패턴분류에서 학습방법 개선

Improvement of Learning Method in Pattern Classification

김 명 찬, 최 종 호
(Myung-Chan Kim and Chong-Ho Choi)

Abstract : A new algorithm is proposed for training the multilayer perceptrion(MLP) in pattern classification problems to accelerate the learning speed. It is shown that the sigmoid activation function of the output node can have deterimental effect on the performance of learning. To overcome this detrimental effect and to use the information fully in supervised learning, an objective function for binary modes is proposed. This objective function is composed with two new output activation functions which are selectively used depending on desired values of training patterns. The effect of the objective function is analyzed and a training algorithm is proposed based on this. Its performance is tested in several examples. Simulation results show that the performance of the proposed method is better than that of the conventional errror back propagation (EBP) method.

Keywords : MLP, backpropagation, objective function, sigmoid activation function, learning speed

I. 서론

최근 20여년 동안 인공 신경회로망의 일종인 다층인식자(multilayer perceptron, MLP)[1-3]에대한 연구가 국내외에서 활발히 진행되어 왔다. 다층인식자는 패턴분류문제, 선형·비선형 함수 근사화, 비선형 시스템 식별 또는 제어, 시계열 예측등의 다양한 분야에 적용된다. 이들 중에서 패턴분류 문제는 다층인식자가 가장 성공적으로 적용되는 분야중의 하나이다.

다층인식자는 입력 데이터들로부터 입출력 관계를 학습한다. 다층인식자의 학습에 가장 많이 사용하는 방법은 오차역전파(error back propagation, EBP)[1-3] 방법이다. 그런데 이 방법을 사용할 때 문제가 복잡해지면 학습을 제대로 못하거나 학습에 소요되는 시간이 매우 길어지는 경우가 자주 발생한다. 따라서 오류역전파의 느린 학습속도를 개선시키려는 연구가 많이 수행되고있다. 학습속도를 향상시키는 방법으로 가변 학습률을 이용하는 방법[4-6], 함수 최적화방법을 도입하여 목적함수의 2차 미분항을 이용하는 방법[4][7], 그리고 이외에도 여러 가지 오차역전파 학습 알고리즘의 변형들이 연구 발표되었다[8-24].

패턴분류문제에서 다층인식자의 출력노드 활성함수(activation function)인 시그모이드 함수의 기울기 특성이 오차역전파 학습속도를 느리게 하는 한 원인이 된다.

특히 학습과정에서 학습 데이터에 대한 다층인식자 출력이 목표출력값의 반대쪽 값을 가지는 경우는 시그모이드 함수 기울기 특성때문에 오차가 매우 큼에도 불구하고 아주 작은 오차분담값이 역전파된다. 이러한 현상을 없애기 위해, Rezgui등[23]은 시그모이드 함수의 기울기도 학습시켜야할 다층인식자의 변수로 보고, 연결값과 함께 학습시키는 방법을 제안하였다. 그러나 이 방법은 표준 오차역전파 학습 알고리즘에서 연결값을 갱신시킬 때 상수배를 해준 것에 불과하며, 계산량이 많아지는 단점이 있다.

Ooyen등[24]은 다층인식자의 출력노드 활성함수로 시그모이드 함수를 그대로 사용하면서, 연결값 갱신식에 시그모이드 함수의 기울기 항을 제거하기 위한 새로운 목적함수를 제안하였다. Ooyen등이 제안한 목적함수는 로그(log)함수를 도입함으로써 각 출력노드에서 오차분담값을 계산할 때 시그모이드 기울기항을 사라지게하여, 목적함수의 기울기가 목표출력값과 다층인식자의 출력값의 차이의 상수배가 되게 한다.

오차역전파 학습 알고리즘은 대표적인 지도학습 알고리즘으로, 학습에 사용되는 모든 입력패턴들과 그들에 대응되는 목표출력값들이 학습전에 정해져 있다. 기존의 오차역전파 방법은 출력노드에서 오차제곱의 합을 목적함수로 보고, 그 목적함수를 계산할 때 목표출력값을 이용하지만, 목표출력값이 가지는 정보를 충분히 활용하지 못한다. 그리고 출력노드 활성함수인 시그모이드 함수의 특성상, 오차역전파 방법은 중반부 이후에 느린 학습구간이 발생하는 원인이 된다. 이러한 목표출력값과 출력노드의 시그모이드 함수 특성은 경우에 따라서 목적함수의 전역최소값(global minimum)을 구하지만 전체 학습 데이터들을 분리하지 못하는 상황이 발생하기도 한다[25].

본 논문에서는 오차역전파 방법이 지도학습이라는 점을 충분히 활용하는 새로운 목적함수와 출력노드의 새로운 활성함수를 제안하며, 이로부터 효율적인 학습방법을 제안한다. 여러 실험에서 제안된 방법이 중반부 이후에 발생하는 느린 학습구간을 단축시킴을 보인다. 제안된 방법에서는 학습 데이터에 대응하는 목표출력값들을 충분히 활용하기위해 각 출력노드마다 목표출력값에 따라서 해당되는 활성함수를 다르게 정한다.

제안된 출력노드 활성함수들로 구성된 목적함수를 사용하면, 학습 데이터가 잘못 분류된 경우에도 시그모이드 함수 때보다 큰 일정한 기울기 성분이 유지된다. 따라서 제안된 목적함수는 학습 데이터가 잘못 분류된 경우에 연결값 갱신성분값이 작아지는 기존의 오차역전파 학습방법의 단점을 보완하여 학습 속도를 개선시킨다. 또한 이 목적함수를 사용하면, Brady등 [25]이 열거한 오차역전파 방법이 패턴분류에 실패하는 경우들을 없앨 수 있다.

본 논문의 구성은 다음과 같다. 오차역전파 학습 알고리즘에 대한 간단한 설명과 각종 부호에 대한 정의를 제 II장에서 한다. 기존의 목적함수의 문제점 및 새로운 목적함수와 그 특징을 제 III장에서 기술한다. 제안된 목적함수의 효율성을 보이기 위한 모의실험은 제 IV장에서 보이고, 결론

접수일자 : 97. 7. 25., 수정완료 : 97. 9. 30.
김명찬 : 삼성 데이터 시스템 (주)
최종호 : 서울대학교 전기공학부
* 본 논문은 1996년도 한국학술진흥재단의 공모과제 연구비에 의하여 연구되었습니다.

을 제 V장에서 내린다.

II. 오차역전파 학습 알고리즘

본 장에서는 가장 많이 사용되는 관성항(momentum)을 포함한 오차역전파 알고리즘 및 기호에 대해서 간단히 설명한다. 각 층을 0(입력층)에서 L(출력층)까지 표시하고 l번째 층의 노드수를 N_l로 표시하자. 또한 p번째 학습 데이터에 의한 l번째 층의 i번째 노드 출력을 O_{ip}^l로 나타내자. 그러면 p번째 학습 데이터에 대한 출력층에서 오차 E_p를 다음과 같이 정의하자.

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

여기서 D_{ip}는 p번째 입력에 대한 출력층에서 i번째 노드의 목표출력값(desired value)이다.

패턴모드(pattern mode) 오차역전파 학습에서는 (1)을 목적함수로 사용한다. 그리고 이 목적함수를 최소로 만드는 연결값을 얻기 위해 매번 새로운 학습 데이터가 입력될 때마다 경사감소법(steepest descent method)를 사용하여 다음과 같이 연결값을 갱신한다[1].

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

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

여기서 k는 연결값이 갱신된 횟수, η는 학습률(learning rate), μ는 관성항을 의미한다. 그리고 (2)와 (3)에서 $W(k)$는 $(k-1)$번째 갱신후의 연결값 벡터로서 $(l-1)$번째 층의 j번째 노드에서 l번째 층의 i번째 노드로의 연결값 $w_{ij}^{l-1}(k)$와 l번째 층 i번째 노드 바이어스 $b_i^l(k)$로 구성된다. 즉 $W(k)=[b_1^1(k), w_{11}^0(k), \cdots, b_i^l(k), \cdots, w_{ij}^{l-1}(k), \cdots, w_{N_lN_{L-1}}^{L-1}(k)]^T$ 이다. 여기서 $[\cdot]^T$는 $[\cdot]$의 전치행렬이다. 또한 $\Delta W(k)$는 k번째 연결값 갱신벡터를 의미한다. 한편 (2)에서 기울기 벡터의 각 성분들은 다음과 같다.

$$\Delta w_{ij}^{L-1}(k) = -\frac{\eta}{N_L}\delta_{ip}^L O_{jp}^{L-1} + \mu\Delta w_{ij}^{L-1}(k-1) \tag{4}$$

$$\Delta w_{ij}^{l-1}(k) = -\frac{\eta}{N_L}\delta_{ip}^l O_{jp}^{l-1} + \mu\Delta w_{ij}^{l-1}(k-1) \tag{5}$$

여기서,

$$\delta_{ip}^L = (D_{ip}-O_{ip}^L)O_{ip}^L(1.0-O_{ip}^L),$$

$$\delta_{jp}^l = \sum_{s=1}^{N_{l+1}}\delta_{sp}^{l+1}w_{sj}^l(k)\,O_{jp}^l(1-O_{jp}^l)$$

이다. 갱신식 (4)과 (5)를 보면, η/N_L은 크기 조정이 된 학습률로 볼 수 있다. 따라서 모의실험시 계산상의 편의를 위해 패턴모드 오차역전파 학습의 경우 η 대신에 $\eta_p(=\eta/N_L)$를 조정한다.

배치모드 오차역전파 학습은 모든 학습 데이터들에 대한 출력결과를 보고, 이에 따라 연결값을 한번에 갱신시켜 나간다. 배치모드 오차역전파 학습에서 사용되는 목적함수 E_{bat}와 연결값 갱신식은 다음과 같다.

$$E_{bat} = \frac{1}{P}\sum_{p=1}^{P}E_p \tag{6}$$

$$\Delta W(n) = -\eta\frac{\partial E_{bat}}{\partial W(n)} + \mu\Delta W(n-1) \tag{7}$$

$$W(n+1) = W(n) + \Delta W(n) \tag{8}$$

여기서 P는 총 학습 데이터 수이며, n는 연결값 갱신횟수(epoch)를 의미한다. 개별적인 연결값의 갱신식을 패턴모드 경우와 마찬가지 방법으로 정리하면 다음과 같다.

$$\Delta w_{ij}^{L-1}(k) = -\frac{\eta}{PN_L}\sum_p\delta_{ip}^L O_{jp}^{L-1} + \mu\Delta w_{ij}^{L-1}(k-1) \tag{9}$$

$$\Delta w_{ij}^{l-1}(k) = -\frac{\eta}{PN_L}\sum_p\delta_{ip}^l O_{jp}^{l-1} + \mu\Delta w_{ij}^{l-1}(k-1) \tag{10}$$

배치모드 오차역전파 경우에도 $\eta/(PN_L)$를 크기 조정이 된 학습률로 볼 수 있다. 모의실험에서 패턴모드 오차역전파 학습의 경우와 마찬가지로 계산상의 편의를 위해 η대신 $\eta_b\ (=\eta/(PN_L))$를 조정하면서 실험한다.

III. 두 형태용 목적함수 (An objective function for binary modes)

목적함수 (1) 또는 (6)을 사용하여 패턴분류 문제를 학습할 때 발생하는 문제에 대해 생각하자. 일반적으로 다층인식자가 오차역전파 학습에의해 패턴분류문제를 학습할 때, 은닉층 및 출력층 노드에서 시그모이드 함수가 활성함수로 사용된다. 시그모이드 함수는 두 개의 극한값(1과 0)을 갖는데, 극한값 주변에서는 시그모이드 함수 기울기가 거의 0인 영역이 존재하는데 이를 포화영역이라 하자. 이러한 시그모이드 함수특성은 오차역전파 학습과 결합하여, 다층인식자가 패턴분류문제를 학습하는데 있어서 때때로 나쁜 영향을 미친다[3].

시그모이드의 두 극한값 0과 1을 목표출력값으로 사용한다면, 극한값 근처의 포화영역 때문에 시그모이드가 목표출력값을 출력하는데 많은 학습횟수를 필요로 한다. 이러한 현상을 방지하기 위해서, 대부분의 경우에 0과 1대신에 0.1과 0.9(또는 0.05와 0.95)를 목표출력값으로 사용한다[22]. 그러나 학습과정 중에 종종 어떤 학습 데이터는 잘못 분류되는 경우(예를 들면 목표 출력값이 0.1(0.9)보다 작은(큰) 경우)가 나타나는데 이는 학습과정에서 매우 바람직하지 않다. 왜냐하면 목적함수에서 오차값은 크지만, 시그모이드 함수의 기울기로인해 연결값을 갱신하는데 필요한 오차분담값들은 아주 작기 때문이다. 따라서 다층인식자의 출력값이 틀린 상태에서 빠져 나오는데 많은 학습횟수가 필요하게 된다.

또한 다음과 같은 경우도 발생하여 학습 중반부 이후에 학습속도가 느려지는 현상이 나타난다. 예를들면 학습 데이터 $P_k(k=1,2)$에 대한 다층인식자 출력이 각각 0.95, 0.84이고, 목표출력값은 모두 0.9라 하자. 그러면 출력값이 0.9보다 큰 P_1에 대한 i번째 출력층 노드에 연결되는 연결값 w_{ij}^{L-1}에 대한 목적함수의 기울기 성분은 다음과 같다.

$$\frac{\partial E_{P_1}}{\partial w_{ij}^{L-1}} = -\frac{1}{N_L}(0.9-0.95)0.95(1-0.95)O_{jP_1}^{L-1} > 0 \tag{11}$$

시그모이드 출력과 기울기항은 항상 양수이므로 P_1에 대한 패턴별 목적함수 E_{P_1}의 기울기가 양수임을 (11)에서 알 수 있다. 따라서 P_1의 역할은 w_{ij}^{L-1}값을 감소시켜서 출력값이 0.9가 되도록 한다. 그러나 출력값이 0.9보다 작은 데이터 P_2에 대한 E_{P_2}의 기울기성분은

$$\frac{\partial E_{P_2}}{\partial w_{ij}^{L-1}} = -\frac{1}{N_L}(0.9-0.84)0.84(1-0.84)O_{jP_2}^{L-1} < 0 \tag{12}$$

으로 데이터 P_2에 대응되는 목적함수 기울기는 음수이다.따라서 P_2는 w_{ij}^{L-1}를 증가시켜 출력값이 0.9에 도달하도록 역할한다.

그런데 배치모드 오차역전파 방법에서는 각 입력에 의해

유발된 목적함수 기울기들의 합이 연결값들이 갱신되어가는 방향과 양을 결정한다. 결국 위의 예에서 P_1에 의해 유발된 기울기 값과 P_2에 의해 유발된 기울기 값이 서로 반대 부호를 가짐으로 P_2가 학습에 끼치는 역할은 감소한다. 따라서 P_2의 경우 목표출력값 0.9를 내보내는 데 많은 학습 횟수가 필요하다. 실용적인 입장에서 보면 P_1은 이미 학습이 완료된 상태이다. 따라서 P_1에 의해 유발된 목적함수 기울기는 P_2의 학습을 위해서 0이 되어도 다층인식자의 학습 성능에는 무관하다. 한편 목표 출력값이 0.1이고 위와 비슷한 경우 (출력값이 0.1 미만인 패턴들과 0.1 이상인 패턴들이 같이 존재하는 상황)에 대해서도 비슷한 해석이 가능하다.

이러한 현상이 발생하는 원인은 출력층에서 시그모이드 함수의 출력 범위가 (0, 1)로서, 목표 출력값인 0.9와 0.1뿐만 아니라 0.9 이상 또는 0.1 이하의 값을 출력할 수 있다는 것이다. 이러한 현상의 극단적인 경우로 Brady등[25]은 어떤 패턴분류문제를 기존 오차역전파 방법으로 학습할 때, E_{bal}의 전역최소값을 주는 연결값 벡터를 찾더라도, 그 벡터는 학습 데이터들을 올바로 분리하지 못하는 예들을 제시하여 오차역전파 방법의 한계를 지적하였다.

학습 중반부에 학습속도가 느려지거나 학습이 다 되더라도 패턴분류를 제대로 못하는 이유는 오차역전파 방법이 지도학습임에도 불구하고, 학습시 각 입력 패턴에 대응하는 출력값을 계산할 때, 대응하는 목표출력값이 갖는 정보를 충분히 사용하지 않고 있기 때문이다. 본 논문에서는 앞서의 문제점들을 없애기 위해 새로운 목적함수를 제안하고 그로부터 학습방법을 유도한다. 제안된 목적함수는 다음과 같은 상황에 그 근거를 두고 있다. 모든 학습 데이터들은 각 출력노드 i 에서 목표출력값이 "high"인 집합 C_{i+}와 "low"인 집합 C_{i-}으로 양분되며, 학습 데이터들과 그들 각각에 대응하는 목표출력값들은 학습 시작전에 정해져 있다. 따라서 각 출력노드별로 목표출력값이 "high"인 집합의 활성함수와 목표출력값이 "low"인 집합의 활성함수를 다르게 정할 수 있다. 즉 지도학습이라는 잇점을 최대한 활용하기 위하여, 각 출력노드에서 매 학습 데이터에 대응하는 출력을 계산할 때, 목표출력값에 따라 서로 다른 두 활성함수를 사용하는 것이다.

배치모드 오차역전파 학습을 위해 다음과 같은 두 형태용 목적 함수 $E_o(n)$을 제안한다.

$$E_o(n)=\frac{1}{2PN_L}\sum_{i=1}^{N_L}\Big\{\sum_{p\in C_{i+}}(D_+-O_{ip+})^2+\sum_{p\in C_{i-}}(D_--O_{ip-})^2\Big\} \quad (13)$$

여기서 D_+와 D_-는 "high"와 "low" 목표출력값을 의미한다. E_o에서는 C_{i+}와 C_{i-}에 속하는 데이터들에 대해서 각각 활성함수 $O_{ip+}(\cdot)$와 $O_{ip-}(\cdot)$를 사용한다. 위의 E_o는 기존 목적함수 E_{bal}의 각 출력노드별로 정리한 것이며, 출력노드 활성함수로는 시그모이드 함수대신 $O_{ip+}(\cdot)$와 $O_{ip-}(\cdot)$를 사용한 것이다. 유사하게 패턴모드 오차역전파 학습을 위한 목적함수로는 다음과 같은 E_{op}를 제안한다.

$$E_{op}(k) = \frac{1}{2N_L}\sum_{i=1}^{N_L}\Big\{(D_+-O_{ip+})^2|_{p\in C_{i+}} + (D_--O_{ip-})^2|_{p\in C_{i-}}\Big\} \quad (14)$$

활성함수 $O_{ip+}(\cdot)$와 $O_{ip-}(\cdot)$는 다음과 같은 성질을 갖도록 한다. 우선 목표출력값 근처에서는 포화영역이 존재하나, 목표출력값의 반대편에는 시그모이드 함수와 달리 포화영역없이 일정한 기울기를 갖는 일차함수 형태를 취한다. 이것은 잘못 분류된 학습 데이터가 있는 경우 기존의 시그모이드 활성함수가 갖는 학습과정의 문제점을 개선시킨다. 출력노드의 활성함수로 시그모이드함수를 사용하여 출력이 D_-와 D_+사이에 있게 되는 경우에는 제안된 활성함수들은 시그모이드와 같게한다.

위와 같은 성질을 갖는 출력층 노드 활성함수 $O_{ip+}(\cdot)$와 $O_{ip-}(\cdot)$로 다음과 같은 함수를 택할 수 있다.

$$O_{ip+}(x)=\begin{cases} D_+ & x\ge a \\ \dfrac{1}{1+e^{-x}} & b<x<a \\ S_+(x-b)+D_- & x\le b, \end{cases} \quad (15)$$

$$O_{ip-}(x)=\begin{cases} S_-(x-a)+D_+ & x\ge a \\ \dfrac{1}{1+e^{-x}} & b<x<a \\ D_- & x\le b \end{cases} \quad (16)$$

여기서,

$$S_+=D_-(1-D_-),\quad S_-=D_+(1-D_+),$$
$$a=-\ln(1/D_+-1),\quad b=-\ln(1/D_--1)$$

이며, x는 p번째 데이터에 대한 i번째 출력노드의 가중치 합(weighted sum)이다. 그리고 S_+와 S_-는 각각 시그모이드 함수값이 D_-와 D_+일 때 기울기이며, a와 b는 시그모이드 함수값이 D_+와 D_-일 때의 가중치 합이다.

위의 (13)과 (14)에서 알 수 있듯이, 제안된 목적함수에서 사용하는 출력층의 활성함수들은 목표출력값에 따라 $O_{ip+}(\cdot)$와 $O_{ip-}(\cdot)$를 선택적으로 사용하므로 목표출력값에 대한 정보를 기존의 목적함수보다 훨씬 효과적으로 이용한다. 시그모이드와 제안된 활성함수 $O_{ip+}(\cdot)$와 $O_{ip-}(\cdot)$를 그림 1 및 그림 2에 보인다. 그림 1과 그림 2를 보면, 시그모이드는 두개의 포화영역을 출력 0과 1근처에서 가지고 있지만, 제안된 활성함수 $O_{ip+}(\cdot)$와 $O_{ip-}(\cdot)$는 하나의 포화영역을 목표출력값 근처에서만 갖는다. 그리고 $a\le x\le b$에서는 시그모이드 함수, 활성함수 $O_{ip+}(\cdot)$와 $O_{ip-}(\cdot)$는 같다. 또한 $O_{ip+}(\cdot)$는 x가 b보다 작으면 x에 비례하여 감소함으로써 x가 b보다 작을수록 그 영향이 목적함수에 크게 작용할 수 있다. 활성함수 $O_{ip-}(\cdot)$도 유사한 특성을 갖는다. 두 형태용 목적함수를 사용함에 있어서, 출력층을 제외한 나머지 은닉층에서는 활성함수로는 시그모이드 함수를 사용한다. 그리고 연결값 갱신식은 오차역전파 방법((4)와 (5) 또는 (9)와 (10))을 사용한다.

제안된 활성함수는 목표출력값을 내보내는 정확한 가중치 합($x=a$ 또는 $x=b$)에서는 미분이 불가능하다. 그러나 x가 a 또는 b와 같은 확률은 거의 0이므로, 이를 무시해도 제안된 목적함수를 사용하는데는 문제가 없다. 따라서 활성함수들의 1차 미분은 다음과 같다.

$$\frac{\partial O_{ip+}(x)}{\partial x}=\begin{cases} 0 & x\ge a \\ O_{ip+}(x)(1-O_{ip+}(x)) & b<x<a \\ S_+ & x\le b \end{cases} \quad (17)$$

$$\frac{\partial O_{ip-}(x)}{\partial x}=\begin{cases} S_- & x\ge a \\ O_{ip-}(x)(1-O_{ip-}(x)) & b<x<a \\ 0 & x\le b \end{cases} \quad (18)$$

제안된 $O_{ip+}(\cdot)$와 $O_{ip-}(\cdot)$의 기울기를 각각 그림 3과 그림 4에 시그모이드 함수의 기울기와 함께 보였다.

시그모이드 기울기는 양쪽 극한으로 갈수록 0에 수렴한다. 그러나 그림 3과 그림 4를 보면 $O_{ip+}(\cdot)$나 $O_{ip-}(\cdot)$는 출력값이 목표출력값과 같은 경우에만 기울기가 0

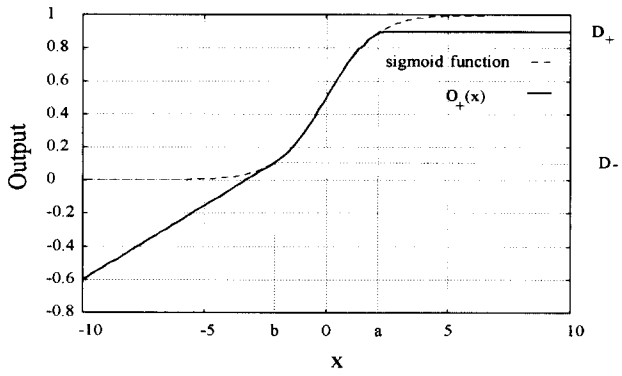

그림 1. $O_{ip+}(\cdot)$와 시그모이드 함수.
Fig. 1. $O_{ip+}(\cdot)$ and sigmoid function.

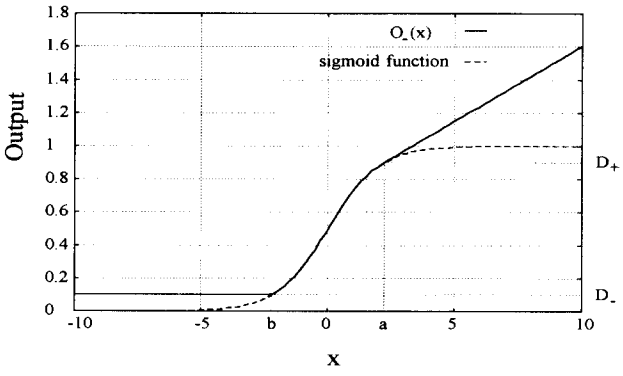

그림 2. $O_{ip-}(\cdot)$와 시그모이드 함수.
Fig. 2. $O_{ip-}(\cdot)$ and sigmoid function.

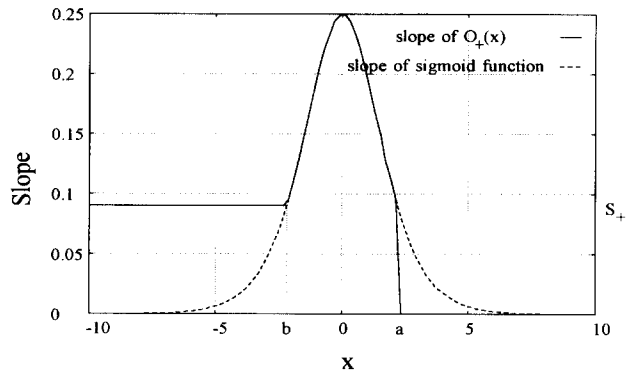

그림 3. $O_{ip+}(\cdot)$와 시그모이드 기울기.
Fig. 3. Slopes of $O_{ip+}(\cdot)$ and sigmoid function.

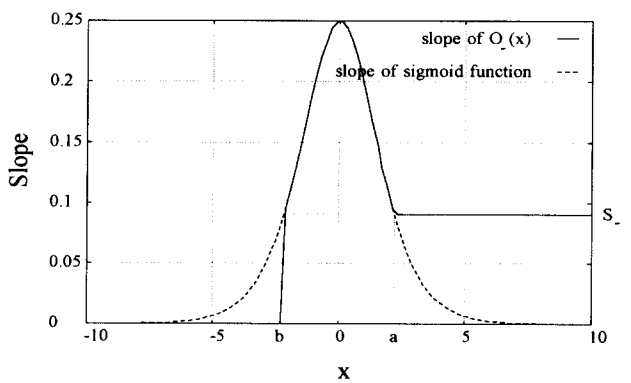

그림 4. $O_{ip-}(\cdot)$와 시그모이드 기울기.
Fig. 4. Slopes of $O_{ip-}(\cdot)$ and sigmoid function.

이고, 그렇지 않은 경우는 각각 기울기가 항상 S_+ 또는 S_-이상이다. 이러한 활성함수의 기울기 특성은, 기존의 오차역전파 학습시 시그모이드 함수의 기울기 특성 때문에 학습이 오래 걸리는 단점을 보완한다.

목적함수 E_o을 앞에서 보인 예에 적용하여, 학습과정에서 어떤 개선점이 발생했는지 알아보자. 앞의 예에서 사용된 데이터들(P_1, P_2)은 C_{i+} 집합에 속해 있다. 따라서 $O_{ip+}(\cdot)$함수만 사용한다. 각 데이터에 대응하는 출력노드에서 가중치 합을 $x_{P_k}(k=1, 2)$ 라 하자. 그러면 데이터 P_1에 대한 목적함수 기울기는

$$\frac{\partial E_{oP_1}}{\partial w_{ij}^{L-1}} = -\frac{1}{N_L}\ (0.9 - O_{iP_1+}(x_{P_1}))\ O_{iP_1+}(x_{P_1})(1 - O_{iP_1+}(x_{P_1}))O_{jP_1}^{L-1} = 0$$

이다. 왜냐하면 P_1에 대응되는 O_{iP_1+}는 0.9이며, 기울기는 0이기 때문이다. 한편 P_2에 대응되는 기울기는

$$\frac{\partial E_{oP_2}}{\partial w_{ij}^{L-1}} = -\frac{1}{N_L}(0.9-0.84)0.84(1-0.84)O_{jP_2}^{L-1} < 0$$

이다. 따라서 제안된 목적함수를 사용하면 P_1에 대응하는 기울기는 0이므로 더 이상 P_2의 학습에 나쁜 영향을 줄 수 없다. 결국 제안된 목적함수는 기존의 목적함수가 가지고 있던 문제점을 해결할 수 있다.

제안된 목적함수를 앞에서 언급한 Brady등[25]이 제시한 예제에 적용하여 오차역전파 방법을 통해 전역최소값을 구할 수 있으면 그것이 학습 데이터들을 올바로 분리해내는 연결값 벡터가 됨을 모의실험상으로 보일 수 있다. 따라서 [25]에서 제시한 예들에 대해 오차역전파 방법으로 제대로 동작하는 해를 찾을 수 있음을 보일 수 있다[26]. 두 형태용 목적함수는 학습속도를 증가시키기 위하여 지금까지 개발된 여러 종류의 학습방법과 결합하여 사용할 수 있다.

제안된 새로운 목적함수는 다층인식자의 학습과정에만 사용한다. 학습후, 다층인식자의 성능시험에서는 두 형태용 목적함수를 사용할 수 없다. 왜냐하면, 성능시험에서는 주어진 데이터에대한 목표출력값을 미리 알 수 없기 때문이다. 이때는 출력노드의 활성함수를 시그모이드 함수로 대체한다. 두 형태용 목적함수를 사용하여 다층인식자를 학습시키는 경우 출력층을 제외한 나머지 은닉층에서 활성함수로는 시그모이드 함수를 사용하기 때문에, 성능시험을 위해 출력층 노드의 활성함수를 시그모이드 함수로 바꾼다 해도 이미 학습된 다층인식자의 성능은 그대로 보존된다. 그 이유는 다음과 같다.

어떤 다층인식자의 연결값들이 두 형태용 목적함수로 학습되었다고 하자. 출력층 i 번째 노드에서 목표출력값이 "high"인 P_k에 대해 O_{iP_k+}가 0.9라 하자. 이때 i 번째 출력노드로 들어가는 가중치 합은

$$\sum_{j=1}^{N_{L-1}} w_{ij}^{L-1} O_j^{L-1} + b_i^L > a$$

이다. 만약 i번째 출력노드 활성함수가 시그모이드 함수로 바뀌었다 하자. 데이터 P_k에 대한 은닉층 출력은 변화가 없으므로, i번째 출력노드로 들어가는 가중치 합은 변화없다. 시그모이드는 단조증가 함수이므로 i번째 출력노드에서 시그모이드는 "high" 목표출력값보다 큰 값을 출력한다. 이는 패턴분류문제의 특성상 P_k는 제대로 학습된 것으로 간주한다. 마찬가지 방법으로 목표 출력값이 "low"인 경우에 대해서도 유사한 설명을 할 수 있다.

한편 제안된 목적함수에 의해 P_k가 잘못 분류되었다고 하자. 즉 P_k에 대한 O_{iP_k+}가 D_-이하인 경우를 생각하자.

이때 i번째 출력노드로 들어가는 가중치 합은

$$\sum_{j=1}^{N_1} w_{ij}^{L-1} O_j^{L-1} + b_i^L < b$$

이다. 출력층 활성함수가 시그모이드 함수로 바뀌어도 i번째 출력노드로 들어가는 가중치 합은 변화가 없고, 시그모이드 함수는 단조 증가함수이기 때문에 시그모이드 함수 역시 D_-값 이하를 출력한다.

그리고 패턴 P_k에 대해 제안된 목적함수에 따른 O_{iP_k+} 또는 O_{iP_k-}가 D_+와 D_-사이일 때 (출력노드로 들어가는 가중치 합이 시그모이드 함수의 a와 b사이에 있으면), 출력층 활성함수가 시그모이드 함수로 바뀌더라도 가중치 합은 변화없으므로, 출력값은 변하지않는다. 따라서 제안된 목적함수로 학습 후 출력노드 활성함수를 시그모이드로 바꾸어도 다층인식자의 분류성능은 그대로 보존된다.

IV. 모의 실험

모의실험에서는 목적함수 E_o나 E_{ot}와 기존의 목적함수 E_{bat}나 E_t를 사용하여 학습하였을 때의 성능을 비교한다. 학습 성공조건은 매우 엄격하게 하여 "high" 목표출력값을 갖는 출력노드 $(D_+ - 0.01)$보다 큰 값을, "low" 목표출력값을 갖는 출력노드는 $(D_- + 0.01)$보다 작은 값을 출력할 경우만 학습이 성공한 것으로 보았다. 나머지 학습 조건인 다층인식자의 구조, 학습률, 관성항 및 학습용 입력 데이터, 그리고 초기 연결값들은 예제마다 다르지만, 같은 예제에서는 동일한 조건을 갖도록 하였다. 초기 연결값들은 모두 그 절대값이 0.1보다 작은 범위에서 무작위로 선택하였다.

1. 이중 나선 (2-spirals) 문제

패턴분류 문제중에서 이중 나선 문제는 학습방법의 성능을 판단하는 대표적인 문제이다[10]. 이 문제를 기본적인 MLP 구조로 학습에 성공한 경우는 거의 없다. 학습에 사용된 다층인식자는 입력노드 2개, 은닉노드 70개, 그리고 출력노드 1개로 구성된다. 배치모드 학습에서 η_t는 0.02, μ는 0.9를 사용한다. 그리고 최대 학습횟수는 50,000번으로 제한하고, 5개의 서로 다른 초기 연결값 집합에 대해 실험하였다. 총 5번의 시도중 기존의 목적함수를 사용한 경우는 학습 제한횟수 이내에 한번도 학습이 안되었고, 제안된 목적함수를 사용하는 경우는 5번 모두 성공적으로 학습되었으며 평균 학습횟수는 39,518회(epoch)였다. 그림 5에 학습곡선이 나타나 있다. 학습 초반부에 두 목적함수는 비슷한 학습곡선을 가지나, 중반부터 두 목적함수 사이에 많은 차이가 나타난다. 이러한 관찰로부터, 제안된 목적함수를 사용함으로써 학습성능이 향상되는 것은 학습 중반이후라는 것을 알 수 있다. 그림 6에서는 학습 진행과정에서 "high" ("low")로 분류될 것이 "low"("high")로 잘못 분류된 데이터들의 평균 갯수들을 나타내었다. 여기서 출력값이(D_+-0.01)과(D_-+0.01) 사이에 있는 것은 학습이 진행중인 것으로 보며 잘못 분류된 데이터에는 포함되지 않았다. 그림 6을 관찰하면, 제안된 목적함수를 사용하는 다층인식자는 5,000번 부근에서 잘못 분류된 데이터들이 발생하기 시작하여, 14,000번 부근에서 잘못 분류된 입력패턴들의 발생갯수가 제일 많고, 이후 점차 감소 추세에 있다. 이는 제안된 활성함수의 기울기 특성에서 그 원인을 찾을 수 있다. 그러나 기존의 목적함수를 사용하는 경우는 5,000번 부근에서 잘못 분류된 입력 데이터들이 발생하여 계속 증가 추세에 있다가, 학습 중반부에 잘못 분류된 입력패턴들의 갯수가 일정하게 유지되고 30,000번 이후에 점차 떨어지는 추세이다. 이는 시그모이드 기울기 특성 때문에 잘못 분류된 데이터의 갯수가 일정하게 유지된 후, 점차 떨어지는 현상이 늦게 나타나기 때문이다.

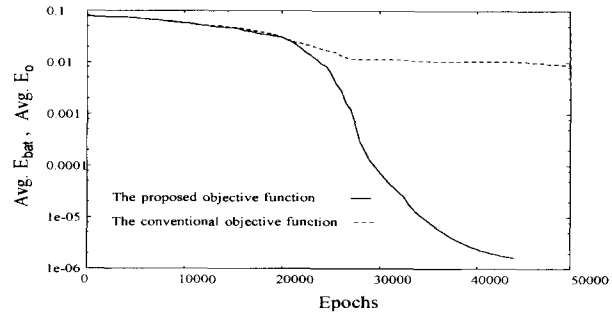

그림 5. 평균 학습곡선.
Fig. 5. Average learning curves.

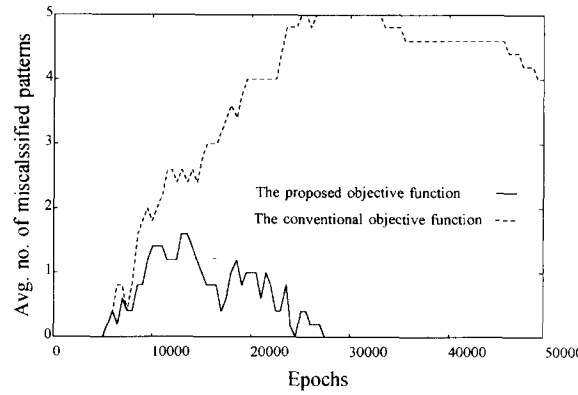

그림 6. 잘못 분류된 패턴수의 평균.
Fig. 6. Number of misclassified patterns.

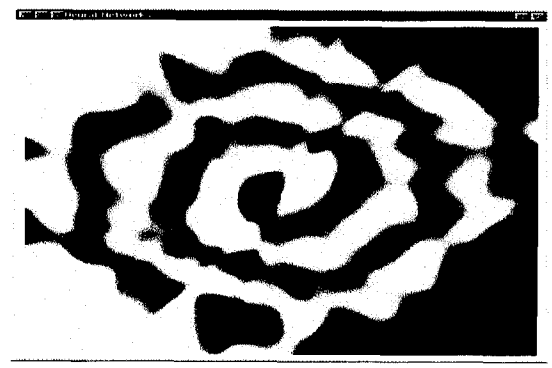

그림 7. 일반화 성능시험의 예.
Fig. 7. A test example of generailization.

제안된 목적함수로 학습된 다층인식자에 대해 출력노드 활성함수를 시그모이드 함수로 바꾸어서 일반화 테스트를 한 결과를 그림 7에 나타냈다.

한편 패턴모드 오차역전파 학습을 같은 문제에 적용해 보았다. 다층인식자의 구조는 배치모드 학습의 경우와 같고, η_p는 0.02, μ는 0.9로 설정하였다. 그러나 두가지 목적함수 둘 다 50,000번 이내에 성공하지 못하였다. 이러한 패턴모드 오차역전파 학습결과는 배치모드의 연결값 갱신방향과 패턴모드의 연결값 갱신방향의 차이에 의한 것으로 보여진다. 그러나 제안된 목적함수를 사용하는 경우에 학습횟수를 더 주면 학습에 성공하는 경우가 발생할 가능성이 있음을 그림 8에서 알 수 있다.

2. 16비트 encoder/decoder 문제

16비트 encoder/decoder 문제를 패턴모드 및 배치모드 오차역전파 학습에대해 모의실험하였다. 다층인식자는 입력노드 16개, 은닉노드 4개, 그리고 출력노드 16개로 구성되었

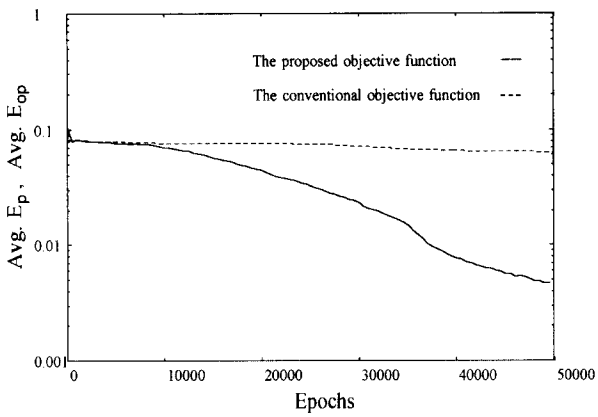

그림 8. 학습곡선의 평균(패턴 모드).
Fig. 8. Learning curves in pattern mode.

다. 그리고 각 학습방법별로 4개의 서로 다른 학습률 및 각 학습률당 5개의 초기 연결값 집합에 대해서 실험하였다. 모든 학습에서 μ는 0.9로 동일하게 설정하고, 최대 학습횟수는 30,000번으로 제한하였다. 배치모드 학습에서 η_b는 각각 0.1, 0.05, 0.025, 0.0125이 사용되었고, 패턴모드 학습에서 η_p는 각각 0.2, 0.1, 0.05, 0.01이 사용되었다.

표 1과 표 2에서 각각 배치모드 학습결과와 패턴모드 학습결과를 나타내었다. 표 1와 표 2에서 괄호안의 숫자는 5번의 학습시도중 학습에 성공한 경우를 의미한다. 표 1에서는 배치모드로 학습하는 경우, 학습에 성공하였을 때의 평균 학습횟수를 나타내었다. 제안된 목적함수를 이용한 배치모드 학습에서는 모든 학습에 성공하였으나 기존의 목적함수를 사용하는 배치모드 학습은 20번의 모든 학습 시도에서 제한된 학습횟수 이내에 성공하지 못했다. 한편 패턴모드 학습결과가 표 2에 나타나 있다. 표 2에서 알 수 있듯이 제안된 목적함수를 사용하는 경우는 η_p가 0.01인 경우를 제외하고는 나머지 모든 학습시도에서 성공적인 결과를 보여주었으나 기존의 목적함수를 사용하는 경우에는 제한된 학습횟수 30,000번 이내에 모두 성공하지 못하였다.

표 1. 배치모드 학습 결과.
Table 1. Results of batch mode learning.

목적함수 종류	학습률 (η_b)			
	0.1	0.05	0.025	0.0125
E_{ba}	- (0/5)	- (0/5)	- (0/5)	- (0/5)
E_o	6945 (5/5)	7469 (5/5)	14096 (5/5)	26707 (5/5)

표 2. 패턴모드 학습결과.
Table 2. Results of pattern mode learning.

목적함수 종류	학습률 (η_p)			
	0.2	0.1	0.05	0.01
E_p	- (0/5)	- (0/5)	- (0/5)	- (0/5)
E_{op}	1757 (5/5)	3580 (5/5)	5893 (5/5)	- (0/5)

16비트 encoder/decoder 문제 학습결과로부터, 제안된 목적함수를 사용한 학습이 넓은 범위의 η_b 또는 η_p에 대해서, 기존의 목적함수를 사용한 경우보다 좋은 성능을 보여줌을 알 수 있다. 그리고 여기에는 제시하지 않았지만 학습곡선을 보면 다중집합 패턴분류 문제에서도 제안된 목적함수는 학습 중반부 이후에 나타나는 느린 학습 구간을 많이 단축시키는 것을 알 수 있다.

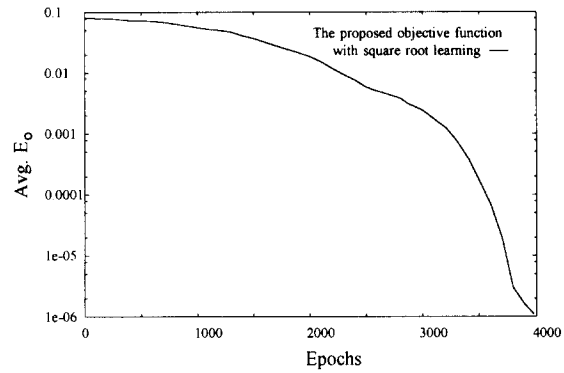

그림 9. 제곱근 학습방법과 병행한 학습곡선.
Fig. 9. Learing curve of square root learning with the proposed objective function.

3. 제곱근 학습 알고리즘[16][19][20]과의 결합

앞의 모의실험에서 보면 두 형태용 목적함수를 이용하는 경우, 학습 초반부는 기존의 목적함수를 사용할 때와 거의 같은 학습곡선을 갖는다는 것을 알 수 있다. 앞의 예에서 기존의 목적함수를 사용하는 경우 학습 초반에 느린 학습구간이 발생하는 원인에 대한 분석이 [20]에 되어 있는데 두 형태용 목적함수를 사용하더라도 같은 분석이 가능하다. 초반 학습속도를 개선하기 위해 가변 학습방법들과 같이 사용하면 효과적이라 생각되어 여기서는 가변 학습방법의 일종인 제곱근 학습방법(square root learning)[20]을 두 형태용 목적함수에 결합하여 사용하였다.

이중 나선 문제를 앞절에서 사용했던 구조를 갖는 다층인식자에 같은 조건하에서 두가지 학습속도 개선방법을 적용하였다. 총 5번의 실험에서 학습이 모두 성공하였고, 평균 학습 횟수는 2,875회였다. 그림 9에 학습곡선의 변화를 보였다. 앞의 이중나선 실험결과와 비교해 보았을 때, 제곱근 학습 알고리즘을 함께 사용함으로써, 학습 초반에 발생하는 느린 학습구간을 단축시켜 많은 학습 속도 개선을 가져왔다. 그리고 학습과정에서 잘못 분류되는 패턴의 갯수 변화를 그림 10에 나타냈다.

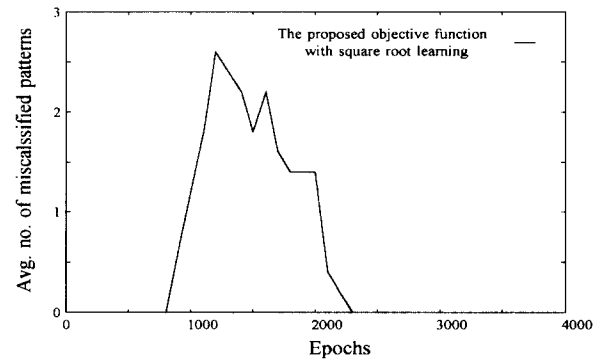

그림 10. 잘못 분류된 패턴수의 평균.
Fig. 10. Average number of misclassified patterns.

IV. 결론

패턴분류 문제에서 오차역전파 학습 알고리즘에 의해 학습된 다층인식자의 응용이 활발하다. 그러나 오차역전파 학

습 알고리즘은 가장 간단한 학습 알고리즘이지만, 학습속도가 느리며, 학습이 실패하는 경우도 많다. 본 논문에서는 패턴분류 문제를 다층인식자가 기존의 배치모드 학습방법에 의해 학습할 때 중반부 이후에 발생하는 느린 학습구간이 발생하는 원인에 대한 분석을 하고, 그 결과를 바탕으로 학습속도 개선방안을 제시하였다.

학습중 특히 중반부 이후에 느린 학습구간이 발생하는 이유는 잘못 분류된 패턴이 발생하는 경우에 시그모이드 함수의 아주 작은 기울기와 목표출력값이 시그모이드 함수의 출력구간에 포함되어 있기 때문이다. 이러한 단점을 없애면서, 지도학습의 잇점을 최대한으로 이용하는 두 형태용 목적함수를 제안하였다. 이 목적함수는 각 출력노드마다 각각의 목표출력값에 대응되는 두 개의 활성함수를 사용하며 목표출력값이라는 정보를 충분히 활용할 수 있게 하였다. 제안된 두 개의 출력층 활성함수는 시그모이드 함수의 간단한 변형으로서, 시그모이드 함수의 단점을 보완해주는 역할을 한다. 여러 가지 모의실험에서 두 형태용 목적함수에서 유도한 학습방법은 학습시 초반부 이후에 나타나는 느린 학습구간을 크게 단축시킴을 보였다. 또한 두 형태용 목적함수는 기존의 목적함수보다 학습률의 크기 변화에 덜 민감한 결과를 보여주고 있다. 그리고 이 목적함수를 사용할 때 학습 초반에 발생하는 느린 학습구간은 가변 학습알고리즘과 함께 사용하게 되면 단축될 수 있음을 모의실험에서 보였다. 두 형태용 목적함수는 가변 학습 방법뿐만 아니라 다른 학습방법과도 같이 사용이 가능하다. 두 형태용 목적함수에 기초한 오차역전파 방법은 기존의 목적함수에 기초한 오차역전파 방법보다 훨씬 유용하여 패턴분류 문제에 널리 쓰일 것으로 기대된다.

참고문헌

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

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

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

[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 rate and momentum adaptation," *Proc. IJCNN*, vol. 1, pp. 696-679, Jan., 1990.

[6] A. A. Minai and R. D. Williams, "Back-propagation heuristics: A study of the extended delta-bar-delta algorithm," *Proc. 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 and F. Palmieri, "MEKA - A fast, local algorithm for training feedforward neural networks," *Proc. IJCNN*, vol. 3, pp. 41-46, June, 1990.

[9] R. S. Scaleo and N. Tepedelenlioglu, "A fast new algorithm for training feedforward neural networks," *IEEE Trans. Signal Processing*, vol. 40, no. 1, pp. 202-210, Jan., 1992.

[10] H. Sawai, A. Waivel, P. affiner, M. Miyatake and K. Shikano, "Parallelism, hierarchy, scaling in time-delay neural networks for spotting Japanese phonemes /CV-syllabes," *Proc. IJCNN*, vol. 2, pp. 81-89, 1989.

[11] K. Lang and M. Witbrok, "Learning to tell two spirals apart," *Proc. 1988 Connectionist Models, Summer School*, Morgan Kaufmann, 1988.

[12] M. Riedmiller and H. Braun, "A direct adaptive method for faster backpropagation learning the RPROP algorithm," *Proc. IJCNN*, vol. 1, pp. 586-591, 1993.

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

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

[15] A. G. Parlos, B. Fernandez, A. F. Atiya, J. Muthusami and W. K. Tsai, "An accelerated learning algoirthm for multilayer perceptron networks," *IEEE Trans. Neural Networks*, vol. 5, pp. 493-497, Sept., 1994.

[16] Myung-Chan Kim and Chong-Ho Choi, "Square root scale and momentum control for BP learning in classification problems," *Proc. ICONIP'94*, vol. 2, pp. 761-766, Oct., 1994.

[17] R. Anand, K. Mehrotra, C. K. Mohan and S. Ranka, "Efficient classification for multiclass problems using modular neural networks," *IEEE Trans. Neural Networks*, vol. 6, no. 1, pp.117-124, Jan., 1995.

[18] S. Ergezinger and E. Thomsen, "An accelerated learning algorithm for multilayer perceptron: optimization layer by layer," *IEEE Trans. Neural Networks*, vol. 6, no. 1, pp. 31-42, Nov. 1995.

[19] 김명찬, 최종호, "신경회로망에서 일괄 학습," 전자공학회논문집, vol. 32-B, no. 3, pp. 100-108, 1995.

[21] 김명찬, 최종호, "배치모드 학습시 다층인식자의 가중치 초기화 방법," 제 1회 항법유도제어 학술대회, pp. 119-124, 1995.

[20] Myung-Chan Kim and Chong-Ho Choi, "Square root learning in batch mode BP for classification problems," *Proc. ICNN*, pp. 2769-2774, Nov. 1995.

[21] 김명찬, 최종호, "배치모드 학습시 다층인식자의 가중치 초기화 방법," 제 1회 항법유도제어 학술대회, pp. 119-124, 1995.

[22] S. Huang and Y. Huang, "Learning algorithms

for perceptrons using back-propagation with selective updates," *IEEE Control System Magazine*, pp. 56-61, April, 1990.

[23] A. Rezgui and N. Tepedelenlioglu, "The effect of the slope of the activation function on the back propagation algorithm," *Proc. IJCNN*, vol. I, pp. 673-678, Jan., 1990.

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

[25] M. L. Brady, R. Raghaven, and J. Slawny, "Back propagation fails to separate where perceptron succeed," *IEEE Trans. Circuit and System*, vol. 30, no. 5, pp. 665-674, May, 1989.

[26] 김명찬, 패턴 분류용 다층인식자의 오차 역전파 학습 알고리즘의 개선, 박사학위 논문, 서울대학교.

김 명 찬

1968년 3월 21일 출생. 1990년 2월 서울대학교 공과대학 제어계측공학과 졸업. 1992년 2월 서울대학교 대학원 제어계측공학과 석사졸업. 1997년 8월 서울대학교 대학원 제어계측공학과 박사졸업. 현재 삼성 SDS 근무중 관심분야는 신경회로망, 패턴인식, 데이터마이닝.

최 종 호

제어·자동화·시스템공학회 논문지 제 2권 제 4호 참조.