

論文95-32B-6-11

역전파 학습의 오차함수 개선에 의한 다층퍼셉트론의 학습성능 향상

(Improving the Error Back-Propagation Algorithm of Multi-Layer Perceptrons with a Modified Error Function)

吳相勳*, 李永稷**

(Sang Hoon Oh, and Young Jik Lee)

요 약

다층퍼셉트론의 역전파 학습 알고리즘 성능을 개선하기 위해서, 새로운 오차함수를 제안하였다. 제안한 오차함수는 출력층의 목표값이 출력값과 차이가 많이 날수록 오차신호를 크게 발생시켜, 신경망의 학습과정에서 출력노드가 부적절하게 포화되는 현상을 줄여준다. 또, 출력층의 목표값이 출력값과 가까워지면 오차신호가 작게 발생하여, 신경망이 학습패턴에 과도하게 학습되는 것을 막아준다. 이 방법은 필기체 숫자인식 문제의 학습을 통하여 그 효용성을 확인하였다.

Abstract

In this paper, we propose a modified error function to improve the EBP(Error Back-Propagation) algorithm of Multi-Layer Perceptrons. Using the modified error function, the output node of MLP generates a strong error signal in the case that the output node is far from the desired value, and generates a weak error signal in the opposite case. This accelerates the learning speed of EBP algorithm in the initial stage and prevents overspecialization for training patterns in the final stage. The effectiveness of our modification is verified through the simulation of handwritten digit recognition.

I. 서 론

패턴인식에 가장 많이 이용되는 신경회로망 모델은 다층퍼셉트론이며, 학습은 역전파 알고리즘을 사용한다. 역전파 알고리즘은 다층퍼셉트론에서 노드 사이를

연결하고 있는 가중치들을 층 별로 오차신호에 따라 바꾼다^[1]. 먼저, 다층퍼셉트론의 가중치들이 임의의 작은 값으로 초기화 된 후 학습패턴이 입력되면 전방향 계산에 의해 출력노드 값들이 계산된다. 그다음, 출력층의 가중치들은 출력노드의 오차신호에 따라 변경되는 데, 출력노드의 오차신호는 목표값과 실제값의 차이에 시그모이드 활성화 함수의 기울기가 곱해진 것이다. 아래층의 가중치들은 출력노드의 오차신호가 출력 가중치들을 통하여 역전파된 신호에 시그모이드 함수의 기울기가 곱해진 오차신호에 따라 변경된다. 그 다음 아래층 역시 같은 방식으로 역전파되는 오차신호에

* 正會員, 韓國電子通信研究所 基礎技術研究部
(ETRI, Research Department)

** 正會員, 韓國電子通信研究所 音聲言語研究室
(ETRI, Spoken Language Processing Section)

接受日字: 1994年10月5日, 수정완료일: 1995年6月1日

따라 변경된다. 이와 같이 역전과 알고리즘은 전방향 계산과 역방향 가중치 변경의 반복에 의해 다층퍼셉트론을 학습시킨다. 그렇지만, 이 방법은 학습속도가 느리다는 단점이 있다^[2]. 다층퍼셉트론의 역전과 학습 과정 중에 출력노드의 목표값과 실제값 간의 차이가 거의 줄지 않는 상태가 나타나는 데, 이 정체상태가 역전과 학습시간의 대부분을 차지하게 된다.

위에서 언급한 역전과 알고리즘 학습의 정체는 노드들의 부적절한 포화 때문이라는 보고가 최근에 있었다^[3,4,5,6,7].

시그모이드 활성화 함수는 기울기가 작은 양 측면의 포화영역과 기울기가 큰 중앙의 활성화영역으로 나눌수 있는 데, 노드의 부적절한 포화란 노드의 가중치 입력값이 목표값의 반대쪽 포화영역에 위치하는 것을 뜻한다.^[7] 이때, 비록 노드의 목표값과 실제값 간의 오차는 크지만 시그모이드 함수의 작은 기울기 때문에 오차신호가 작게 발생하여, 그 노드에 연결된 가중치는 조금만 변하게 된다. 즉, 목표값과 실제값 간의 오차가 크더라도 노드의 오차신호가 작은 경우에 역전과 학습의 정체현상이 나타난다. 이 문제를 해결하기 위하여, Rezgui와 Tepedelenlioglu는 시그모이드 활성화 함수의 기울기를 변경하는 방법을 제안하였다^[4]. 그렇지만, 이 방법은 역전과 학습의 향상은 저조한 반면 부가적인 계산이 필요하다. 또, 다층퍼셉트론의 오차함수를 cross-entropy 함수로 제안하여 출력노드의 오차신호가 목표값과 실제값의 차이가 되도록 한 시도가 있었다^[5,6]. 이 방법은 역전과 학습속도가 매우 빨라졌지만, hidden 층의 포화 때문에 역전과 학습의 정체현상이 나타났으며 학습패턴에 과하게 학습되는 경향이 있다. 한편, 학습초기에 출력층 노드의 부적절한 포화가 일어날 확률을 계산하고, 작은 가중치로 다층퍼셉트론을 초기화 하면 역전과 학습의 정체를 피할 수 있다는 보고가 있었다^[7]. 그렇지만, 출력노드의 부적절한 포화는 역전과 학습 과정 중에 항상 존재하며, 이 부적절한 포화가 없어져야만 다층퍼셉트론이 제대로 학습되었다고 할 수 있다.

이 논문에서는, 오차함수의 수정을 통하여 출력노드가 역전과 학습과정 중에 부적절하게 포화되는 가능성을 줄이는 알고리즘을 제안하였다. 2장에서는 mean-squared error(m.s.e.) 함수와 cross-entropy 오차 함수를 이용한 역전과 알고리즘을 설명하고, 새로운 오차 함수와 그에 따른 역전과 알고리즘을 설명하였다. 3장에서 이 3가지 방법을 필기체 숫자인식 문제의 학습 시뮬레이션을 통하여 비교하였으며, 4장에서 결론을 맺었다.

II. 오차함수의 수정

L층으로 이루어진 다층퍼셉트론을 고려하자. 각 L층은 N_l 개의 노드로 구성되어 있으며, 그 상태벡터는 $x^{(l)} = [x_1^{(l)}, x_2^{(l)}, \dots, x_{N_l}^{(l)}], l = 0, 1, 2, \dots, L$, 로 표시한다. 그리고, $x^{(0)}$ 와 $x^{(L)}$ 은 각각 다층퍼셉트론의 입력 상태 벡터와 출력상태 벡터를 나타낸다. 또한, $t = [t_1, t_2, \dots, t_{N_L}]$ 는 임의의 입력벡터에 대한 목표벡터를 나타낸다. 이와 같은 구조의 다층퍼셉트론에 학습패턴 x_p 가 입력되면, 전방향 계산에 의해 각 층의 노드 상태가 시그모이드 변환에 의해 $[-1, +1]$ 사이에서

$$x_j^{(l)} = \mathcal{A}(a_j^{(l)}) = \frac{2}{1 + \exp\{-a_j^{(l)}\}} - 1, j = 1, 2, \dots, N_l \quad (1)$$

와 같이 결정된다. 여기서

$$a_j^{(l)} = w_j^{(l)} + \sum_{i=1}^{N_{l-1}} w_{ji}^{(l)} x_i^{(l-1)} \quad (2)$$

이며, $w_{ji}^{(l)}$ 은 $x_i^{(l-1)}$ 과 $x_j^{(l)}$ 사이의 연결 가중치, $w_j^{(l)}$ 은 $x_j^{(l)}$ 의 bias를 나타낸다.

이렇게 $x_{k(L)}$ 이 구해지면, 다층퍼셉트론의 학습정도는 일반적으로 m.s.e. 함수

$$E_m(x_p) = \frac{1}{2} \sum_{k=1}^{N_L} (t_k - x_k^{(L)})^2 \quad (3)$$

에 의해 측정된다^[11]. $E_m(x_p)$ 를 최소화 하기 위해서, 각 가중치들은 $E_m(x_p)$ 의 가중치에 대한 편미분 값에 따라 변경시킨다.

출력층 가중치 $w_k^{(L)}$ 은

$$\Delta w_k^{(L)} = -\eta \frac{\partial E_m \partial a_k^{(L)}}{\partial a_k^{(L)} \partial w_k^{(L)}} = \eta \delta_k^{(L)} x_j^{(L-1)} \quad (4)$$

에 따라 변한다. 여기서,

$$\delta_k^{(L)} = -\frac{\partial a_k^{(L)}}{\partial} E_m = (t_k - x_k^{(L)}) \frac{(1 - x_k^{(L)})(1 + x_k^{(L)})}{2} \quad (5)$$

은 출력노드 $x_k^{(L)}$ 에서 발생하는 오차신호이며, η 는 학습률이다.

아래층 가중치 $w_{ji}^{(l)}$ 는

$$\Delta w_{ji}^{(l)} = -\eta \frac{\partial E_m \partial a_k^{(L)}}{\partial a_k^{(L)} \partial w_{ji}^{(l)}} = \eta \delta_k^{(L)} x_j^{(l-1)} \quad (6)$$

에 의해 변하며, 이때

$$\delta_j^{(l)} = -\frac{\partial E_m}{\partial a_j^{(l)}} = \frac{(1 - x_j^{(l)})(1 + x_j^{(l)})}{2} \sum_{k=1}^{N_{l+1}} w_{kj}^{(l+1)} \delta_k^{(l+1)} \quad (7)$$

은 상위층의 가중치를 통하여 역전파된 오차신호이다. 이 과정을 모든 학습패턴에 대하여 한번씩 수행한 것을 sweep 단위로 표시한다. 전체패턴에 대한 오차는

$$E_m = \frac{1}{P} \sum_{p=1}^P E_m(x_p) \quad (8)$$

으로 정의된다.

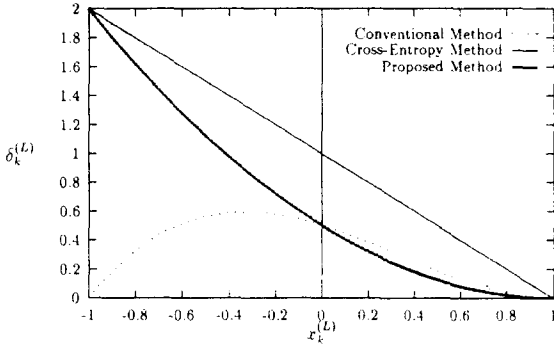


그림 1. 학습방법에 따른 출력노드의 오차신호 비교 ($t_k = 1$)

Fig. 1. The error signals in each method when $t_k = 1$

위에서 설명한 역전파 알고리즘에서, 식 (5)의 오차신호 $\delta_k^{(L)}$ 은 목표값과 실제값의 차이에 시그모이드 활성화 함수의 기울기가 곱해진 형태이다. 만약, $x_k^{(L)}$ 이 -1 혹은 +1에 가까운 값이면, 기울기에 대한 항 때문에 $\delta_k^{(L)}$ 은 아주 작은 값이 된다. 즉, $t_k = 1$ 이고 $x_k^{(L)} \approx -1$ 인 경우 혹은 그 반대인 경우에, $x_k^{(L)}$ 은 연결된 가중치들을 조정하기에 충분히 강한 오차신호를 발생시키지 못한다(그림 1). 이와 같은 출력노드의 부적절한 포화가 역전파 학습에서 E_m 의 최소화를 지연시키게 한다.

학습시간의 지연을 해결하기 위해서는 출력노드의 부적절한 포화 시 오차신호가 작아져서는 안된다. 이를 위해서, cross-entropy 오차함수가

$$E_c(x_p) = - \sum_{k=1}^{N_k} \{ (1-t_k) \ln(1-x_k^{(L)}) + t_k \ln(x_k^{(L)}) \} \quad (9)$$

와 같이 제안되었다^{15,61}. 이 오차함수를 이용하면 식 (5)의 오차신호는

$$\delta_k^{(L)} = (t_k - x_k^{(L)}) \quad (10)$$

로 바뀐다. 그 이외의 다른 식들은 E_m 을 이용한 역전

파 알고리즘과 동일하다. 이 방법을 이용하면, $x_k^{(L)}$ 이 부적절하게 포화된 경우 $\delta_k^{(L)}$ 이 커져서 부적절한 포화를 빠져나오게 된다.

한편, 학습패턴은 여러개이므로 역전파 학습에서한 학습패턴에 의한 가중치의 변화방향이 학습패턴 전체에 대한 오차를 줄이는 방향과 일치하거나 경쟁할 것이다¹²¹. 예를 들어, 역전파 학습과정 중의 경쟁에 의해 어떤 학습패턴의 입력시 부적절하게 포화되는 상태로 밀려나는 출력노드가 있을 수 있다. 따라서, 부적절하게 포화되는 출력노드는 부적절한 포화상태를 벗어나기 위해서 cross-entropy 방법처럼 강한 오차신호를 발생시켜야 한다. 반면에 목표값과 같은 방향으로 포화된 출력노드는 약한 오차신호를 발생시켜야 한다. 이렇게 하면, 한 패턴에 대한 가중치들의 역전파 학습이 학습패턴 전체의 역전파 학습에 의해 결정된 가중치들을 흐트러뜨리는 정도를 최소화 시킬 것이다. 위와 같이 출력노드의 상태에 따른 오차신호의 강,약은 역전파 학습과정의 경쟁을 줄일 것이다. 또한, 학습패턴에 대한 과도한 학습에 따른 일반화 성능 저하를 방지하기 위해서도 적절하게 포화된 출력노드는 약한 오차신호를 발생시켜야 한다¹⁸¹.

이러한 관점에서, 역전파 학습을 위한 오차함수를

$$E(x_p) = - \sum_{k=1}^{N_k} t_k \left[-x_k^{(L)} + \frac{1+t_k^2}{2} \ln \frac{1+x_k^{(L)}}{1-x_k^{(L)}} + t_k \ln(1-x_k^{(L)})(1+x_k^{(L)}) \right] \quad (11)$$

로 수정하였다. 이 수정된 오차함수를 이용하면 출력노드의 오차신호는

$$\delta_k^{(L)} = \frac{t_k(t_k - x_k^{(L)})^2}{2} \quad (12)$$

이 된다. 역전파 학습을 위한 다른 수식은 E_m 을 이용한 역전파 알고리즘과 동일하다. 따라서, 수정된 오차함수를 이용한 역전파 알고리즘은 출력노드가 목표값에 가까우나에 따라 적절한 오차신호를 발생시켜서 경쟁에 의한 출력노드의 부적절한 포화를 줄여준다. 그림 1은 인 경우에 $t_k = 1$ 인 경우에 $x_k^{(L)}$ 에 따른 오차신호를 비교한 것이다. 식 (12)는 극한치가 ± 1 인 시그모이드 변환을 사용한 다층퍼셉트론의 출력노드 오차신호를 제안한 것이다. 극한치가 0,1인 시그모이드 변환을 사용한 다층퍼셉트론을 학습 시 같은 효과를 얻기 위해서는 출력노드의 오차신호가

$$\delta_k^{(L)} = (2t_k - 1)(t_k - x_k^{(L)})^2 \quad (13)$$

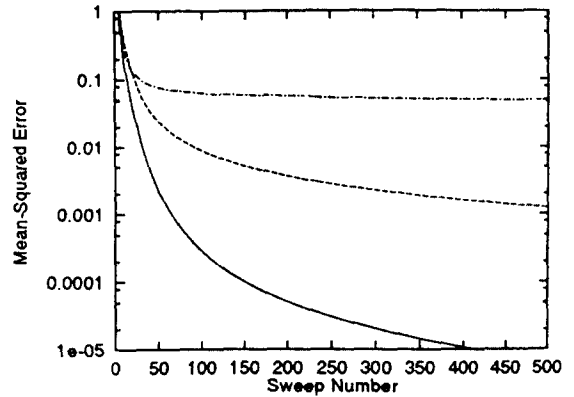
이 되도록 하면된다.

M.s.e.와 cross-entropy를 이용한 역전파 학습에서, 오차함수의 기대치는 $x_k^{(l)} = E[t_k/x_p]$ 일 때 최소가 된다¹⁹⁾. 물론, 여기서 제한한 오차함수도 같은 성질을 지닌다.

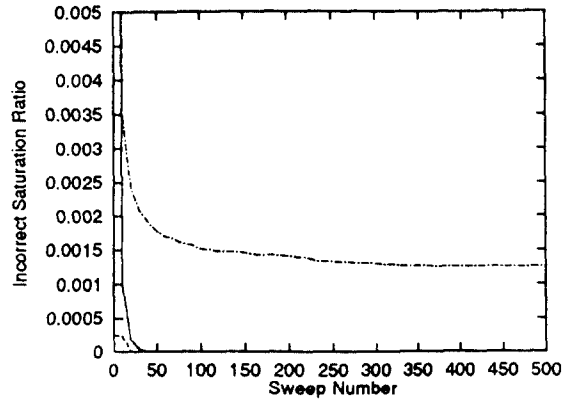
III. 시뮬레이션

2장에서 설명한 3가지 오차함수를 이용한 역전파 알고리즘을 비교하기 위해서, 필기체 숫자인식 문제를 다층퍼셉트론에 학습시켰다. 우편봉투에 적힌 숫자를 모아둔 CEDAR 데이터베이스¹⁰⁾로부터 2,308개의 필기체 숫자영상을 임의로 선택하여 정규화 후 역전파 학습에 사용하였다. 한 숫자영상은 12×12 pixel로 이루어져 있으며, 각 pixel은 0에서 15까지의 정수값을 가진다. 다층퍼셉트론은 입력 144, 중간층 30, 출력층 10개의 노드로 구성되어 있으며, 목표값이 출력되도록 학습시켰다. 일반화를 측정하기 위한 test 패턴은 2,213개를 사용하였다.

이 3가지 학습방법을 비교할 때, 각각의 방법에서 $x_k^{(l)}$ 이 같더라도 $\delta_k^{(l)}$ 의 기대치가 다르므로 같은 학습률을 사용하는 것은 공평하지 못하다¹⁶⁾. 또, 매 학습 sweep에서 $x_k^{(l)}$ 의 분포를 알 수 없으므로, $x_k^{(l)}$ 이 -1과 +1 사이에서 균일분포라는 가정 하에 $\eta\delta_k^{(l)}$ 의 기대치가 같아지도록 3가지 학습방법에서의 학습률 η 를 정하였다. 따라서, E_c 를 이용한 역전파 알고리즘의 학습률은 0.02, E_i 인 경우는 0.03, E_m 인 경우는 0.06을 학습률로 정하고서 3가지 학습방법을 비교하였다. 그리고, 각 학습방법에서 사용한 오차함수가 다르므로, 학습정도를 나타내는 오차의 기준은 일반적인 m.s.e.를 사용하였다.



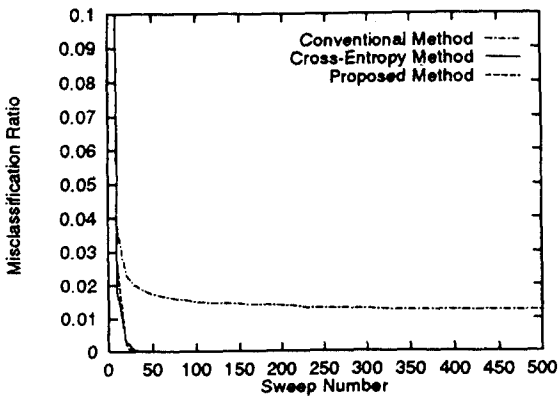
(b)



(c)

그림 2. 학습방법에 따른 학습패턴에 대한 (a) 오인식률, (b) m.s.e., 및 (c) 출력노드의 부적절한 포화비율 비교(초기가중치 범위: $[-1 \times 10^{-4}, 1 \times 10^{-4}]$).

Fig. 2. (a) the misclassification ratio, (b) mean-squared error, and (c) incorrect saturation ratio of output nodes for training patterns using the three learning methods with initial weights in $[-1 \times 10^{-4}, 1 \times 10^{-4}]$.



(a)

다층퍼셉트론은 초기가중치들이 -1×10^{-4} 에서 1×10^{-4} 의 범위에서 균일분포를 가지도록 초기화 하였다. 그리고, 초기치가 다르게 9번 시뮬레이션을 수행하여 그 결과의 평균을 그림으로 그렸다. 그림 2는 학습패턴에 대한 인식률, m.s.e., 및 출력노드의 부적절한 포화비율을 보여준다. 점선은 m.s.e.를 이용한 방법, 실선은 cross-entropy를 이용한 방법, 그리고, 쇠선은 이 논문에서 제안한 방법이다. 먼저, 그림 2(a)를

보던 제안한 방법과 cross-entropy 방법은 학습패턴을 100% 인식하였지만, m.s.e. 방법은 학습패턴 모두를 제대로 인식하지 못하였다. 학습과정의 오차를 나타내는 그림 2(b)에서도 m.s.e. 방법은 오차가 다른 방법에 비해 높은 값을 유지한다. 그 이유는 그림 2(c)에서 보는 바와 같이 m.s.e. 방법은 출력노드의 부적절한 포화비율이 '0'이 되지 못하였기 때문이다. 그리고, 이 논문에서 제안한 방법은 학습초기의 경쟁에 의한 부적절한 포화현상이 cross-entropy를 이용한 방법보다 훨씬 작다.

한편, cross-entropy 방법은 그림 2(b)에서와 같이 학습패턴에 대한 m.s.e.가 아주 급격히 줄어들지만, 이것은 오차가 작은 경우에도 그림 1에서와 같이 출력노드의 강한 오차신호 때문에 학습패턴에 대하여 필요 이상으로 강하게 학습되는 것을 뜻한다.

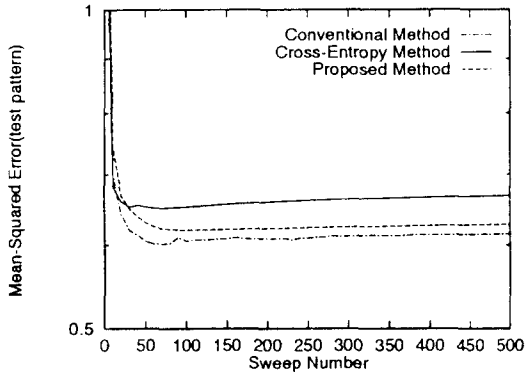
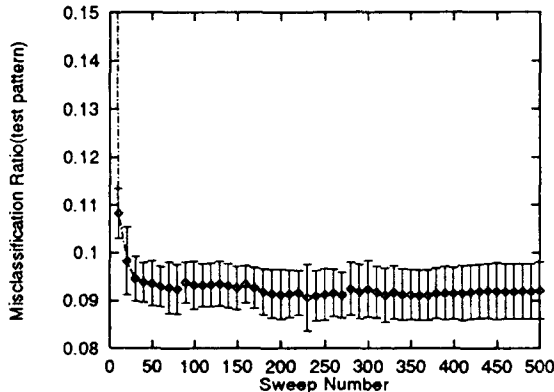
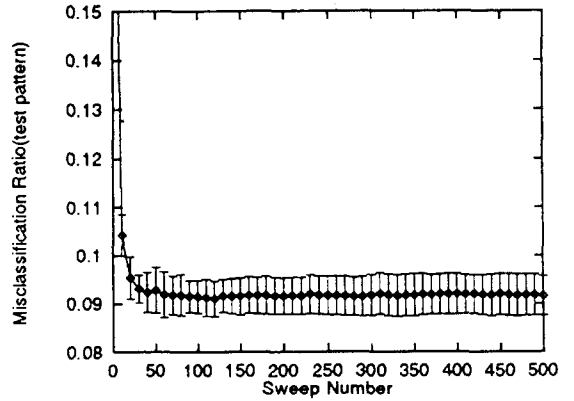


그림 3. 학습방법에 따른 test 패턴에 대한 m.s.e.의 비교(초기가중치 범위: $[-1 \times 10^{-4}, 1 \times 10^{-4}]$).

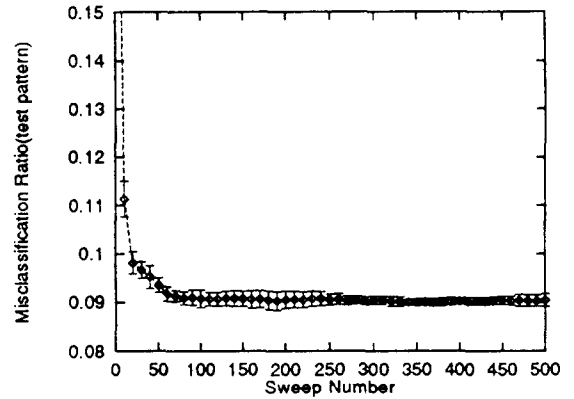
Fig. 3. The m.s.e. for test patterns using the three learning methods with initial weights in $[-1 \times 10^{-4}, 1 \times 10^{-4}]$.



(a) Conventional Method



(b) Cross-Entropy Method

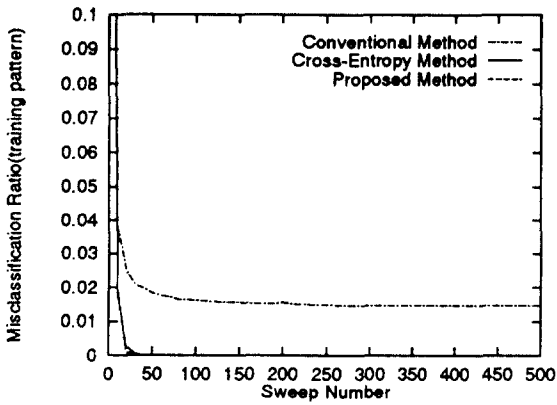


(c) Proposed Method

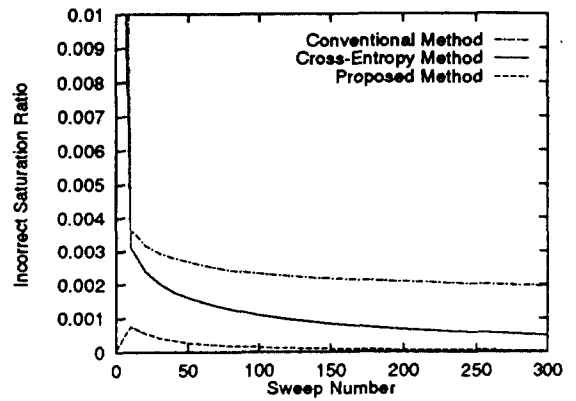
그림 4. 학습방법에 따른 test 패턴에 대한 오인식률의 비교(초기가중치 범위: $[-1 \times 10^{-4}, 1 \times 10^{-4}]$).

Fig. 4. The mean and standard deviation of misclassification ratio for test patterns (initial weight range: $[-1 \times 10^{-4}, 1 \times 10^{-4}]$).

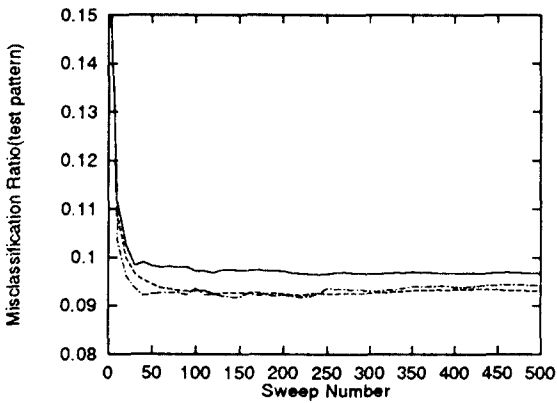
그 결과 그림 3에서 보는 바와 같이 test 패턴에 대한 m.s.e.가 제일 나쁘다. 이에 반해, 제안한 학습방법은 출력노드 값이 목표값에 가까워지면 오차신호가 매우 약해져서 학습패턴에 대한 과도한 학습을 감소시킨다. 그림 4는 test 패턴에 대한 인식률의 평균값과 표준편차를 그린 것이다. 이 값들 역시 각 학습방법에 대해 9번 시뮬레이션한 데이터에서 구한 것이다. 인식률의 평균값은 제안한 방법이 다른 방법에 비해 조금 좋지만, 표준편차는 제안한 방법이 다른방법에 비해 훨씬 작다.



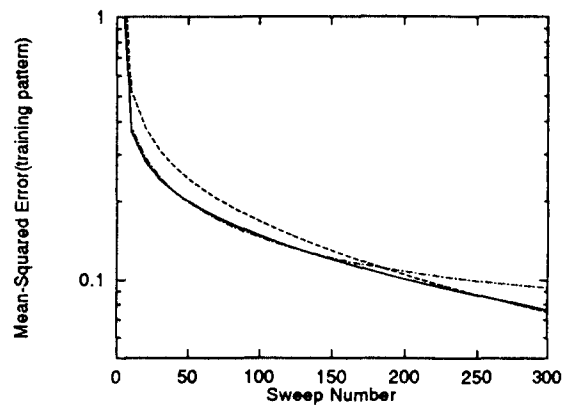
(a)



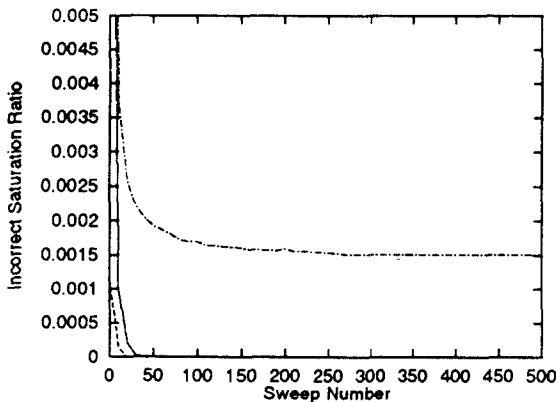
(a)



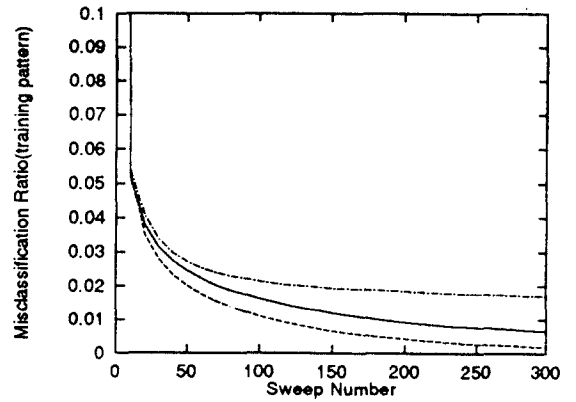
(b)



(b)



(c)



(c)

그림 5. 학습방법에 따른 오인식률과 부적절한 포화 비율 비교(초기가중치 범위: $[-1 \times 10^{-4}, 1 \times 10^{-4}]$)

Fig. 5. The misclassification ratio and incorrect saturation ratio using the three learning methods with initial weights in $[-1 \times 10^{-4}, 1 \times 10^{-4}]$.

그림 6. 학습방법에 따른 부적절한 포화 비율과 학습패턴에 대한 오차 및 오인식률 비교(18,468 학습패턴)

Fig. 6. The incorrect saturation ratio, m.s.e., and misclassification ratio for training patterns using the three learning methods with 18,468 training patterns.

즉, 제안한 방법이 초기 가중치에 대한 학습의 민감도가 제일 작은 데, 그 이유는 학습과정의 경쟁을 최소화 시켰기 때문이다. 그림 5는 초기가중치 범위를 $[-1 \times 10^{-1}, 1 \times 10^{-1}]$ 으로 하고서 시물레이션 한 결과이다. 초기가중치가 $[-1 \times 10^{-4}, 1 \times 10^{-4}]$ 인 경우와 나타나는 양상은 비슷하다.

필기체 숫자 인식같은 실제적인 문제의 학습에서 2,308개의 학습패턴은 너무 적은 것이며, 따라서 학습이 쉬웠을 것이다. 복잡한 문제에서 제안한 방법의 효용성을 확인하기 위해서 CEDAR 데이터 베이스의 총 18,468개 학습패턴을 모두 학습시켰다.

그림 6, 7은 초기가중치 범위 $[-1 \times 10^{-4}, 1 \times 10^{-4}]$ 에서 학습률 0.002(cross-entropy), 0.003(proposed), 0.006 (conventional)을 사용하여 9번 학습 후 그 평균을 그린 것이다. 그림 6(a)에서 보는 바와 같이 문제가 복잡하므로 어느 학습방법이든 부적절한 포화비율이 '0'이 되지 못하였지만, 제안한 방법이 제일 작은 값을 나타냈다. 한편, 학습패턴에 대한 m.s.e.는 제안한 방법이 cross-entropy 방법과 비슷하나, 학습패턴에 대한 인식률은 제안한 방법이 더 좋다. 그리고, 그림 7에서와 같이 test 패턴에 대한 인식률 역시 제안한 방법이 제일 좋다. Cross-entropy 방법은 학습초기에 test 패턴에 대한 인식률이 향상되었으나 학습패턴에 대한 과도한 학습 때문에 test 패턴의 인식률이 제일 낮게 된다. 제안한 방법 역시 과도한 학습에 의한 test 패턴의 인식률 저하가 나타난다. 즉, 출력노드의 값이 목표값과 같아지도록 과도하게 학습되면 일반화 성능은 저하된다. 이것은 제안한 방법이 일반화 성능의 저하를 방지하기 위해서 적절하게 포화된 출력노드는 약한 오차신호를 발생시키는 것의 타당성을 입증한다.

학습방법의 보다 더 공정한 비교를 위하여 학습률이 매 sweep 마다 $t_k - x_k^{(L)}$ 과 $\delta_k^{(L)}$ 의 기대치 비율에 따라

$$\eta(n) = \eta_0 \frac{\sqrt{E[(t_k(n) - x_k^{(L)}(n))^2]}}{E[\delta_k^{(L)}(n)]} \quad (14)$$

과 같이 변화도록 하였다.

여기서, $E[(t_k(n) - x_k^{(L)}(n))^2]$ 과 $E[\delta_k^{(L)}(n)]$ 은 n sweep에서 모든 학습패턴과 출력노드를 고려한 기대치이다.

그러면, 가중치의 변화량을 결정하는 $\eta(n)\delta_k^{(L)}(n)$ 의 세기는 $\delta_k^{(L)}$ 의 종류에 상관없이

$$E[\eta^2(n)\delta_k^{(L)}(n)] = \eta_0^2 E[(t_k(n) - x_k^{(L)}(n))^2] \quad (15)$$

이 된다. 즉, 이 가변학습률을 이용하면 $\delta_k^{(L)}$ 의 모양에 따른 특성을 살리면서 $\eta(n)\delta_k^{(L)}(n)$ 의 세기는 같은 효과를 얻을 수 있다.

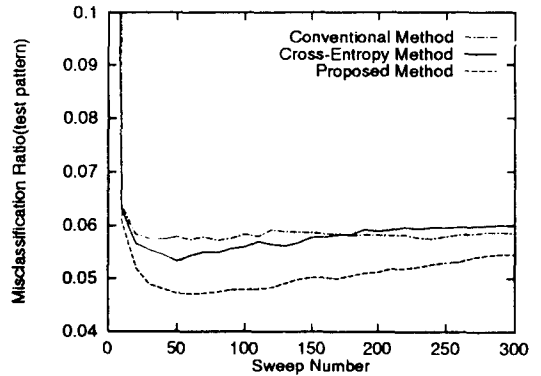


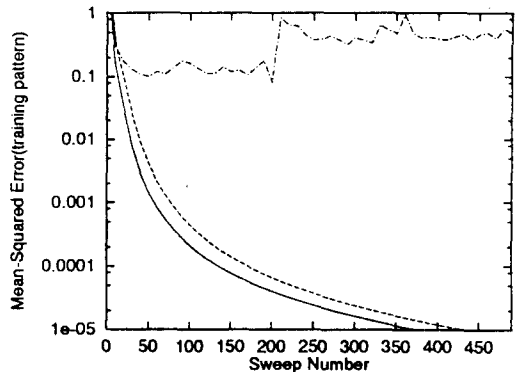
그림 7. 학습방법에 따른 test 패턴의 인식률 비교 (18,468 학습패턴)

Fig. 7. The misclassification ratio for test patterns using the three learning methods with 18,468 training patterns.

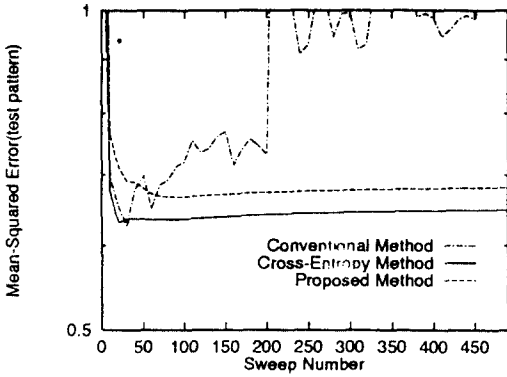
시물레이션에서 $\eta(n)$ 을 결정 시, n sweep 초기에 $E[\delta_k^{(L)}(n)]$ 와 $E[(t_k(n) - x_k^{(L)}(n))^2]$ 를 알 수 없으므로 $(n-1)$ 번째 sweep에서 얻은 기대치들을 이용하였다.

그림 8은 $\eta_0=0.02$ 인 가변학습률을 이용한 역전과 학습의 결과를 비교한 것이다.

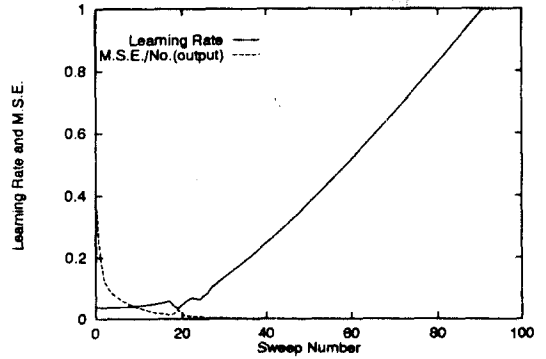
여기서, 다층퍼셉트론의 초기가중치는 $[-1 \times 10^{-4}, 1 \times 10^{-4}]$ 에서 균일분포를 가지도록하였으며, 초기치가 다르게 4번 시물레이션한 결과의 평균치를 그림 8으로 그렸다.



(a)



(b)



(a)

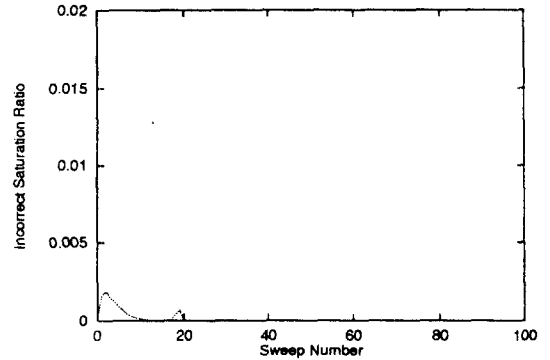
그림 8. 가변 학습률을 이용한 학습방법에 따른 m.s.e.의 비교(초기가중치 범위: $[-1 \times 10^{-4}, 1 \times 10^{-4}]$)

Fig. 8. The m.s.e. for training patterns and test patterns of the three learning methods with adaptive learning rates.

먼저, 학습패턴에 대한 결과를 보면, cross-entropy 방법은 식 (14)에서 학습률이 η_0 으로 고정되어 있으므로 가변학습률에 의해 영향을 받지 않는다. 이 논문에서 제안한 학습방법의 학습패턴에 대한 오차 수렴속도는 가변학습률에 의해 빨라졌지만(그림 8(a)), 오차가 작은 경우 출력노드의 오차신호는 가변 학습률에 의해 증가되어서 test 패턴에 대한 m.s.e.는 고정학습률인 경우보다 매우 나빠졌다(그림 8(b)). 즉, 학습패턴에 대한 과도한 학습이 일반화 성능을 저하시킨다는 앞의 주장을 뒷받침 한다. m.s.e.를 이용한 학습방법에서는 학습패턴 및 test 패턴에 대한 m.s.e.가 sweep에 따라 매우 심하게 변동하였다. 결국, 가변 학습률이 학습에 나쁜 영향을 미친다. 그 이유는 여기서 제안한 학습방법의 한 시뮬레이션 결과를 그린 그림 9를 보면 알 수 있다. 그림 9는 가변학습률을 적용한 학습과정에서 매 sweep에서 측정된 $\eta(n)$, 학습패턴에 대한 출력노드의 평균 m.s.e., 그리고 출력노드의 부적절한 포화비율을 보여준다. $\eta(n)$ 은 학습패턴에 대한 m.s.e.가 줄어들수록 식(14)에 따라 증가한다.

이때, 학습률이 커지면 한 번에 바뀌는 가중치의 양이 크므로 경쟁에 의해 출력노드가 부적절하게 포화될 가능성이 증가한다. 이렇게 되면, m.s.e.가 증가하여 η_n 은 감소한다.

이 과정을 반복하다가 출력노드의 부적절한 포화가 발생하지 않으면 m.s.e.는 계속 감소하며 $\eta(n)$ 은 증가하게 된다.



(b)

그림 9. 제안한 학습방법의 가변학습률 적용 시 학습률, m.s.e., 출력노드의 부적절한 포화비율(초기가중치 범위: $[-1 \times 10^{-4}, 1 \times 10^{-4}]$)

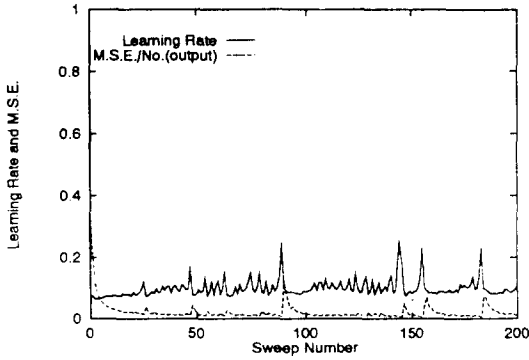
Fig. 9. A simulation result for training patterns using the proposed method with adaptive learning rate.

m.s.e.를 이용한 학습 방법은 출력노드의 부적절한 포화상태를 벗어나기가 힘들므로 그림 10에서와 같이 m.s.e.가 계속 증가, 감소를 반복하게 된다. 즉, 학습률이 커지면 경쟁에 의해 부적절한 포화현상이 일어날 가능성이 커지므로 오히려 학습에 나쁜 영향을 미친다. 이제까지 출력노드의 오차신호가 출력값과 목표값 차의 2차 함수인 경우의 학습성능을 시뮬레이션으로 확인하였다. 만약 출력노드의 오차신호가

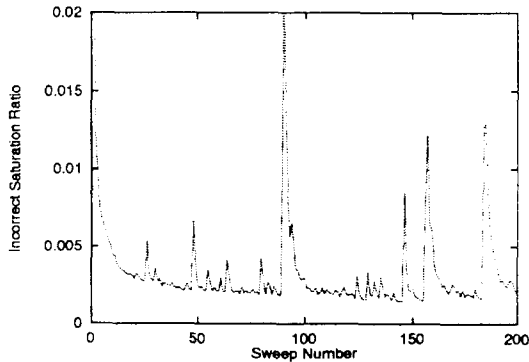
$$\delta_k^{(L)} = \frac{t_k(t_k - x_k^{(L)})^4}{8} \quad (16)$$

와 같이 출력값과 목표값 차의 4차 함수로 주어지면, 출력노드의 부적절한 포화 시 오차신호의 강해지는 정

도는 2차 오차함수의 경우보다 클 것이며 적절한 포화 노드에서 발생하는 오차신호는 2차 오차함수보다 작을 것이다. 즉, 4차 오차함수의 부적절한 포화노드 감소 및 학습패턴에 대한 과도한 학습방지 능력이 2차 오차함수 보다 뛰어날 것이다.



(a)



(b)

그림 10. m.s.e.를 이용한 EBP 알고리즘의 가변 학습률 적용 시 학습률, m.s.e., 출력노드의 부적절한 포화비율(초기가중치 범위: $[-1 \times 10^{-4}, 1 \times 10^{-4}]$)

Fig. 10. A simulation result for training patterns using the conventional EBP algorithm with adaptive learning rate.

IV. 결 론

이 논문에서는, 다층퍼셉트론의 역전과 알고리즘에 의한 학습 시 나타나는 출력노드의 부적절한 포화에 의한 학습의 정체현상을 줄이는 학습방법을 제안하였다. 이 방법은 출력노드의 값이 목표값과 멀어지면 강한 오차신호를 발생시켜 부적절한 포화에서 벗어나게

한다. 출력노드의 값이 목표값과 가까우면 약한 오차신호를 발생시켜, 경쟁에 의한 출력노드의 부적절한 포화가 최소화 되도록 하며 학습패턴에 대한 과도한 학습은 방지한다.

제안한 학습방법은 CEDAR 데이터베이스의 필기체 숫자를 인식하는 문제의 시뮬레이션을 통하여, 일반적으로 사용하는 m.s.e. 역전과 알고리즘 및 cross-entropy 방법과 비교하였다. 시뮬레이션 결과, m.s.e. 역전과 알고리즘은 학습패턴에 대한 인식을 및 오차가 출력노드의 부적절한 포화 때문에 줄지않았다. Cross-entropy 방법은 학습패턴에 대하여 아주 잘 학습하였지만, 학습패턴에 대하여 너무 과도하게 학습되어 일반화 성능은 오히려 좋지 못하다. 제안한 방법은 학습패턴에 대하여 출력노드의 부적절한 포화를 최소화 시키면서 잘 학습하였으며, 일반화 성능도 다른 방법보다 뛰어났다. 특히, 복잡한 문제에서 제안한 방법이 효율적이었다.

감사의 글

이 논문은 정보통신부와 한국통신의 출연금에 의하여 수행한 연구의 결과입니다. 이 연구를 수행하는 데 많은 도움을 주신 이일항 박사께 감사드립니다.

참 고 문 헌

[1] D.E. Rumelhart and J.L. McClelland, *Parallel Distributed Processing*. MIT Press, Cambridge, MA, 1986.

[2] R. K. Cheung, I. Lustig, and A.L. Kornhauser, "Relative effectiveness of training set patterns for back propagation," *Proc., IJCNN Jan. 15-19, 1990 Washington, DC, USA*, vol. I, pp. 673-678.

[3] J.R. Chen and P. Mars, "Stepsize variation methods for accelerating the backpropagation algorithm," *Proc. IJCNN Jan. 15-19, 1990, Washington, DC, USA*, vol. I, pp. 601-604.

[4] A. Rezgui and N. Tepedelenlioglu, "The effect of the slope of the activation function on the back propagation algorithm," *Proc. IJCNN Jan. 15-19, Washington, DC, USA, vol. I, pp.*

- 707-710.
- [5] A. Krzyzak, W. Dai, and C. Y. Suen, "Classification of large set of handwritten characters using modified back propagation model," *Proc. IJCNN June 17-21, 1990, San Diego, USA*, vol. III, pp. 225-232
- [6] A. van Ooyen and B. Nienhuis, "Improving the convergence of the back-propagation algorithm," *Neural Networks*, vol. 5, pp. 465-471, 1992.
- [7] Y. Lee, S.-H. Oh, and M. W. Kim, "An analysis of premature saturation in back-propagation learning," *Neural Networks*, vol. 6, pp. 719-728, 1993.
- [8] J.B. Hampshire II and A. H. Waibel, "A novel objective function for improved phoneme recognition using time-delay neural networks," *IEEE Trans. Neural Networks*, vol. 1, pp. 216-228, June 1990.
- [9] M.D. Richard and R.P. Lippmann, "Neural network classifier estimate Bayesian a posteriori probabilities," *Neural Computation*, vol. 3, pp. 461-483, 1991.
- [10] J.J. Hull, "A database for handwritten text recognition research," *IEEE Trans. Pat. Ana. Mach. Int.*, accepted to appear as a correspondence.

----- 저 자 소 개 -----

吳 相 勳(正會員) 第 28卷 B編 第 4號 參照.

현재 한국전자통신연구소 기초기
술연구부 선임연구원

李 永 稷(正會員) 第 28卷 B編 第 3號 參照.

현재 한국전자통신연구소 통신시스템
연구단 음성언어 연구실장