論文98-35C-6-8

오류 역전파 알고리즘의 n차 크로스-엔트로피 오차신호에 대한 민감성 제거를 위한 가변 학습률 및 제한된 오차신호

(Adaptive Learning Rate and Limited Error Signal to Reduce the Sensitivity of Error Back-Propagation Algorithm on the n-th Order Cross-Entropy Error)

吳 相 勳 *, 李 壽 永 **

(Sang-Hoon Oh and Soo-Young Lee)

요 약

다층퍼셉트론의 학습에서 나타나는 출력노드의 부적절한 포화를 해결하기 위해서 n차 크로스-엔트로피 오차함수가 제안되었으나, 이 오차함수를 이용한 학습성능은 오차함수의 차수에 민감하여 적절한 차수를 결정해야 하는 문제점이 있다. 이 논문에서는, 학습의 진행에 따라 학습률을 가변시키는 새로운 방법을 제시하여 다층퍼셉트론의 학습성능이 n차 크로스-엔트로피 오차함수의 차수에 덜 민감하도록 한다. 또한, 가변학습률이 매우 커지는 경우에 학습이 불안정해지는 것을 방지하기 위해서 오차신호의 크기를 제한하는 방법을 제시한다. 마지막으로, 필기체 숫자 인식 문제와 갑상선 진단 문제의 시뮬레이션으로 제안한 방법의 효용성을 검증한다.

Abstract

Although the nCE(n-th order cross-entropy) error function resolves the incorrect saturation problem of conventional EBP(error back-propagation) algorithm, the performance of MLP's (multilayer perceptrons) trained using the nCE function depends heavily on the order of the nCE function. In this paper, we propose an adaptive learning rate to make the performance of MLP's insensitive to the order of the nCE error. Additionally, we propose a limited error signal of output node to prevent unstable learning due to the adaptive learning rate. The effectiveness of the proposed method is demonstrated in simulations of handwritten digit recognition and thyroid diagnosis tasks.

Ⅰ. 서 론

신경회로망 모델 중 실제적인 응용문제의 학습에 많이 이용되는 다층퍼셉트론은 각 층을 이루는 노드들과 그 노드들을 연결하는 가중치들로 구성되어 있다. 일반적으로 이 다층퍼셉트론은 MSE(mean-squared error)를 줄이는 형태로 EBP(error back-propagation) 알고리즘에 따라 학습된다[1]. 이 EBP 알고리즘에서 출력층 노드의 오차신호는 그 노드의 목표값과 실제 값의 차이에 시그모이드 함수의 기울기가 곱해진 형태이다. 그 결과, 비록 출력노드의 실제값이 목표값과 차이가 크더라도 시그모이드 활성화 함수의 기울기가 작은 포화영역에 위치한 경우-즉, 그 출력노드가 부적절하게 포화된 경우-에는 학습 지연 현상이

* 正會員, 韓國電子通信研究院, 移動通信技術研究團
(Mobile Communications System Division, ETRI)
** 正會員, 韓國科學技術員, 電氣電子工學科
(Electrical Engineering, KAIST)
接受日字:1997年7月3日, 수정완료일:1998年5月8日

나타나거나 혹은 출력노드의 부적절한 포화를 유발시키는 학습패턴이 아예 학습되지 않는다[2, 3, 4, 5].

이러한 문제의 해결책으로 n차 크로스-엔트로피(nCE: n-th order cross-entropy) 오차함수가 제안되었다[6]. 이 오차함수를 이용하여 EBP 알고리즘에 따라 다층퍼셉트론을 학습시킬 경우, 출력층 노드의 오차신호는 시그모이드 함수의 기울기 항이 없이 목표값과 실제 출력값 간의 차이의 n차 함수로 나타난다. 이 nCE 방법은 MSE를 이용한 경우에 나타나는 출력층 노드의 부적절한 포화 문제와 크로스-엔트로피 오차함수를 이용한 경우 학습패턴에 대한 과도한 학습 때문에 나타나는 일반화 성능의 저하 문제를 해결하였다[6]. 그렇지만, nCE 오차함수를 미분함으로써 얻어지는 출력층 노드의 오차신호는 n이 증가할수록 출력노드의 부적절한 포화와 학습패턴에 대한 과도한 학습을 방지하는 성질을 강화시키는 반면에 학습속도를 지연시키는 단점이 있다. 결국, 다층퍼셉트론의 학습성능이 nCE 오차함수의 차수 n에 민감하게 변하므로 좋은 학습결과를 빠른 학습속도로 얻기 위해서는 nCE 오차함수의 적절한 차수를 구해야 하는 문제가 있다.

이 논문에서는 앞에서 지적한 nCE 함수의 차수 선정 문제에 대한 해결책으로 학습률을 학습의 진행에 따라 가변시키는 새로운 방법을 제시한다. 즉, 학습률이 목표값과 실제값 차이의 기대치와 출력층 노드 오차신호의 기대치의 비율에 따라 변하도록 하여, nCE 오차함수의 차수 n에 따라 오차신호의 모양이 변하더라도 가중치 변경량을 결정하는 학습률과 오차신호의 곱은 같은 기대치를 지니도록 한다. 그 결과 nCE 오차함수의 차수가 증가하더라도 오차신호의 부적절한 포화 및 과도한 학습방지 효과는 유지시키면서 학습속도가 느려지는 단점을 개선시켜, 다층퍼셉트론의 학습성능이 nCE 오차함수의 차수에 민감하지 않도록 한다. 한편, 여기서 제시하는 가변학습률은 학습의 진행시 매우 큰 값을 지닐 수 있으며, 이 경우 가중치의 급격한 변화로 인해 학습이 매우 불안정한 상태가 나타난다. 이를 방지하기 위해서 출력노드 오차신호의 크기를 제한하는 방법을 제시한다.

2장에서는 MSE와 nCE 오차함수를 간략히 알아보고 nCE 오차함수의 차수에 따라 다층퍼셉트론의 학습성능이 민감하게 변하는 이유를 설명하며, 3장에서 이 문제에 대한 해결책으로 새로운 가변학습률과 제한된 오차신호를 제안한다. 4장에서 필기체 숫자 인식 문제와 갑상선 진단 문제의 시뮬레이션으로 제안한 방법의 효용성을 검증하고, 마지막으로 5장에서 결론을 맺는다.

Ⅱ. MSE와 nCE 오차함수

L 층으로 이루어진 다층퍼셉트론을 고려하자. 각 l층은 N_l개의 노드로 구성되어 있으며, 그 상태벡터는 $x^{(l)} = [x_1^{(l)}, x_2^{(l)}, \cdots, x_{N_l}^{(l)}]$, $l = 0, 1, 2, \cdots, L$, 로 표시한다. 여기서, $-1 < x_j^{(l)} < 1$이며, $\mathrm{x}^{(0)}$와 $\mathrm{x}^{(L)}$은 각각 다층퍼셉트론의 입력상태 벡터와 출력상태 벡터를 나타낸다. 또한, $\mathrm{t} = [t_1, t_2, \cdots, t_{N_L}]$는 임의의 입력벡터에 대한 목표벡터를 나타낸다. 이와 같은 구조의 다층퍼셉트론에 임의의 학습패턴 x_p가 입력되면, 출력층에서 MSE 함수가

$$E_m(\mathrm{x}_p) = \frac{1}{2}\sum_{j=1}^{N_L}(t_j - x_j^{(L)})^2 \tag{1}$$

로 계산된다[1]. 모든 학습패턴 $\mathrm{X}_p (p = 1, 2, \cdots, P)$에 대하여 $E_m(\mathrm{x}_p)$를 최소화 하기 위해서, 다층퍼셉트론의 가중치 $w_{ji}^{(l)}$는 EBP 알고리즘에 따라

$$\Delta w_{ji}^{(l)} = \eta \delta_j^{(l)} x_i^{(l-1)} \tag{2}$$

에 의해 변경된다. 여기서,

$$\delta_j^{(l)} = \begin{cases} (t_j - x_j^{(L)})\dfrac{(1 - x_j^{(L)})(1 + x_j^{(L)})}{2}, & \text{when } l = L, \\ \dfrac{(1 - x_j^{(l)})(1 + x_j^{(l)})}{2}\sum_{k=1}^{N_{l+1}} w_{kj}^{(l+1)}\delta_k^{(l+1)}, & \text{when } 1 \le l \le L-1, \end{cases} \tag{3}$$

은 $x_j^{(l)}$의 오차신호이고, η는 학습률이다. 이러한 가중치 변경 과정을 모든 학습패턴에 대하여 한번씩 수행한 것을 epoch 단위로 표시한다.

위에서 설명한 EBP 알고리즘에서, 식 (3)의 출력층 노드 오차신호 $\delta_j^{(L)}$은 목표값과 실제값의 차이에 시그모이드 활성화 함수의 기울기가 곱해진 형태이다. 만약, $x_j^{(L)} \approx \pm 1$이면, $\delta_j^{(L)}$은 시그모이드의 기울기 항 때문에 아주 작은 값이 된다. 즉, $t_j = 1$ 인데 $x_j^{(L)} \approx -1$이거나 혹은 그 반대인 상황으로 $x_j^{(L)}$이 부적절하게 포화되면, $x_j^{(L)}$은 연결된 가중치들을 조정하기에 충분히 강한 오차신호를 발생시키지 못한다(그림 1). 이와 같은 출력노드의 부적절한 포화가 EBP 알고리즘에 따른 다층퍼셉트론의 학습에서 MSE의 최소화

를 지연시킨다.

한편, 학습패턴은 여러개이므로 EBP 알고리즘에서 한 학습패턴에 의한 가중치의 변화방향이 학습패턴 전체에 대한 오차를 줄이는 방향과 일치하거나 경쟁할 것이다[7]. 예를 들어, EBP 학습과정 중의 경쟁에 의해 어떤 학습패턴의 입력시 부적절하게 포화되는 상태로 밀려나는 출력노드가 있을 수 있다. 따라서, 부적절하게 포화되는 출력노드는 부적절한 포화상태를 벗어나기 위해서 강한 오차신호를 발생시켜야 한다. 반면에 적절하게 포화되어 목표값에 근접한 출력노드는 약한 오차신호를 발생시켜야 한다. 이렇게 하면, 한 패턴에 대한 가중치들의 변경이 학습패턴 전체의 EBP 학습에 의해 결정된 가중치들을 흐트러뜨리는 정도를 최소화 시킬 것이다. 위와 같이 출력노드의 상태에 따른 오차신호의 강·약은 역전파 학습 과정에서 출력노드가 부적절하게 포화될 가능성을 줄일 것이다. 또한, 학습패턴에 대한 과도한 학습에 따른 일반화 성능 저하를 방지하기 위해서도 적절하게 포화된 출력노드는 약한 오차신호를 발생시켜야 한다[8].

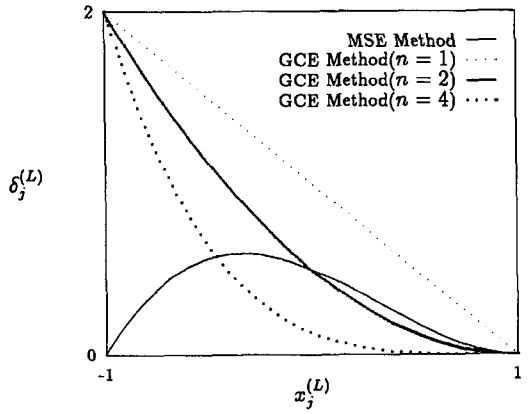


그림 1. 목표값이 1인 출력노드의 오차신호(n: nCE 함수의 차수)

Fig. 1. The error signal of output node with $t_j = 1$ (n: the order of nCE function).

이러한 관점에서, 역전파 학습을 위한 nCE 오차함수가

$$E_f(x) = -\sum_{k=1}^{N_L} \int \frac{t_j^{n+1}(t_j - x_j^{(L)})^n}{2^{n-2}(1-x_j^{(L)2})} dx_j^{(L)}, \text{ where } t_j = \pm \text{ and } n = 1, 2, \cdots \tag{4}$$

로 제안되었다[6]. 이 nCE 오차함수를 이용하면 출력노드의 오차신호는

$$\delta_j^{(L)} = \frac{t_j^{n+1}(t_j - x_j^{(L)})^n}{2^{n-1}} \tag{5}$$

이며, EBP 학습을 위한 다른 수식은 E_m을 이용한 경우와 동일하다. 따라서, nCE 오차함수($n \geq 2$)를 이용한 역전파 알고리즘은 출력노드가 목표값과의 거리에 따라 적절한 오차신호를 발생시켜서 경쟁에 의한 출력노드의 부적절한 포화를 줄여준다.

그림 1은 $t_j = 1$인 경우에 $x_j^{(L)}$에 따른 오차신호를 비교한 것이다. 여기서, $n = 1$인 경우가 크로스-엔트로피 오차함수를 이용한 오차신호 $(t_j - x_j^{(L)})$에 해당된다[5]. 이 그림에서 보는 바와 같이 nCE를 이용한 오차신호는 MSE 방법과 달리 출력노드가 부적절하게 포화될수록 강하게 발생되어, 출력노드가 학습의 진행에 따라 부적절한 포화 상태에 이르지 않도록 한다. 특히, nCE 오차신호 곡선은 n이 증가할수록 더욱 급격히 변하여, n이 큰 오차신호를 학습에 사용할수록 출력노드의 부적절한 포화를 더욱 더 효과적으로 방지한다. 그렇지만, 이 경우 오차신호가 목표값 근처에서 너무 작아져서 학습패턴에 대한 과도한 학습 방지 효과는 좋아지는 반면에 학습속도가 너무 느리게 된다. 이 때문에 nCE를 이용한 EBP 학습 시 다층퍼셉트론의 학습성능은 nCE 오차함수의 차수 n에 따라 심하게 변하므로, 출력노드의 부적절한 포화와 과도한 학습을 방지하면서도 빠른 학습결과를 얻기 위해서는 n을 적절히 정해야 한다.

Ⅲ. 가변학습률과 제한된 오차신호

앞에서 지적한 문제-즉, nCE를 이용한 EBP 학습이 n에 민감함-를 해결하기 위해서 학습률을 매 epoch 마다 $(t_j - x_j^{(L)})^2$과 $\delta_j^{(L)2}$의 기대치 비율에 따라

$$\eta(s) = \eta_o \sqrt{\frac{E[t_j(s) - x_j^{(L)}(s))^2}{E[\delta_j^{(L)2}(s)]}} \tag{6}$$

과 같이 변하도록 하였다. 여기서, $E[(t_j(s) - x_j^{(L)}(s))^2]$과 $E[\delta_j^{(L)2}(s)]$은 s번째 epoch에서 모든 학습패턴과 출력노드를 고려한 기대치이고, η_o는 첫번째 epoch의 학습률이다. 그러면, 가중치의 변화량을 결정하는 $\eta(s)\delta_j^{(L)}(s)$의 세기는 $\delta_j^{(L)}$의 종류에 상관없이 기대치가

$$E[\eta^2(s)\delta_j^{(L)^2}(s)] = \eta_o^2 E[(t_j(s)-x_j^{(L)}(s))^2] \quad (7)$$

이 된다. 즉, 이 가변학습률을 이용하면 $\delta_j^{(L)}$의 모양에 따른 특성을 살리면서 $\eta(s)\delta_j^{(L)}(s)$의 세기는 같은 기대치를 가지는 효과를 얻을 수 있다. 그리고, $n=1$인 경우 $\delta_j^{(L)} = t_j - x_j^{(L)}$이므로 $\eta(s) = \eta_o$이다. 실제 문제에 대한 학습 시뮬레이션에서 $\eta(s)$을 결정 시, s epoch 초기에 $E[\delta_j^{(L)^2}(s)]$와 $E[(t_j(s)-x_j^{(L)}(s))^2]$를 구할 수 없으므로 $(s-1)$번째 epoch에서 얻은 기대치들을 이용한다.

학습률을 식 (6)과 같이 변경시키는 방법은 n에 상관없이 $E[\eta^2(s)\delta_j^{(L)^2}(s)]$을 같게하여, nCE를 이용한 학습에서 n의 변화 시 오차신호의 부적절한 포화 및 과도한 학습방지 효과는 유지시키면서 학습 속도가 느려지는 단점을 개선시킨다. 그 결과 다층퍼셉트론의 학습성능이 nCE 오차함수의 차수 n에 민감하지 않게 된다. 그렇지만, 학습이 진행되어 $x_j^{(L)}$이 t_j 근처에 도달하면 $n\neq 1$인 경우 $E[(t_j(s)-x_j^{(L)}(s))^2]/E[\delta_j^{(L)^2}(s)]$은 매우 커짐을 그림 1에서 추정할 수 있다. 결국 $\eta(s)$가 매우 큰 값을 지녀, 가중치가 급격히 변하는 불안정한 학습상태가 나타난다. 이것을 방지하기 위해서

$$\delta_j^{(L)} = \begin{cases} \delta_j^{(L)}, & \text{if } -3\sqrt{E[\delta_j^{(L)^2}(s)]} < \delta_j^{(L)} < 3\sqrt{E[\delta_j^{(L)^2}(s)]} \\ \mathrm{sgn}(\delta_j^{(L)}) \times 3\sqrt{E[\delta_j^{(L)^2}(s)]}, & \text{otherwise}, \end{cases} \quad (8)$$

와 같이 $\delta_j^{(L)}$의 크기를 제한하였다. 여기서

$$\mathrm{sgn}(x) = \begin{cases} +1, & \text{if } x \geq 0 \\ -1, & \text{otherwise}, \end{cases} \quad (9)$$

이다. 이와 같이 하면 출력층 가중치의 변경량은 $-1 < x_i^{(L-1)} < 1$ 이므로 식 (8)에 의해

$$|\Delta w_{ji}^{(L)}| \leq \eta(s) \times |\delta_j^{(L)}(s)|_{max} \times |x_i^{(L-1)}|_{max} \quad (10)$$

$$< 3\eta_o \sqrt{E[(t_j(s)-x_j^{(L)}(s))^2]} \quad (11)$$

이 되어 가중치가 급격히 변하는 것이 방지된다. 제한된 $\delta_j^{(L)}$의 역전파된 값에 따라 변경되는 중간층 가중치 역시 급격히 변하지 않으므로, 학습률을 식 (6)처럼 가변시키는 EBP 학습은 제한된 오차신호에 의해 안정적인 특성을 보인다.

만약, $\delta_j^{(L)}(s)$가 Gaussian 분포를 가지며 평균이 0에 가까운 값이라면, 표준편차 $\sigma \approx \sqrt{E[\delta_j^{(L)^2}(s)]}$이고 $\delta_j^{(L)}(s)$의 99.7%가 $\pm 3\sigma$내에 존재할 것이다. 이럴 경우, 식 (8)과 같은 제한기준에 의해 $\delta_j^{(L)}$이 변경되는 부분은 극히 작으므로 제한된 오차신호가 학습특성을 왜곡시키는 부분은 아주 미미하다. 실제 문제에서 $\delta_j^{(L)}(s)$의 분포는 Gaussian이 아니므로 위와 같이 주장할 수 없다. 그렇지만, 시뮬레이션에서 $\delta_j^{(L)}(s)$의 확률밀도함수를 조사하여 $E[\delta_j^{(L)}(s)] \approx 0$이고 $\pm 3\sqrt{E[\delta_j^{(L)^2}(s)]}$내에 $\delta_j^{(L)}(s)$의 대부분이 존재하는 것을 확인하기로 한다.

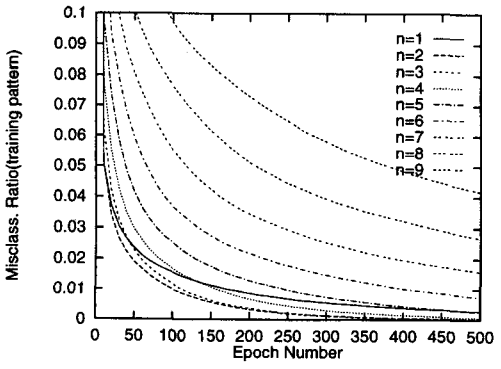
여기서 제안한 가변 학습률은 기본적으로 출력노드의 부적절한 포화 문제를 해결한 nCE 오차함수에서 차수 n이 증가할수록 EBP 학습의 속도가 느려지는 문제를 $E[\eta^2(s)\delta_j^{(L)^2}(s)]$이 n에 따라 변하지 않게하여 해결한다. 만약, 이 가변 학습률을 MSE를 이용한 EBP 알고리즘과 같이 출력노드의 부적절한 포화 현상이 심하게 나타나는 학습방법에 적용시키면, 학습속도가 빨라지는 효과를 기대할 수 없다. 한편, 다층퍼셉트론의 EBP 학습속도를 개선시키기 위해 제안되는 여러 오차함수 혹은 오차신호들을 공정하게 비교하기 위해서는 학습률을 제대로 정해야 한다[5]. 여기서 제안한 가변 학습률은 $\delta_j^{(L)}$이 어떤 형태이든 EBP 학습에서 가중치의 변화량을 결정하는 $\eta(s)\delta_j^{(L)}(s)$의 세기가 같은 기대치를 가지도록 하므로, 이와 같은 오차함수들의 학습속도 비교 시뮬레이션 시 사용하기에 적합하다.

가변 학습률은 여기서 제안한 것 이외에도 MSE를 이용한 EBP 알고리즘의 학습속도를 빠르게 하기 위해 여러 형태로 제안되었다[3] [9, pp. 268-272]. 이 기존의 가변 학습률들을 nCE를 이용한 EBP 알고리즘에 적용시키면 경우에 따라 학습속도를 빠르게 할 수도 있겠지만, 근본적으로 가중치의 변경량을 결정하는 $E[\eta^2(s)\delta_j^{(L)^2}(s)]$이 nCE의 차수 n에 따라 다른 값을 지닐 것이다.

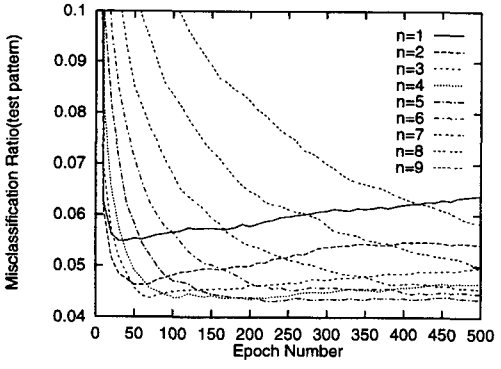
따라서, MSE를 이용한 EBP 알고리즘의 학습속도를 빠르게 하기 위해 제안된 기존의 가변 학습률들은 nCE 오차함수의 차수에 따른 EBP 학습성능의 민감성 문제를 해결할 수 없다.

Ⅳ. 시뮬레이션

앞에서 설명한 방법의 효용성을 확인하기 위해서 필기체 숫자인식 문제를 다층퍼셉트론에 학습시켰다. 우편봉투에 적힌 숫자를 모아둔 CEDAR 데이타베이스[10]로 부터 18,468개의 필기체 숫자영상을 임의로 선택하여 정규화 후 학습패턴으로 사용하였으며, 일반화 성능을 측정하기 위한 시험 패턴으로 2,213개를 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. 초기 가중치는 $[-1\times10^{-4}, 1\times10^{-4}]$ 에서 균일분포를 가지도록 임의로 선정하였으며, 각각의 학습방법에 따라 초기 가중치를 다르게 하면서 9번 시뮬레이션한 결과의 평균을 그림에 나타내었다.



(a)



(b)

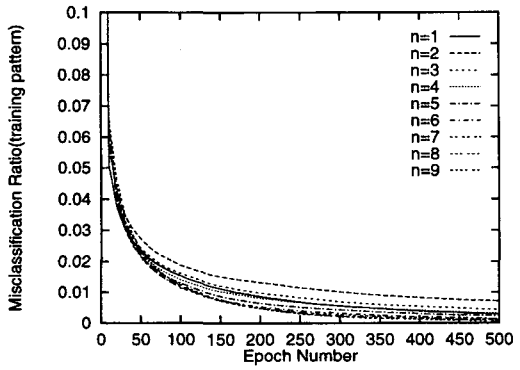
그림 2. nCE를 이용한 학습 시뮬레이션 결과 (n: nCE 함수의 차수)(a) 학습패턴에 대한 오인식률 (b) 시험패턴에 대한 오인식률

Fig. 2. Simulation results based on the nCE function (n: the order of nCE function). (a) Misclassification ratio for the training patterns (b) Misclassification ratio for the test patterns

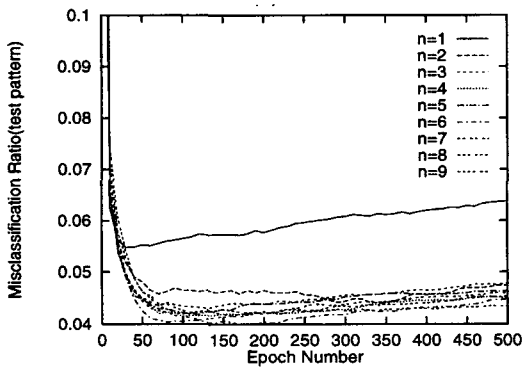
먼저 학습률을 고정시키고 nCE를 이용한 EBP 알고리즘에 따라 MLP를 학습시킨 결과를 그림 2에 그렸다. 이때, $x_j^{(L)}$이 $[-1, +1]$ 에서 균일분포라는 가정 하에 nCE 오차함수를 이용한 EBP 학습에서 $E[\eta\delta_j^{(L)}]$이 같은 값을 지니도록 하는 학습률을 계산하였다. 즉, n차 nCE 오차함수를 위한 학습률 $\eta = 0.001\times(n+1)$이다. 학습패턴에 대한 시뮬레이션 결과를 그림 2(a)에서 보면 $n=1$인 경우보다 $n=2,3$인 경우 학습패턴에 대한 오인식률이 빠르게 감소하나, $n\geq4$ 에서는 n이 증가할수록 학습속도가 느려짐을 알 수 있다. 이러한 현상이 나타나는 이유는 다음과 같다. MSE를 이용한 EBP 학습 시 $\delta_j^{(L)}$이 시그모이드의 기울기에 비례하는 특성 때문에 출력노드의 부적절한 포화현상이 심하여, 다층퍼셉트론이 제대로 학습되지 않아 500번째 epoch에서 학습패턴에 대한 오인식률이 0.02임을 [6]의 시뮬레이션에서 볼 수 있다. nCE를 이용한 학습 시 $n=1$인 경우 $\delta_j^{(L)} = (t_j - x_j^{(L)})$이므로 출력노드의 부적절한 포화가 MSE를 이용한 학습의 경우 보다 감소하여, 학습패턴에 대한 오인식률이 크게 감소하는 것도 [6]의 시뮬레이션에 설명되었다. 그렇지만 그림 1에서 보는 바와 같이 $x_j^{(L)}\approx t_j$일 때 $\delta_j^{(L)}(n=1)$이 MSE 혹은 nCE ($n\geq2$) 경우보다 상대적으로 크므로, 제대로 학습된 패턴들에 대한 가중치 변경이 다른 학습패턴에 대한 가중치 변경과 경쟁하는 현상이 심하여 학습에 나쁜 영향을 미칠 것이다. n이 증가하면 nCE 오차신호는 출력노드가 부적절하게 포화될수록 급격히 증가하여 출력노드가 부적절한 포화 상태를 빨리 벗어나도록 하며, 적절하게 포화된 출력노드에 대해서는 작은 값을 지녀 학습패턴 간의 경쟁 현상을 감소시킨다. 결국, 그림 2(a)에서 $n=2,3$인 경우 학습패턴에 대한 오인식률 곡선이 $n=1$인 경우보다 좋아진다. 그렇지만, n이 더욱 커져 $n\geq4$로 되면 출력노드의 오차신호가 목표값 근처에서 너무 작게된다. 즉, 목표값 근처에 위치한 출력 노드에 연결된 가중치들의 변경량이 미미하여 학습속도가 느려지게 된다.

시험패턴에 대한 결과를 그림 2(b)에서 보면, $n=1$인 경우 출력노드의 오차신호가 목표값 근처에서 크므로 학습패턴에 대한 과도한 학습이 발생하고 그 결과 일반화 성능이 나쁘게 나타난다. n이 증가할수록 출력노드의 오차신호가 목표값 근처에서 약하게

되어 학습패턴에 대한 과도한 학습을 방지하므로 일반화 성능은 좋아진다. 그렇지만, $n \geq 6$인 경우 출력노드의 오차신호가 목표값 근처에서 너무 작으므로 학습속도가 느려져서 시험패턴에 대한 오인식률도 매우 늦게 감소함을 볼 수 있다. 여기서 시뮬레이션한 문제에서는 $n = 3, 4$가 학습속도 및 일반화 성능의 관점에서 최적의 nCE 차수라고 볼 수 있다. 이 그림에서 보는 바와 같이 학습성능이 n에 따라 매우 많이 변하므로, 응용문제에 맞는 적절한 차수 n의 선정이 중요하다.



(a)



(b)

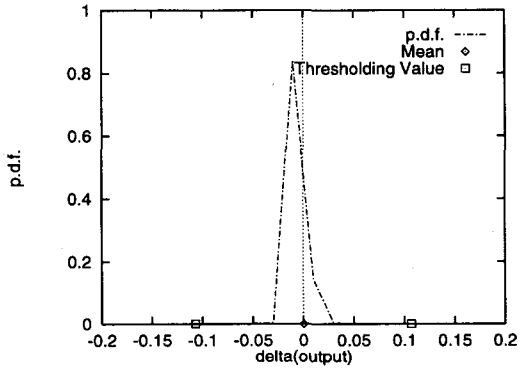
그림 3. 가변 학습률과 제한된 오차신호를 적용한 시뮬레이션 결과(n: nCE 함수의 차수)
(a) 학습패턴에 대한 오인식률 (b) 시험패턴에 대한 오인식률

Fig. 3. Simulation results with the adaptive learning rate and the limited error signal (n: the order of nCE function).
(a) Misclassification ratio for the training patterns (b) Misclassification ratio for the test patterns

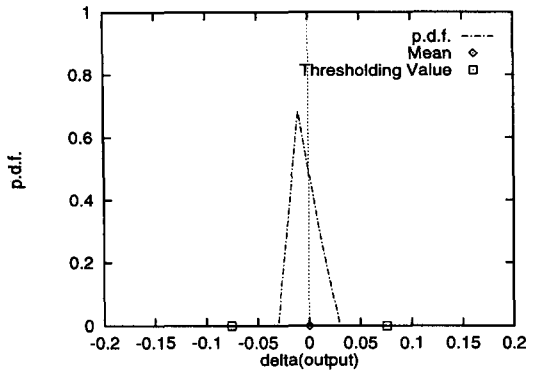
다음으로 $\eta_o = 0.002$로 두고서 가변학습률과 제한된 오차신호를 nCE를 이용한 EBP 알고리즘에 적용한 결과를 그림 3에 그렸다. $n = 1$인 경우 $\eta(s) = \eta_o$이므로, 그림 2와 3에서 $n = 1$ 곡선은 같은 것이다. $n \geq 2$ 인 경우 그림 3을 2와 비교해 보면, n의 증가에 따른 학습성능의 변화가 매우 감소한 것을 알 수 있다. 먼저, 그림 3(a)에서 학습패턴에 대한 결과를 보면, $n = 2, 3$인 경우 학습속도가 다소 느려진 것 이외에는 n의 변화에 따른 학습속도의 차이가 크지 않다. 시험패턴에 대한 결과를 그림 3(b)에서 보면, $n \geq 2$인 경우는 $n = 1$일 때보다 일반화 성능이 훨씬 좋으면서 비슷한 결과들을 나타낸다. 즉, $n \geq 2$인 경우 출력노드의 오차신호가 부적절한 포화와 학습패턴 간의 경쟁 및 과도한 학습을 방지하는 곡선적 특성은 유지하면서도, 가변 학습률에 의해 다른 n에 대해서도 $E[\eta^2(s)\delta_j^{(L)^2}(s)]$을 같게하여 학습속도가 n에 민감하지 않도록 한 것이다.

다음으로, 가변학습률을 이용한 학습이 안정적인 특성을 지니도록 하기 위해서 사용한 제한된 오차신호가 학습의 특성을 크게 변형시키는 지를 알아보기 위해서, 학습과정에서 오차신호의 분포와 그 크기를 제한하는 한계값을 조사하여 그림 4에 나타내었다. 이 그림에서 보는 바와 같이 $E[\delta_j^{(L)}(s)] \approx 0$ 이며 한계값 $\pm 3\sqrt{E[\delta_j^{(L)^2}]}$은 $\delta_j^{(L)}$의 확률밀도함수의 끝부분에 위치한다. 따라서, 학습과정에서 한계값에 의해 제한되는 $\delta_j^{(L)}$은 아주 작은 부분이므로 제한된 오차신호가 학습의 특성을 변형시키는 부분은 극히 미미하다.

3장에서 주장한 바와 같이 제한된 오차신호가 안정적인 학습에 기여하였는 지를 확인하기 위해서 오차신호를 제한하지 않고서 가변학습률 만을 적용시킨 시뮬레이션 결과를 그림 5에 그렸다.



(a)



(b)

그림 4. $\delta_j^{(L)}$의 확률밀도함수와 $3\sqrt{E[\delta_j^{(L)^2}]}$
(a) 150번째 epoch (b) 450번째 epoch

Fig. 4. The probability density function of $\delta_j^{(L)}$ and $3\sqrt{E[\delta_j^{(L)^2}]}$.
(a) the 150th epoch (b) the 450th epoch

이 그림에서 보는 바와 같이 $n=1, 2$인 경우는 학습이 안정적이었으나, $n=4$인 경우 학습패턴에 대한 학습이 잘 이루어지다가 갑자기 불안정해지는 것을 볼 수 있다. 그 이유는 앞에서 지적한 바와 같이, 모든 학습패턴에 대하여 $x_j^{(L)} \approx t_j$일 때 $\eta(s)$가 매우 커져서 가중치가 급격히 변하기 때문이다. 즉, $n=4$인 경우 nCE 오차신호의 출력노드에 대한 부적절한 포화방지 효과가 뛰어나서 모든 학습패턴에 대하여 $x_j^{(L)} \approx t_j$가 된다. 이러한 경우 제한된 오차신호가 가중치 변경량을 제한시켜 그림 3과 같은 안정적인 특성을 얻게 된다.

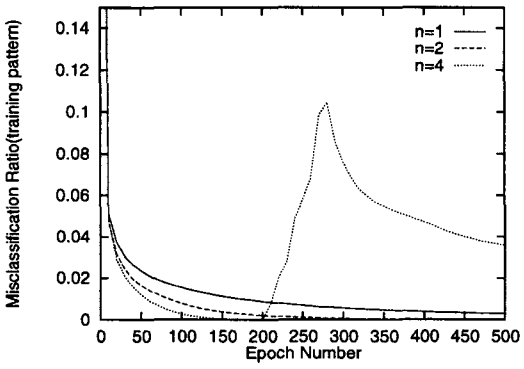
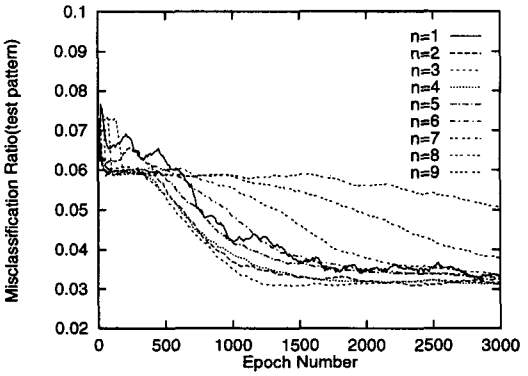
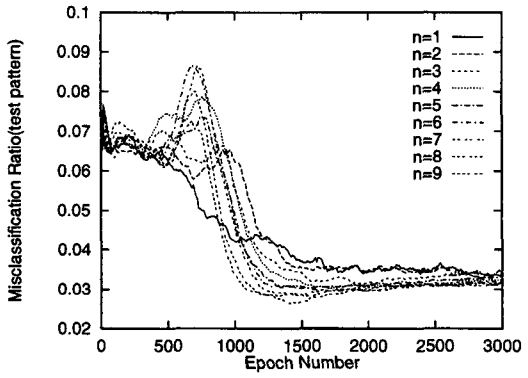


그림 5. 가변학습률을 적용한 학습에서 학습패턴에 대한 오인식률(n: nCE 오차함수의 차수)

Fig. 5. Misclassification ratio for the training patterns through the EBP learning with the adaptive learning rate (n: the order of the nCE function).



(a)



(b)

그림 6. 갑상선 진단 문제의 시험패턴에 대한 오인식률 (n: nCE 함수의 차수)
(a) 고정 학습률을 사용한 학습결과 (b) 가변 학습률과 제한된 오차신호를 사용한 학습결과

Fig. 6. Misclassification ratio of test patterns for the problem of diagnosing thyroid function (n: the order of nCE function).
(a) EBP learning with fixed learning rate
(b) EBP learning with the proposed learning rate and limited error signal

제시한 가변 학습률 및 제한된 오차신호의 효용성을 확인하기위해 PROBEN1 데이타 베이스의 갑상선 진단 문제에 대해서도 시뮬레이션하였다[11]. 이 갑상선 데이타는 UCI 데이타 베이스[12]의 갑상선 데이타를 근거로 만들어진 것으로 3,600개의 학습패턴, 1,800개의 확인(validation) 패턴, 및 1,800개의 시험패턴으로 구성되어 있다. 이 패턴들의 클래스(class) 분포는 3 클래스에 각각 5.1%, 92.6%, 2.3%이다. 다층퍼셉트론은 입력 노드 21, 중간층 노드 16, 출력 노드 3으로 구성되며, 초기 가중치는 $[-1\times10^{-3}, 1\times10^{-3}]$에서 균일분포를 가지도록 한다.

이 문제에서 nCE 오차함수를 근거로 학습률이 고정된 EBP 학습의 학습률은 앞의 필기체 숫자 인식 문제 시뮬레이션에서와 같이 $E[\eta\delta_j^{(L)}]$ 이 같은 값을 지니도록 $\eta = 0.01 \times (n+1)$로 한다. 그리고, 제시한 가변 학습률 및 제한된 오차신호에 따른 EBP 학습시 $\eta_o = 0.02$로 한다. 그림 6은 각각의 학습 방법에 대해 초기 가중치를 다르게 설정 후 EBP 학습을 수행하는 과정을 9번 실시한 결과 중 시험패턴에 대한 오인식률의 평균을 그린 것이다.

먼저, 고정된 학습률을 사용한 결과를 그림 6(a)에서 보면, nCE 오차함수의 차수 n=1인 경우보다 n=2,3, 및 4 인 경우 좋은 특성을 보인다. 한편, $n \geq 5$이면 학습속도가 느려져서 오인식률이 천천히 감소한다. 다음으로 가변 학습률과 제한된 오차신호에 대한 결과를 그림 6(b)에서 보면, 비록 학습과정의 과도기적 현상이 1,000 epoch 미만에서 심하게 나타나지만 어느 정도 학습이 진행된 후의 곡선이 n에 대하여 민감한 정도는 (a) 보다 훨씬 감소한 것을 볼 수 있다. 그림 6(a)의 n=1인 경우와 (b)의 모든 n에서 학습곡선의 굴곡이 심한 이유는 학습패턴 간의 경쟁현상 때문이다. 즉, 학습패턴의 분포가 특정 클레스에 집중되어 있어, 상대적으로 패턴 수가 적은 클래스에 속하는 패턴들은 패턴 수가 많은 클래스에 속하는 패턴들의 학습에 의해 학습이 안된 상태로 밀려나기 쉽다. nCE($n \geq 2$)에 근거한 고정 학습률 EBP 학습의 경우 이러한 경쟁 현상을 완화시켜 그림 6(a)에서는 학습곡선의 굴곡이 미약하다. 가변 학습률을 사용한 경우에는 학습이 제대로 진행되지 않은 상태에서 학습률이 증가하여 가중치의 변경량이 커진다. 이 상태에서 학습패턴 간의 경쟁이 심화되어 그림 6(b)에서 보는 바와 같이 학습곡선의 굴곡이 심한 현상이 나타난다. 그렇지만, (b)에서도 1,000 이상의 epoch에서는 학습이 제대로 진행되어 학습곡선의 굴곡은 미미하다.

갑상선 진단 문제의 시험패턴에 대한 학습 결과(그림 6(a)와 (b))를 필기체 숫자 인식 문제의 결과(그림 2(b) 및 그림 3(b))와 비교해 보면, 학습률이 고정된 경우에는 그림 2(b)와 그림 6(a)에서 보는 바와 같이 시험패턴에 대한 오인식률이 nCE의 차수 n에 민감하다. 제시한 가변 학습률 및 제한된 오차신호를 적용한 경우에는 그림 3(b)와 그림 6(b)에서 보는 바와 같이 오인식률의 n에 대한 민감성이 크게 감소하였다. 한편, 그림 3(b)에서는 $n=1$인 경우와 $n \geq 2$인 경우의 오인식률이 확연히 차이났으나, 그림 6(b)에서는 n에 따른 오인식률의 차이가 급격히 변하지 않았다. 오히려, 그림 6(a)와 (b)를 비교해 보면 $n=2,3$인 경우 가변학습률에 의해 오인식률 곡선이 나빠진 것을 알 수 있다. 그렇지만, 가변 학습률및 제한된 오차신호를 적용시킨 경우 필기체 숫자와 갑상선 진단 문제에서 $n \geq 3$이면 시험패턴에 대한 오인식률이 n에 크게 민감하지 않으면서 좋은 특성을 보인다. 따라서, 실제적인 패턴인식 문제의 학습 시 nCE 오차함수를 이용한 EBP 알고리즘에 여기서 제안한 가변 학습률과 제한된 오차신호를 적용시키면, 최적의 nCE 오차함수 차수를 구하지 않더라도 $n \geq 3$로 하면 학습의 지연현상 없이 출력노드의 부적절한 포화현상, 학습패턴 간의 경쟁현상, 및 과도한 학습을 방지한 좋은 학습결과를 얻을 수 있을 것이다.

V. 결 론

이 논문에서는 n차 크로스-엔트로피 오차함수를 이용한 다층퍼셉트론의 EBP 학습에서 학습성능이 오차함수의 차수에 따라 크게 바뀌는 문제를 해결하기 위해서 가변학습률과 제한된 오차신호를 제시하고, 필기체 숫자 인식문제와 갑상선 진단 문제의 시뮬레이션으로 효용성을 확인하였다.

가변학습률은 n차 크로스-엔트로피 함수에 의한 오차신호의 특성을 살려서 출력노드의 부적절한 포화 방지와 학습패턴에 대한 과도한 학습방지 효과는 유지시키면서, 오차신호의 차수가 큰 경우 학습속도가 느려지는 문제를 해결하였다. 제한된 오차신호는 출력노드 오차신호의 아주 작은 부분 만을 한계값으로 변경시키므로, 학습을 크게 변형시키지 않으면서도 안정적인 학습결과를 얻도록 하였다.

따라서, 다층퍼셉트론을 n차 크로스-엔트로피 오차함수를 이용하여 학습시킬 때 여기서 제안한 가변학습률과 제한된 오차신호를 사용하면 nCE 오차함수의 최적 차수를 구하지 않고 고차의 오차함수를 사용하면 된다.

참 고 문 헌

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

Press, Cambridge, MA, 1986.

[2] Y. Lee and S.-H. Oh and M. W. Kim, "An analysis of premature saturation in back-propagation learning", *Neural Networks*, vol. 6, pp. 719-728, 1993.

[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, 1990.

[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, 1990, Washington, DC, USA*, vol. I, pp. 707-710, 1990.

[5] A. van Ooyen and B. Nienhuis, "Improving the convergence of the back-propagation algorithm," *Neural Networks*, vol. 5, pp. 465-471, 1992.

[6] S.-H. Oh, "Improving the error back-propagation algorithm with a modified error function," *IEEE Trans. Neural Networks*, vol. 8, 799-803, 1997.

[7] 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, 1990.

[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] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford Univ. Press, New York, 1995.

[10] J. J. Hull, "A database for handwritten text recognition research," *IEEE Trans. Pat. Ana. Mach. Int.*, vol. 16, no. 5, pp. 550-554, May 1994.

[11] L. Prechelt, "Probenl-a set of neural network benchmark problems and benchmarking rules," *Technical Report 21/94*, Karlsruhe university, Germany 1994.

[12] C. J. Merz and P. M. Murphy, *UCI Repository of machine learning database*. University of California, Irvine, Dept. of Information and Computer Sciences, 1998. http://www.ics.uci.edu/~mlearn/MLRepository.html.

저 자 소 개

吳 相 勳(正會員)

1986년 부산대학교 전자공학과 학사. 1988년 부산대학교 전자공학과 석사. 1988년 ~ 1989년 금성반도체(현 LG 반도체) 근무. 1990년 ~ 현재 ETRI 선임연구원. 1995년 ~ 현재 KAZST 전기 및 전자공학과 박사과정 재학중. 주관심분야는 연산기능의 이론·응용, 및 구현

李 壽 永(正會員) KITE Journal of Electrical Engineering, vol. 3, no. 1, pp. 77-78, 1992 參照