활성화함수의 기울기를 이용한 수렴속도 개선 알고리듬

○김 대국*, 이 상희*, 김 백섭**, 권 호열*
* 강원대학교 전자공학과, **한림대학교 전자계산학과

Improved algorithm for learning speed by using the slope of activation function

D.K.Kim, S.H.Lee, B.S.Kim, H.Y.Kwon,
* Dept. of Electronics Engineering, Kang Weon National Univ
** Dept. of Computer Science, Hallym Univ.

ABSTRACT

Although the back-propagation(BP) algorithm is widely used for its simple structure and easy learning method, it has a drawback of slow convergence rate.

In this paper, we propose an algorithm to improve this problem by manipulating the slope parameter of the activation function.

The steepest descent method is used in learning the slope parameter, as in the case of weight.

The simulation shows that the learning rates of the proposed algorithm is faster than the conventional BP algorithm.

I. 서 론

다층 퍼셉트론은 입력패턴을 받아들이는 입력노드들과 출력을 나타내는 출력층으로 이루어져 있고, 그 사이에 이들 입력층과 출력층에 모두 연결되어 있는 중간층들이 있는 구조이다.

이렇게 중간층들을 삽입함으로써 단층퍼셉트론이 가졌던 문제, 즉 비선형 판별 경우에 학습이 불가능 하다는 문제를 해결하여 supervised 패턴인식, 예측, machine learning 등 여러분야에 이용되고 있다. 다층퍼셉트론의 학습방법으로는 오류역전파 방법이 제안되어 있는데, 이는 학습데이타를 입력층에 인가하여 나오는 출력값과 주어진 원하는 출력값사이의 평균오차를 최소화 하기 위해 gradeint descent search 방법을 사용한다.

인접층 노드간의 가중치값들을 작은 난수의 값으로 초기화 한 후에, 훈련데이타들을 반복적으로 입력층에 적용시킴으로서 학습을 수행한다.

이러한 과정을 통해서 계산된 실제출력값과 이상적인 출력값과의 차이, 즉, 오차를 이전의 층에 전파시킴으로서 가중치를 변화시켜, 전체 입력패턴들에 대한 오차가 매우 작은 값을 가질때까지 위 과정을 반복한다.

역전파(BACK-PROPAGATION) 학습방법은 많은 적용분야에서 널리 사용되고 있는 반면에 오차의 수렴속도가 느리다는 것과 초기치에 따라 조기 포화 상태에 들어 간다는 등의 문제가 있다.

수렴속도를 빠르게 하기 위해 학습률이나 관성항를 이용한 연구[2], 초기가중치를 결정하는 방법에 대한 연구[3], 2차 미분치를 사용하는 방법[4], 학습률을 동적으로 변경시키거나, 오차함수의 수정을 통한 알고리듬, 개선 등의 연구를 통하여 학습속도를 개선하고 있다.

또한 최적화기법중[3]의 하나로 지역최소값에 수렴되는 것을 방지하기 위한 방법중의 하나로 시뮬레이티이드 어닐링방법이 이용되는데, 이는 액체상태로 부터 온도를 서서히 낮추는 과정을 활성화함수에 적용하였는데, 적용된 온도 파라미타는활성화함수에서 기울기 역활을 하고 있음을 알 수 있다.

본 논문에서는 활성화 함수에 기울기 파라미타를 도입[5]하여 이를 학습시킴으로써 조기 포화 상태에서 쉽게 벗어나고 학습속도를 빠르게 하는 방법을 제안하고자 한다. 또한 제안된 알고리듬을 기존의 것[1]과 비교하기 위해 잘알려진 xor, parity, encoder 문제등에 적용한 결과를 보인다.

II. 제안된 BP 신경망의 학습 알고리듬

1. 제안된 알고리듬의 구조

오류역전파 학습알고리듬은 supervised 학습으로써 외부교사가 존재하여 입력패턴 X에 대해 원하는 출력패턴 D 를 출력하도록 연결가중치를 조정하여 실제출력 O 와 원하는 출력 D 와의 오차 E 를 최소로 하는 것이다. BP 신경망에서 뉴런간의 가중치를 W라 하고 학습패턴을 P=1,2,3,...,N 라고 하면, 입력패턴 P에 대응하는 에러는

$$E_P = \sum_k 1/2(D_{kp} - O_{kp})^2 \qquad (1)$$

로 표시된다. 여기서 D_{kp} 는 k 번째 출력 뉴론에 대해 원하

는 출력, 그리고 O_{kp} 는 그 뉴론의 실제 출력이다.

전체패턴에 대응하는 오차는

$$E = \sum_{P} E_P \qquad (2)$$

로 주어진다.

신경망을 학습시킨다는 것은 E 를 최소로 하는 파라미타를 찾는 문제로 귀착된다. 이를 위하여 제안된 알고리듬은 다음과 같다.

뉴런 j 의 총 입력량을 net_j 라고 표시하면 이는

$$net_j = \sum_{i} W_{ji} * O_{pi} + \theta_j \qquad (3)$$

이다. 여기서 W_{ji}는 뉴런 i 에서 뉴런 j 로 가는 가중치이고 θ_j 는 뉴런 j 의 bias 를 나타낸다.

마찬가지로 중간층에서 출력층으로의 총 입력량도 식(3)와 같은 방법으로 구할 수 있다.

뉴런망 전달함수에 식(3)에서 구한 총 입력량을 인가하는 출력 O_j 는 다음과 같다.

$$O_j = F_j(net_j) = 1/(1+e^{-net_j}) \qquad (4)$$

하지만, 제안된 알고리듬에서는 활성화 함수에 기울기 T 를 도입함으로써 아래식과 같이 변형하였다.

$$O_j = F_j(T_j*NET_j)=1/(1+e^{-T_j*net_j}) \qquad (5)$$

steepest decent 방법을 이용하여 기울기 T_j 와 가중치 W_{ji} 들은 다음과 같이 학습되어 진다..

$$\Delta W_{kj}(n+1) = \eta*\delta_k*O_j + \alpha*\Delta W_{kj}(n) \qquad (6)$$

$$\delta_k = -(D_k-O_k)*F_k'(NET_k)$$

$$\Delta W_{ji}(n+1) = \eta*\delta_j*O_i + \alpha*\Delta W_{ji}(n) \qquad (7)$$

$$\delta_j = (\sum \delta_k*W_{kj})*F_j'(NET_j)$$

$$\Delta T_k = \eta*(D_k-O_k)*NET_k*F_k'(NET_k) \qquad (8)$$

$$\Delta T_j=\eta*(D_k-O_k)*F_k'(NET_k)*(\Sigma_j W_{kj}*NET_j*F_j'(NET_j)). \qquad (9)$$

윗 식들에서 k 는 출력층의 k 째 뉴런, j 는 중간층의 j 번째 뉴런을 나타낸다.

2. 활성화함수에서의 적용된 기울기의 역활

역전파 학습 신경회로망은 가중치들을 변화시켜 에러가 감소되는 방향으로 학습되는데, 가중치들을 변화시키는 요소들은 에러, 활성화함수의 기울기에 의해 결정된다.

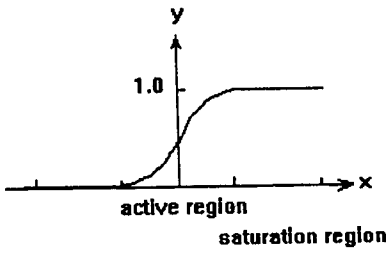

그림 1. 활성화 함수

활성화 함수는 그림 1 에서와 같이 기울기가 큰 활성화영역과 기울기가 작은 포화영역으로 구분될 수 있다.

어떤 뉴런의 출력값이 포화 영역에 속하게 되면 활성화함수에서의 작은 기울기로 인해 가중치의 변화량은 아주작다.

이러한 포화상태로 부터 벗어나기 위해서는 활성화 영역을 넓혀줄 필요가 있는데, 활성화함수에서의 기울기 파라미타 T 를 도입하고 가중치와 같이 이를 기존의 학습알고리듬에서와 마찬가지로 역전파학습시켜 포화상태를 빠져 나갈수 있는 효과를 얻고자 한다.

또한 가중치와 활성화 함수에서의 기울기 파라미타를 동시에 역전파 시킴으로서 에러에 영향을 받은 T 파라미타는 가중치에 곱해져서 오류역전파 학습중에 경사값의 변화정도에 따라서 학습률을 동적으로 변경시킴으로서 수렴속도가 개선되는 동작을 기대할 수 있다.

III. 실험결과 및 토의

제안된 학습 알고리듬과 기존의 알고리듬[1]의 비교에 가장 많이 이용되는 XOR, PARITY, ENCODER 문제에 적용하여 비교 하였다.

실험에 이용된 가중치 변경방법으로 하나의 패턴에 대해 가중치를 바꾸지 않고 모든 훈련 패턴을 수행한 가중치를 변경해나가는 배치형태로 실험하였다.

초기 가중치는 수렴조건에 중요한 역활을 하는데, 일반적으로 사용되는 $-0.3<W<0.3$을 만족하는 난수를 설정하였으며 출력단 노드에서의 T 값은 초기치 1.0 값을 설정하였고 중간층에서의 기울기 초기치는 1.2 로 주었다.

각 문제에 대한 중간층 노드의 갯수는 학습에 많은 영향을 주는데 노드의 갯수가 많으면 학습시간이 오래걸리며 적을경우에 구분을 하지 못하므로 일반적으로 사용되어온 노드 수로서 실험을 하였으며, 오차가 0.1 보다 작아지면 학습이 끝나도록 하였다.

1. XOR 문제

가중치및 기울기는 각각 값은 $-0.3<W<0.3$, T=1.0 으로 초기화 하였고 학습률과 관성항은 여러가지로 바꾸어가며 실험을 하였으나 대체로 0.8 이상의 큰값이 주어졌을때 좋은결과를 얻었다.

실험결과를 표-1 에 보인다. 여기서 기존의 알고리듬 경우는 반복횟수가 가장적게되는 파라미타 값에 대한 경우이고 제안된 알고리듬 경우는 몇 개의 파라미타들에 대한 결과를 보인다. 대체적으로 기존의 것에 비해 3-9배 정도 빠른 결과를 얻었다.

그림 2 a. 에 반복횟수에 따른 오차의 변화를 보이고 그림 2 b. 에 출력층 뉴런의 기울기 파라미타의 변화를 보인다.

X O R	학습률	관성항	반복횟수	ERROR
제안된 알고리듬	0.98 0.98 0.98 0.92	0.92 0.7 0.6 0.8	100 78 46 35	0.08 0.06 0.06 0.09
기존 알고리듬	0.8	0.8	312	0.09

<표-1> XOR 문제에 대한 수렴속도 비교

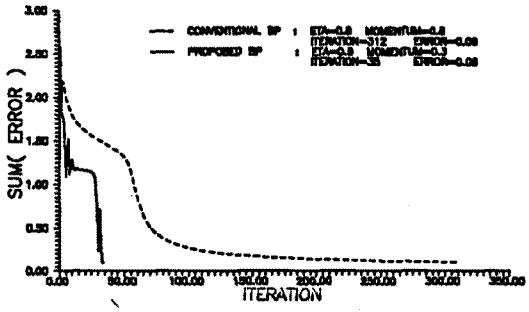

그림 2 a. XOR 문제에 대한 학습곡선

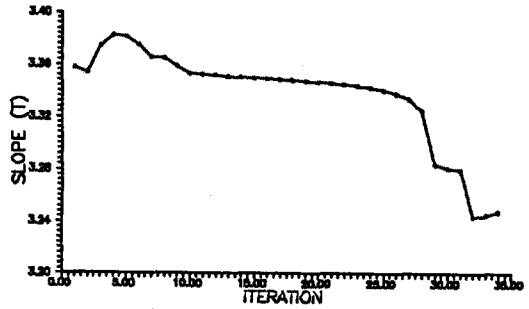

그림 2 b. 기울기 파라미타의 변화 <XOR>

2. PARITY 문제

입력층과 중간층의 노드수는 4개, 출력층은 1개의 출력을 갖는 5비트 홀수패리티 문제로, 중간층 및 출력층의 기울기 파라메타는 각각 4개, 1개의 파라미타를 사용하였다.
실험결과를 표-2 에 보인다. 기존의 방법에 비해 3 배 가량 빠른 결과를 얻었으나 파라미타 값에 따라서는 3000 번의 반복횟수에도 원하는 오차값을 얻지 못한 경우도 있었다.

그림 3 a. b. 는 각각 학습곡선과 출력층 뉴런의 기울기 파라미타의 변화를 보인다.

PARITY	학습률	관성항	반복횟수	ERROR
제안된 알고리듬	0.6 0.7 0.6	0.3 0.4 0.5	1010 888 3000	0.09 0.09 0.63
기존 알고리듬	0.8	0.8	2880	0.09

<표-2> PARITY 문제에 대한 수렴 속도 비교

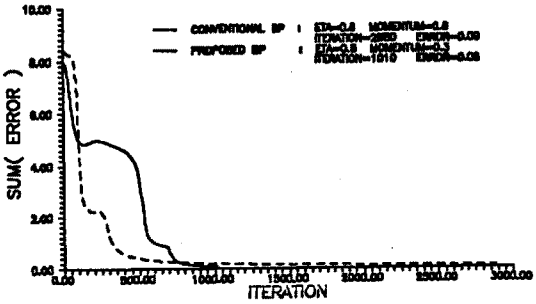

그림 3 a. PARITY 문제에 대한 학습곡선

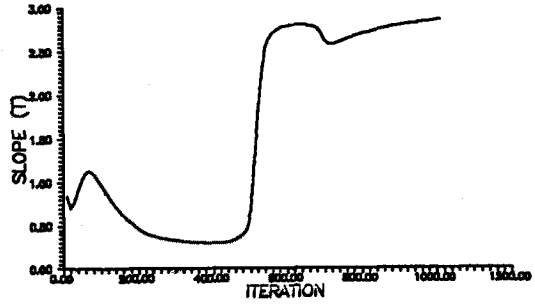

그림 3 b. 기울기 파라미타의 변화 <PARITY>

3. 5-3-5 ENCODER 문제

XOR 문제가 어떤 패턴의 분별력을 구분하는 기준으로서 사용되는 반면 ENCODER 문제는 알고리듬의 일반화 정도를 측정하는 문제로 많이 이용되고 있으므로 CONVENTIONAL BP와 제안된 BP의 수렴속도를 비교하여 보았다.

표-3 에서 알 수 있듯이 파라미타 값을 여러가지로 변화시켜도 기존의 방법에 비해 빠른 결과를 얻었다.

그림 4. 에 학습곡선을 보인다.

ENCODER	학습률	관성항	반복횟수	ERROR
제안된 알고리듬	0.7 0.6 0.7 0.8 0.9	0.4 0.3 0.3 0.4 0.4	1939 2464 2111 1707 1515	0.09 0.09 0.09 0.09 0.09
기존 알고리듬	0.8	0.8	5000	0.14

<표-3> ENCODER 문제에 대한 수렴속도 비교

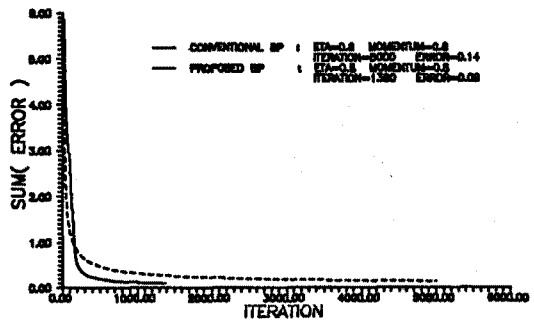

그림 4. ENCODER 문제에 대한 학습곡선

4. 조기포화 상태

초기 선택된 가중치의 영향으로 조기포화 상태에 걸려 학습이 잘 이루어지지 않는 CONVENTIONAL BP 알고리듬의 문제를 기울기 파라미타를 씀으로서 조기포화 상태를 벗어 날수 있음을 보이기 위하여 실험 1 에 사용된 XOR 문제에 초기 가중치의값을 $-3.5 < W < 3.5$ 줌으로써 조기포화 상태들어가도록 하였다.

이 경우 학습곡선의 비교를 그림 4 a. 에 보인다. 기존의 방법으로는 오차가 거의 변화지 않았지만 제안된 방법에서는 학습이 일어남을 알 수 있다. 각 뉴런의 출력값의 변화를 알기위해 입력패턴 (0,1) 이 인가되었을때 현재 반복 횟수에서의 출력값과 바로 전 반복횟수에서의 출력값의 변화량의 절대값을 그림 5 a. b. 에 보인다.

그림 5 a. 는 기존의 알고리듬 경우로써 값의 변화가 거의 없었으나 그림 5 c. 의 제안된 알고리듬 경우는 값의 변화를 알 수 있다.

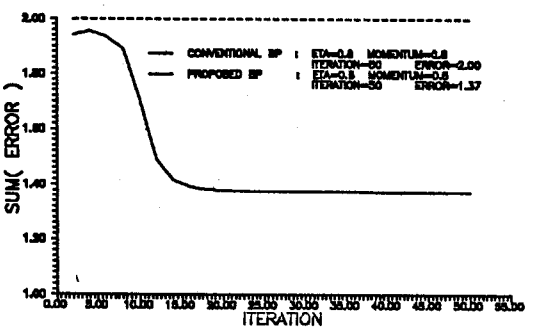

그림 5 a. 조기 포화의 경우의 학습곡선

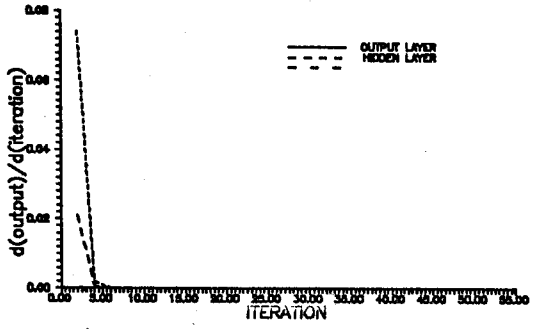

그림 5 b. 기존 방법의 뉴런 출력값의 변화

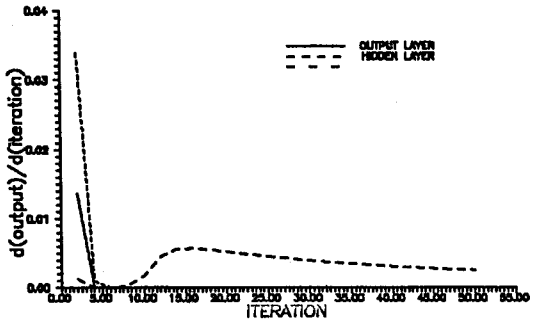

그림 5 c. 제안된 방법의 뉴런 출력값의 변화

IV. 결 론

기존의 오류역전파 학습방법이 가중치에 대해서만 학습했던데에 반해, 활성화 함수에 기울기 파라미타를 도입함으로써 경사재추적기법을 사용 기울기 및 가중치의 값들이 서로 상호작용함으로서 보다 빠른 학습속도를 얻었다.

또한 초기가중치들의 영향으로 조기포화상태에 머물러 있어 오차가 변하지 않는 상태가 계속되는 현상을 피하기 위해 동적인 기울기 파라미타를 사용함으로서 가중치값의 변화에 영향을 줌으로서 이러한 조기포화 현상을 벗어날 수 있는 역활을 해 주고 있다.

앞으로 연구해야할 과제로서 제안된 알고리듬에서의 기울기 및 가중치들의 상호작용에 대한 더 나은 이론적 분석을 통해 이러한 알고리듬의 보편성를 입증하는것이 가장 큰 문제점으로 제시되었다.

참고문헌

[1] Rumelhart, Mc Clelland and the PDP research Group, Parallel Distributed Processing, MIT Press, vol.1, p. 318, 1986.

[2] Jacobs, R.A., "Increased rates of convergence through learning rate adatation," Neural Networks, vol.1 pp. 295-308, 1988

[3] D. Nguyen and B. Widrow, "improving the learning speed of 2-layer neural networks by choosing initial value of the adaptive weights," Proc. int. Joint conf. Neural Networks, vol.III, pp 21-26, 1990

[4] Parker, D, B., "Optimal Algorithms for Adative Networks: Second Order Direct Propagation, and Second Order Hebbian Learning," In Processings of the IEEE International Conference on Neural Networks, vol. II, pp.593-600, 1987

[5] A. Rezgui et al., "The effect of the slope of the activation function on the back propagation algorithm," Proc. Int. Joint Conf. Neural Networks, vol. I, pp 707-710, jan. 1990