論文95-32B-3-11

신경회로망에서 일괄 학습

(Batch-mode Learning in Neural Networks)

金 明 潔 *, 崔 悰 鎬 *

(Myung-Chan Kim, and Chong-Ho Choi)

요 약

본 논문에서는 패턴 분류문제에 있어서 오류역전달 학습방법의 학습속도 향상을 위해서 학습률과 관성변수가 학습시마다 조정되는 새로운 일괄 학습 (batch-mode learning) 알고리즘을 제안한다. 제안된 알고리즘에서는 학습 패턴수와 출력 노드수에 대하여 평균을 취한 목적함수를 사용하고, 오차역전달시 연결가중치 개개의 오차분담값을 크게하기 위해 목적함수의 기울기 놈 (norm) 으로 기울기를 정규화시킨다. 또한 적용문제에 따라 학습률과 관성변수가 민감하지 않도록 설정하기 위해, 이들을 목적함수 기울기 놈과 연결가중치 갯수의 함수 형태로 하여, 매 연결가중치 갱신때마다 바꾼다. 이때 학습률은 목적함수 기울기 놈의 제곱근, 관성변수는 기울기 놈에 따라 변화시킨다. 두가지 대표적인 패턴분류 모의실험 결과 기존의 일괄 학습이나 패턴별 오류역전달 알고리즘보다 제안된 일괄 학습알고리즘이 학습속도가 향상된 것으로 나타났다.

Abstract

A batch-mode algorithm is proposed to increase the speed of learning in the error backpropagation algorithm with variable learning rate and variable momentum parameters in classification problems. The objective function is normalized with respect to the number of patterns and output nodes. Also the gradient of the objective function is normalized in updating the connection weights to increase the effect of its backpropagated error. The learning rate and momentum parameters are determined from a function of the gradient norm and the number of weights. The learning rate depends on the square root of the gradient norm while the momentum parameters depend on the gradient norm. In the two typical classification problems, simulation results demonstrate the effectiveness of the proposed algorithm.

I. 서 론

신경회로망은 그 사용범위가 넓어 패턴인식과 같은 분류문제(classification problem), 비선형 시스템 식별 및 제어, 시계열 예측등에 활발히 응용되고 있다. 신경회로망은 그 적용문제에 따라 적절한 학습 알고리즘들이 개발되어 사용되고 있으나, 그들중 오류역전달 (error backpropagation, EBP, BP) 알고리즘이 가장 널리 사용되고 있다[1][2][3]. 신경회로망 기본 구조의 하나인 다층인식자 (multilayer perceptron, MLP) 를 학습시키는 대표적인 학습방법인 BP 알고리즘의 기본 개념은 출력층에서의 기대값과 실제 출력

* 正會員, 서울大學校 制御計測工學科
(Dept. of Control and Instr. Eng., ERC-ACI, ASRI, Seoul National Univ.)

※ 본 연구는 일주학술문화재단의 지원에 의해 수행되었습니다.

接受日字 : 1994年 10月 8日

값 사이의 오차를 감소시켜 나가는 방향으로 연결가중치를 조정해 나가는 것이다 [1] [3]. BP 알고리즘은 그 기본 개념의 단순성에도 불구하고 MLP 와 결합하여 강력한 문제 해결 능력을 보임으로써 많은 응용분야에서 적용되고 있다[2]. 그러나 MLP를 학습시키는데 많은 시간이 걸리며 문제가 복잡해지면 학습 시키기가 어렵다는 단점을 가지고 있다.

현재까지 연구된 BP 알고리즘의 학습속도 향상에 대한 연구를 보면 크게 두가지로 분류할 수 있다. 첫째는 학습률을 경험적인 판단에 의해 변화시키는 방법이다. 예를들면 연속적인 연결가중치 변화분이 서로 반대부호가 되면, 진동을 하고 있는 것으로 판단하여 그 연결가중치에 대한 학습률을 감소시키며, 반대로 연속적인 연결가중치 변화분이 같은 부호이면 그 연결가중치에 대한 학습률을 증가시켜 주는 것이다[4] [5] [6]. 두번째는 비선형 프로그래밍 방법 및 함수 최적화 문제에서 사용되는 방법들을 도입하는 것으로서 주로 오차함수의 1차 및 2차 미분항을 이용하는 Newton's Method, Conjugate Gradient Method 등이 있다 [4] [7]. 그러나 이러한 방법은 신경회로망 구조가 커질수록 계산량이 급격히 증가하는 단점이 있다. 이외에도 Kalman Filter 이론을 적용하여 학습속도를 향상시키는 연구도 있으며, 학습이 안되는 패턴만을 학습시키는 방법이 제안되기도 하였다[8] [9] [10].

알고리즘에는 크게 두가지 학습모드가 있다. 하나는 패턴 (pattern, on-line-mode) 학습이며, 다른 하나는 일괄 (batch-mode) 학습이다. 패턴 학습은 하나의 입력패턴이 학습에 사용될 때마다 연결가중치들이 갱신되는 방법인데 반하여, 일괄 학습은 전체 패턴이 학습에 한번씩 사용된 후 (1 epoch) 연결가중치들을 갱신하는 방법이다. 초기 BP 는 일괄 및 패턴 학습이 동시에 사용되었으나, 일괄 학습은 학습속도가 매우 느려서 패턴 학습이 주로 사용되었다. 현재는 패턴 학습에 관성항 부분을 첨가해서 쓰는 형태가 널리 사용되고 있는 실정이다. 그러나 패턴 학습은 하나의 입력 패턴으로 인한 오차함수의 기울기 (패턴별 경사정보) 를 이용하므로 학습속도가 느려지기 쉬우며, 학습 데이타의 구성 순서가 학습속도에 큰 영향을 미치게 된다. 이런 측면에서 보면 일괄 학습은 학습 데이타 구성 순서와는 무관하게 전체패턴에 대한 오차함수 기울기 정보를 연결가중치 갱신에 이용할 수 있다는 장점을 가지고 있어 이를 잘 이용하면 학습도 더 잘 시킬 수 있을 것이다. 한편 이 두가지 BP 학습 알고리즘들은 학습률 및 관성항 변수 설정에 따라 학습속도가 영향을 많이 받는다. 그러나 적용문제에 따른 두가지 변수 설정에 대한 정해진 규칙이 없다. 단지 신경회로망 구성자의 경험에 의존하여 적용문제에 따라 시행착오로 학습률 및 관성항 변수를 설정하는 것이 보통이다.

일괄 학습의 잇점인 전체패턴에 대한 오차함수 기울기 정보를 이용하면서 느린 BP 학습속도를 향상시키려는 연구가 있었다[11] [12]. 그 연구에서는 목적함수 기울기항만으로 연결가중치 갱신식을 구성하였으며, 목적함수 기울기 놈의 제곱으로 기울기 항을 나누어 주었다. 그리고 학습률은 오차함수 값과 미리 설정된 스텝크기의 곱을 사용하였다. 그러나 이방법은 여전히 스텝크기 설정에 대한 기준이 없는 상태이고, 문제에 따라 다른 값을 사용하여야 하며, 국부최소값 근처에서는 매우 큰 점프가 일어날 가능성이 많다.

본 논문에서는 일괄 학습의 잇점을 이용하면서, 될 수 있으면 문제에 대해서 영향을 덜 받도록 학습률 및 관성항 변수를 설정하여 기존의 오류역전달 학습 알고리즘의 학습속도를 향상시키는 방법을 제안한다. 따라서 제안된 알고리즘은 Atiya[11] 와는 달리 목적함수 기울기항 및 관성항을 연결강도 갱신식에 사용한다. 제안된 알고리즘은 일괄 학습이므로 연결가중치가 갱신되어갈 올바른 방향을 사용할 수 있고, 느린 일괄 학습속도를 향상시키기 위해 목적함수 기울기 놈과 개별 연결가중치와의 관계를 이용한다. 우선 제안된 알고리즘은 학습패턴 수 및 출력노드 갯수에 대하여 정규화 (normalize) 를 시킨 목적함수를 사용하여 여러가지 적용문제에 있어서 학습 성공여부 및 연결가중치 갱신에 대한 공통적인 판단근거를 제공한다. 그리고 정규화된 목적함수의 기울기를 정규화시켜서 연결가중치 오차분담값을 키워주는 형태를 취하여 연결가중치 갱신값을 크게만든다. 또한 매 연결가중치 갱신시에 적용문제에 덜영향을 받기 위해 목적함수 기울기놈과 전체 연결가중치 갯수의 함수 형태로 학습률 및 관성항 변수값을 조정한다.

본 논문의 구성은 다음과 같다. II 장에서는 BP 알고리즘에 대해 간단히 설명을 하고, III 장에서는 본 논문에서 제안하는 알고리즘을 설명한다. 그리고 IV 장에서는 제안된 알고리즘의 성능평가에 관한 두가지 대표적인 모의실험 결과를 제시하고, 결론을 V 장에서 맺는다.

II. 기존의 BP 학습 알고리즘

본 논문에서 고려하고 있는 신경회로망 구조는, 초기 신경회로망 모델인 단층인식자 (perceptron) 의 한정된 함수사상 구현이라는 단점을 극복한 모델인 다층인식자로서, 은닉층이 1개이상의 다층인식자는 임의의 함수사상이 가능하다고 알려져 있다[13]. 다층인식자를

학습시키는 대표적인 알고리즘으로는 비교적 간단한 BP 알고리즘이 널리 알려져 있는데 여기서 간단히 살펴보자.

각 계층을 0 에서 L 까지 표시하되 0 번째 계층은 입력층, L 번째 계층은 출력층이라고 하자. 그리고 l 번째 계층의 노드수는 N_l 로 표시하고, p 번째 입력패턴에 의한 l 번째 계층의 i 번째 노드의 출력은 O_{pi}^{l} 로 나타내자. 그러면 p 번째 입력패턴에 대한 출력층에서 오차 E_p 는 다음과 같이 정의된다.

$$E_p = \frac{1}{2N_L}\sum_{i=1}^{N_L}(D_{pi}-O_{pi}^{L})^2 \tag{1}$$

여기서, D_{pi} 는 p 번째 입력패턴에 대한 L 번째 층의 i 번째 노드의 목표 출력값이다. 위의 (1)을 가지고 연결가중치를 수정해 나가는 패턴 학습 연결가중치 갱신식은 다음과 같다.

$$\Delta W(k) = -\eta\frac{\partial E_p}{\partial W(k)} + \mu\Delta W(k-1)(k-1) \tag{2}$$

$$W(k+1) = W(k) + \Delta W(k) \tag{3}$$

여기서 k 는 패턴 학습에서의 시간지표로서 연결가중치가 갱신된 횟수를 나타내며, η 는 학습률, μ 는 관성변수를 나타낸다. $W(k)$ 는 $(k-1)$ 번째 연결가중치 갱신후의 연결가중치 벡터로서 $(l-1)$ 번째 층의 j 번째 노드에서 l 번째 층의 i 번째 노드로의 연결가중치 성분 $w_{ij}^{l-1}(k)$ 과 l번째 층의 i번째 노드의 바이어스 성분 $\theta_i^l(k)$로 구성되어 있다. 즉 $W(k) = [\, w_{11}^{0}(k), \cdots, \theta_1^l(k), \cdots, w_{ij}^{l-1}(k), \cdots, \theta_i^l(k), \cdots, w_{N_l N_{l-1}}^{L}(k), \theta_{N_l}^{L}]^{T}$ 이다. 여기서 $[\,\cdot\,]^{T}$ 는 $[\,\cdot\,]$ 의 전치행렬 (transpose) 을 의미한다. 또한 $\Delta W(k)$ 는 k 번째 갱신에서의 연결가중치 갱신값 벡터이다. 위의 (2)에서 기울기 항목의 각 성분들은 다음과 같이 나타낼 수 있다.

$$\frac{\partial E_p}{\partial w_{ij}^{L-1}(k)} = \delta_{pi}^{L} O_{pj}^{L-1} \quad \text{L번째 층에서} \tag{4}$$

$$\frac{\partial E_p}{\partial w_{jm}^{l-1}(k)} = \delta_{pj}^{l} O_{pm}^{l-1} \quad l\in\{1,\cdots,(L-1)\}\text{번째 층에서} \tag{5}$$

$$\delta_{pi}^{L} = -(D_{pi}-O_{pi}^{L})O_{pi}^{L}(1.0-O_{pi}^{L}) \tag{6}$$

$$\delta_{pj}^{l} = -\sum_{m=1}^{N_{l+1}}\delta_{pm}^{l+1} w_{mj}^{l}(k) O_{pj}^{l}(1.0-O_{pj}^{l}) \tag{7}$$

기존의 일괄 BP 학습에서의 목적함수 E 와 연결가중치 갱신식은 다음과 같다.

$$E = \sum_{p=1}^{P} E_p \tag{8}$$

$$\Delta W(n) = -\eta\frac{\partial E}{\partial W(n)} + \mu\Delta W(n-1) \tag{9}$$

여기서 P 는 학습에 사용된 총 입력패턴 수이며, n 는 일괄 학습에서 연결가중치 갱신 횟수 (epochs) 를 의미한다.

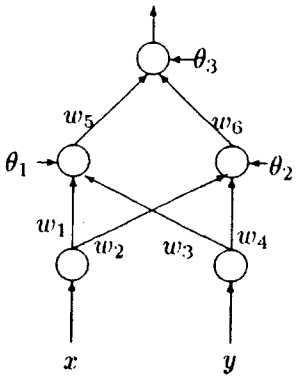

그림 1. XOR 문제 해결에 쓰인 신경회로망 구조
Fig. 1. Neural Network for XOR Problem.

연결가중치 갱신식 (2)와 (9)에서, BP 알고리즘은 경사감소법의 일종이며 학습속도는 학습률 η 의 영향을 크게 받음을 알 수 있다. 또한 (2) 와 (9) 에서, 일괄 BP 연결가중치 갱신성분 값이 패턴별 BP 의 그것보다 작아질 수 있음을 알 수 있다. 왜냐하면 일괄 BP 는 패턴별 BP 가 사용하는 기울기 성분들의 합을 연결가중치 갱신에 사용하기 때문이다. 예를들어, 그림 1과 같은 신경회로망에 XOR 문제를 BP 방법으로 학습해 나갈 때 연결가중치 w_1 과 w_5 에 대한 오차함수 기울기가 어떻게 변화하는지를 그림 2에 나타냈다. 그림 1에서 w_i 는 연결가중치, θ_i는 바이어스를 의미한다. 그림 2에 의하면, (9) 에서 쓰이는 기울기 $\frac{\partial E}{\partial w_i}$ 는 초반부에 거의 0 근처임을 알 수 있다. 그러나 (2)에서 쓰이는 기울기 (각 입력패턴에 의한 기울기) $\frac{\partial E_p}{\partial w_i}$ 는 0 근처 값은 아니다. 이러한 현상에 대하여서는 다음과 같이 설명할 수 있다. 출력노드에서 목적 출력값을 0 또는 1 이라 하자. 첫번째 학습에서 연결가중치 w_5 경우를 보자. 초기 연결가중치 값들은 0 근처이므로 각 노드 출력은 0.5 근처가 된다. 아래의 대략적인 계산식은 그림 2에서 연결가중치 w_5 가 초반에 각 패턴별로 0.0625 와 −0.0625 근처의 기울기 성분을 갖고 있음과 일치한다.

$$\frac{\partial E_1}{\partial w_5} = (0.0-0.5)\cdot 0.5\cdot(1.0-0.5)\cdot 0.5 = -0.0625 \quad (10)$$

$$\frac{\partial E_2}{\partial w_5} = (1.0-0.5)\cdot 0.5\cdot(1.0-0.5)\cdot 0.5 = 0.0625 \quad (11)$$

$$\frac{\partial E_3}{\partial w_5} = (1.0-0.5)\cdot 0.5\cdot(1.0-0.5)\cdot 0.5 = 0.0625 \quad (12)$$

$$\frac{\partial E_4}{\partial w_5} = (0.0-0.5)\cdot 0.5\cdot(1.0-0.5)\cdot 0.5 = -0.0625 \quad (13)$$

$$\frac{\partial E}{\partial w_5} = \sum_{p=1}^{4}\frac{\partial E_p}{\partial w_5} = 0.0 \quad (14)$$

위와같은 설명은 그림 1의 모든 연결가중치에 대해서 할 수 있다.

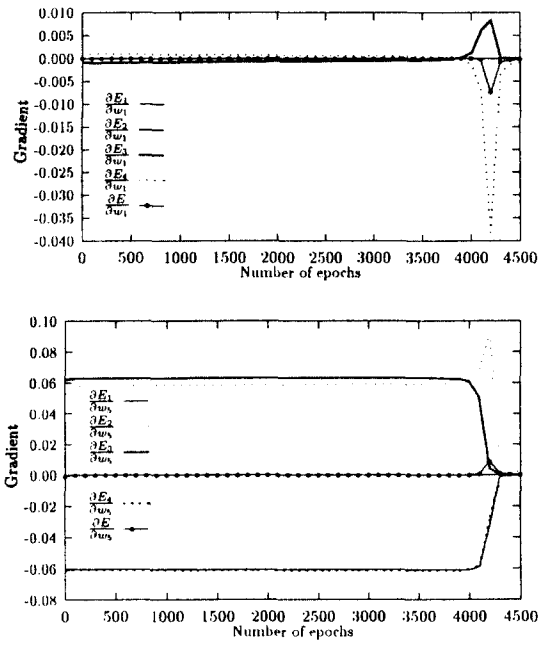

그림 2. w_1, w_5에 대한 오차함수 기울기 변화
Fig. 2. Curves of $\frac{\partial E}{\partial w_1}$ and $\frac{\partial E}{\partial w_5}$.

위와같은 현상은 훈련 데이타 구성, 연결가중치 초기화 및 시그모이드 함수특성에 의한 대칭 (symmetric) 때문이다. 일반적으로 패턴분류 문제에서 학습 데이타는 각 패턴별로 고루 분포시키는데, XOR 경우도 출력 1 에 해당하는 패턴 2 개, 출력 0 에 해당되는 패턴 2 개로 분포되어 있다. 그리고 초기 연결가중치는 조기포화 (premature saturation) 상태를 방지하기 위해 0 근처 값으로 설정된다. 또한 각층 노드가 사용하는 활성함수 (activation function) 는 시그모이드 함수로서, 그 특성상 0.5 를 중심으로 하여 대칭인 형태를 이루고 있다. 이 세가지 요인으로 인해 학습 초반부에 각 입력패턴별 목적함수 기울기는 대칭이 되며 그 부호가 달라 기울기의 합은 거의 0 이 되어 그림 2와 같은 현상이 나타나는 것이다. 이러한 대칭성때문에 일괄 BP 가 패턴 BP 보다 연결가중치 갱신값이 매우 작게 되며, 학습에 걸리는 시간이 길어지게 된다.

Ⅲ. 오차함수 기울기 놈 (norm) 을 이용한 일괄 BP 학습방법

일반적으로 BP 알고리즘을 이용하여 학습하는 경우, 학습곡선에서 오차가 급격하게 떨어지는 구간 (빠른 학습구간) 및 오차가 거의 변화가 없는 구간 (느린 학습구간) 이 번갈아 나타나면서 학습이 진행되는 현상을 관찰할 수 있다. 두구간을 비교해보면, 느린 학습구간에 비교적 많은 연결가중치 갱신이 일어난다. 때로는 느린 학습구간이 매우 길어 학습 진행이 안되는 것처럼 보이기도 한다. 따라서 느린 학습구간을 없애거나 또는 얼마나 많이 단축시키느냐가 일괄 BP 알고리즘을 효율적으로 만드는 중요한 관건이 된다.

느린 학습구간이 발생하게 되는 원인은 여러가지 요인이 존재한다고 볼 수 있다. 그중에서 목적함수가 빗물받이 홈통 (rain gutter[2]) 과 평원지대 (flat plateau[2]) 들을 가지고 있어서 느린 학습구간이 존재한다고 여겨진다. 느린 학습구간에서 오차함수 기울기 놈 $\|\frac{\partial E}{\partial W(n)}\|$ 는 때때로 매우 작은 값을 가지게 된다. 경우에 따라서는 오차함수 기울기가 10^{-7} 이하가 되기도 한다. 이런 경우 한 epoch 에 $\|\Delta W\| = 0.1$ 정도 변하게 된다고 하여도 $\Delta E = 0.01$ 이 되게 하기 위해서는 10^6 번의 학습이 필요하다고 볼 수 있다. 느린 학습구간은 입력층과 바로 윗 은닉층 사이의 연결가중치들로 구성되는 입력영역 분할경계함수 (discriminant boundary function) 의 기울기 조정이 일어나는 학습초기에 많이 발생하는 경향이 있다. 입력영역 분할 경계함수 기울기 조정이 끝난 후, 은닉층과 출력층 사이 연결가중치 갱신이 신속하게 이루어져 급격한 오차감소 현상을 보이는 빠른 학습구간이 나타난다.

학습속도 향상을 위해서 느린 학습구간을 올바른 방향으로 빠르게 지나가는 것이 본 논문에서 제시하는 알고리즘이 성취하고자 하는 목적이다. 제안된 알고리즘이 가지는 기본 생각은 올바른 방향을 선택하기 위해 일괄 학습방법을 사용하며 학습속도의 향상을 위해 기울기 $\frac{\partial E}{\partial W(n)}$ 의 각 요소값을 바로 자신의 놈 (euclidean norm) 을 이용하여 정규화시켜 학습시 각 연결가중치들의 오차분담의 크기를 키워주는 것이다. 우선 훈련 데이타 갯수에 상관없이 연결가중치 갱신에

있어서 일관된 판단 조건을 주기 위해서 오차함수를 패턴수 P와 출력층 노드 갯수 N_L 에 대하여 정규화시켰다. 여기서 사용한 오차함수 E_b 는 다음과 같다.

$$E_b = \frac{1}{P}\sum_{p=1}^{P} E_p = \frac{1}{2PN_L}\sum_{p=1}^{P}\sum_{i=1}^{N_L}(D_{p,i} - O_{b,i}^{L})^2 \quad (15)$$

기존의 (8) 과 같이 표현된 오차함수는 훈련용 패턴수에 따라 그 값이 변하나 제안된 오차함수 (15) 는 패턴수 및 출력수에 대해서 정규화되어 있어 대부분의 적용문제에 있어서 학습 성공여부에 대한 일관된 판단을 내릴 수 있다. E_b 가 취하는 최대값과 최소값은 각각 0.5 와 0 이다. 학습시 초기 연결 강도값들은 주로 0 근처값으로 설정 $(w_{ij}(0) \approx 0, \forall i,j)$ 되므로 각층의 노드들의 출력은 시그모이드 함수 특성상 0.5 근처 값을 갖는다. 이경우 제안된 오차함수 (15)는 적용문제에 관계없이 학습초기에 대략 $(0.5)^2/2 = 0.125$근처값을 취하게 된다. 이런 관점에서 볼 때 제안된 오차함수는 훈련 패턴수에 무관하게 학습 진행정도에 대한 일관된 판단근거를 제공할 수 있다.

위의 오차함수 E_b 를 이용한 제안된 연결가중치 갱신식은 다음과 같다.

$$\Delta w_{ij}^{l}(n) = -\eta(n)\frac{\partial E_b}{\partial w_{ij}^{l}(n)} + \mu_{ij}(n)\Delta w_{ij}^{l}(n-1) \quad (16)$$

$$\eta(n) = a(z)\sqrt{z}/\|\frac{\partial E_b}{\partial W(n)}\| \quad (17)$$

$$a(z) = \max[\ 0.8,\ -7\sqrt{z} + 1.5\ ] \quad (18)$$

$$\mu_{ij}(n) = (1.0 - b(z)\frac{|\frac{\partial E_b}{\partial w_{ij}^{l}(n)}|}{\|\frac{\partial E_b}{\partial W(n)}\|}(1.0 - z) \quad (19)$$

$$b(z) = \min[\ 1.0, 9\sqrt{z} + 0.1\ ] \quad (20)$$

$$z = \frac{|\frac{\partial E_b}{\partial W(n)}|}{\sqrt{N_w}} \quad (21)$$

여기서 N_w 는 연결가중치 총 갯수이다. 그리고 $a(z) \in (0.8, 1.5), b(z) \in (0.1, 1.0)$ 이 됨을 쉽게 알 수 있다.

오차함수 기울기 $\frac{\partial E_b}{\partial W(n)}$ 벡터는 (16) 와 (17)에서 알 수 있듯이 자신의 기울기 놈(norm)에 의해 정규화된다. 위식들을 다시 정리하여 쓰면 다음과 같다.

$$\Delta w_{ij}^{l}(n) = -\eta_{new}(n)\frac{\frac{\partial E_b}{\partial w_{ij}^{l}(n)}/\sqrt{N_w}}{z} + \beta_{ij}(n)\Delta w_{ij}^{l}(n-1) \quad (22)$$

$$\eta_{new}(n) = a(z)\sqrt{z} \quad (23)$$

$$\beta_{ij}^{l}(n) = (1.0 - b(z)\frac{|\frac{\partial E_b}{\partial w_{ij}^{l}(n)}|/\sqrt{N_w}}{z})(1.0 - z) \quad (24)$$

위의 (21)와 (22) 에 의하면, z 는 목적함수를 $E_b/\sqrt{N_w}$ 로 하였을 때의 기울기 놈으로 볼 수 있다. 그리고 $\frac{\partial E_b}{\partial W(n)}/\sqrt{N_w}$ 의 각각의 성분은 연결가중치가 아주 큰 경우를 제외하고는 1 보다 작은 값을 가진다. 따라서 목적함수 기울기항을 정규화시킨다는 것은 학습시 연결가중치의 오차분담값이 커지는 효과를 가져다 주는데 이는 기존의 BP 알고리즘이 입력층으로 내려갈수록 오차분담값이 작아져 학습시간이 오래 걸리는 단점을 개선시켜주는 요인이다. 또한 z 를 사용하는 이유는 적용문제와 학습 패턴수 및 연결가중치 총수에 어느정도 덜 민감한 연결가중치 갱신식을 만들기 위해서이다.

한편 (22) 로 구성된 $\Delta W(n)$ 의 놈을 취하면 관성항을 무시할 수 있는 경우는 $\|\Delta W(n)\| = \eta_{new}(n)$ 이 됨을 알 수 있다. 그리고 (23)에 따르면 학습률이 새로운 목적함수의 기울기 놈의 제곱근에 따라 변하는 것을 알 수 있다. 여기서 제곱근을 사용하기 때문에 기울기 놈이 매우 작아지더라도 학습률이 급격히 줄어드는 것을 방지하는데, 이는 제곱근 함수의 기울기 특성상에서 오는 잇점으로 학습률 조정의 중요한 요소이다. 기울기 항을 정규화시켜주는 방법은 다른 논문에서도 찾아볼 수 있으나 [11] 학습률을 기울기 놈의 제곱근에 비례하게 하여 설정하는 방법은 여기서 처음 시도된다. 또한 (18) 의 역할은 $\eta_{new}(n)$ 가 z 의 값에 따라 너무 크거나 작은 값이 되는 것을 보상하는 것이다. 다음 Ⅳ장의 모의실험 결과에서 대부분의 기울기 놈은 $10^{-7} \sim 10^{-2}$ 사이에 존재하였는데, 이러한 경우에 $\eta_{new}(n)$ 은 대략 계산하여 약 $1.5 \cdot 10^{-3.5} \sim 0.8 \cdot 10^{-1}$ 사이의 값을 취하게 된다. 그런데 이범위는 기존의 BP 알고리즘이 택하는 학습률의 선택범위에 들어간다고 볼 수 있다.

제안된 알고리즘에서 관성변수는 각 연결강도별로 변화되도록 하였다. 위의 (24) 에서 알 수 있듯이 $(1.0 - z)$ 항에 의해 연결가중치 전체에 대한 일차적인 관성항 영향을 조정한다. 그리고 일률적인 관성변수 적용을 하는 기존의 오류역전달 알고리즘과는 달리 연결가중치가 개별적으로 정규화된 목적함수 기울기 성분의 절대값 $\frac{|\frac{\partial E_b}{\partial w_{ij}^{l}(n)}|/\sqrt{N_w}}{z}$ 과 (20) 를 이용하여 자신의 관성항 적용정도를 결정하는 이차적인 조정을 한다. 이

러한 두차례 관성항 조정은 일률적인 관성항의 적용보다는 학습에 좀 더 효율적이라 기대된다.

표 1. 각 알고리즘의 epoch 당 학습에 필요한 계산량

Table 1. Computational requirement for weight update per epoch.

알고리즘	곱셈
패턴별 BP	$4N_wP+(s+2)N_nP+N_2N_1P$
일괄BP	$2N_w(P+1)+(s+2)N_nP+N_2N_1P$
제안된 일괄 BP	$N_w(2P+5)+(s+2)N_nP+N_2N_1P+2r$

기존의 일괄 BP, 패턴별 BP 및 제안된 일괄 학습 방법들 사이의 곱셈 계산량을 대략 비교한 결과를 표 1에 나타내었다. 표 1에서는 은닉층이 하나있는 MLP 경우로서, 학습 패턴을 전부 학습에 사용한 (1 epoch 에 해당) 계산량이다. 여기서 N_n 는 은닉층 노드 수 N_1 과 출력층 노드 수 N_2 의 합이다. 그리고 r는 제곱근 계산, s 는 시그모이드 함수 계산에 필요한 곱하기 연산량을 의미한다. 전방향 과 역방향 계산은 세가지 알고리즘이 모두 같은 정도의 계산량을 필요로하는 계산이다. 따라서 계산량이 차이가 나는 부분은 연결가중치 갱신부분이다. 표 1에서 알 수 있듯이, 매 epoch 당 계산량은 제안된 알고리즘이 패턴별 BP보다 작다.

관성항을 무시한 (16) (또는 (22)) 의 첫번째 항 (기울기 항) 만 가지고 연결가중치를 갱신한 후의 목적함수 변화분을 테일러 시리즈 1차항만 고려 하여 나타내면 다음과 같다.

$$\Delta E_b \propto \| \frac{\partial E_b}{\partial W(n)} \|^{\frac{3}{2}} \tag{25}$$

그리고 기존의 일괄 BP 알고리즘의 경우 연결가중치를 갱신한 후의 목적함수 변화분을 같은 방식으로 계산하면 다음과 같다.

$$\Delta E \propto \| \frac{\partial E}{\partial W(n)} \|^2 \tag{26}$$

두식 (25) 과 (26) 의 표현에서 보면, 만약 기울기 놈이 매우 작은 경우 ($10^{-4} \sim 10^{-8}$) 에는 ΔE_b 가 ΔE 보다 큰 값을 가짐을 알 수 있다. 따라서 본 논문에서 제안된 알고리즘이 기울기 절대값이 작은 오차영역에서는 기존의 일괄 BP 알고리즘보다 훨씬 효과적이라 할 수 있다. 또한 제안된 알고리즘은 일괄 학습이므로 데이타의 학습 순서는 고려할 필요가 없으며, 매 연결가중치 갱신시 계산되는 기울기 방향은 연결가중치가 갱신되어가야 할 올바른 방향이라는 잇점도 가지게 된다. 따라서 데이타 구성순서에 영향을 받으며, 또한 지그재그로 연결가중치를 갱신해 나가는 패턴별 BP 알고리즘보다도 여기서 제시한 방법이 빠른 학습 속도를 보여줄 수 있는 가능성이 크다고 하겠다.

Ⅳ. 모의 실험

제안된 알고리즘은 2가지 대표적인 문제, 간단한 2 비트 패리티(XOR) 문제와 복잡한 two spirals 문제, 에 대해서 모의실험을 하였다[1] [2] [3] [14] [15] [16]. XOR 문제에 대해서는 10개의 서로 다른 초기 연결가중치에 대해서 모의실험을 실시하였고, two spirals 문제에 대해서는 5가지 서로 다른 초기 연결가중치에 대해서 모의실험을 실시하였다. 두가지 문제에 대해서 학습 성공조건으로는 공통으로 모든 학습패턴이 제대로 분류가 되었을 경우로 하였다. 그리고 목적출력값은 패턴 학습에서는 보통 많이 사용하는 0.1과 0.9 를, 일괄 학습에서는 0 과 1 를 사용하였다. 각 문제에 대해서 기존의 일괄 및 패턴별 BP 와 제안된 알고리즘을 비교하였다.

1. XOR 문제

XOR 문제를 위한 MLP 구조는 2 개의 입력 노드를 가진 입력층, 2개의 노드를 가진 은닉층 및 1 개의 노드로 구성된 출력층으로 이루어졌다. 패턴 학습을 위해서는 학습률 η는 0.2, 관성항 μ 는 0.9 로 놓고 학습을 하였다. 그림 3에서 제안된 알고리즘과 패턴별 BP 알고리즘의 학습곡선의 평균을 나타냈다.

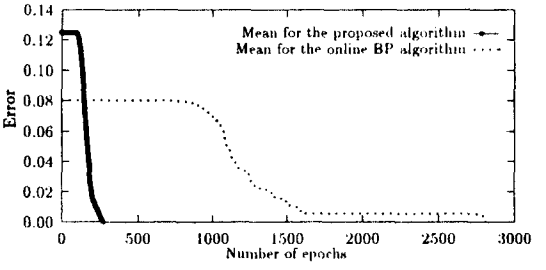

그림 3. 제안된 알고리즘과 패턴별 BP 알고리즘 학습곡선 비교

Fig. 3. Learning curves for the on-line BP algorithm and the proposed algorithm.

그림 3에서 알 수 있듯이 패턴별 BP 알고리즘은 학습 초반부에 일종의 느린 학습구간을 보이고 있다. 그러나 제안된 알고리즘은 예상대로 초기 느린 학습구간을 단축시켰다고 볼 수 있다. 기존의 일괄 BP 알고리즘은 패턴 학습보다 더 학습이 안되었다. 총 10 번의

모의실험에서 학습 성공횟수, 학습이 끝날 때까지 평균 연결가중치 갱신횟수, 성공적인 학습을 보인 최소한의 연결가중치 갱신 횟수 등에 대하여 표 2에서 패턴별 BP 알고리즘과 제안된 알고리즘을 비교하였다. 제안된 알고리즘은 기존의 BP 알고리즘보다는 적어도 7배 빨리 학습이 됨을 알 수 있었다.

표 2. XOR 문제의 모의실험 결과(10번 시도, 각 3,000 epochs 까지 수행)

Table 2. Simulation results for XOR problem(10 trials with 3,000 epochs trial)

알고리즘	학습 성공회수	최단 학습시간 (epochs)	평균 학습시간 (epochs)
패턴별 BP	10/10	1,070	1,421
일괄 BP	2/10	1,862	2,009
제안된 일괄 BP	10/10	156	195

2. Two Sprials 문제

제안된 알고리즘의 성능을 판단하기에 XOR 문제는 너무 쉬운 문제이다. 그래서 기존의 BP 알고리즘을 가지고서는 학습시키기가 어렵다고 알려진 two spirals 문제에 대해서 모의실험을 수행하였다. 훈련 데이타는 총 194개의 패턴으로 구성되는데 *xy* 평면상에서 2개의 나선을 그리는 점들의 집합이다. 각 나선은 각각 1과 0 을 나타내는 총 97개의 점들로 이루어졌다.

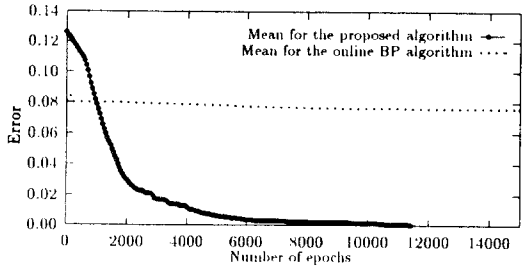

그림 4. 제안된 알고리즘과 패턴별 BP 와의 학습 곡선 비교

Fig. 4. Learning curves for the on-line BP algorithm and the proposed algorithm.

Lang 과 Witbrock [14] 은 일반적인 다층인식자 구조가 아닌 특수한 구조의 신경회로망을 사용하였다. 이들의 연구결과 에 따르면 Increasing learning rate 와 관성항을 이용한 BP 알고리즘의 경우 세번의 서로 다른 초기 연결가중치의 실험 평균 약 20,000번의 연결가중치 갱신 횟수가 걸렸고, 비선형 cross-entropy 목적함수를 이용한 경우 평균 약 11,000 번이 걸렸다. 또한 Quickprop 알고리즘을 이용한 경우에는 약 7,900 번의 연결가중치 갱신 횟수가 걸렸다 [14] [15]. 그러나 이와 같은 결과를 얻기 위해서는 arctangent 오차 함수와 망각인자 (forgetting factor) 가 필요하였다.

본 연구에서 사용한 MLP 구조는 은닉층이 하나인 구조로써 위의 다른 연구와 같은 5 층 MLP 가 아니다. 입력층 2 개, 은닉 층은 65 개, 출력층은 1 개의 노드로써 각 계층이 구성된다. 패턴별 BP 알고리즘을 위해서는 학습률은 0.02, 관성항은 0.9를 사용하였다.

기존 일괄 BP 학습이나 패턴 학습방법으로는 학습에 성공하지 못하였다. 그림 4에 나타난 바와 같이 패턴별 BP 알고리즘은 일종의 매우 긴 느린 학습구간을 보여주는 경우가 발생한다. 그러나 제안된 학습알고리즘은 5번의 서로 다른 초기 연결가중치에서 시작하여 모두 12,000 연결가중치 갱신횟수 이전에 학습을 마쳤다. 제안된 학습 알고리즘의 학습곡선을 보면 초기에 급격한 오차감소 현상을 볼 수 있다. 5 번의 모의 실험결과 평균 약 8,902번의 연결가중치 갱신횟수가 학습에 소요되었다. 표 3에 패턴 학습과 제안된 알고리즘을 비교하였다. 이상의 결과로 제안된 학습 알고리즘은 기존의 일괄, 패턴별 BP 알고리즘보다 그리고 현재까지 알려진 일부 다른 학습 알고리즘 보다 빠른 학습속도를 보였다.

표 3. Two Spirl 문제의 모의실험 결과(5번 시도, 각 15,000 epochs 까지 수행)

Table 3. Simulation results for two spirals problem (5 trials, 15,000 epochs/trial)

알고리즘	학습 성공회수	최단 학습시간 (epochs)	평균 학습시간 (epochs)
패턴별 BP	0/5	-	-
일괄 BP	0/5	-	-
제안된 일괄 BP	5/5	5,057	8,902

V. 결 론

일괄 BP 알고리즘의 단점인 느린 학습구간을 단축시키기 위해 학습시 학습률 및 관성변수를 변하게 하는 일괄 학습방법을 제안하였다. 제안된 알고리즘은 학습용 데이타수에 대해서 정규화된 목적함수를 사용하며, 전체 패턴에 대한 목적함수 기울기 정보를 이용한다. 그리고 이러한 목적함수의 기울기를 전체 연결가중치 갯수로 나눔으로써 적용문제들에 대하여 덜 민감한 연결가중치 갱신식이 되도록 하였다. 즉 정규화된 목적함수 기울기 놈과 전체 연결가중치 갯수의 함수로 학습률과 관성변수를 매 학습시 마다 설정한다. 이때 학

습률은 기울기 놈의 제곱근의 함수이며, 관성변수는 기울기 놈의 함수이다. 이러한 일괄 학습은 목적함수 기울기 놈이 매우 작은 느린 학습구간에서 빠른 이동을 가능하게 한다. 제안된 알고리즘의 성능 평가를 위해 두가지 대표적인 패턴분류 문제에 대해서 모의실험을 하였다. 모의실험 결과 제안된 알고리즘은 초반에 일어나는 느린 학습 구간을 단축시킬 수 있었으며 좋은 학습결과를 얻었다. 이 결과는 제안된 일괄 학습방법이 다른 패턴분류 문제에 대해서도 좋은 학습결과를 줄 가능성이 높음을 의미하며, 기울기 놈과 전체 연결가중치 갯수의 함수로 제안된 학습률과 관성변수 설정이 효과적이라 볼 수 있다. 제안된 알고리즘은 패턴별 BP보다 훨씬 성능이 좋으며 계산량도 상대적으로 작아 분류문제에서 많이 쓰이는 학습방법이 되리라 기대한다.

참 고 문 헌

[1] R. P. Lippmann, "An introduction to computing with neural nets," *IEEE ASSP Magazine*, pp. 4-22, April 1987.

[2] J. Hertz, A. Krogh, and R. G. Palmer, *Introduction to the theory of neural computation*, vol. 1, Addison-Wesley Publishing Co., 1991.

[3] D. E. Rumelhart and J. L. McClelland, *Parallel Distributed Processing*, vol.1, vol. 2, MIT Press, 1986.

[4] R. A. Jacobs, "Increased Rates of Convergence Through Learning Rates Adaptation," *Neural Networks*, vol.1, pp. 325-334, 1988.

[5] A. A. Minai and R. D. Williams, "Acceleration of BackPropagation through Learning Rates and Momentum Adaptation," *Proc. of IJCNN*, vol. 1, pp. 676-679, Jan. 1990.

[6] A. A. Minai and R. D. Williams, "Back-Propagation Heuristics: A Study of the Extended Delta-Bar-Delta Algorithm," *Proc. of IJCNN*, vol. 1, pp. 595-600, July 1990.

[7] S. Becker and Y. Le Cun, "Improving the convergence of backpropagation learning with second order methods," *Proc. 1988 Connectionist Models Summer School*, pp. 29-37, 1988.

[8] S. Shah, F. Palmieri, and M. Datum, "Optimal Filtering Algorithms for Fast Learning in Feedforward Neural Networks," *Neural Networks*, vol. 5, pp. 779-787, 1992

[9] R. S. Scaleo and N. Tepedelenlioglu, "A Fast New Algorithm for Training Feedforward Neural Networks," *IEEE Trans. on Signal Processing*, vol. 40, no. 1, pp.202-210, Jan. 1992.

[10] H. Sawai, A. Waivel, P. Haffiner, M. Miyatake, and K. Shikano, "Parallelism, Hierachy, Scaling in Time-Delay Neural Networks for Spotting Japanese Phonemes /CV-Syllabes," *Proc. of IJCNN*, vol. 2, pp. 81-89, 1989.

[11] A. Atiya, A. Parlos, J. Muthusami, B. Fernandez, and W. Tsai, "Accelerated learning in multilayer networks," *Proc. of IJCNN*, vol. 3, pp. 925-929, 1992.

[12] A. G. Parlos, B. Fernandez, A. F. Atiya, J. Muthusami, and W. K. Tsai, "An Accelerated Learning Algorithm for Multilayer Perceptron Networks," *IEEE Trans. on Neural Networks*, vol. 5, pp. 493-497, 1994.

[13] R. Hecht-Nielsen, *Neurocomputing*, Addison-Wesley Publishing Co., 1989.

[14] K. Lang and M. Witbrock, "Learning to tell two spirals apart," *Proceedings of 1988 Connectionist Models Summer School*, Morgan Kaufmann, 1988.

[15] M. Riedmiller and H. Braun, "A Direct Adaptive Method for Faster Backpropagation Learning The RPROP Algorithm," *Proc. of IJCNN*, vol. 1, pp. 586-591, 1993.

[16] J. M. Zurada, "Lambda learning rule for feedforward neural networks," *Proc. of IJCNN*, pp. 1808-1811, 1993.

저 자 소 개

金 明 潔(正會員)

1968년 3월 21일생. 1990년 2월 서울대학교 제어계측공학과 졸업(학사). 1992년 2월 서울대학교 대학원 제어계측공학과 석사. 1992년 3월 ~ 현재 서울대학교 대학원 제어계측공학과 박사과정. 연구분야 : 신경회로망 이론, 패턴인식

崔 悰 鎬(正會員) 第 27卷 B編 第 2號 參照

현재 서울대학교 제어계측공학과 교수.