한국퍼지 및 지능 시스템학회 논문지 1998, Vol. 8, No. 5. (1998) 98~106

확률적 근사법과 공액기울기법을 이용한 다층신경망의 효율적인 학습

An Efficient Training of Multilayer Neural Networks Using Stochastic Approximation and Conjugate Gradient Method

조 용 현
Yong-Hyun Cho

대구효성가톨릭대학교 전자 · 정보공학부

요 약

본 논문에서는 신경망의 학습성능을 개선하기 위해 확률적 근사법과 공액기울기법에 기초를 둔 새로운 학습방법을 제안하였다. 제안된 방법에서는 확률적 근사법과 공액기울기법을 조합 사용한 전역최적화 기법의 역전파 알고리즘을 적용함으로써 학습성능을 최대한 개선할 수 있도록 하였다. 확률적 근사법은 국소최소점을 벗어나 전역최적점에 치우친 근사점을 결정해 주는 기능을 하도록 하며, 이점을 초기값으로 하여 결정론적 기법의 공액기울기법을 적용함으로써 빠른 수렴속도로 전역최적점으로의 수렴확률을 높였다. 제안된 방법을 패리티 검사와 패턴 분류에 각각 적용하여 그 타당성과 성능을 확인한 결과 제안된 방법은 초기값을 무작위로 설정하는 기울기하강법에 기초를 둔 기존의 역전파 알고리즘이나 확률적 근사법과 기울기하강법에 기초를 둔 역전파 알고리즘에 비해 최적해로의 수렴확률과 그 수렴속도가 우수함을 확인할 수 있었다.

ABSTRACT

This paper proposes an efficient learning algorithm for improving the training performance of the neural network. The proposed method improves the training performance by applying the back-propagation algorithm of a global optimization method which is a hybrid of a stochastic approximation and a conjugate gradient method. The approximate initial point for fast global optimization is estimated first by applying the stochastic approximation, and then the conjugate gradient method, which is the fast gradient descent method, is applied for a high speed optimization. The proposed method has been applied to the parity checking and the pattern classification, and the simulation results show that the performance of the proposed method is superior to those of the conventional backpropagation and the back-propagation algorithm which is a hybrid of the stochastic approximation and steepest descent method.

1. 서 론

신경망은 단순한 처리능력을 가지는 뉴런들이 대규모로 상호 연결되어 분산지식표현, 일반화, 그리고 학습 등의 속성을 가지고 있어 패턴 인식[1-3]과 같은 과제들을 능률적으로 해결하는데 많이 이용되고 있다. 그러나 복잡하고 계산집약적인 문제들을 풀기 위해서는 대규모의 뉴런들을 이용하여 신경망을 구성하기 때문에 아직까지는 해를 구하는 데에 있어서 비현실적인 학습시간이 요구되거나 국소최적해로의 수렴과 같은 문제가 발생되고 있다[1-15].

신경망을 학습에 이용할 때는 인식이나 분류의 능력을 높이기 위하여 입력층과 출력층 사이에 한개 이상의 은닉층(hidden layer)을 갖는 다층(multilayer) 구조를 흔히 사용한다[4]. 이때 흔히 쓰이는 역전파(backpropagation) 알고리즘은 기울기하강(gradient descent)법에 해당하는 결정론적(deterministic) 알고리즘으로 최적해로의 수렴속도는 빠르나 국소최적해를 만났을 때 이를 벗어나기가 어렵다[5-7]. 이러한 역전파 알고리즘에서는 학습율(learning rate)과 모멘트(momentum) 등의 학습 파라미터, 초기의 연결가중치(synapse weight), 또는 신경망의 구조 등에 따라 그 성능이 달라진다.

Rumelhart 등[7]은 역전파 알고리즘에서 연결가중치를 경신(update)할 때에 학습 파라미터를 변화시켜 그 학습성능을 개선하였으며, 이때 학습율과 모멘트는 경험적으로 설정하였다. Pedone 등[8]은 은닉층만의 뉴런 활성화분포의 표준편차에 따라 은닉층과 출

력층 각각의 연결가중치 경신에 이용되는 학습율과 모멘트를 적응 조정하여 학습속도를 개선하였다. 한편 Charalmbous[9]는 고전적인 순차적(iterative) 최적화 방법 중에서 가장 빠른 수렴속도의 속성을 가진 것으로 알려진 공액기울기(conjugate gradient)법을 이용하여 신경망의 학습속도를 개선하였다. 그러나 이들 연구들에는 국소최적해로 수렴될 가능성이 여전히 남아 있다.

Hirose 등[10]은 은닉층의 뉴런수를 에너지함수의 증감에 적응 조정하였고, Wang 등[11]은 자기성장학습(self growing learning) 알고리즘을 제안하여 전역최적해로의 수렴확률을 개선하였다. 그러나 이들 방법에서도 은닉층 뉴런의 추가 및 삭제에 따른 알고리즘 및 시스템의 복잡도가 증가되었다. 한편, Cho 등[14]은 확률적 근사법(stochastic approximation)[15]과 기울기하강법에 기초를 둔 전역최적화의 역전파 알고리즘을 제안하여 초기값에 다소 무관하게 빠른 수렴속도로 전역최적화를 수행함으로써 학습성능을 더욱 개선시켰다.

이제까지의 연구에서는 주로 전역최적해로의 수렴확률을 개선하거나 또는 학습속도를 개선하기 위한 연구 중 어느 하나에 역점을 두었을 뿐, 두 가지를 동시에 만족시키기 위한 연구는 찾아보기 힘들다. 또한 전역최적해로의 수렴확률이나 또는 학습속도의 개선을 위하여 취한 방법으로는 주로 학습율과 모멘트의 조정이나 은닉층 뉴런 수와 같은 신경망의 구조 변경 등을 이용하였다. 특히 신경망에 있어서 초기 연결가중치는 학습성능을 결정하는 중요한 요소이나 대부분의 연구에서는 이를 경험적으로 또는 무작위로 설정하고 있으며, 초기 연결가중치를 전역최적해 가까이에 설정함으로써 전역최적해로의 수렴확률과 학습속도를 동시에 개선[14]하려는 체계적인 방법은 거의 찾아볼 수 없다.

본 논문에서는 확률적 근사법과 공액기울기법을 조합 사용한 역전파 알고리즘을 이용함으로써 다층신경망의 학습성능을 개선할 수 있는 효율적인 방법을 제안하였다. 제안된 방법에서는 확률적 근사법과 공액기울기법에 기초한 전역최적화 기법의 역전파 알고리즘을 다층신경망의 학습 알고리즘으로 사용함으로써 최적해로의 수렴확률과 그 수렴속도를 최대한 개선할 수 있도록 하였다. 확률적 근사법은 국소최소점을 벗어나 전역최적점에 치우친 근사적인 초기치를 결정해 주는 기능을 하도록 하고, 이점을 초기값으로 하여 공액기울기법을 적용함으로써 빠른 수렴속도로 전역최적화가 가능하도록 하였다. 제안된 방법을 패리티 검사와 패턴 분류에 각각 적용하여 초기값을 무작위로 설정하는 기울기하강법에 기초를 둔 기존의 역전파 알고리즘이나 확률적 근사법과 기울기하강법에 기초를 둔 역전파 알고리즘을 이용한 결과들과 비교 고찰하였다.

2. 확률적 근사법과 공액기울기법을 이용한 전역최적화 기법

함수 평활화(function smoothing) 기능을 갖는 확률적 근사법은 다극점함수(multiextremal function)의 전역최소점(global minimum)을 찾는 최적화 알고리즘이다[15]. 이때 최적화는 다극점함수 $f(x)$에 평활화 함수(smoothing function) $h(\gamma, \beta)$를 상승적분(convolution)하여 식 (1)과 같은 평활된 함수(smoothed function) $f(x, \beta)$를 구하고, $h(\gamma, \beta)$의 분산 즉, $f(x)$의 평활정도를 제어하는 파라미터 β를 순차적으로 감소시켜 가며 근사화를 반복함으로써 이루어진다.

$$f(x,\beta)=\int_{R^n} h(\gamma,\beta)\, f(x-\gamma)\, d\gamma=\int_{R^n} h(x-\gamma,\beta)\, f(\gamma)\, d\gamma \qquad (1)$$

여기서 x는 상태변수이고 γ는 무작위교란벡터(random perturbation vector)이다. 평활화는 다극점함수를 단일극점함수로 변형하기 위한 목적으로 상승적분을 근사화하여 사용하며 평활화 함수 $h(\gamma, \beta)$를 이용하여 상태 x 주위의 제한된 공간에 대해 다극점함수 $f(x)$를 평균하여 구한다. 이때 γ를 생성하는 평활화 함수 $h(\gamma, \beta)$로는 가우스(Gaussian), 균일(uniform), 또는 코오시(Cauchy) 확률밀도함수(probability density function)가 이용된다. 따라서 확률적 근사법에 의한 상태변수 x의 경신은 식 (2)와 같다. 즉,

$$\begin{aligned}
&x(k+1)=x(k)-\pi(k)\,d(k)\\
&d(k)=(1-\rho(k))\,d(k-1)+\rho(k)\,\xi(k),\ 0\le\rho\le 1\\
&\rho(k)=\frac{\rho(k-1)}{1+\rho(k-1)-R},\ 0<R<1\\
&\xi(k)=\nabla_x f(x(k),\beta)\\
&\pi(k)=\frac{STEP}{\|\xi(k)\|}
\end{aligned} \qquad (2)$$

이다. 여기서 k는 반복회수이고 $STEP$은 경신방향에 대한 비례상수로 실험에서는 0.01로 하였다. 또한 $\tau(k)$와 $d(k)$는 각각 반복회수 k에서의 경신방향과 step size이다.

예로서 연속이고 미분 가능한 2개의 다른 최소점을 가지는 함수 $f(x)$

$$f(x)=x^4-16x^2+5x \qquad (3)$$

에서 평활된 함수 $f(x, \beta)$는

$$f(x,\beta)=\left(\frac{1}{2}\right)[f(x+\beta\gamma)+f(x-\beta\gamma)] \quad (4)$$

와 같이 상태 x 점을 중심으로 $\pm\beta\gamma$ 범위내에서의 평균치를 근사값으로 사용한다.

그림 1은 함수 $f(x)$에 분산제어 파라미터 β의 순차값(sequence)을 $\{\beta\}=\{5.0, 4.0, 3.0, 2.0, 1.0, 0.001\}$로, 평활화 함수 $h(\gamma, \beta)$를 균일 확률밀도함수로 하고, 초기상태를 $x_0=4.0$으로 할 때에, 각 β값에 대한 $f(x, \beta)$의 변화를 나타낸 것이다. 그림에서 보는 바와 같이 확률적 근사법을 이용하여 전역최적해를 구하는 데는 β의 순차값 $\{\beta\}$와, 각 β값에 대한 제어 파라미터들을 적절하게 설정해야 하는 번거로움이 있을 뿐만 아니라 알고리즘의 확률적 동작으로 수렴속도가 느리다는 등의 문제점이 있다. 한편 그림에서 β값이 5.0, 4.0, 그리고 3.0일 때를 보면 이 정도의 큰 β값에 대해서는 $f(x, \beta)$가 단일극점 함수로 단봉특성을 가짐을 알 수 있다. 이때 각 $f(x, \beta)$의 최소점은 원래 다극점함수 $f(x)$의 전역최적해에 대한 근사값이다. 그러므로 초기에 큰 β값으로 평활화를 통한 확률적 근사법을 적용한다면 다극점함수 $f(x)$의 전역최적해로의 수렴이 용이한 초기상태를 구할 수 있을 것으로 보인다. 또한 β의 값이 클수록 함수 $f(x)$의 전역최소점 x^*와 평활된 함수 $f(x, \beta)$가 최소가 되는 상태 x_l 사이의 거리 DSA는 증가한다. 그러나 상태 x_l는 국소최소점 x_m보다는 전역최소점 x^*쪽 경사면에 가깝게 위치한다. 따라서 이 x_l을 초기상태로 하여 결정론적 방법으로 수렴속도가 빠른 공액기울기법을 적용한다면 직접 전역최소점 x^*로 빠르게 수렴시킬 수 있을 것이다.

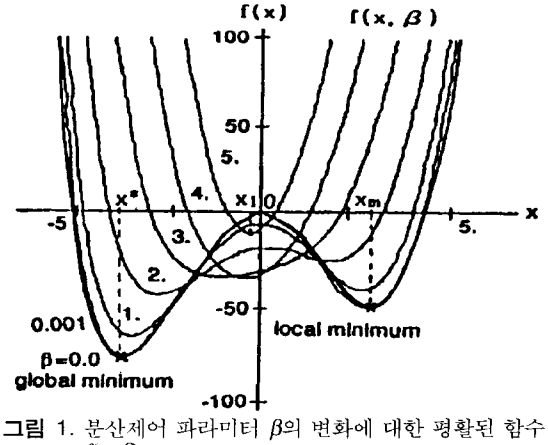


그림 1. 분산제어 파라미터 β의 변화에 대한 평활된 함수 $f(x, \beta)$.

Fig. 1. Smoothed function $f(x, \beta)$ to dispersion control parameter β.

따라서 학습을 위한 다층신경망에서 연결가중치벡터 $\boldsymbol{w}$와 학습패턴 p에 대한 평균자승 오차함수(mean square error function) $E_p(\boldsymbol{w})$의 최소값을 찾는 과정은 확률적 근사법으로 상태벡터 $\boldsymbol{x}$에 대한 비용함수 $f(\boldsymbol{x})$의 최소값을 찾는 과정으로 대응시킬 수 있다. 그러므로 다층신경망의 학습에 확률적 근사법을 이용하여 평활된 평균자승 오차함수의 값이 최소인 연결가중치를 구하고, 구해진 연결가중치를 초기값으로 하여 공액기울기법의 빠른 수렴특성을 갖는 역전파 알고리즘을 이용하면 빠른 시간내에 전역최소의 평균자승 오차를 보장하는 학습을 시킬 수 있음을 알 수 있다.

먼저, 다층신경망에 확률적 근사법을 적용하기 위해서 전체 오차함수 $E(\boldsymbol{w})$ 및 학습패턴 p에 대한 평균자승 오차함수 $E_p(\boldsymbol{w})$를 각각 식 (5)와 같이

$$\begin{aligned} E(\boldsymbol{w}) &= \textstyle\sum_p E_p(\boldsymbol{w}) \\ E_p(\boldsymbol{w}) &= (1/2)\textstyle\sum_i (x_{ip}(\boldsymbol{w})-d_{ip})^2 \end{aligned} \quad (5)$$

정의한다. 여기서 $x_{ip}(\boldsymbol{w})$와 d_{ip}는 각각 p번째 패턴에 대한 출력층 뉴런 i의 실제 출력과 원하는 출력이다. 또한 평균자승 오차함수 $E_p(\boldsymbol{w})$의 평활된 함수 $E_p(\boldsymbol{w}, \beta)$ 및 평활된 함수의 기울기 $\nabla_W E_p(\boldsymbol{w}, \beta)$는 각각 식 (6)과 같이

$$\begin{aligned} E_p(\boldsymbol{w},\beta) &= (1/2)[E_p(\boldsymbol{w}+\beta\gamma)+E_p(\boldsymbol{w}-\beta\gamma)] \\ &= (1/4)\textstyle\sum_i [2d_{ip}^2 - 2d_{ip}(x_{ip}(\boldsymbol{w}+\beta\gamma) \\ &\quad + x_{ip}(\boldsymbol{w}-\beta\gamma)) + x_{ip}^2(\boldsymbol{w}+\beta\gamma) + x_{ip}^2(\boldsymbol{w}-\beta\gamma)] \\ \nabla_W E_p(\boldsymbol{w},\beta) &= (1/4)\textstyle\sum_i [-2d_{ip}(\nabla_W x_{ip}(\boldsymbol{w}+\beta\gamma) \\ &\quad + \nabla_W x_{ip}(\boldsymbol{w}-\beta\gamma)) + \nabla_W x_{ip}^2(\boldsymbol{w}+\beta\gamma) \\ &\quad + \nabla_W x_{ip}^2(\boldsymbol{w}-\beta\gamma)] \end{aligned} \quad (6)$$

구할 수 있다. 이때 출력층과 은닉층의 평활된 함수의 기울기는 각각 다음과 같이 계산된다. 즉, 출력층 연결가중치 경신시 평활된 함수의 기울기는

$$\begin{aligned} \nabla_W E_p(\boldsymbol{w},\beta) = (1/2)\textstyle\sum_i [(f(\sum_j (w_{ij}+\beta\gamma) y_{jp}(\boldsymbol{w})) - d_{ip}) \\ f(\sum_j (w_{ij}+\beta\gamma) y_{jp}(\boldsymbol{w}))(1-f(\sum_j \\ (w_{ij}+\beta\gamma) y_{jp}(\boldsymbol{w}))) + (f(\sum_j \\ (w_{ij}+\beta\gamma) y_{jp}(\boldsymbol{w})) - d_{ip}) f(\sum \\ (w_{ij}-\beta\gamma) y_{jp}(\boldsymbol{w}))(1-f(\sum_j \\ (w_{ij}-\beta\gamma) y_{jp}(\boldsymbol{w})))] y_{jp}(\boldsymbol{w}) \end{aligned} \quad (7)$$

와 같으며, 은닉층 연결가중치 경신시 평활된 함수의

기울기는

$$\begin{aligned}\nabla_w E_p(\boldsymbol{w},\beta) = (1/2)\sum_i [(f(\sum_j w_{ij} f(\sum_m \\ (w_{jm}-\beta\gamma) o_{mp}))-d_{ip}) f(\sum_j w_{ij} f(\sum_m \\ (w_{jm}-\beta\gamma) o_{mp}))(1-f(\sum_j w_{ij} f(\sum_m \\ (w_{jm}-\beta\gamma) o_{mp}))) \sum_j w_{ij} f(\sum_m \\ (w_{jm}+\beta\gamma) o_{mp}) + (f(\sum_j w_{ij} f(\sum_m \\ (w_{jm}-\beta\gamma) o_{mp}))-d_{ip}) f(\sum_j w_{ij} f(\sum_m \\ (w_{jm}-\beta\gamma) o_{mp}))(1-f(\sum_j w_{ij} f(\sum_m \\ (w_{jm}-\beta\gamma) o_{mp}))) \sum_j w_{ij} f(\sum_m \\ (w_{jm}-\beta\gamma) o_{mp})] o_{mp}(1-f(\sum_m \\ (w_{jm}+\beta\gamma) o_{mp})) \qquad (8)\end{aligned}$$

과 같이 계산된다. 여기서 i, j 및 m은 각각 출력층, 은닉층 및 입력층의 뉴런번호이다. $y_{jp}(\boldsymbol{w})$는 입력패턴 p에 대한 은닉층 뉴런 j의 출력이며, o_{mp}는 입력패턴 p에 대한 입력층이나 혹은 은닉층이 여러개이면 그 아래층 뉴런 m의 출력이다.

한편, 기존의 순차적 최적화 방법중에서 수렴속도가 대단히 빠른 기법중의 하나인 공액기울기법[1]은 선형이나 비선형 방정식의 해를 구하는 문제에 적용될 수 있다. 일반적인 선형방정식

$$A\boldsymbol{x}(k) = b, \quad k = 0, 1, 2, \cdots \qquad (9)$$

에서 $\boldsymbol{x}(k)$와 b는 각각 $n\times 1$ 벡터이고, A는 $n\times n$ 행렬로 표현되는 방정식 계수로서 positive definite이며 invertible하다고 가정할 때 이를 풀기 위한 비용함수 $f(\boldsymbol{x}(k))$는

$$f(\boldsymbol{x}(k)) = -\frac{1}{2}\boldsymbol{x}(k)^T A\boldsymbol{x}(k) + \boldsymbol{x}(k)^T b \qquad (10)$$

로 정의될 수 있다. 여기서 T는 전치행렬을 의미한다. 식 (10)에서 A가 positive definite이면 $f(\boldsymbol{x}(k))$는 완전 볼록(strictly convex) 함수로 오직 하나의 최소값을 갖게 된다. 또한 $A\boldsymbol{x}(k)=c(c=2b)$이면 식 (10)은 영의 값을 가져 선형방정식을 푸는 문제는 비용함수 $f(\boldsymbol{x}(k))$를 최소화하는 문제로 변환될 수 있다. 그러므로 $\nabla f(\boldsymbol{x}(k)^*)=0$이면 $\boldsymbol{x}(k)^*$는 비용함수를 최소로 한다. 즉, 식 (10)을 미분하면

$$\nabla f(\boldsymbol{x}(k)) = \frac{\partial f(\boldsymbol{x}(k))}{\partial \boldsymbol{x}(k)} = -A\boldsymbol{x}(k) + b = 0 \qquad (11)$$

이 되어, $\boldsymbol{x}(k)$를 구하면 $A\boldsymbol{x}(k)=b$의 최적해가 된다. 이 때에 $\boldsymbol{x}(k+1)$은 반복에 의하여

$$\boldsymbol{x}(k+1) = \boldsymbol{x}(k) + \boldsymbol{r}(k)\boldsymbol{s}(k) \qquad (12)$$

로 구해진다. 여기서도 k는 반복수이며 $\boldsymbol{s}(k)$는 경신방향이고, $\boldsymbol{\gamma}(k)$는 $f(\boldsymbol{x}(k))+\boldsymbol{\gamma}(k)\boldsymbol{s}(k)=\min_{\gamma\in R} f(\boldsymbol{x}(k)+\gamma\boldsymbol{s}(k))$에 의해 정의되는 스칼라 step size이다. 여기서 $\boldsymbol{s}(k)$와 $\boldsymbol{\gamma}(k)$는

$$\begin{aligned}&\boldsymbol{g}(k) = \nabla f(\boldsymbol{x}(k)) = -A\boldsymbol{x}(k) + b \\ &\boldsymbol{g}(k-1) = \nabla f(\boldsymbol{x}(k-1)) = -A\boldsymbol{x}(k-1) + b \\ &\boldsymbol{s}(k) = -\boldsymbol{g}(k) + \boldsymbol{q}(k)\boldsymbol{s}(k-1) \\ &\quad = A\boldsymbol{x}(k) - b + \boldsymbol{q}(k)\boldsymbol{s}(k-1) \\ &\boldsymbol{s}(0) = -\boldsymbol{g}(0) \\ &\boldsymbol{q}(k) = \frac{(\boldsymbol{g}(k)^T\boldsymbol{g}(k))}{(\boldsymbol{g}(k-1)^T\boldsymbol{g}(k-1))} \\ &\boldsymbol{\gamma}(k) = -\frac{(\boldsymbol{s}(k)^T\boldsymbol{g}(k))}{\boldsymbol{s}(k-1)^T\Gamma\boldsymbol{s}(k-1)} \qquad (13)\end{aligned}$$

에 의해 구해지며 식 (13)을 공액기울기법이라 한다. 여기서 Γ는 Hessian matrix이다. 공액기울기법에서는 비용함수값이 단조 감소하는 결정론적 동작특성으로 기울기하강법에 비해 수렴속도는 현저히 증가되나 국소최소해로의 수렴 속성은 그대로 가진다.

따라서 학습을 위한 다층 신경망에서 연결가중치 벡터 $\boldsymbol{w}$와 학습패턴 p에 대한 평균자승 오차함수 $E_p(\boldsymbol{w})$의 최소값을 찾는 과정은 공액기울기법에서 상태벡터 $\boldsymbol{x}$에 대한 비용함수 $f(\boldsymbol{x})$의 최소값을 찾는 과정으로 대응시킬 수 있다. 결국 평균자승 오차함수 $E_p(\boldsymbol{w})$에 공액기울기법을 적용할 때 p번째 패턴에 대한 연결가중치 벡터 $\boldsymbol{w}$의 변화는

$$\begin{aligned}&\boldsymbol{w}(k+1) = \boldsymbol{w}(k) + \boldsymbol{r}(k)\boldsymbol{s}(k) \\ &\boldsymbol{g}(k) = \nabla_w E_p(\boldsymbol{w}) \qquad (14)\end{aligned}$$

로 구해진다. 식 (14)에서 $\boldsymbol{r}(k)$와 $\boldsymbol{s}(k)$는 각각 반복회수 k에 따른 최적의 step size와 경신방향이며, $\boldsymbol{g}(k)$는 $E_p(\boldsymbol{w})$의 공액기울기이다. 이때 $\boldsymbol{r}(k)$ 및 $\boldsymbol{s}(k)$는 각각 식 (13)에 의하여 계산된다.

따라서 확률적 근사법과 공액기울기법을 이용하여 다층신경망의 학습성능을 개선하기 위해 제안한 역전파의 학습 알고리즘은 다음과 같이 정리될 수 있다.

단계 1: 다층신경망의 전체 오차함수 $E(\boldsymbol{w})$ 및 학습패턴 p에 대한 평균자승 오차함수 $E_p(\boldsymbol{w})$를 정의한다.

단계 2: 출력층 및 은닉층의 평활된 함수 $E_p(\boldsymbol{w}, \beta)$와 그 기울기 $\nabla_w E_p(\boldsymbol{w}, \beta)$를 각각 계산한다.

단계 3: 분산제어 파라미터 β를 크게 하여 출력층과 은닉층의 평활된 함수의 기울기를 계산하여 확률적 근사법을 수행한다.

단계 4: 단계 3에서 구해진 각 층의 연결가중치를

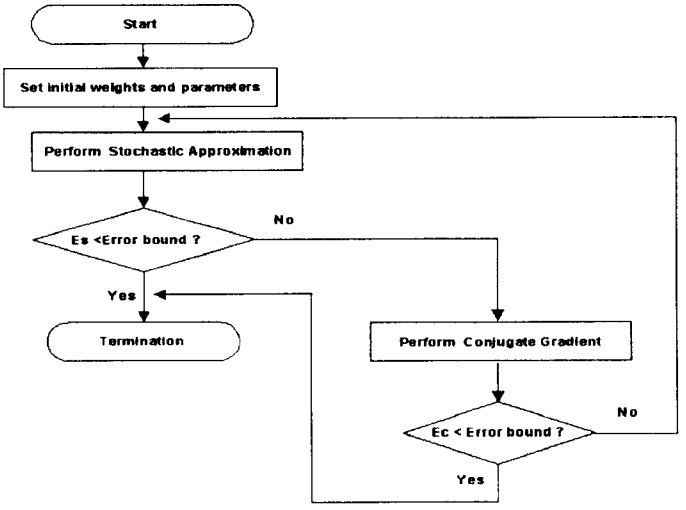


그림 2. 제안된 학습 알고리즘의 전체 흐름도.
Fig. 2. Overalll flow chart of the proposed learning algorithm.

초기값으로 하여 공액기울기법의 역전파 알고리즘을 수행한다.

단계 5: 단계 4에서 구해진 전체 오차함수가 허용치 이하이면 종료하고, 아니면 단계 3으로 간다.

이와 같이 제안된 방법에서는 확률적 근사법을 이용하여 초기 연결가중치를 설정하고 설정된 연결가중치를 초기값으로 하여 공액기울기법의 역전파 알고리즘을 사용한다. 따라서 기존의 기울기하강법에 기초를 둔 역전파 알고리즘이 그 초기값에 따라 국소최적해에 수렴되는 문제를 해결할 수 있으며, 공액기울기법의 빠른 수렴속도도 그대로 살릴 수 있다. 만일 확률적 근사법과 공액기울기법에 기초를 둔 역전파 알고리즘을 적용하여 구해지는 해가 확률적 근사법의 1회 적용에 의한 초기치 설정에 따라 국소최적해로 빠질 경우에는 알고리즘 단계 3에서 단계 5까지를 반복 적용함으로써 결국 전역최적해에 도달하게 된다.

그림 2는 확률적 근사법과 공액기울기법을 조합한 제안된 학습알고리즘의 전체 흐름도를 나타낸 것이다. 그림에서 E_s와 E_c는 각각 확률적 근사법과 공액기울기법의 역전파 알고리즘을 이용할 때 계산되는 전체 오차함수 값이다.

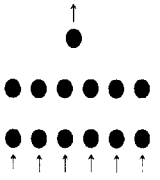
3. 응용 예 및 시뮬레이션 결과고찰

제안된 방법의 타당성과 학습성능을 확인하기 위한 응용실험 대상으로 패리티 검사와 패턴 분류를 택하였다. 패리티 검사는 가장 유사한 패턴들이 서로 다른 출력들을 요구하는 특징을 가진 여러 개의 국소최적해가 존재하는 문제 중의 하나이다. 이 문제는 출력층 뉴런의 수가 1개인 구조의 신경망을 이용하며, XOR 문제는 2비트 패리티 검사의 대표적인 문제에 속한다. 또한 패턴 분류는 학습을 통하여 불완전한 입력패턴을 분류해 내는 특징을 가진 문제로서 영상이나 음성의 분류에 널리 응용되고 있다. 이 문제에서는 출력층 뉴런의 수가 학습패턴의 개수와 동일한 구조의 신경망이 이용된다. 이 두 문제들은 신경망을 이용한 학습에서 알고리즘의 성능을 평가하는 대표적인 문제로 이용되고 있다[7]. 따라서 제안된 방법을 이들 문제에 각각 적용하여 그 타당성을 확인하였으며, 무작위로 설정한 초기의 연결가중치를 이용하는 기존의 역전파 알고리즘 및 확률적 근사법과 기울기하강법에 기초를 둔 역전파 알고리즘의 결과와 그 성능을 비교 고찰하였다.

학습에 이용된 다층신경망의 구조로는 층 사이의 뉴런간에 완전한 연결을 갖는 입력층, 은닉층 및 출력층으로 구성된 삼층 구조를 택하였으며, 입력층과 은닉층의 뉴런 개수는 동일하게 설정하였다. 초기 연결가중치들의 범위는 −0.5와 +0.5 사이의 값으로 하였으며 랜덤시드(random seed)의 변경으로 연결가중치들을 변화시킬 수 있도록 하였다. 알고리즘의 종료는 계산 반복횟수(number of iterations)가 20000번 이상이거나, 전체 오차함수의 값이 주어진 값 PEV(permissible error value) 이하일 때로 하였다. 여기서 계산 반복횟수는 모든 입력패턴이 연결가중치를 경신하기 위해서 한 번씩 다 이용된 때를 1회로 하였다.

Traing patterns	Desired outputs
000000	1
010000	0
000010	0
101000	1
000101	1
010100	1
001010	1
101010	0
010101	0
111111	1

(a)



(b)

그림 3. 10개의 6비트 패턴에 대한 패리티 검사 문제의 (a) 학습패턴과 (b) 신경망의 구조.

Fig. 3. The training patterns for 10 patterns of (a) 6-bit parity check and (b) it's neural network structure.

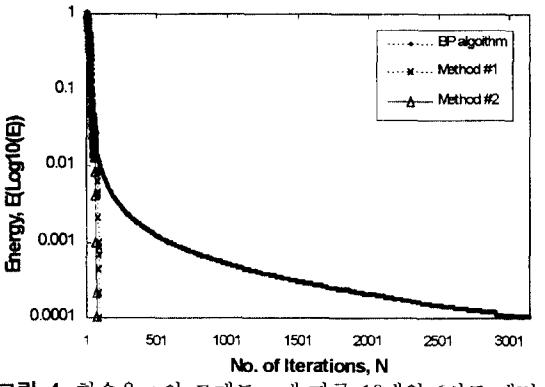


그림 4. 학습율 η와 모멘트 α에 따른 10개의 6비트 패턴에 대한 패리티 검사 문제의 전체 오차함수 E.

Fig. 4. Energy function E of the parity check for the 10 of the 6-bit patterns for learning rate η and momentum α.

특히 역전파 알고리즘이나 제안된 방법에서 초기 학습율 η_o와 초기 모멘트 α_o는 각각 그 조합으로 실험한 결과 중에서 가장 우수한 조합의 값으로 설정하였다. 제안된 방법에서 분산제어 파라미터 β의 값은 3.0으로 하여 확률적 근사법을 1회 적용하였으며, 무작위교란벡터 γ를 생성하는 평활화 함수 $h(\gamma, \beta)$로는 균일 확률밀도함수를 이용하였다. 앞으로의 설명에서 방법 1은 확률적 근사법과 기울기하강법의 역전파 알고리즘을 이용한 학습방법이며, 방법 2는 확률적 근사법과 공액기울기법의 역전파 알고리즘을 이용한 제안된 학습방법이다. 여기서 방법 2는 방법 1에서의 기울기하강법 대신 더 빠른 수렴속성을 가진 공액기울기법을 이용한 경우이다.

3.1 패리티 검사

패리티 검사는 각 학습패턴 내에 포함된 "1"의 개수에 따라 출력을 결정하는 문제로서, 실험에서는 일부분(partial) 및 우수(even) 패리티로 "1"의 개수가 짝수 개이면 "1", 홀수 개이면 "0"이 출력되도록 학습시켰다.

실험은 학습패턴의 수와 크기가 각각 6과 10인 그림 3에서와 같은 문제를 대상으로 하였으며, 허용오차함수 PEV는 0.0001로 하였다.

그림 4는 3의 학습 문제에서 랜덤시드를 10, 학습율 η와 모멘트 α를 각각 0.5와 0.9로 하여 역전파 알고리즘과 방법 1 및 방법 2를 각각 적용할 때 반복횟수 N에 따른 전체 오차함수 E의 변화를 나타낸 것이다. 그림에서 보는 바와 같이 방법 1과 방법 2는 기존의 기울기하강법을 이용하는 역전파 알고리즘보다 최적해로의 수렴속도가 빠르며, 특히 방법 2의 수렴속도는 방법 1의 수렴속도보다도 더 빠름을 알 수 있다. 이는 방법 2에 이용되는 공액기울기법이 방법 1에서 이용되는 기울기하강법보다 빠른 수렴속도의 속성을 가지기 때문이다. 그러나 그림에서는 랜덤시드의 단 한 가지에 대해서만 얻은 것으로 전체적인 특징의 차이를 보기는 어렵다.

표 1은 그림 3의 패리티 검사 문제에 대해 학습 파라미터와 랜덤시드를 변화시키면서 실험한 결과들을 나타낸 것이다. 그 결과는 각 학습 파라미터에 대해 100개의 서로 다른 초기 연결가중치로 실험한 결과들 중에서 최적해로 수렴된 경우들의 평균값이다. 표에서 N_s, N_{bp} 및 N_{cg}는 각각 확률적 근사법, 기울기하강법 및 공액기울기법의 반복횟수이며, t_{bp}, t_{p1} 및 t_{p2}는 각각 역전파 알고리즘, 방법 1 및 방법 2에서의 N_{bp},

표 1. 10개의 6비트 패턴에 대한 패리티 검사의 실험결과

Table 1. Results of the parity check for 10 of the 6-bit patterns.

η, α	BP algorithm			Method #1			Method #2		
	N_{bp}	E	t_{bp}	N_s, N_{bp}	E	t_{p1}	N_s, N_{cg}	E	t_{p2}
1.0, 0.0	20000	0.000176	13.9	3, 405.9	< 0.0001	0.3	3, 374.3	< 0.0001	0.3
0.8, 0.8	8553.5	< 0.0001	5.9	3, 113.4	< 0.0001	0.1	3, 96.6	< 0.0001	0.1
0.5, 0.9	6971.7	< 0.0001	4.9	3, 91.2	< 0.0001	0.1	3, 81.8	< 0.0001	0.1
0.3, 0.7	18952.3	< 0.0001	13.2	3, 427.4	< 0.0001	0.4	3, 397	< 0.0001	0.3

t_i: CPU time in [sec].

N_{bp}와 N_s 및 N_{cg}와 N_s에 각각 소요된 CPU 시간의 합이다. E는 종료시의 전체 오차함수 값이다. 표에서 나타난 바와 같이 방법 1과 2는 기존의 역전파 알고리즘보다 그 수렴속도가 빠름을 알 수 있다. 특히, η=1.0과 α=0.0의 경우에 기존의 역전파 알고리즘에서는 최적해로 수렴되지 못하였으나 방법 1과 2는 모두 최적해로 수렴되었다. 이는 연결가중치의 초기값이 국소최적해 근처에 설정된 경우이며 최적해에 치우친 근사값을 결정해 주는 확률적 근사법의 역할을 보여주는 것이다. 또한 방법 2에서는 공액기울기법의 적용으로 기존의 기울기하강법을 적용한 방법 1보다는 그 수렴속도가 약 1.1배 정도 개선되었으나 수렴확률은 동일하다. 이는 두 방법 모두에 확률적 근사법이 적용되었기 때문이다. 표에서 나타난 바와 같이 방법 1의 수렴속도는 기존의 역전파 알고리즘에 비해 약 53배 정도 개선되었다.

표 2에서는 20개의 7비트 패턴에 대한 패리티 검사에 있어서 역전파 알고리즘 및 방법 1과 방법 2를 각각 적용하여 랜덤시드를 다르게 100번씩 시도한 실험결과, 최적해로 수렴된 경우들에 대한 반복횟수 및 CPU 시간의 각각 평균 $\bar{x}$와 표준편차 σ를 나타낸 것이다. 여기서 학습율 η와 모멘트 α는 각각 0.5와 0.9로 하였다. 표에서 보는 바와 같이 방법 1에서는 역전파 알고리즘에 비하여 최적해로의 수렴확률이 약 1.6배 정도 개선되었으며, 수렴속도에서는 약 63.4배 정도 개선되었다. 또한 방법 1과 2의 표준편차는 역전파 알고리즘의 표준편차보다 약 86.2배와 약 128.2배 정도 적은 값이다. 이는 역전파 알고리즘에 비해 확률적 근사법이 적용된 방법 1과 2의 성능이 초기 연결가중치에 덜 의존함을 보여준다.

한편, 표 1에서 처럼 10개의 6비트 패턴에 대한 패리티 검사에 있어서 η=0.5와 α=0.9의 경우 100번 시도에 따른 결과가 모두 최적해로 수렴되었으나, 표 2의 20개 7비트 패턴의 경우는 최적해로의 수렴율이 약 0.63으로 이는 문제의 규모가 커질수록 초기 연결가중치의 설정이 더욱 어려우며 이에 따른 학습성능의 의존성을 보여준 것이다. 그러나 방법 1과 방법 2의 경우에는 이 정도 규모의 문제에 대해서 모두 최적해로 수렴되었으며 문제의 규모가 커질수록 그 성능의 개선정도가 우수함을 확인할 수 있다.

표 2. 20개의 7비트 패턴에 대한 패리티 검사의 100번 시도에 따른 실험결과

Table 2. Results for the 20 of the 7-bit patterns parity check of 100 trials

	BP algorithm		Method #1		Method #2	
	$\bar{x}$	σ	$\bar{x}$	σ	$\bar{x}$	σ
N	5084.1	1473.8	80.2	17.7	71.3	11.5
t	8.7	2.6	0.3	0.5	0.3	0.3
Pr	63%		100%		100%	

$\bar{x}$: Mean, σ: Standard deviation, P_r: Convergence ratio.

신경망이 가지는 우수한 속성 중에 하나는 일반화이다. 이를 확인하기 위하여 랜덤시드를 0으로 하고 허용오차함수 PEV는 0.0001로 하여 20개 7비트의 패턴을 학습패턴으로 나머지 108개 7비트 패턴은 시험패턴으로 사용하였다. 실험결과, 총 128개 패턴 중에서 역전파 알고리즘의 경우는 75개, 확률적 근사법이 적용된 방법 1과 2의 경우는 각각 75개와 74개의 패턴이 정확하게 분류되었다. 여기서 기존의 역전파 알고리즘과 방법 1이 방법 2보다는 약간 우수한 일반화성능을 가짐을 확인할 수 있다. 그러나 이 정도의 경우는 수렴속도와 수렴율과 같은 성능 및 초기 연결가중치 설정 등을 고려할 때, 제안된 방법 2가 다른 두 방법에 비하여 평균적으로 우수한 학습성능을 가진 알고리즘임을 알 수 있다.

3.2 패턴 분류

패턴 분류에서는 각 학습패턴에 대해 해당되는 출력 뉴런만 "1"이 되고, 나머지 출력 뉴런들은 "0"을 출력하도록 학습시킨다. 실험에서는 첫 번째 학습패턴에 대해 첫 번째 출력 뉴런만 "1"을 출력하고 나머지 출력 뉴런은 "0"을 출력하는 즉, 각 패턴의 입력순서와 같은 번호에 해당되는 출력 뉴런만 반응하도록 학습시켰다.

실험은 그림 5에서와 같이 학습패턴의 크기가 9비트인 5개의 패턴에 대하여 실시하였으며, 학습의 종료는 그 반복횟수가 20000번 이상이거나, 오차함수값이 설정된 오차값 PEV=0.1 이하일 때, 혹은 각 출력 뉴런들이 원하는 출력을 나타낼 때로 하였다.

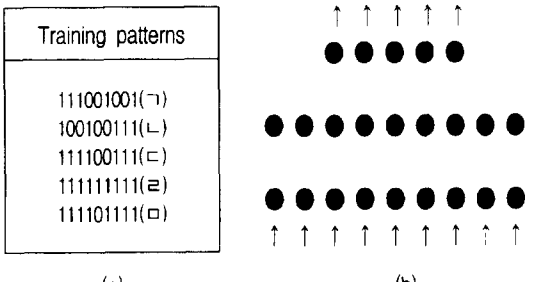


그림 5. (a) 5개의 9비트 학습패턴과 (b) 이 패턴 분류 문제를 풀기 위한 신경망의 구조.

Fig. 5. 5 of the 9-bit training patterns (a) and it's neural network structure (b) for pattern classification.

표 3. 5개의 9비트 패턴에 대한 패턴 분류의 실험결과
Table 3. Results of the pattern classification for 5 of the 9-bit patterns

η, α	Random seed	BP algorithm		Method #1		Method #2	
		N_{bp}	E	N_s, N_{bp}	E	N_s, N_{cg}	E
0.3, 0.7	0	1142	<0.1	8, 720	<0.1	8, 658	<0.1
	10	20000	0.4*	3, 568		3, 471	
	20	20000	0.4*	4, 639		4, 492	
	50	1029	<0.1	4, 572		4, 507	
0.5, 0.9	0	20000	0.6*	8, 531	<0.1	8, 410	<0.1
	10	20000	0.8*	3, 294		3, 203	
	20	2257	<0.1	4, 387		4, 276	
	50	3134	<0.1	4, 552		4, 435	
0.8, 0.8	0	1873	<0.1	8, 613	<0.1	8, 504	<0.1
	10	20000	0.7	3, 379		3, 287	
	20	1891	<0.1	4, 445		4, 339	
	50	3095	<0.1	4, 604		4, 483	

t_i: CPU time in [sec].

표 3은 3개의 학습 파라미터 조합 각각에 대해 랜덤시드를 4가지로 변화시켜 가며 실험한 결과를 나타낸 것이며, *는 제약조건을 만족하지 못하는 결과이다. 표에서 나타난 것과 같이 역전파 알고리즘에서는 3개의 학습 파라미터 모두에 대해 랜덤시드가 10일 때는 그 종료조건을 만족하지 못하였으므로 원하는 출력을 얻을 수 없었다. 이와 같은 상황이 발생되는 것은 초기 연결가중치가 국소최적해 근처에 설정된 경우로, 기존의 기울기하강법에 기초를 둔 역전파 알고리즘은 그 동작 특성상 가장 가까운 국소최적해로 수렴하기 때문으로 추측된다. 또한 랜덤시드가 50인 경우에는 학습율과 모멘트에 따라 학습시간이 줄어들고 패턴을 분류해 낼 수 있는 최적해로도 수렴될 수 있어, 학습 파라미터가 학습성능에 크게 영향을 미침을 확인할 수 있다. 그러나 방법 1과 2에서는 4개의 랜덤시드 모두에 대하여 패턴을 분류해 낼 수 있는 최적해로 수렴되었다. 표에서는 방법 1과 2는 기존의 역전파 알고리즘에 비해 최적해로의 수렴확률이 약 3.0배 정도 개선되었고, 그 수렴속도는 각각 약 3.7배와 약 4.6배 정도 개선되었다. 또한 방법 1보다는 공액기울기법을 이용하는 방법 2가 평균적으로 약 1.2배 정도 더 빠른 수렴속도를 갖는다.

4. 결 론

본 논문에서는 전역최적화 기법을 다층신경망의 학습 알고리즘으로 사용함으로써 학습성능을 개선할 수 있는 효율적 방법을 제안하였다. 제안된 방법에서는 학습에 확률적 근사법과 공액기울기법을 혼합 사용한 전역최적화 기법을 적용함으로써 학습성능을 최대한 개선할 수 있도록 하였다. 확률적 근사법은 국소최소점을 벗어나 전역최적 점에 치우친 근사점을 결정해 주는 기능을 하도록 하고, 이점을 초기값으로 하여 결정론적 공액기울기법의 역전파 알고리즘을 적용함으로써 빠른 수렴속도로 전역최적화가 가능하도록 하였다.

제안된 방법을 20개의 7비트 및 10개의 6비트 패턴에 대한 패리티 검사와 5개의 9비트 패턴에 대한 패턴 분류에 각각 적용하여 그 타당성과 성능을 확인하였다. 20개의 7비트 패턴 패리티 검사에 있어서 기존의 역전파 알고리즘에 비해 최적해로의 수렴확률이 약 2배 정도 개선되었고, 수렴속도에서는 기존의 역전파 알고리즘 및 확률적 근사법과 기울기하강법을 조합한 방법에 비해 각각 약 71배와 약 2배 정도 개선되었다. 패턴 분류에서도 제안된 방법은 이들 두 방법들에 비해 우수한 학습성능을 가짐을 확인할 수 있다.

향후 제안된 혼성방법을 고속병렬 시스템으로 구현하여 규모가 큰 문제에 대한 신속한 최적학습에 대하여 원활한 연구 수단을 마련하는 일이 남아있다.

참고문헌

[1] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation Numerical Method*, Prentice-Hall, London, pp. 1-50, 1989.

[2] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Anaysis*, John Wiley & Sons, Inc., London, pp. 10-43, 1973.

[3] J. T. Tou and R. C. Gonzalez, *Pattern Recognition*

Principles, Addison-Wesley Pub. Co., London, pp. 158-242, 1974.

[4] S. Akaho and S. Amari, "On the capacity of three-layernetworks", *International Joint Conference on Neural Networks*, Vol. 3, pp. 1-6, June 1990.

[5] A. J. Maren, C. T. Harston, and R. M. Pap, *Handbook of Neural Computing Application*, Academic Press, pp. 107-250, 1990.

[6] J. A. Freeman and D. M. Skapura, *Neural Networks: Algorithms, Applications, and Programming Techniques*, Addison Wesley, London, pp. 89-168, 1991.

[7] D. E. Rumelhart and J. L. McClelland, *Parallel Distributed Processing*, MIT press, Cambridge, MA., pp. 282-362, 1986.

[8] R. Pedone and D. Parisi, "Learning the learning parameters", *International Joint Conference on Neural Networks*, Vol. 3, pp. 2033-2037, Nov. 1991.

[9] C. Charalmbous, "Conjugate gradient algorithm for efficient training of artificial neural networks", *IEE., Proceeding-G*, Vol. 139, No. 3, June 1992.

[10] Y. Hirose, K. Yamashita, and S. Hijiya, "Back-propagation algorithm which varies the number of hidden units", *Neural Networks*, Vol. 4, No. 1, pp. 61-66, 1991.

[11] S. D. Wang and C. H. Hsu, "A self growing learning algorithm for determining the appropriate number of hidden units", *International Joint Conference on Neural Networks*, Vol. 2, pp. 1098-1104, Nov. 1991.

[12] N. Baba, "A new approach for finding the global minimum of error function of neural networks", *Neural Networks*, Vol. 2, No. 5, pp. 367-373, 1989.

[13] J. Sun, W. I. Grosky, and M. H. Hassoun, "A fast algorithm for finding global minima of error functions in layered neural networks", *Proceedings of International Joint Conference on Neural Networks*, Vol. 1, pp. 585-588, Jan. 1990.

[14] Y. H. Cho and H. M. Choi, "Improving the training performances of the multilayer neural networks by SAS-based optimal estimation of initial weights", *JTC-CSCC*, pp. 475-478, July 1992.

[15] M. A. Styblinski and T. S. Tang, "Experiments in nonconvex optimization: stochastic approximation with function smoothing and simulated annealing", *Neural Networks*, Vol. 3, No. 4, pp. 467-483, 1990.

조용현(Yong-Hyun Cho)

1979년 2월 : 경북대학교 전자공학과 졸업(공학사)

1981년 2월 : 동대학원 전자공학과 졸업(공학석사)

1993년 2월 : 동대학원 전자공학과 졸업(공학박사)

1983년 9월~1984년 2월 : 삼성전자(주)

1984년 3월~1987년 2월 : 한국전자통신연구소

1987년 3월~1997년 2월 : 영남전문대학 전자과 부교수

1997년 3월~현재 : 대구효성가톨릭대학교 공과대학 전자정보공학부 조교수

주관심분야 : 신경망, 병렬분산처리, 신호처리, 컴퓨터구조, 교환기 등