기울기 조정에 의한 다층 신경회로망의 학습효율 개선방법에 대한 연구
-A Study on the Learning Efficiency of Multilayered Neural Networks using Variable Slope-

이 형 일[1]
Yih Hyeong-il
남 재 현[2]
Nam Jai-hyun
지 선 수[3]
Ji Seon-su

Abstract

A variety of learning methods are used for neural networks.Among them, the backpropagation algorithm is most widely used in such image processing, speech recognition, and pattern recognition. Despite its popularity for these application, its main problem is associated with the running time, namely, too much time is spent for the learning. This paper suggests a method which maximize the convergence speed of the learning. Such reduction in the learning time of the backpropagation algorithm is possible through an adaptive adjusting of the slope of the activation function depending on total errors, which is named as the variable slope algorithm. Moreover experimental results using this variable slope algorithm is compared against conventional backpropagation algorithm and other variations; which shows an improvement in the perfomance over pervious algorithms.

1. 서론

신경회로망은 인간의 뇌 구조를 모방함으로써 학습능력과 병렬처리능력 등을 이용하여 최적화 문제, 화상인식, 음성인식, 비젼, 연상기억등 여러 분야에 사용되어 왔으며, 신경회로망의 구조나 사용 목적에 따라 여러가지 학습 알고리즘이 개발되어 있다([1], [2]). 일반적으로 신경회로망 모델들은 학습을 통하여 가중치 또는 임계치를 적절히 조정한다. 학습 방법에는 여러가지가 있으나 그 중 대표적인 것으로 역전파 알고리즘이 있는데, 학습에 소요되는 시간이 많이

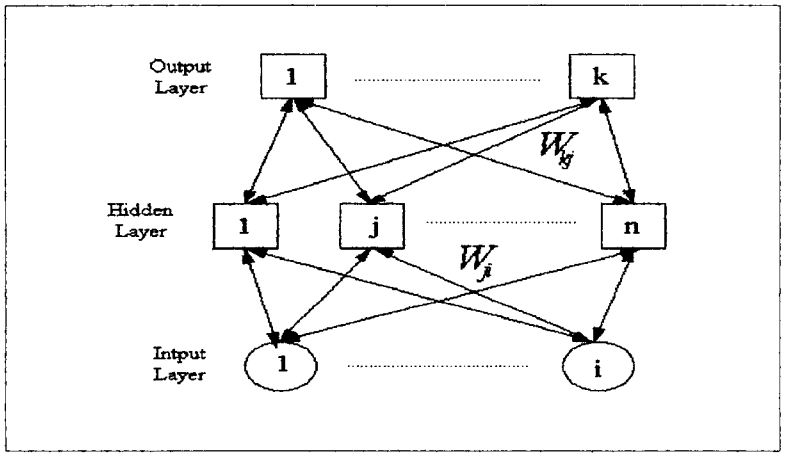
1) 명지대학교 대학원 컴퓨터공학과 박사과정 수료
2) 영월공업전문대학 사무자동화과
3) 원주전문대학 사무자동화과

걸린다는 단점을 가지고 있다([3], [8]).

본 논문에서는 다층 신경회로망의 대표적인 역전파 알고리즘을 살펴보고, 학습 시간을 줄이기 위한 방법을 제안하고, 기존의 알고리즘과 비교 분석하여 학습의 효율이 향상되었음을 보였다.

2. 역전파 알고리즘

다층 구조를 갖는 신경회로망에서 대표적인 학습방법으로는 역전파 알고리즘이 있다. 역전파 알고리즘은 각 노드에 제시된 패턴에 대하여 활성화 함수(Activation Function)를 이용하여 출력(O_j)을 하고, 요구되는 출력(T_j)과 실제 출력(O_j)과의 차이를 계산하여 이 차이를 출력층에서 입력층으로 역전도시키면서 층과 층 사이의 가중치를 조절한다. <그림 1>은 다층구조로 된 신경회로망을 표현하고 있다([1], [7], [9]). <그림 1>에서 W_{ji} 는 중간층과 입력층 사이의 가중치를 나타내고 W_{kj} 는 출력층과 중간층 사이의 가중치를 나타낸다.



< 그림 1 > 다층구조 신경회로망

활성화 함수는 각 뉴런의 결과치를 규격화하는 데 사용되는 데, 다층 구조의 신경회로망에서 사용하는 활성화 함수로는 Hardlimiter, Threshold, Sigmoid 함수등이 있다. 역전파 알고리즘에 쓰이는 활성화 함수는 다음과 같은 sigmoid 함수를 사용한다([1], [7], [5]).

$$f_j(net_{pj}) = \frac{1}{1+e^{-net_{pj}}} \tag{1}$$

$$net_{pj} = \sum_i W_{ji} O_{pi} + \theta_j \tag{2}$$

$$O_{pj} = f_i(net_{pj}) \tag{3}$$

여기서 W_{ji}는 i층과 j층 사이의 가중치(weight)를 나타낸다. O_{pj}는 p 번째 패턴의 j번째 뉴런의 출력이다. θ_j는 Threshold를 나타낸다([1], [2], [4], [7]). 출력층에서 계산된 실제 출력값과 요구되는 출력값 사이의 오류는 다음과 같이 구할 수 있다.

$$E_p = \frac{1}{2} \sum_j (T_{pj} - O_{pj})^2 \tag{4}$$

$$E^n = \sum_p^n E_p = \frac{1}{2} \sum_p^n \sum_j (T_{pj} - O_{pj})^2 \quad (5)$$

식 (4)는 출력층에서 발생하는 p 번째 패턴의 오류값을 나타내고 식 (5)는 각 패턴의 오류를 합산한 전체 오류를 나타내고 있다. T_{pj}는 출력층에서 p 번째 패턴의 j 번째 뉴런의 목표값이다. 계산된 전체 오류가 정의된 오류(예:10^{-4}, 10^{-5}, ...)보다 작을 때 학습을 마치게 된다([1], [7], [9]). 출력층과 중간층 사이의 가중치 변화량 δ_{pj}는 다음과 같다.

$$\delta_{pj} = -\frac{\sigma E_p}{\sigma net_{pj}} = -\frac{\sigma E_p}{\sigma O_{pj}} \frac{\sigma O_{pj}}{\sigma net_{pj}} = (T_{pj} - O_{pj}) f_j'(net_{pj}) \quad (6)$$

중간층과 입력층 사이의 가중치 변화량 δ_{pj}는 다음과 같다.

$$\begin{aligned} \delta_{pj} &= \frac{\sigma O_{pj}}{\sigma net_{pj}} \sum_k \frac{\sigma E_p}{\sigma net_{pk}} \frac{\sigma net_{pk}}{\sigma O_{pj}} \\ &= f_j'(net_{pj}) \sum_k \frac{\sigma E_p}{\sigma net_{pk}} W_{kj} \\ &= f_j'(net_{pj}) \sum_k \delta_{pk} W_{kj} \end{aligned} \quad (7)$$

식 (6)과 (7)를 이용하여 수정되는 가중치의 변화량은 다음과 같다.

$$\Delta W_{ji}(n+1) = \eta \delta_{pj} O_i + \alpha \Delta W_{ji}(n) \quad (8)$$

여기서 η는 학습율을 나타내고, α는 모멘텀율을 나타낸다([1] ,[7], [9]).

전체적인 학습과정을 살펴보면 다음과 같다([1], [7]).

1 단계 : 입력층에서 출력층까지 식 (1), (2), (3)에 의해 각 뉴런의 출력값을 구한다.
2 단계 : 식 (4), (5)로 전체 오류를 구하여, 정의된 오류보다 작으면 학습을 마치게 된다.
3 단계 : 식 (6), (7)로 층과 층 사이의 가중치 변화량을 계산하고, 식 (8)을 적용하여 새로운 가중치를 구한다.
4 단계 : [1 단계]로 가서 반복 수행한다.

3. 가변 기울기를 이용한 학습 알고리즘

기존에 언급된 바와같이 역전파 알고리즘의 단점은 학습이 종료되는 데 많은 시간이 걸린다는 것이다([3], [8]). 이와같은 문제점을 해결하기 위하여 본 논문에서는 활성화 함수의 기울기를 발생하는 전체 오류양에 따라 적절하게 변화시킴으로써 학습 수행 시간을 최소화하려고 하였다.

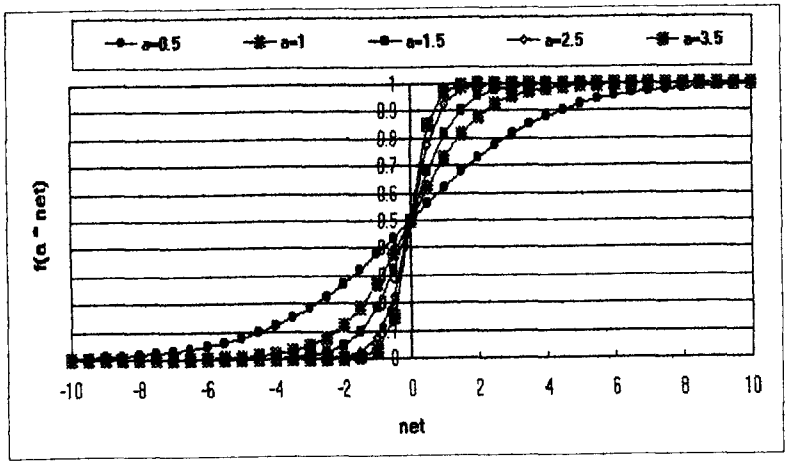
<그림 2>는 기울기의 변화에 따른 활성화 함수의 모양을 보여주고 있다. 그림에서 보는 바와 같이, 학습이 진행됨에 따라 활성화 함수의 기울기가 커진다면 출력층에서의 출력이 요구되는 출력에 더욱 가깝게 되어 수렴하는 속도가 빨라질 것이다([3], [6], [8]). 따라서 기울기를 <그림 2>와 같이 변화 시키기 위해 다음의 수식을 제안한다.

$$\lambda = -\log_{10}(E), \quad E \leq 0.1 \text{일때} \tag{9}$$

$$\lambda = -\log_{10}(E) + 0.5, \quad 0.3 < E < 1.0 \text{일때} \tag{10}$$

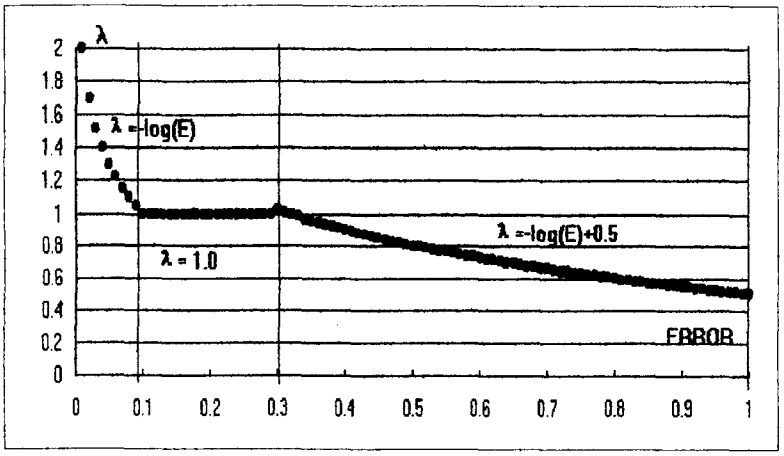
식 (10)에서 0.5를 더해주는 이유는 오류가 0.3에서 1.0 사이에 있을 때 $-\log_{10}(E)$의 범위가 0.507에서 0.0046이 되기 때문에 활성화 함수의 출력값 변화폭이 아주 미세하게되어 학습 효율을 저하시킨다. 그러므로 적당한 기울기를 정해주기 위해 0.5을 더해주었다. 식 (9), (10)에 따라 sigmoid 함수에 λ를 추가하면 다음의 식과 같이 된다.

$$fj(net_{pj}) = \frac{1}{1 + e^{-\lambda net_{pj}}} \tag{11}$$



< 그림 2 > sigmoid 함수의 기울기

<그림 3>은 오류값과 λ 사이의 관계를 그래프로 표현한 것이다.



< 그림 3 > 전체오류에 따른 활성화 함수의 기울기 변화

4. 실험

4.1 가변 모멘텀율을 사용한 기법

학습 시간을 줄이기 위한 다른 방법중 하나는 모멘텀율과 학습율을 오류의 변화에 따라서 변화시키는 방법이 있다. 식 (8)의 η 또는 α를 오류 값에 따라 변화시켜 학습수렴속도를 향상시킨다. 가변 모멘텀(α)율을 사용한 방법과 가변 학습율(η)을 사용한 방법 중 가변 모멘텀율을 사용한 방법이 효율이 좋기 때문에 비교의 대상으로 삼는다([1], [2], [3]). 전체 오류값에 따라 변화하는 모멘텀율은 다음과 같다([1], [4], [10]). 여기에서E 는 전체 오류의 양을 K 는 모멘텀율의 초기치를 나타낸다. 학습초기에 전체오류가 0.1이 될 때까지는 모멘텀율을 0.6으로 수행하고, 0.1이하에서는 다음 식 (12)에 의해 모멘텀율 변화시킨다.

$$\alpha = 1 - \frac{1-K}{\log(E)^2+1}, \qquad E<0.1\text{일때} \tag{12}$$

4.2 실험 조건

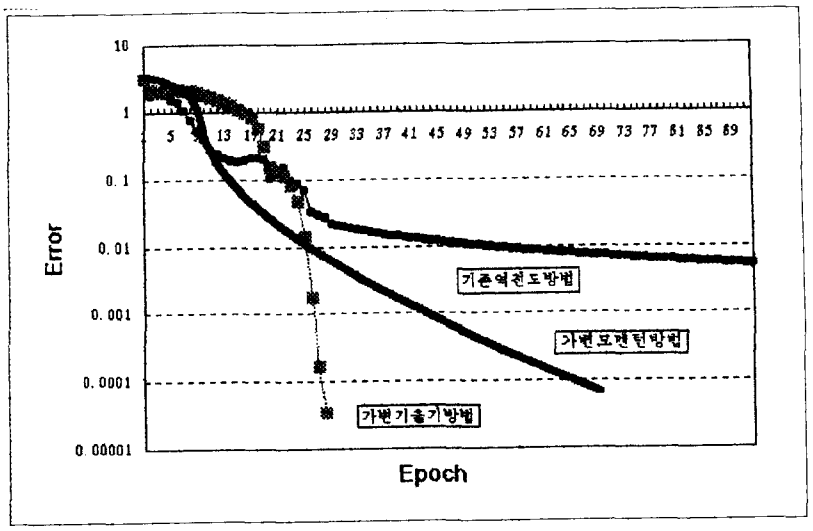
학습에 쓰인 데이타는 입력 패턴의 뉴런 갯수가 5개 미만 일때는 패턴의 갯수를 입력 뉴런수와 같게 하였으며, 입력 뉴런수가 5개 이상일 때는 패턴 갯수를 10개로 하였다. 또 학습에 쓰이는 계수의 값들로 α=0.9로, η=0.7로 하였으며, 최고 허용 오류값은 0.0001로 하였다. 학습이 수행되는 과정에서, 변화하는 오류의 양을 자세하게 그래프로 표현하여 각 알고리즘의 성능을 비교하였다.

학습이 끝난 뒤에는, 임의의 패턴을 입력하여 출력 패턴을 구한 뒤, 학습에 의해 저장된 패턴과의 해밍거리를 구하여 각 알고리즘의 효율을 평가하였다. 해밍거리가 최소인 패턴으로 수렴하는 경우를 올바른 답(Correct Answer)으로 간주하였으며, 저장되지 않은 패턴으로 수렴하는 경우를 모호한 패턴(Spurious Answer)으로 간주하였다.그리고 저장된 패턴과의 해밍거리가 최소가 아닌 패턴으로 수렴하는 경우는 틀린 답(Wrong Answer)으로 간주하였다([7]). 비교의 대상으로는 기존의 역전파 알고리즘과 가변 모멘텀율을 사용한 알고리즘과 비교하였다.

4.3 실험 결과

<그림 4>는 학습 횟수에 따라 변화하는 오류의 양을 나타낸다. 그림에서 Epoch는 학습 횟수를 나타낸다. 가변기울기를 이용한 학습 알고리즘은 계속적으로 오류가 줄어들어 기존의 역전파 알고리즘보다도 훨씬 빠르게 수렴한다.

다음의 표에서는 가변 기울기를 이용한 알고리즘이 수렴속도뿐만 아니라 효율면에서도 향상되었음을 보여주고 있다. <표 1>부터 <표 8>에서 3*3은 입력 뉴런 3개와 입력 패턴 3개의 경우 평가시를 나타내며 이후의 것도 마찬가지로 뉴런과 패턴 수를 점차 증가시켜 평가한 경우이다. 또한 epoch는 학습 횟수를 나타내며, C는 올바른 패턴으로 수렴하는 경우, S는 모호한 패턴으로 수렴하는 경우, W는 틀린 패턴으로 수렴하는 경우를 의미한다. 그리고 X는 학습이 끝난 뒤 임의의 입력 패턴을 입력하여 구한 출력 패턴 중에서 0.4와 0.6 사이에 존재하는 값들의 개수로 특히 이진화하기 어려운 출력값으로 모호한 패턴이다. 그리고, VSR는 가변 기울기를 이용하여 학습을 마친 뒤 활성화 함수 기울기의 값을 기존의 역전파 알고리즘처럼 1로 고정하여 평가한 경우이며, MRR은 가변 모멘텀율을 이용하여 학습을 마친 뒤, 효율을 평가 할 때 가변 기울기를 이용하여 학습이 끝났을 때의 활성화 함수 기울기 값을 사용한 경우이다. 계산된 수치는 입력 뉴런의 갯수가 같은 것들의 평균치이다.



< 그림 4 > 학습 횟수에 대한 오류의 변화

<표 1>부터 <표 8>까지는 각 알고리즘의 성능 비교를 한 것으로 표에서 사용한 기호의 의미는 다음과 같다.

BP : 기존의 역전파 알고리즘의 평가

MR : 가변 모멘텀율을 이용한 알고리즘의 평가

VS : 가변 기울기 알고리즘의 평가

VSR : 가변 기울기 알고리즘를 이용하여 학습한 경우+활성화 함수 기울기를 1로 고정여 평가

MRR : 가변 모멘텀율 알고리즘을 이용하여 학습한 경우 + 가변 기울기를 이용하여 학습이 끝났을 때의 활성화 함수 기울기로 평가

<표 1> 알고리즘의 성능 비교(1)

3*3	epoch	C	S	W	X
BP	1931.0	8.0	0.0	0.0	0.0
MR	40.5	6.5	0.5	1.0	0.5
VS	27.5	7.5	0.0	0.5	0.0
VSR	27.5	6.5	0.5	1.0	0.5
MRR	40.5	7.0	0.0	1.0	4.0

<표 2> 알고리즘의 성능 비교(2)

4*4	epoch	C	S	W	X
BP	4043.0	8.7	6.3	1.0	4.0
MR	44.7	11.0	4.3	0.7	1.3
VS	37.0	10.0	4.7	1.3	1.3
VSR	37.0	8.3	6.7	1.0	6.7
MRR	44.7	11.3	3.3	1.3	5.0

<표 3> 알고리즘의 성능 비교(3)

5*10	epoch	C	S	W	X
BP	4258.2	16.2	13.5	2.2	4.0
MR	44.2	17.0	11.8	3.2	3.8
VS	38.5	20.8	6.8	4.5	2.8
VSR	38.5	18.5	10.0	3.5	6.0
MRR	44.2	19.2	9.5	3.2	6.0

<표 4> 알고리즘의 성능 비교(4)

6*10	epoch	C	S	W	X
BP	6205.0	23.0	39.2	1.8	11.0
MR	43.6	23.4	38.2	2.4	11.2
VS	31.6	31.4	26.0	6.6	3.8
VSR	31.6	25.6	34.8	3.6	19.6
MRR	43.6	29.6	31.6	2.8	7.0

<표 5> 알고리즘의 성능 비교(5)

7*10	epoch	C	S	W	X
BP	7783.2	33.3	93.2	1.5	29.3
MR	74.5	43.0	77.8	7.2	17.7
VS	32.3	42.0	78.0	8.0	12.2
VSR	32.3	34.3	89.3	4.3	47.8
MRR	74.5	48.5	71.3	8.2	8.0

<표 6> 알고리즘의 성능 비교(6)

8*10	epoch	C	S	W	X
BP	8249.8	75.2	177.3	3.5	77.2
MR	57.7	74.2	175.8	6.0	66.8
VS	36.3	81.2	162.2	12.7	30.3
VSR	36.3	63.3	184.7	8.0	108.7
MRR	57.7	93.3	149.5	13.2	9.0

<표 7> 알고리즘의 성능 비교(7)

9*10	epoch	C	S	W	X
BP	11122.2	133.1	371.8	7.1	172.5
MR	67.0	135.0	362.6	14.4	143.4
VS	46.2	143.2	344.4	24.4	68.8
VSR	46.2	106.2	392.2	13.5	247.8
MRR	67.0	160.0	329.1	22.9	10.0

<표 8> 알고리즘의 성능 비교(8)

10*10	epoch	C	S	W	X
BP	11361.1	206.8	806.9	10.3	403.8
MR	67.0	213.6	797.0	13.4	375.0
VS	36.1	215.1	781.4	22.3	182.0
VSR	36.1	172.9	831.8	19.3	519.6
MRR	67.0	264.7	729.3	30.0	85.4

<표 1>에서 <표 8>까지의 결과를 분석하면, VS가 다른 학습 방법에 비해 빠르게 학습을 마치며, 올바른 답(correct answer)으로 수렴하는 패턴의 수가 약간 증가하였다. 그리고, VSR의 경우에는 올바른 답(correct answer)으로 수렴하는 패턴의 수가 감소하였으며, VRR의 경우는 학습 횟수는 VS보다 많지만 올바른 답(correct answer)으로 수렴하는 패턴의 수가 증가하였다. 기존의 역전파 알고리즘에 대한 제안한 학습 방법의 성능은 다음의 표로 나타낼 수 있다.

<표 9> 기존의 역전파 알고리즘에 대한 제안한 학습 방법의 성능

CA : correct answer
SA : spurious answer
WA : wrong answer

	Epoch	CA	SA	WA
3*3	70배 감소	1.1배 증가	1.4배 증가	2배 증가
4*4	98배 감소	1.02배감소	1.1배 증가	1.6배 증가
5*10	132배 감소	1.3배 증가	1.4배 증가	1.9배 증가
6*10	146배 감소	1.3배 증가	1.4배 증가	4배 증가
7*10	153배 감소	1.2배 증가	1.2배 증가	3.7배 증가
8*10	267배 감소	1.2배 증가	1.1배 증가	4.2배 증가
9*10	291배 감소	1.3배 증가	1.1배 증가	2.8배 증가
10*10	340배 감소	1.1배 증가	1.1배 증가	2.4배 증가

<표 9>에서 보는 바와 같이 입력 뉴런이나 입력 패턴의 수가 증가할 수록, 활성화 함수의 기울기를 가변시키면서 학습하는 방법이 기존의 역전도 알고리즘보다 약 70배에서 340배 정도 학습횟수가 줄어들었으며, 올바른 답(correct answer)으로 수렴하는 경우도 약 1.1배에서 1.3배 정도 증가하였다. 반면, 모호한 패턴으로 수렴(spurious answer)하는 경우는 약 1.1배에서 1.4배정도 감소하였으며, 틀린 답(wrong answer)으로 수렴하는 경우는 약 1.6배에서 4.2배 정도 증가하였다. 따라서, 제안한 학습 방법이 효율을 저하시키지 않으면서, 학습 횟수를 크게 줄일 수 있다.

5. 결론 및 향후전망

각 알고리즘의 성능 분석 결과, 본 논문에서 제안한 가변 기울기를 이용한 학습방법이 역전파 알고리즘의 단점인 학습 수행 시간이 많이 걸린다는 것을 극복하였으며, 가변 모멘텀율을 사용했을 때 나타나는 오류의 진폭도 나타나지 않았다. 효율면에서도 패턴의 갯수가 많을 수록 다른 알고리즘보다 우수한 것으로 나타났다. 따라서 본 논문을 패턴의 갯수가 많은 실제의 응용을 한다면 기존의 알고리즘의 단점을 효과적으로 극복할 수 있을 것이다.
앞으로 연구 해야할 과제로는 학습을 수행하기 전에 임의의 값으로 가중치를 지정하지 않고, 효과적인 방법으로 지정하는 학습 알고리즘 등에 관한 연구가 이루어져야 할 것이다.

참 고 문 헌

[1] J. L. McClelland, D. E. Rumelhart and the PDP Research Group, "Parallel Distributed Processing", Vol 1, p318-p362, *MIT Press*, 1986.

[2] Richard P. Lippmann, "An Introduction to Computing with Neural Nets", *IEEE*, 1987.

[3] A. Rezgui, et al., "The Effect of the Slope of the Activation Function on the Back Propagation Algorithm", *Proc.Int.Joint Conf. Neural Networks*, 1990.

[4] Paul J. Werbos, "Backpropagation Through Time :What It Does and HoW to Do it", *IEEE*, 1990.

[5] Y. Lacouture, "Mean-variance Back-Propagation:A Connectionist Learning Algorithm with a Selective Attention Mechanism", *University Laval, IJCNN*, VOL 2, 1991.

[6] Chi-Cheng Jou, "Fuzzy Activation Functions", *National Chiao Tung University, IJCNN*, 1991

[7] J.A.Freeman, D.M.Skapura, "Neural Networks - Algorithms, Applications, and Program ming Techiques", p89-p102, *Addison Wesley Press*, 1991.

[8] Y.Lee, S.Oh, M.Kim, "The Effect of Initial Weights on Premature Saturation in Back-Propagation Learning", *Electronics & Telecommunications Research Institute, IJCNN*, VOL 1, 1991.

[9] R.C.Lacher, "Artificial Neural Networks", p1-p37, p79-p92, *Florida State University Press*, 1992.

[10] Y.I Yih, J. H. Nam, S. S. Ji, "An Improvementing Method of the Learning Time of Multilayered Neural Networks using Predicted Weight" *submitted for publication* 1996