다층퍼셉트론의 은닉노드 근사화를 이용한 개선된 오류역전파 학습

(Modified Error Back Propagation Algorithm using the Approximating of the Hidden Nodes in Multi-Layer Perceptron)

곽 영 태 † 이 영 직 †† 권 오 석 †††
(Young Tae Kwak) (Young Gik Lee) (Oh Seok Kwon)

요 약 본 논문은 학습 속도가 계층별 학습처럼 빠르며, 일반화 성능이 우수한 학습 방법을 제안한다. 제안한 방법은 최소 제곱법을 통해 구한 은닉층의 목표값을 이용하여 은닉층의 가중치를 조정하는 방법으로, 은닉층 경사 벡터의 크기가 작아 학습이 지연되는 것을 막을 수 있다. 필기체 숫자인식 문제를 대상으로 실험한 결과, 제안한 방법의 학습 속도는 오류역전파 학습과 수정된 오차 함수의 학습보다 빠르고, Ooyen의 방법과 계층별 학습과는 비슷했다. 또한, 일반화 성능은 은닉노드의 수에 관련없이 가장 좋은 결과를 얻었다. 결국, 제안한 방법은 계층별 학습의 학습 속도와 오류역전파 학습과 수정된 오차 함수의 일반화 성능을 장점으로 가지고 있음을 확인하였다.

Abstract This paper proposes a novel fast layer-by-layer algorithm that has better generalization capability. In the proposed algorithm, the weights of the hidden layer are updated by the target vector of the hidden layer obtained by least squares method. The proposed algorithm improves the learning speed that can occur due to the small magnitude of the gradient vector in the hidden layer. This algorithm was tested in a handwritten digits recognition problem. The learning speed of the proposed algorithm was faster than those of error back propagation algorithm and modified error function algorithm, and similar to those of Ooyen's method and layer-by-layer algorithm. Moreover, the simulation results showed that the proposed algorithm had the best generalization capability among them regardless of the number of hidden nodes. The proposed algorithm has the advantages of the learning speed of layer-by-layer algorithm and the generalization capability of error back propagation algorithm and modified error function algorithm.

1. 서 론

전방향 신경회로망의 대표적인 모델인 다층퍼셉트론(multi-layer perceptron)은 학습 기능에 의해 임의의 비선형 함수를 근사화 할 수 있기 때문에, 여러 분야에서 응용되고 있다. 다층퍼셉트론의 학습은 오류역전파(error back propagation) 학습을 많이 사용하며, 오류역전파 학습은 학습 방법이 간단하고, 구현이 쉬운 장점이 있다. 그러나, 오차 함수를 가중치로 1차 미분하는 경사 강하법(gradient descent method)을 이용하기 때문에, 학습 시간이 오래 걸리고 지역 최소점(local minimum)에 빠질 수 있는 단점이 있다[1,2,3].

오류역전파 학습의 학습 시간을 단축시키기 위해, Ooyen은 오차 함수를 평균제곱(mean squared) 함수 대신 교차엔트로피(cross-entropy) 함수를 사용했다[4]. 이 방법은 학습 속도는 빠르지만, 일반화 성능이 낮은 문제점이 있다. 또한, Chen은 오차 함수를 학습률로 미분하여 학습률을 변경시켰다[5]. 이런 학습은 학습률에 대한 추가적인 미분을 요구한다. 그리고, 학습하는 동안 출력층의 부적절한 포화를 방지하기 위해, 오차 함수로 평균제곱 함수와 엔트로피 함수를 사용하는 수정된 오차 함수(modified error function)를 이용한 학습도 제

† 정 회 원 : 충남대학교 컴퓨터공학과
ytkwak@ce.cnu.ac.kr
†† 비 회 원 : 한국전자통신연구원 연구원
ylee@etri.re.kr
††† 비 회 원 : 충남대학교 컴퓨터공학과 교수
oskwon@ce.cnu.ac.kr
논문접수 : 2000년 8월 10일
심사완료 : 2001년 7월 2일

안되고 있다[6].

최근에는 다층퍼셉트론의 학습을 위해 오차 함수의 2차 미분을 이용한 공액 경사법(conjugate gradient method)과 뉴톤 방법(Newton's method)등이 이용되고 있다[7]. 공액 경사법은 이전의 가중치 변경 벡터와 오차 함수의 경사 벡터를 직교하도록 하여 가중치를 변경한다. 하지만, 공액 경사법은 가중치 변경 벡터의 크기를 결정하는데, 많은 계산량이 필요하다. 그리고, 뉴톤 방법은 가중치의 경사도에 헤시언(Hessian) 역행렬을 곱하여 가중치를 변경한다. 따라서, 뉴톤 방법도 헤시언 행렬과 헤시언 역행렬의 계산에 많은 시간이 필요하다. 이러한 2차 미분을 사용하는 방법은 오류역전파 학습보다 학습 속도는 빠르지만, 계산량이 많은 단점이 있다.

계산량이 많은 공액 경사법이나 뉴톤 방법을 대신하여, 다층퍼셉트론의 가중치를 계층별(layer-by-layer)로 학습하는 방법들이 제안되고 있다[8,9,10,11]. Wang은 출력층의 가중치를 목표 벡터와 은닉층의 출력 벡터를 이용하여 최소 제곱법(least squares method)으로 구한다. 또한, 은닉층의 목표 벡터도 최소 제곱법으로 구한 후, 은닉층의 평균제곱 오차 함수를 정의하고, 최소 제곱법으로 은닉층의 가중치를 구한다[8]. 이런 계층별 학습은 최소 제곱법을 사용하기 때문에, 학습 속도가 빠르다. 또한, 다층퍼셉트론의 크기가 작고, 학습 패턴의 수가 적은 응용에 적합하다. 그러나, 은닉층의 목표 벡터가 선형 분리가 불가능한 경우, 일정한 학습 시간 후에도 더 이상 학습이 진행되지 않는 단점이 있다[9]. 일반화 성능이 낮은 Wang의 방법을 개선하기 위해, Yam은 출력층의 가중치를 최소 제곱법으로 구하고, 은닉층의 가중치는 오류역전파 학습을 했다[9]. 이 방법도 최소 제곱법을 사용하여 학습 속도는 빠르지만, 아직도 일반화 성능이 낮은 문제점이 있다.

따라서, 본 논문에서는 학습 속도는 계층별 학습처럼 빠르며, 일반화 성능은 오류역전파 학습이나 수정된 오차 함수의 학습처럼 우수한 학습 방법을 제안한다. 제안한 방법은 평균제곱 오차 함수에서 은닉층 출력 벡터의 경사 벡터를 구한다. 그리고, 출력층 가중치의 의사 역행렬(pseudoinverse matrix)을 이용하여, 은닉층 목표 벡터를 구한다[12,13]. 이렇게 얻은 은닉층 경사 벡터와 은닉층 목표 벡터를 합성하여 새로운 은닉층 경사 벡터를 정의하여, 은닉층의 가중치를 조정한다.

제2절은 오류역전파 학습과 학습 속도를 단축시키는 기존 방법과 계층별 학습에 대해 간단히 소개한다. 그리고, 제안한 방법에 대한 개념과 알고리즘이 있다. 제3절의 실험에는 기존 방법과 제안한 방법을 필기체 숫자(CEDAR 데이타베이스)[14]를 대상으로 실험한 결과가 있다. 실험 결과, 제안한 방법이 계층별 학습의 학습 속도와 오류역전파 학습이나 수정된 오차 함수의 일반화 성능을 나타내는 결과를 얻었다. 제4절의 결론은 실험 결과를 분석하여 제안한 방법의 장점 및 응용성을 제시한다.

2. 이 론

학습 시간을 단축하기 위해, 본 논문에서 사용한 다층퍼셉트론은 그림1과 같은 은닉층이 하나 있는 2층퍼셉트론이다.

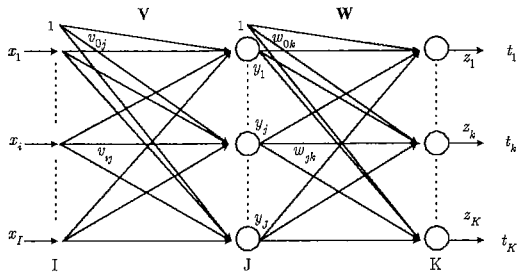


그림 1 2층퍼셉트론의 구조

그림 1에서 입력 벡터는 $\mathbf{x}_p = [1, x_{p1}, \cdots, x_{pi}, \cdots, x_{pI}]^T$, 은닉층 출력 벡터는 $\mathbf{y}_p = [1, y_{p1}, \cdots, y_{pj}, \cdots, y_{pJ}]^T$, 출력 벡터는 $\mathbf{z}_p = [z_{p1}, \cdots, z_{pk}, \cdots, z_{pK}]^T$ 이며, 목표 벡터는 $\mathbf{t}_p = [t_{p1}, \cdots, t_{pk}, \cdots, t_{pK}]^T$ 이다. 여기서, p는 학습 패턴의 인덱스이다. 그리고, 입력층에서 j번째 은닉노드로 연결된 가중치 벡터는 $\mathbf{v}_j((I+1)\times 1) = [v_{0j}, \cdots, v_{ij}, \cdots, v_{Ij}]^T$, 은닉층에서 k번째 출력노드에 연결된 가중치 벡터는 $\mathbf{w}_k((J+1)\times 1) = [w_{0k}, \cdots, w_{jk}, \cdots, w_{Jk}]^T$ 이며, v_{0j}와 w_{0k}는 각각 은닉노드와 출력노드의 바이어스이다. 또한, 가중치 행렬은 각각 $\mathbf{V}((I+1)\times J) = [v_1, \ldots, v_j]$, $\mathbf{W}((J+1)\times K) = [\mathbf{w}_1, \cdots, \mathbf{w}_k, \cdots, \mathbf{w}_K]$이다.

기본적인 오류역전파 학습은 식(1)과 같은 평균제곱 오차 함수를 가중치로 1차 미분하여 오차가 감소하는 방향으로 학습한다[1,2,3]. 이런 오류역전파 학습은 식(2)와 식(3)과 같은 경사 강하법을 이용하기 때문에 학습 속도가 느리다. 식(2)와 식(3)의 δ는 각각 출력노드와 은닉노드의 오차 신호이다. 그리고, 각 노드의 출력을 계산하는 활성화 함수는 식(4)와 같은 −1~1의 시그모이드 함수를 사용한다. 식(2)와 식(3)은 각 학습 패턴마다 가중치를 조정하는 패턴모드(pattern mode) 학

습이며, t는 학습 시간을, η는 학습률을 나타낸다.

$$E = \frac{1}{2}\sum_{p=1}^{P}\sum_{k=1}^{K}(t_{pk} - z_{pk})^2 \quad (1)$$

$$w_{jk}(t+1) = w_{jk}(t) + \eta\Delta w_{jk} \quad (2)$$

$$\Delta w_{jk} = -\frac{\partial E}{\partial w_{jk}} = (t_{pk} - z_{pk})f'(net_{pk})y_{pj} = \delta_{pk}y_{pj}$$

$$v_{ij}(t+1) = v_{ij}(t) + \eta\Delta v_{ij} \quad (3)$$

$$\Delta v_{ij} = -\frac{\partial E}{\partial v_{ij}} = \sum_{k=1}^{K}\delta_{pk}w_{jk}f'(net_{pj})x_{pi} = \delta_{pj}x_{pi}$$

$$f(net) = \frac{2}{1+e^{-\lambda net}} - 1 \quad (4)$$

2.1 관련 연구

2.1.1 Ooyen의 방법

Ooyen과 Nienhuis는 출력노드의 오차 신호가 시그모이드 함수의 포화 영역에서 정체되는 것을 방지하기 위해, 오차 함수를 평균제곱 함수 대신 식(5)와 같은 교차 엔트로피 함수를 사용했다[4]. 이 방법은 출력노드의 오차 신호가 목표값에 선형적으로 변하므로, 학습 속도는 빠르다. 그러나, 학습 말기에 오차 신호가 너무 강하게 작용하여, 오류역전파 학습보다 일반화 성능이 낮은 단점이 있다.

$$E = -\sum_{p=1}^{P}\sum_{k=1}^{K}\left[(1+t_{pk})\ln(1+z_{pk}) + (1-t_{pk})\ln(1-z_{pk})\right] \quad (5)$$

$$\Delta w_{jk} = -\frac{\partial E}{\partial w_{jk}} = (t_{pk} - z_{pk})y_{pj} = \delta_{pk}y_{pj} \quad (6)$$

2.1.2 수정된 오차 함수의 학습

수정된 오차 함수의 학습은 교차엔트로피 함수의 단점을 보완하기 위해 제안되었다[6]. 출력층의 오차 신호가 목표값에 대해 선형적으로 변하지 않도록, 식(7)과 같은 오차 함수를 사용하고, 가중치는 식(8)과 같이 조정한다. 이 방법에서 학습 속도는 오류역전파 학습보다 빠르며, 일반화 성능은 교차엔트로피 함수보다 우수하다.

$$E = -\sum_{p=1}^{P}\sum_{k=1}^{K}t_{pk}\left[-z_{pk} + \frac{1+t_{pk}^2}{2}\ln\frac{1+z_{pk}}{1-z_{pk}} + t_{pk}\ln(1-z_{pk})(1+z_{pk})\right] \quad (7)$$

$$\Delta w_{jk} = -\frac{\partial E}{\partial w_{jk}} = \frac{t_{pk}(t_{pk} - z_{pk})^2}{2}y_{pj} = \delta_{pk}y_{pj} \quad (8)$$

2.1.3 Wang의 방법

Wang과 Chen은 다층퍼셉트론의 계층별 학습을 제안하였다[8]. 우선, 전체 학습 패턴에 대한 은닉층의 출력 벡터($\mathbf{y}_{p,p=1,\cdots,P}$)를 구한다. 그리고, 하나의 학습 패턴($q$)에 대한 최적의 출력층 가중치($\mathbf{W}_q^*$)를 식(9)와 같이 최소 제곱법으로 구한다. 여기서, 학습 패턴(q)에 대한 은닉층의 출력 벡터는 제외되어 인덱스가 $P-1$이 되고, $\mathbf{d}_p = [f^{-1}(t_{p1}),\cdots,f^{-1}(t_{pk}),\cdots,f^{-1}(t_{pK})]^T$는 목표 벡터를 시그모이드 역함수로 변환한 벡터이다.

$$\mathbf{W}_q^* = \left(\sum_{p=1}^{P-1}\mathbf{y}_p\mathbf{y}_p^T\right)^{-1}\left(\sum_{p=1}^{P-1}\mathbf{y}_p\mathbf{d}_p^T\right) \quad (9)$$

식(9)에서 얻은 출력층의 가중치로 학습 패턴(q)에 대한 은닉층의 목표 벡터($\mathbf{y}_q^*$)를 식(10)처럼 최소 제곱법으로 구한다. 이 때, 은닉층의 목표 벡터가 활성화 함수의 범위를 넘어가면 절삭한다. 여기서, $(\mathbf{W}_q^*)^+$는 의사 역행렬이며, $\mathbf{d}_p = [f^{-1}(t_{p1}),\cdots,f^{-1}(t_{pk}),\cdots,f^{-1}(t_{pK})]^T$이다.

$$\mathbf{y}_q^* = (\mathbf{W}_q^*)^+\mathbf{d}_q \quad (10)$$

식(10)의 은닉층 목표 벡터를 이용하여 은닉층의 평균제곱 오차 함수를 정의하고, 정의된 평균제곱 오차 함수를 이용하여 은닉층 가중치($\mathbf{V}^*$)를 다시 최소 제곱법으로 식(11)처럼 구한다. 식(11)의 $\mathrm{s}p = [f^{-1}(y_{p1}^*),\cdots, f^{-1}(y_{pj}^*),\cdots,f^{-1}(y_{pJ}^*)]^T$는 식(10)에서 구한 은닉층 목표 벡터를 시그모이드 역함수에 입력한 것이다.

$$\mathbf{V}^* = \left(\sum_{p=1}^{P}\mathbf{x}_p\mathbf{x}_p^T\right)^{-1}\left(\sum_{p=1}^{P}\mathbf{x}_p\mathbf{s}_p^T\right) \quad (11)$$

계층별 학습의 최소 제곱법은 가중치를 목표 벡터에 대해 선형적으로 조정한다. 따라서, 학습 속도는 빠르다. 그러나, 은닉층의 목표 벡터가 활성화 함수의 범위를 벗어나는 경우, 그 값을 절삭하는 변환 함수를 사용한다. 이런 변환 함수는 은닉층 목표 벡터를 선형 분리가 불가능한 패턴으로 만들기 때문에, 일정한 학습 시간 후, 더 이상 학습이 진행되지 않는 현상이 나타난다. 이것은 계층별 학습의 일반화 성능을 떨어지게 한다[9].

2.1.4 Yam의 방법

Yam과 Chow는 계층별 학습의 일반화 성능을 높이기 위해, 최소 제곱법과 오류역전파 학습을 접합했다[9]. 즉, 출력층의 가중치는 식(12)와 같은 최소 제곱법으로 조정한다. 그리고, 은닉층의 가중치는 식(13)과 같은 오류역전파 학습을 한다. 이 방법도 출력층의 가중치를 최소 제곱법으로 학습하기 때문에, 학습 속도는 빠르다. 그러나, 학습 패턴의 과다 학습으로 인해 일반화 성능이 낮다.

$$\mathbf{W}^* = (\mathbf{Y}^T)^+\mathbf{D}^T \quad (12)$$

$$\Delta v_{ij} = -\frac{\partial E}{\partial v_{ij}} = \sum_{k=1}^{K}(d_{pk} - \mathbf{y}_p^T\mathbf{w}_{*k}^*)w_{jk}^*y_j'x_{pi} \quad (13)$$

여기서, $\mathbf{Y} = [\mathbf{y}_1,\cdots,\mathbf{y}_p,\cdots\mathbf{y}_P]$는 전체 학습 패턴에 대한 은닉층의 출력 행렬이며, $\mathbf{D} = [\mathbf{d}_1,\cdots,\mathbf{d}_p,\cdots\mathbf{d}_P]$는 시그모이드 역함수에 목표 벡터를 입력하여 얻은 행렬이다.

2.2 제안한 방법

오류역전파 학습은 오차 함수의 표면을 따라 가중치를 조정하는 경사 강하법을 이용하기 때문에, 학습이 오차 함수의 지역 최소점에 빠지거나, 출력층의 오차 신호가 조기포화 되면, 가중치의 변화량이 작아 학습 속도가 느려진다[15]. 따라서, 본 논문은 오류역전파 학습에 의한 은닉층의 경사 벡터와 최소 제곱법에 의한 은닉층의 목표 벡터를 이용하는 학습 방법을 제안한다.

우선, 오류역전파 학습에서 은닉층 출력 벡터와 오차 함수의 관계를 알아보자. 그림2는 하나의 학습 패턴(p)을 입력했을 때, 은닉층 출력 벡터($\mathbf{y}_p$)에 대한 오차 함수(E_p)를 나타낸다. 여기서, y_1-y_J평면은 J차원의 은닉층 공간을 나타내고, E_p축은 학습 패턴(p)에 대한 오차를 나타내며, t는 학습 시간(epoch)을 나타낸다. 또한, $\mathbf{y}_p^*$는 최소 제곱법으로 구한 현재의 학습 패턴(p)에 대한 은닉층 목표 벡터이다.

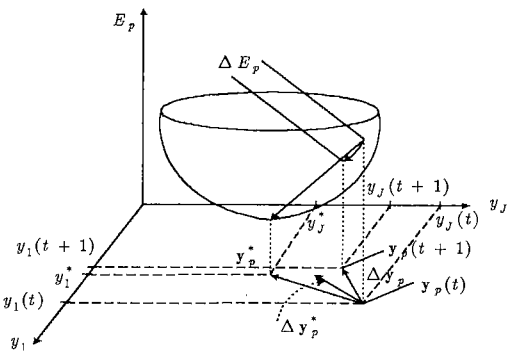


그림 2 은닉층 출력 백터에 대한 오차 함수

현재의 은닉층 출력 벡터를 $\mathbf{y}_p(t)$라고 할 때, 오류역전파 학습은 식(14)와 같은 은닉층 경사 벡터를 이용하여, $\mathbf{y}_p(t)$를 $\mathbf{y}_p(t+1)$로 조정한다. 이런 $\mathbf{y}_p(t+1)$는 오류역전파 학습에서 $\mathbf{y}_p(t)$에 대한 은닉층 목표 벡터라고 할 수 있고, $\Delta\mathbf{y}_p$는 현재의 은닉층 출력과 은닉층 목표 벡터사이의 오차 벡터라고 할 수 있다. 이 때, $\Delta\mathbf{y}_p$의 크기는 오차 함수의 기울기에 따라 변한다. 즉, 오차 함수의 기울기가 작을 때, 학습 속도가 느려진다.

$$\Delta y_{pj} = -\frac{\partial E}{\partial y_{pj}} = \sum_{k=1}^{K} \delta_{pk} w_{jk}$$
$$\Delta \mathbf{y}_p = \left[\Delta y_{p1}, \cdots, \Delta y_{pj}, \cdots, \Delta y_{pJ}\right]^T \tag{14}$$

이런 학습 정체를 방지하기 위하여, 제안한 학습에서는 오류역전파 학습의 은닉층 목표 벡터($\mathbf{y}_p(t+1)$) 외에 계층별 학습의 은닉층 목표 벡터($\mathbf{y}_p^*$)를 사용한다. 계층별 학습에서 현재의 은닉층 출력 벡터와 은닉층 목표 벡터사이의 오차 벡터는 $\mathbf{y}_p^*-\mathbf{y}_p(t)$이다. 따라서, 제안한 학습의 새로운 경사 벡터는 $\Delta\mathbf{y}_p$와 $\mathbf{y}_p^*-\mathbf{y}_p(t)$를 합성한 $\Delta\mathbf{y}_p^*$를 사용한다. 이와 같이, 제안한 방법은 현재 은닉층 출력 벡터($\mathbf{y}_p(t)$)가 은닉층 목표 벡터($\mathbf{y}_p^*$)에 점진적으로 근사하도록 학습하는 것이다.

학습 알고리즘에서, 출력층의 가중치는 식(2)와 같은 오류역전파 학습을 하고, 은닉층의 경사 벡터는 식(14)에서 구한다. 그리고, 현재의 학습 패턴(p)에 대한 은닉층 목표 벡터는 식(15)에서 최소 제곱법으로 구할 수 있다. 식(15)의 은닉층 목표 벡터($\mathbf{y}_p^*$)는 식(16)에서 얻는다. 이 때, 출력층의 가중치 중, 바이어스에 대한 은닉노드의 값은 1로 고정되어 있다. 따라서, 바이어스에 대한 은닉노드는 은닉층 목표 벡터에서 제외된다. 여기서, $\mathbf{W}_{off_bias}^T$는 바이어스가 제외된 가중치 행렬이고, $\mathbf{W}_{bias}^T$는 바이어스 벡터이다. 또한, $\mathbf{d}_p = [f^{-1}(t_{p1}), \cdots, f^{-1}(t_{pk}), \cdots, f^{-1}(t_{pK})]^T$이다.

$$\mathbf{W}^T\mathbf{y}_p = \mathbf{d}_p \tag{15}$$
$$\mathbf{y}_p^* = (\mathbf{W}_{off\_bias}^T)^+(\mathbf{d}_p - \mathbf{w}_{bias}^T) \tag{16}$$

식(16)의 $(\mathbf{W}_{off_bias}^T)^+$는 의사 역행렬로 식(17)과 같다. 이런 의사 역행렬은 QR-Decomposition이나 Single Value Decomposition으로 구한다. 구현에 있어서 householder reflection을 사용하는 QR-Decomposition이 간단하다[12,13].

$$(\mathbf{W}^T)^+ = \begin{cases} \mathbf{W}(\mathbf{W}^T\mathbf{W})^{-1} & \mathrm{J} \geq \mathrm{K} \\ (\mathbf{W}\mathbf{W}^T)^{-1}\mathbf{W} & \mathrm{J<K} \end{cases} \tag{17}$$

은닉층의 새로운 경사 벡터인 $\Delta\mathbf{y}_p^*$는 $\Delta\mathbf{y}_p$와 $\mathbf{y}_p^*-\mathbf{y}_p(t)$를 합성하여 만든다. 식(18)에서 은닉층 목표 벡터에 대한 각 성분(y_{pj}^*)이 시그모이드 함수내에 존재하면, 그 노드는 학습이 올바르게 된 것이므로, 오류역전파 학습의 $\Delta y_{pj}(t)$를 사용한다. 이런 상태는 은닉노드의 학습이 충분히 진행된 상태이다. 그러나, 은닉층 목표 벡터의 성분(y_{pj}^*)이 시그모이드 함수의 범위를 벗어나면, 학습 초기에 학습이 올바르게 진행되지 않은 상태이므로, 새로운 경사 벡터의 성분을 만든다.

$$\Delta y_{pj}^* = \begin{cases} \Delta y_{pj}(t) & \text{if } \left|y_{pj}^*\right| < 1 \\ \Delta y_{pj}(t) + yn_{pj}^* & \text{otherwise} \end{cases} \tag{18}$$

새로운 경사 벡터를 합성하기 전에, 은닉층의 목표 벡터는 시그모이드 함수의 범위를 벗어날 수 있으므로, 식(19)와 같이 정규화 한다. 식(19)의 정규화 과정은 $\mathbf{y}_p^*-\mathbf{y}_p(t)$로 현재의 은닉층 출력 벡터가 학습해야 할 방향성을 결정하고, $\left\|\mathbf{y}_p^*-\mathbf{y}_p(t)\right\|$로 나누어 벡터 크기를 정규화

한 다음, 은닉층 경사 벡터와 크기가 같게 하기 위하여 $\|\Delta \mathbf{y}_p\|$를 곱한다. 이렇게 정규화된 yn_{pj}^*와 $\Delta y_{pj}(t)$를 식(18)과 같이 합성하여, 은닉층의 새로운 경사 벡터의 성분을 만들고, 이 벡터를 이용하여 식(20)처럼 은닉층의 가중치를 조정한다.

$$\mathbf{yn}_p^* = \|\Delta \mathbf{y}_p\| \frac{\mathbf{y}_p^* - \mathbf{y}_p(t)}{\|\mathbf{y}_p^* - \mathbf{y}_p(t)\|} \tag{19}$$

$$\Delta v_{ij} = \Delta y_{pj}^* f'(net_{pj}) x_{pi} \tag{20}$$

제안한 방법은 학습 초기에 은닉노드의 출력이 매우 작으면, 은닉층 목표 벡터의 성분이 시그모이드 함수의 범위를 벗어나게 된다. 따라서, 오류역전파 학습과 계층별 학습의 합성에서 계층별 학습을 중심으로 학습하게 되어, 학습 속도를 가속화 한다. 그리고, 계층별 학습이 진행된 후, 은닉층 목표 벡터의 성분이 시그모이드 함수의 범위 내에 존재하면, 오류역전파 학습을 수행한다. 이런 오류역전파 학습은 계층별 학습의 단점인 선형 분리가 불가능한 은닉층 목표 벡터를 비선형 분리가 가능하게 하여, 일반화 성능을 향상시킨다. 이와 같이, 제안한 방법은 오류역전파 학습의 비선형적인 학습과 최소 제곱법의 선형적인 학습을 이용한다. 따라서, 계층별 학습의 단점인 낮은 일반화 성능을 극복하고, 오류역전파 학습의 학습 속도를 향상시킬 수 있다.

3. 실험 및 결과

실험은 필기체 숫자인식 문제를 대상으로 실험하였다. 필기체 숫자는 CEDAR 데이타베이스[14]의 숫자중, 각 숫자에 대해 100개씩 모두 1,000개를 학습에 사용하였고, 시험 패턴은 각 숫자에 대해 50개씩 500개의 숫자를 이용하였다. 한 숫자의 영상은 12×12 픽슬(pixel)로 구성되며, 각 픽슬은 16 레벨의 그레이 이미지로 바꾼 후 [−1∼+1]로 정규화 한 다음, 다층퍼셉트론에 입력시켰다. 그림 3은 학습에 사용된 패턴을 그레이 이미지로 나타낸 것이다.

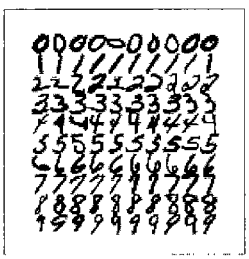


그림 3 학습 패턴

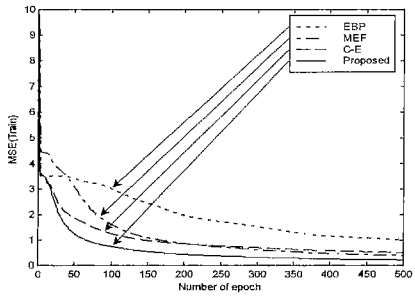
다층퍼셉트론은 입력노드 144개, 출력노드 10개, 은닉노드를 10, 15, 20, 25개로 변경하면서 구성하였다. 학습 방법은 오류역전파 학습, 수정된 오차 함수의 학습, Ooyen의 방법, Wang의 방법, Yam의 방법과 제안한 방법을 사용했다. 각 학습 방법은 은닉노드의 수에 따라, 가중치의 초기화를 3회씩 다르게 하여 총 12회를 학습했다. 학습의 완료 조건은 학습 시간을 500 epoch로 했다. 학습 방법은 C로 구현하고 의사 역행렬에 대한 계산은 MATLAB과 접합시켜 구현했다. 학습 결과는 식(21)과 같은 평균제곱 오차 함수로 표현하며, 이렇게 표현된 학습 결과는 3회에 대한 평균값으로 각 은닉노드 수에 대한 대표 학습 결과를 나타냈다.

$$E = \frac{1}{P}\sum_{p=1}^{P}\sum_{k=1}^{K}(t_{pk} - z_{pk})^2 \tag{21}$$

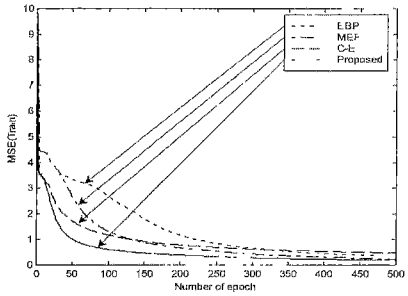
그림 4와 그림 5는 오류역전파 학습, 수정된 오차 함수의 학습, Ooyen의 방법, 제안한 방법의 학습 결과로서, 학습 패턴과 시험 패턴에 대한 오차 함수의 변화를 나타낸다. 그리고, 그림 6과 그림 7은 Wang의 방법, Yam의 방법, 제안한 방법의 학습 결과로서, 학습 패턴과 시험 패턴에 대한 오차 함수의 변화를 나타낸다. 여기서, 수평축은 학습 시간의 단위인 epoch이며, 수직축은 식(21)에서 정의한 평균제곱 오차 함수이다. 그리고, 'Proposed'는 제안한 학습 방법을 나타내고, 'C-E'는 Ooyen의 교차엔트로피 함수를 나타내며, 'MEF'는 수정된 오차 함수의 학습을 나타낸다.

먼저, 그림4와 그림5에서 오류역전파 학습은 은닉노드의 수가 증가할수록, 학습 속도와 일반화 성능이 우수해진다. 그러나, 은닉노드의 수가 적은 경우, 학습이 제대로 이루어지지 않는다. 또한, 학습 초기에 학습 속도가 다른 학습 방법에 비해 느린 결과를 얻었다. 오류역전파 학습의 단점을 보안한 Ooyen의 방법은 학습 속도가 오류역전파 학습보다 빠르다. 그러나, 은닉노드의 수가 증가할수록, 출력층의 오차 신호가 목표 벡터에 대해 선형적으로 학습하기 때문에, 일반화 성능이 오류역전파 학습보다 낮게 나타났다.

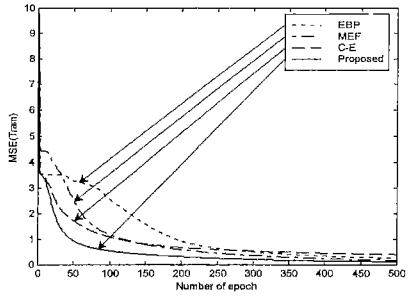
수정된 오차 함수의 학습은 오류역전파 학습과 Ooyen 방법의 단점을 보완한 결과를 보여준다. 초기 학습 속도는 오류역전파 학습보다 빠르고, Ooyen의 방법보다 느리다. 그러나, 학습 말기에, Ooyen의 방법은 오차가 더 이상 축소되지 않는 상황에서도 수정된 오차 함수의 학습은 오차를 축소할 수 있었다. 수정된 오차 함수의 일반화 성능은 은닉노드의 수에 상관없이, 오류역전파 학습이나 Ooyen의 방법보다 우수하게 나타났다.



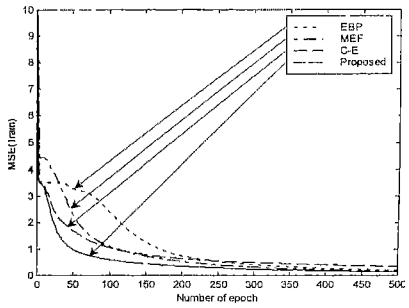
(a) 10개의 은닉노드



(b) 15개의 은닉노드

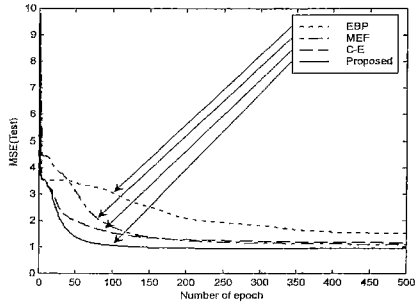


(c) 20개의 은닉노드

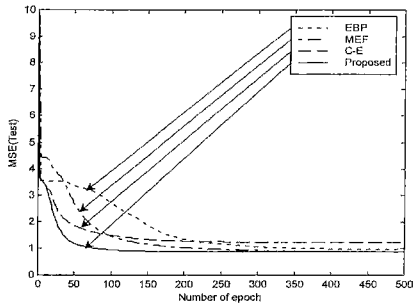


(d) 25개의 은닉노드

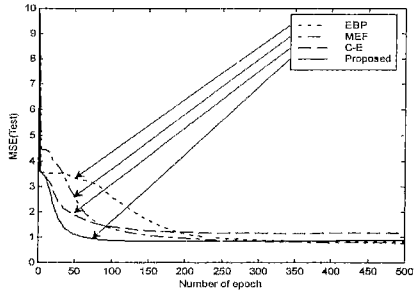
그림 4 학습 패턴의 오차 함수(오류역전파 학습, 수정된 오차 함수의 학습, Ooyen의 방법, 제안한 방법)



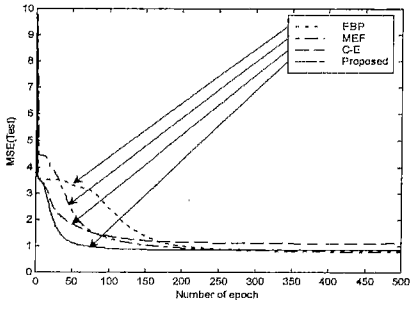
(a) 10개의 은닉노드



(b) 15개의 은닉노드

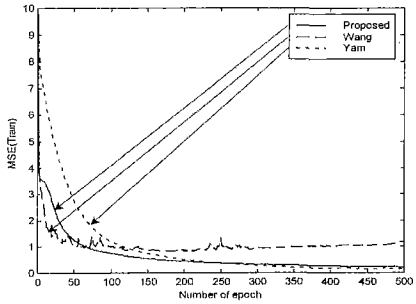


(c) 20개의 은닉노드

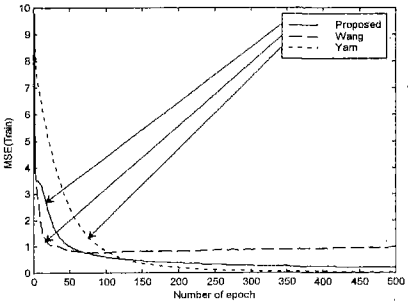


(d) 25개의 은닉노드

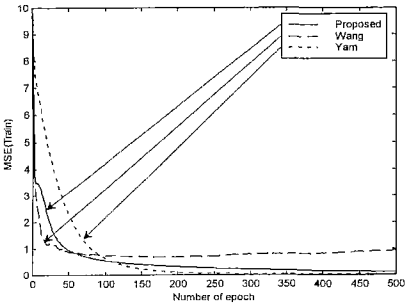
그림 5 시험 패턴의 오차 함수(오류역전파 학습, 수정된 오차 함수의 학습, Ooyen의 방법, 제안한 방법)



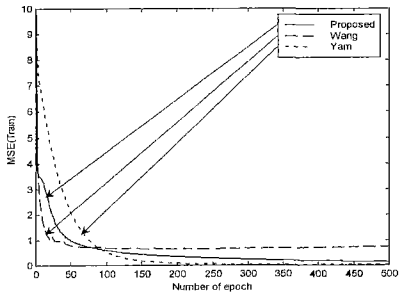
(a) 10개의 은닉노드



(b) 15개의 은닉노드

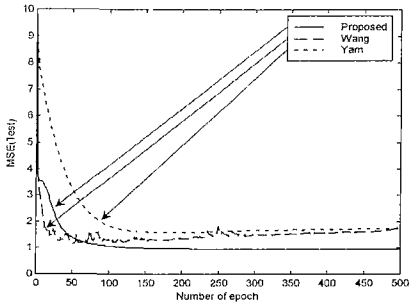


(c) 20개의 은닉노드

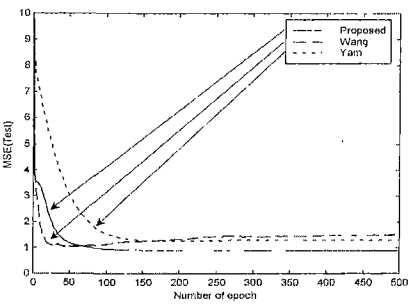


(d) 25개의 은닉노드

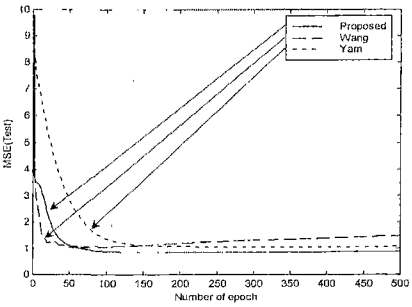
그림 6 학습 패턴의 오차 함수(Wang의 방법, Yam의 방법, 제안한 방법)



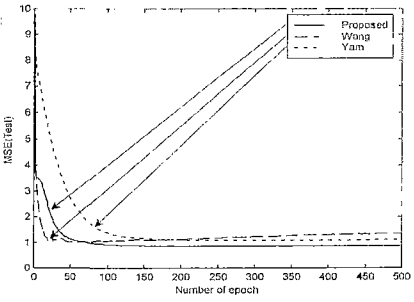
(a) 10개의 은닉노드



(b) 15개의 은닉노드



(c) 20개의 은닉노드



(d) 25개의 은닉노드

그림 7 시험 패턴의 오차 함수(Wang의 방법, Yam의 방법, 제안한 방법)

제안한 학습 방법은 학습 초기에 가장 빠른 학습 속도를 나타냈다. 이것은 계층별 학습의 선형적인 특징때문이다. 즉, 계층별 학습이 선형 분리가 가능한 오차까지 빠른 속도로 학습하는 것이다. 그리고, 학습 말기에는 오류역전파 학습의 비선형적인 학습을 하여 일반화 성능을 향상시킨다. 이런 일반화 성능은 은닉노드의 수에 상관없이, 가장 좋은 결과를 나타냈다.

그림6과 그림7은 기존 계층별 학습과 제안한 학습 방법을 비교한다. Wang의 계층별 학습은 학습 초기에, 학습 속도가 가장 빠르나, 일정한 시간 후, 더 이상 학습이 진행되지 않았다. 따라서, 수렴성이나 일반화 성능이 가장 낮게 나타났다. Wang의 방법에서 학습 속도가 빠른 원인은 출력층에서 선형적인 최소 제곱법으로 학습하기 때문이며, 학습이 더 이상 진행되지 않는 것은 은닉층의 목표 벡터를 비선형 분리할 수 없기 때문이다. 즉, 계층별 학습은 은닉층 목표 벡터가 선형 분리가 불가능할 때까지, 빠르게 선형적으로 학습한다.

Yam의 방법은 출력층의 가중치를 최소 제곱법으로 학습하기 때문에, 학습 패턴에 대해서 빠른 속도를 보인다. 특히, 은닉노드의 수가 증가할수록, 다른 학습 방법에 비해, 학습 패턴에 대한 수렴성이 우수하게 나타났다. 그리고, 은닉층의 가중치는 오류역전파 학습을 하기 때문에, 일반화 성능이 Wang의 방법보다 우수하게 나타났다. 전체적으로, Yam 방법의 일반화 성능은 Ooyen의 방법과 비슷하거나 약간 낮게 나타났다.

제안한 방법의 학습 속도는 학습 초기에 계층별 학습을 하지만, 동시에 오류역전파 학습도 수행하기 때문에, Wang의 계층별 학습보다 느리다. 그리고, 출력층의 선형 학습을 하는 Yam의 방법보다 빠르다. 그러나, 일반화 성능면에서는 다른 두 계층별 학습보다 우수한 결과를 보여준다. 제안한 방법은 계층별 학습이 선형 분리가 가능한 오차까지 학습한 후에, 오류역전파 학습에 의해 비선형 학습을 수행하여 일반화 성능을 향상시킨다. 따라서, 은닉노드의 수에 관련없이, 가장 좋은 일반화 성능을 얻었다. 이것은 제안한 방법이 학습 속도에서는 계층별 학습의 장점을 취하고, 일반화 성능에서는 오류역전파 학습이나 수정된 오차 함수의 장점을 취하고 있기 때문이다.

4. 결 론

본 논문은 학습 속도가 계층별 학습처럼 빠르며, 오류역전파 학습과 수정된 오차 함수의 일반화 성능을 가진 학습 방법을 제안한다. 출력층의 오차 신호에 선형적인 특성을 포함한 Ooyen의 방법, Wang의 방법, Yam의 방법은 학습 속도는 빠르지만, 일반화 성능이 낮다. 또한, 오류역전파 학습과 수정된 오차 함수의 학습은 일반화 성능이 높지만, 학습 초기에 수렴 속도가 느린 단점이 있다.

제안한 방법은 오차 함수에 대한 은닉층 출력 벡터의 경사 벡터를 구하고, 출력층 가중치의 의사 역행렬을 이용하여 은닉층 목표 벡터를 구한다. 이렇게 구해진 은닉층 경사 벡터와 은닉층 목표 벡터를 합성하여 새로운 은닉층 경사 벡터를 정의한다. 이런 새로운 은닉층 경사 벡터를 이용하여 은닉층의 가중치를 변경한다. 따라서, 제안한 방법은 오류역전파 학습에서, 은닉층의 경사 벡터의 크기가 작아 학습이 지연되는 것을 방지할 수 있다.

필기체 숫자인식 문제를 대상으로 한 실험에서, 제안한 방법의 학습 속도는 오류역전파 학습과 수정된 오차 함수의 학습보다는 빠르고, Ooyen의 방법, Wang의 방법, Yam의 방법보다는 비슷하거나 우수하였다. 또한, 일반화 성능은 은닉노드의 수에 관련없이 가장 좋은 결과를 얻었다. 이것은 제안한 방법이 학습 속도에서는 계층별 학습의 장점을 취하고, 일반화 성능에서는 오류역전파 학습과 수정된 오차 함수의 장점을 취하고 있기 때문이다. 또한, 제안한 방법은 구현이 간단하고 학습 속도가 빠르며, 일반화 성능이 우수한 장점을 가지고 있다.

참 고 문 헌

[1] D. E. Rumelhart and J. L. McClelland, *Parallel Distributed Processing,* MIT Press, Cambridge, MA, pp.318-362, 1986.

[2] R. P. Lippmann, "An Introduction to Computing with Neural Nets," *IEEE ASSP Magazine,* pp. 4-22, April, 1987.

[3] Dan Hammerstrom, "Working with Neural Networks," *IEEE Spectrum,* pp.46-53, July, 1993.

[4] A. Van Ooyen and B. Nienhuis, "Improving the convergence of the back-propagation algorithm," *Neural Networks,* vol. 78, pp.465-471, 1992.

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

[6] Sang-Hoon Oh and Youngjik Lee, "A modified error function to improve the Error Back-Propagation algorithm of Multi-Layer Perceptrons," *ETRI Journal,* vol. 17, pp.11-22, April, 1995.

[7] C. M. Bishop, *Neural Networks for Pattern Recognition,* Clarendon Press, Oxford, 1997.

[8] G.-J. Wang and C.-C. Chen, "A Fast Multilayer

Neural-Network Training Algorithm Based on the Layer-By-Layer Optimizing Procedures," *IEEE Trans. Neural Networks*, vol. 7, pp.768-775, May, 1996.

[9] Jim. Y. F. Yam and Tommy W. S. Chow, "Extended Least Squares Based Algorithm for Training Feedforward Networks," *IEEE Trans. Neural Networks*, vol. 8, pp.806-810, May, 1997.

[10] R. S. Scalero and N. Tepedelenlioglu, "A Fast New Algorithm for Training Feedforward Neural Networks," *IEEE Trans. Signal Processing*, vol. 40, pp.202-210, Jan. 1992.

[11] S. Ergezinger and E. Thomsen, "An Accelerated Learning Algorithm for Multilayer Perceptrons: Optimization Layer by Layer," *IEEE Trans. Neural Networks*, vol. 6, pp.31-42, Jan. 1995.

[12] David J. Winter, *Matrix Algebra*, Macmillan Publishing Company, 1992.

[13] James M. Ortega, *Matrix Theory*, Plenum Press, 1987.

[14] J. J. Hull, A database for handwritten text recognition research, *IEEE Trans. Pattern and Machine Intell.*, vol. 16, pp.550-554, 1994.

[15] 오상훈, 이영직, 김명원, "역전파 학습시 초기 가중치가 학습의 조기 포화에 미치는 영향", 전자공학회 논문지, 제28권 4호, pp.90-97, 1991.

곽 영 태

1993년 충남대학교 컴퓨터공학과 공학사. 1995년 충남대학교 컴퓨터공학과 공학석사. 2001년 충남대학교 컴퓨터공학과 공학 박사. 관심분야는 패턴 인식, Neural Networks, 컴퓨터 비젼, 영상 처리

이 영 직

1979년 서울대학교 전자공학과 공학사. 1981년 한국과학기술원 산업전자과 공학석사. 1981년 3월 ~ 1984년 6월 (주) 삼성전자 근무. 1989년 Polytechnic University (Brooklyn) 공학 박사. 1989년 1월 ~ 현재 한국전자통신연구원 책임연구원. 관심분야는 음성인식, Neural Networks, Array Signal Processing 등

권 오 석

1977년 서울대학교 전자공학과 공학사. 1980년 한국과학기술원 전기 및 전자공학과 공학 석사. 1994년 한국과학기술원 전기 및 전자공학과 박사수료. 1980년 ~ 현재 충남대학교 컴퓨터공학과 교수. 관심 분야는 퍼지 이론 및 응용, Neural Networks, 디지털 회로, 컴퓨터 비젼 등