solution W^1 such that $\partial E(W^1)/\partial W = 0$, and assume that there exists only one element $W = W^*$ such that $E(W) = 0$ inside $B1 \cap B$ and that Hessian matrix $H(\xi)$ is bounded so that there exists a positive constant M satisfying

$$\left|\left(\frac{\partial E(k)}{\partial W(k)}\right)^T H(\xi)\frac{\partial E(k)}{\partial W(k)}\right| \leq M\left(\frac{\partial E(k)}{\partial W(k)}\right)^T \frac{\partial E(k)}{\partial W(k)}. \quad (5)$$

If $W \in B1$, then W can escape from local minimum by update rule (2). Moreover if $W \in B$, then W will converge to the global solution in B when we use a small learning rate λ.

proof: Near a local solution W^1, from (2), we have

$$\|\Delta W\| = \lambda\frac{E(W)}{\sqrt{C}} = \lambda\frac{E(W^1) + \frac{\partial E(\xi)}{\partial W}(W - W^1)}{\sqrt{C}},$$

where $\min\{w_i, w_i^1\} < \xi_i < \max\{w_i, w_i^1\}$. If $\|W - W^1\|$ is sufficiently small, then $E(W^1)$ is much greater than either of $\|\partial E(W)/\partial W\|\|(W - W^1)\|$ and $\|\partial E(\xi)/\partial W\|\|(W - W^1)\|$, thus $\|\Delta W\| \approx \lambda E(W^1)/\sqrt{C}$. If $W \to W^1$, then $\|\Delta W\| \to \infty$. This implies that W can escape from the local solution.

Near the global solution W^*, by (3) and (5), we have

$$E(k+1) \leq E(k)(1-\lambda) + \frac{1}{2}\lambda^2 \cdot \frac{ME^2(k)}{C} \leq E(k)\left(1 - \lambda + \frac{1}{2}\lambda^2 \cdot \frac{ME(k)}{C}\right). \quad (6)$$

W can converge to W^* if

$$-1 < 1 - \lambda + \frac{1}{2}\lambda^2 \cdot \frac{ME(k)}{C} < 1 \quad (7)$$

holds, which gives the condition on the learning rate λ. The first inequality in (7) is equivalent to $0 < \lambda < \lambda_0$, while the second inequality implies that λ can take on any value if $4ME(k)/C \geq 1$, but λ must be in $(0, \lambda_1) \cup (\lambda_2, +\infty)$ if $4ME(k)/C \leq 1$, in which λ_0 and λ_1 and λ_2 are as follows,

$$\lambda_0 = \frac{2C}{ME(k)}, \ \lambda_1 = \frac{1 - \sqrt{1 - ME(k)/C}}{ME(k)/C}, \ \lambda_2 = \frac{1 + \sqrt{1 - ME(k)/C}}{ME(k)/C}. \quad (8)$$

Since $ME(k)/C$ is bounded inside $B(\epsilon)$ by Lemma 1, if we choose learning rate such that belongs to $(0, \lambda_0) \cap ((0, \lambda_1) \cup (\lambda_2, +\infty))$, then $\lim_{k\to\infty} E(k) = 0$. This implies that W can converge to the global solution inside B. □

So far we have studied the global minimization property of the learning method (2). Now we discuss a bound on learning rate for stability of the learning method.

If we assume $\lambda \geq 1$ in (3), we have

$$\frac{E^2(k)}{C^2}\left(\frac{\partial E(k)}{\partial W(k)}\right)^T H(\xi)\frac{\partial E(k)}{\partial W(k)} \geq 0. \quad (9)$$

If the error function decreases as the learning proceeds, i.e. $E(k+1) - E(k) \leq 0$, then (3) together with (9) gives bounds on λ as

$$1 \leq \lambda \leq 2\frac{C^2}{E(k)\left(\frac{\partial E(k)}{\partial W(k)}\right)^T H(\xi)\frac{\partial E(k)}{\partial W(k)}}. \quad (10)$$

We can obtain another set of bounds on the learning rate. Substituting (5) to (3) yields

$$\begin{aligned} E(k+1) &= E(k)(1-\lambda) + \frac{1}{2}\lambda^2\frac{E^2(k)}{C^2}\left(\frac{\partial E(k)}{\partial W(k)}\right)^T H(\xi)\frac{\partial E(k)}{\partial W(k)} \\ &\leq E(k) - E(k)\lambda + M\frac{1}{2}\lambda^2\frac{E^2(k)}{C}, \end{aligned} \quad (11)$$

and thus

$$E(k+1) - E(k) \leq -E(k)\lambda + M\frac{1}{2}\lambda^2\frac{E^2(k)}{C}. \quad (12)$$

Therefore, inequality

$$-E(k)\lambda + M\frac{1}{2}\lambda^2\frac{E^2(k)}{C} \leq 0 \quad (13)$$

is a sufficient condition for decrease in error function. This inequality gives a bound on λ as

$$\lambda \leq 2\frac{C}{E(k)M}. \quad (14)$$

Since $E(k+1) \geq 0$, we have from (11)

$$E(k) - E(k)\lambda + M\frac{1}{2}\lambda^2\frac{E^2(k)}{C} \geq 0. \quad (15)$$

This quadratic inequality holds for arbitrary value of λ as far as (5) is satisfied. So, we have

$$E^2(k) - 4\cdot\frac{1}{2}\cdot M\frac{E^2(k)}{C}E(k) \leq 0, \quad (16)$$

and therefore

$$M\frac{E(k)}{C} \geq \frac{1}{2}. \quad (17)$$

This gives a constant bound 4 on the right hand side of (14). This implies that λ should belong to interval (0,4) for stability of the learning. These bounds are constant and thus are good guideline for choice of the learning rate before starting the learning. But note that this is a necessary condition.

We can obtain stricter necessary conditions on λ for learning convergence. Let us denote $E^2(k)/C^2 \cdot (\partial E(k)/\partial W(k))^T \cdot H(\xi) \cdot \partial E(k)/\partial W(k)$ by $G(k)$. By (3), we have

$$\begin{aligned}
&E(k+1)\\
&= E(k)(1-\lambda) + \frac{1}{2}\lambda^2\cdot G(k)\\
&= E(k-1)(1-\lambda)^2 +\\
&\quad \frac{1}{2}\lambda^2\cdot G(k-1)(1-\lambda) + \frac{1}{2}\lambda^2 G(k)\\
&= (E(k-2)((1-\lambda) + \frac{1}{2}\lambda^2 G(k-2))(1-\lambda)^2\\
&\quad + \frac{1}{2}\lambda^2 G(k-1)(1-\lambda) + \frac{1}{2}\lambda^2 G(K)\\
&= E(k-2)((1-\lambda)^3 + \frac{1}{2}\lambda^2 G(k-2)(1-\lambda)^2\\
&\quad + \frac{1}{2}\lambda^2 G(k-1)(1-\lambda) + \frac{1}{2}\lambda^2 G(K)\\
&= \cdots\\
&= E(1)(1-\lambda)^k + \frac{1}{2}\lambda^2 G(1)(1-\lambda)^{k-1}\\
&\quad + \frac{1}{2}\lambda^2 G(2)(1-\lambda)^{k-2} + \cdots + \frac{1}{2}\lambda^2 G(K).
\end{aligned}$$

Assume that $G_{\max} = \max_j\{G(j)\}$ and $G_{\min} = \min_j\{G(j)\}$ are bounded except for at a few points which are jumping points from a local minima point to other points. Then we have

$$\begin{aligned}
&E(1)(1-\lambda)^k + \frac{1}{2}\lambda^2 G_{\min}\left\{\frac{1-(1-\lambda)^k}{1-(1-\lambda)}\right\}\\
&\leq E(k+1)\\
&\leq E(1)(1-\lambda)^k + \frac{1}{2}\lambda^2 G_{\max}\left\{\frac{1-(1-\lambda)^k}{1-(1-\lambda)}\right\}.
\end{aligned} \quad (18)$$

From (18), when $|1-\lambda| \leq 1$, $E(k)$ is bounded as $k \to \infty$. When $E(1) = G_{\max}/(2\lambda)$, $E(k)$ is also bounded but in this case for any positive value of λ. Thus we have the following theorem.

Theorem 2 : Assume that $G_{\max}$ and $G_{\min}$ are bounded during learning (2) except for at a few points which are jumping points from a local minima point to other points. If leaning is convergent, then leaning rate λ satisfies either of the following:

1. $|1-\lambda| \leq 1$,
2. $E(1) = G_{max}/(2\lambda)$.

Now we discuss the bounds on λ using fixed point theorem. Form (3), $E(k)$ can be considered as self-iterates. By fixed point theorem, if $|\partial E(k+1)/\partial E(k)| \leq 1$, this iteration is convergent. Applying this condition to (3), we have

$$\left|1 - \lambda + \lambda^2\frac{E(k)}{C^2}\left(\frac{\partial E(k)}{\partial W(k)}\right)^T H(\xi)\frac{\partial E(k)}{\partial W(k)}\right| \leq 1. \quad (19)$$

By defining $L(k) = E(k)/C^2 \cdot (\partial E(k)/\partial W(k))^T \cdot H(\xi)\cdot \partial E(k)/\partial W(k)$, the condition is rewritten as $|1-\lambda+\lambda^2 L(k)| \leq 1$, which is equivalent to

$$\begin{aligned}
L(k)\lambda(\lambda-\lambda_1) &\leq 0,\\
L(k)(\lambda-\lambda_2)(\lambda-\lambda_3) &\geq 0,
\end{aligned} \quad (20)$$

where $\lambda_1 = 1/L(k)$, $\lambda_2 = (1-\Theta^{1/2})/(2L(k))$, $\lambda_3 = (1+\Theta^{1/2})/(2L(k))$ and $\Theta = 1 - 8L(k)$. If $L(K) \leq 0$, then λ_1 and λ_3 are negative value, but λ_2 is positive, so solution to this inequality belongs to $(0,\lambda_2)$. In the case that $L(K) \geq 0$, the solution belongs to $(0,\lambda_2) \cup (\lambda_3,\lambda_1)$ if $L(k) \leq 1/8$ or $(0,\lambda_2)$ if $L(k) \geq 1/8$.

Finally, let us discuss how to choose a suitable learning rate. In (11), let us consider making the right hand side equal to $\theta E(k)$ $(0 < \theta < 1)$. This is equivalent to making $E(k+1)$ smaller than $\theta E(k)$ which implies convergent learning. Then we have

$$\theta E(k) = E(k) - E(k)\lambda + M\frac{1}{2}\frac{\lambda^2 E^2(k)}{C}. \quad (21)$$

Let us define

$$\Delta = 1 - 4 \cdot (1-\theta)\frac{1}{2}M\frac{E(k)}{C}. \tag{22}$$

If $E(k) > 0$ and $\Delta \geq 0$, then the solution to (40) is

$$\frac{1-\sqrt{\Delta}}{ME(k)/C} \text{ or } \frac{1+\sqrt{\Delta}}{ME(k)/C}. \tag{23}$$

From (17) and the fact that $\Delta \geq 0$, we have

$$\frac{1}{2(1-\theta)} \geq \frac{E(k)M}{C} \geq \frac{1}{2},$$

or equivalently

$$\frac{C}{2(1-\theta)E(k)} \geq M \geq \frac{C}{2E(k)}. \tag{24}$$

From (23), λ that assures convergence can be calculated for a given large M satisfying (5) and (24). For example, first we can select a large M, say $M = 100$, then select θ by (24) and calculate λ by (23). If $E(k+1) \leq \theta E(k)$, M needs no change, but otherwise M is increased and θ is selected again by (24) until $E(k+1) \leq \theta E(k)$.

3. Learning Method Using Prior Information

A learning method using local information and global information is reported in [2]. Here, we will discuss neural network learning methods that incorporate different kinds of prior information, i.e. transformation of functions, equilibrium points of dynamical systems and partial derivatives of functions.

3.1 Prior Information Based on Transformation of Function

In this subsection, we will introduce how to use prior information based on transformations $F_i(\cdot)$, $i = 1, \cdots, m$ of function to be approximated $f_j(X)$. If we know transformed values of the target values $F_i(f_j(X))$, then we expect our neural network to have the same transformed values. Therefore, we can define error functions in addition to the regular error function (1) as

$$E_i(W) = \sum_{j=1}^{N} \left[F_i(f_j(X)) - F_i(g_j(X))\right]^2. \tag{25}$$

Transformation $F_i(\cdot)$ can be a maximum Lyapunov exponent operator for chaotic teach functions. Other possible choices of $F_i(\cdot)$ include fractional dimension operators, Fourier transform operators, mean-value operators and wavelet transformation operators. Also, we can adopt any monotonic function as F_i to distort error surface and assist the solution to escape from local optima. By minimizing sum of these additional error functions $E_i(W)$ and the regular error function $E(W)$ defined by (1), we have the update rule

$$\Delta W(k+1) = \begin{cases} -\lambda G(W) & \text{if } E \neq 0, \\ 0 & \text{otherwise,} \end{cases} \tag{26}$$

where

$$G(W) = \frac{(E(k) + \sum_{i=1}^{m} E_i(k))}{\|\frac{\partial(E(k)+\sum_{i=1}^{m} E_i(k))}{\partial W(k)}\|^2} \cdot \frac{\partial(E(k) + \sum_{i=1}^{m} E_i(k))}{\partial W(k)}.$$

3.2 Prior Information Based on Equilibrium Points of Dynamical Systems

Suppose identifying a dynamical system $y(k+1) = f[y(k), y(k-1), \cdots, y(k-n), u(k)]$ by a neural network. The network is to approximate function f. Because for any dynamic system with self-equilibrium, $y(\infty)$=constant for constant input $u(k) = 1$, by fixed point theorem, we have $(\partial f/\partial y(\infty))^2 \leq 1$. So the neural network also must have this characteristic after learning. Here the neural network with single output is assumed to be

$$g(X) = \sum_{k=1}^{m} \alpha_k \varphi(W_k^T X + \theta_k). \tag{27}$$

If the neural network has this characteristic, then

$$g1 = \left(\frac{\partial g}{\partial y(\infty)}\right)^2 - 1 = \left[\sum \alpha_k \cdot \varphi^{(1)}(y(\infty))(\sum_{i=1}^{n+1} w_{i,k})\right]^2 - 1 \leq 0.$$

Then we can define a error function by introducing a penalty term with a Lagrange multiplier s as

$$E1(W) = E(W) + s \cdot (g^+)^2, \tag{28}$$

where $g^+ = \max\{0, g1\}$. Because $g^+ = 0$ for $g1 \leq$, and $g^+ = g1$ for $g1 \geq 0$, derivative of $(g^+)^2$ exists except for $g1 = 0$, and is equal to $2g^+(\partial g1/\partial W)\dot{W}$. Here let us assume $E1(W)$ is a function of continuous time t, then

$$\frac{dE1}{dt} = \frac{\partial E}{\partial W}\dot{W} + \dot{s}(g^+)^2 + 2sg^+\frac{\partial g1}{\partial W}\dot{W}$$
$$= (\frac{\partial E}{\partial W} + 2sg^+\frac{\partial g1}{\partial W})\dot{W} + \dot{s}(g^+)^2.$$

If we take $\dot{W}$ as the following,

$$\dot{W} = -\frac{E + \dot{s}(g^+)^2}{\|\frac{\partial E}{\partial W} + 2s(g^+)\frac{\partial g}{\partial W}\|^2} \cdot \left(\frac{\partial E}{\partial W} + 2s(g^+)\frac{\partial g1}{\partial W}\right), \quad (29)$$

then we have

$$\frac{dE1}{dt} = -E, \quad (30)$$

where Lagrange multiplier s is assumed as $\dot{s} = \epsilon g^+ E$ or $\dot{s} = \epsilon g^+$ so that g^+ becomes zero as fast as possible, and ϵ is a small positive number.

From this, $dE1/dt = 0$ if and only if $E = 0$. This indicates that learning is convergent. Thus, we can get a difference equation for learning:

$$\Delta W(k+1) = -\frac{\lambda E + \dot{s}(g^+)^2}{\|\frac{\partial E}{\partial W} + 2s(g^+)\frac{\partial g}{\partial W}\|^2} \cdot \left(\frac{\partial E}{\partial W} + 2s(g^+)\frac{\partial g1}{\partial W}\right). \quad (31)$$

3.3 Prior Information Based on Taylor Expansion

Prior information in the form of partial derivatives of teach function at origin is treated here. We assume $f(X) \in C^\infty$ and define output error as $\Delta EF(X) = f(X) - \sum_{k=1}^m \alpha_k \varphi(W_k^T X + \theta_k)$ (here the bias term θ_k is set to 1). Condition $\Delta EF(0) = 0$ is equivalent to letting coefficients of each term of Taylor expansion of ΔEF at $X = 0$ equal to zero, then we have

$$f(0,\cdots,0) = \sum_{k=1}^m \alpha_k \varphi(1),$$
$$f^{(j)}_{x_{i1}}(0,\cdots,0) = \sum_{k=1}^m \alpha_k w^j_{i1,k} \varphi^{(j)}(1),$$
$$f^{(j+1)}_{x_{i1},x^j_{i2}}(0,\cdots,0) = \sum_{k=1}^m \alpha_k w^j_{i2,k} w_{i1,k} \varphi^{(j+1)}(1), \quad (32)$$

where $f^{(j+1)}_{x_{i1},x^j_{i2}}(0,\cdots,0)$ is the j-th order partial derivative with respect to x_{i2} after the first partial derivative with respect to x_{i1}. Equation (32) can be written in a matrix form:

$$A_1 B_1 = C_1 \quad (33)$$

where

$$A_1 = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ w_{i1,1} & w_{i1,2} & \cdots & w_{i1,m1} \\ \vdots & \vdots & \cdots & \vdots \\ w^{m1-1}_{i1,1} & w^{m1-1}_{i1,2} & \cdots & w^{m1-1}_{i1,m1} \end{bmatrix},$$
$$B_1 = [\alpha_1\ \alpha_2\ \cdots\ \alpha_{m1}]^T,$$
$$C_1 = \begin{bmatrix} \frac{f(0,\cdots,0)}{\varphi(1)} - \sum_{k=m1+1}^m \alpha_k \\ \frac{f^{(1)}_{x_{i1}}(0,\cdots,0)}{\varphi^{(1)}(1)} - \sum_{k=m1+1}^m \alpha_k w_{i1,k} \\ \vdots \\ \frac{f^{(m1-1)}_{x_{i1}}(0,\cdots 0)}{\varphi^{(m1-1)}(1)} - \sum_{k=m1+1}^m \alpha_k w^{m1}_{i1,k} \end{bmatrix}.$$

Assuming that inverse of matrix A_1 exists for $m1 = m$, we have

$$B_1 = A_1^{-1} C_1. \quad (34)$$

From (34), α_k can be expressed by $w_{i1,k1}$, $k1 = 1, 2, \cdots, m$.

To incorporate the constraint (34) in the learning algorithm, let us define

$$T1^{(k)}_{w_{(i1,j1)}} = \frac{\partial \alpha_k}{\partial w_{i1,j1}}, \quad (35)$$

and expand the error function $E(W)$ in a Taylor series in terms of W and $\alpha = [\alpha_1\ \cdots\ \alpha_m]$:

$$\Delta E(W,\alpha) = \sum_{i=1}^n \sum_{k=1}^m \frac{\partial E}{\partial w_{i,k}} \Delta w_{i,k} + \sum_{k=1}^n \frac{\partial E}{\partial \alpha_k} \Delta \alpha_k + \cdots. \quad (36)$$

If

$$w_{i1,k1} \neq w_{i1,k2}, \quad (37)$$

is satisfied for a certain $i1$ where $k1 \neq k2, k1 = 1, 2, \cdots, m, k2 = 1, 2, \cdots, m$, then we have

$$\Delta E(W,\alpha) = \sum_{i=1}^n \sum_{k=1}^m \frac{\partial E}{\partial w_{i,k}} \Delta w_{i,k} + \sum_{j=1} \sum_{k=1}^m \frac{\partial E}{\partial \alpha_k} T1^{(k)}_{w_{(i1,j)}} \Delta w_{i1,j} + \cdots$$
$$= \sum_{i\neq i1}^n \sum_{k=1}^m \frac{\partial E}{\partial w_{i,k}} \Delta w_{i,k} + \sum_{k=1}^m \left(\sum_{j=1} \frac{\partial E}{\partial \alpha_k} T1^{(k)}_{w_{(i1,j)}} + \frac{\partial E}{\partial w_{i1,k}} \right) \cdot \Delta w_{i1,k} + \cdots. \quad (38)$$

According to (2), we have the update rules

$$\Delta w_{i1,j} = -\lambda \frac{E}{\|g2\|^2 + \|\frac{\partial E}{\partial w_{i,k}}\|^2} (g2_{i1,j} + \frac{\partial E}{\partial w_{i1,k}}), \quad (39)$$

$$\Delta w_{i,k} = -\lambda \frac{E}{\|g2\|^2 + \|\frac{\partial E}{\partial w_{i,k}}\|^2} \frac{\partial E}{\partial w_{i,k}}, \quad (40)$$

where $i \neq i1$, λ is the learning rate, $i1$ is the indexing number which satisfies (37), and $g2_{i1,j} = \sum_{j=1}(\partial E/\partial \alpha_k) T1^{(k)}_{w(i1,j)}$. Weights $w_{i1,j}$ and $w_{i,j}$ can be adjusted by (39) and (40). Then, we get $\Delta\alpha$ to update α by using (35).

4. Simulations

A system identification problem is adopted as an example to evaluate the learning methods proposed in the paper. The system to be identified is $y_p(k+1) = P[y_p(k), y_p(k-1), y_p(k-2), u(k), u(k-1)]$, where $P[x_1, x_2, x_3, x_4, x_5] = \{x_1 \cdot x_2 \cdot x_3 \cdot x_5(x_3 - 1) + x_4\}(1 + x_3^2 + x_2^2)$, and $u(k) = 1$, $k = 1, \cdots, 100$. The neural network has five input, 60 hidden and one output nodes, and the node function is $(1 - e^{-x})/(1+e^{-x})$. Initial values of hidden layer weights and output layer weights are generated from uniform distributions over $(-0.25, 0.25)$ and $(-0.1, 0.1)$, respectively. Evaluated learning methods are listed in **Table 1**, where Δ_1 is the proposed learning rule that does not use prior information, Δ_2 uses transformation-based prior information, Δ_3 incorporates Taylor expansion prior information, Δ_4 is for dynamical system with equilibrium points and Δ_5 is the well-known backpropagation learning rule with momentum term.

The final error values and required learning iterations are listed in **Table 2**. The results show learning rates greater than 1 give better results by the method Δ_1. The learning can be improved by using prior information, especially when prior information based on Taylor expansion or equilibrium points is available (Δ_3 or Δ_4), the results are much better. The backpropagation with momentum term Δ_5 gives much worse result than any of the proposed methods. The results indicate that proposed methods are very useful.

5. Conclusions

In the paper, a learning method is proposed using prior information to constrain or decrease the numbers of local minima points of error function in order to speed up learning or obtain a solution of global optimum. Also a bound of learning rate for learning is given. The effectiveness and applicability have been demonstrated by simulations.

Table 1 Learning Methods.

Δ_1	$\Delta W = -\lambda \frac{E(k)}{C} \frac{\partial E(k)}{\partial W(k)}$
Δ_2	$\Delta W = -\lambda \frac{E(k)+\sum_{j=1}^{3} E_j(k)}{\|\frac{\partial E(k)}{\partial W(k)} + \sum_{j=1}^{3} \frac{\partial E_j(k)}{\partial W(k)}\|^2} \cdot \left(\frac{\partial E(k)}{\partial W(k)} + \sum_{j=1}^{3} \frac{\partial E_j(k)}{\partial W(k)}\right)$, $F_1(x) = \frac{1-e^{-x}}{1+e^{-x}}$, $F_2(x) = x^3$, $F_3(x) = \frac{1}{1+e^{-x}}$
Δ_3	$\Delta W = -\lambda \frac{E(k)}{C} \frac{\partial E(k)}{\partial W(k)}$, $\sum_{j=1} \Delta\alpha_j = 0$
Δ_4	$\Delta W = -\lambda \frac{E1(k)}{C} \frac{\partial E1(k)}{\partial W(k)}$, $\sum_{j=1} \Delta\alpha_j = 0$, $E1(k) = E(k) + s \cdot (g^+)^2, g^+ = \max[0, g]$, $g = [\sum \alpha_k \varphi^{(1)}(y(\infty)) \cdot (w_{1,k} + w_{2,k} + w_{3,k})]^2 - 1 \leq 0$, $\dot{s} = \epsilon g^+$, ϵ =0.00000001
Δ_5	$-\lambda \frac{\partial E(k)}{\partial W(k)} + m\Delta w(k-1)$

Table 2 Simulation Results.

Learning method	Learning rate	Iteration number	Final error
Δ_1	λ=1.6	28472	0.000033
Δ_1	λ=0.6	49239	0.000108
Δ_1	λ=1.6 and 0.6 are used alternately	49988	0.000033
Δ_1	λ=random in (0,2)	49921	0.000057
Δ_1	λ=0.01 for large M	49964	0.004019
Δ_2	λ=1.6	49963	0.000059
Δ_2	λ=0.8	48812	0.000578
Δ_3	λ=1.6	49778	0.000001
Δ_4	λ=1.6	44470	0.000001
Δ_5	λ=0.0005, $m = 0.8$	49999	0.002873

6. Acknowledgement

This work was partly supported by the Shanghai Education Science Foundation under grant 03AK014.

References

1) M. Gori and A. Tesi, "On the problem of local minima in backpropagation", *IEEE Trans. on Neural Networks*, **14**,76–85, 1997.

2) S. H. Lee and R. M. Kil, "Inverse mapping of continuous functions using local and global information", *IEEE Trans. on Neural Networks*, **5**, 409-423, 1992.