

## 신뢰도 추정을 위한 분산 학습 신경 회로망

### A Variance Learning Neural Network for Confidence Estimation

오 조 영빈\*, 권 대갑\*, 윤 한종\*\*, 이 경래\*\*

\*한국과학기술원 기계공학과(Tel:042-869-3265 Fax:042-869-5225 s\_ybcho@cais.kaist.ac.kr)

\*\*LG 전자 생산기술센터(Tel:0551-60-3679 Fax:0551-89-8116)

**Abstracts** Multilayer feedforward networks may be applied to identify the deterministic relationship between input and output data. When the results from the network require a high level of assurance, considering of the stochastic relationship between the data may be very important. The variance is one of the useful parameters to represent the stochastic relationship. This paper presents a new algorithm for a multilayer feedforward network to learn the variance of dispersed data without preliminary calculation of variance. In this paper, the network with this learning algorithm is named as a variance learning neural network(VALEAN). Computer simulation examples are utilized for the demonstration and the evaluation of VALEAN.

**Keywords** Stochastic Process, Gaussian Distribution, Variance Learning, Confidence Estimation, Multilayer Feedforward Network

#### 1. 서론

신경회로망은 병렬 분산 처리 기능을 가진 비선형 연산 소자인 뉴런을 기본단위로 하여 구성되며 미지의 입출력 관계를 학습할 수 있다. 신경 회로망에는 연상기억 메모리 (Associative Memory), 경쟁 학습 신경망 (Competitive Network), 다층 신경망 (Multilayer Feedforward Network) 등과 같은 여러 가지 형태가 있으며 각각 특정한 학습 규칙을 가지고 있다. Hetch-Nielsen(1987)은 임의의 연속 함수가 무한개의 뉴런과 2 개의 은닉층으로 구성된 다층 신경망으로 구현될 수 있음을 증명하였고 Hornik (1989)등은 충분히 많은 뉴런과 3 개의 은닉층을 가진 다층 신경망으로 임의의 함수가 근사화 될 수 있음을 보였다.

다층 신경망의 함수 근사화 기능은 학습 이론의 발전으로 패턴 인식, 매개변수 추정, 인공지능과 같은 분야에 널리 적용되었다. 이러한 다층 신경망의 학습은 입출력 변수 사이에 존재하는 비확률적 관계를 근사화 하는 것이 주 목적이었다. 그러나, 인식 결과의 신뢰도가 중요한 검사 작업과 같은 특정한 응용 분야의 경우 입출력 변수 사이의 확률적 관계를 무시하기 어려운 경우가 있다. 이런 경우, 확률적 영향을 고려하기 위하여 인식 결과의 오차에 대한 분산을 추정하는 것이 유용할 수 있다. 본 연구에서는 다층 신경회로망을 분산 추정에 사용하기 위하여 새로운 학습 알고리즘을 개발하였고 이 알고리즘을 이용한 다층 신경회로망을 분산 학습 신경 회로망 (VALEAN : VAriance LEARNING neural Network)으로 명명하였다. 이 알고리즘은 가우시안 분포의 분산값을 추정하는 것을 목표로 한다.

#### 2. 확률과정과 분산학습신경망

$n$  차원 입력 변수  $x$ 와 출력 변수  $y$  사이의 관계  $y(x)$ 는 다음과 같이 비확률적 관계식  $y_d(x)$ 와 확률적 관계식  $y_s(x)$ 로 표현할 수 있다.

$$y(x) = y_d(x) + y_s(x) \quad (1)$$

여기서 비확률적 관계는 기존의 다층 신경망 학습법칙에 의해 구해지지만 확률적 관계는 기술할 수 없으며 학습 에러로 나타난다. 만일 확률적 관계가 평균이 0 이고 임의의 분산을 가진 가우시안 분포라면 다음과 같이 표현할 수 있다.

$$E[y_s|x] = 0 \quad (2)$$

$$P(y_s|x) = \frac{1}{\sqrt{2\pi} \sigma(x)} \exp\left(-\frac{y_s^2(x)}{2\sigma^2(x)}\right) \quad (3)$$

여기서  $\sigma^2(x)$ 은 확률적 관계식  $y_s(x)$ 의 분산이며 입력 변수  $x$ 의 함수로 기술된다. 위식에서 보는 바와 같이 확률적 관계식은 분산  $\sigma^2(x)$ 만으로 완전히 기술할 수 있다. 만일 학습 결과에 대한 높은 신뢰도가 요구된다면 다음과 같이 결과의 하한 값  $y_L(x)$ 을 정할 수 있다.

$$y_L(x) = y_d(x) - z_\alpha \sigma(x) \quad (4)$$

본 연구에서는 이와 같은 응용을 위하여 분산을 학습하는 신경회로망을 개발하고자 하였으며 그 구조적 형태는 Figure 1과 같다. 기존의 다층 신경망 NNI은 비확률적 함수를 학습하기 위한 것이며 동시에 사용된 분산 학습 신경망 NN2는 학습되지 못하고 남은 확률적 함수  $y_s(x)$ 를 학습하기 위한 것이다. 위에서 가정한 것과 마찬가지로 오차가 평균이 0인 가우시안

분포를 가지며 기존의 신경회로망의 학습이 성공적이라면 비확률적 함수는 NNI에 의해 충분히 기술될 수 있다. 이 경우 확률적 관계를 추정하기 위해 분산 학습 신경회로망이 학습을 시작한다. 기존의 오류역전파 알고리즘은 입력패턴과 그에 대응되는 출력 패턴을 가지고 있어야 하는데 분산을 학습하고자 하는 경우에는 주어진 입력패턴  $x$ 에 대하여 출력값인 분산  $\sigma^2(x)$ 는 주어지지 않는다. 대신 주어진 입력패턴  $x$ 에 대하여 여러개의  $y_s(x)$ 만이 존재할 뿐이다. 만일 다차원 입력 패턴일 경우에는 일일이 분산값  $\sigma^2(x)$ 을 구하는 것은 매우 지루한 일이다. 여기서는 입력 패턴  $x$ 와  $y_s(x)$ 를 직접 이용하여 자동으로 분산  $\sigma^2(x)$ 를 구하는 학습 알고리즘을 소개하고자 한다.

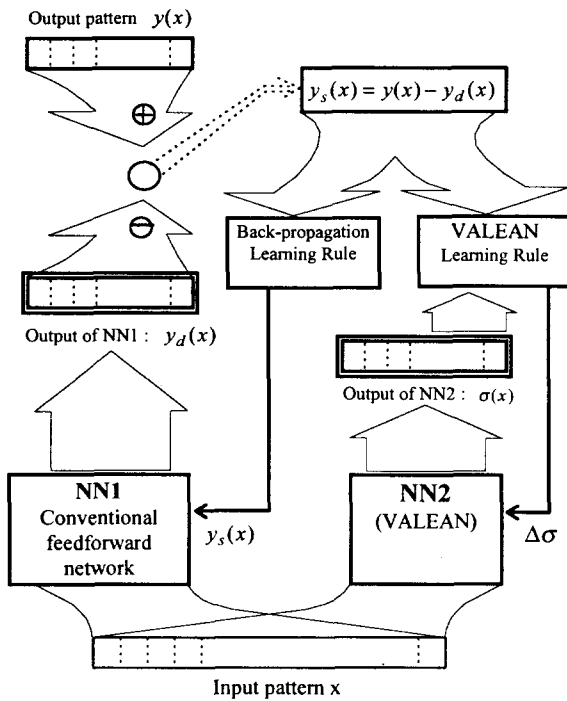


Figure 1. Schematic Diagram of Variance Learning

### 3. 분산 추정을 위한 학습 법칙

가우시안 분포를 가지는 데이터의 분산을 학습하는 알고리즘에 설명을 집중하기 위하여 본 논문에서는 Figure 1의 NNI의 학습이 끝난 상태에서 시작하기로 한다. 주어진 입력  $x$ 에 대한 분산값의 추정치는 학습시간  $t$ 와  $t+1$ 에서 아래와 같이 표현될 수 있다.

$$\sigma_N^{(t+1)}(x) = \sigma_N^{(t)}(x) + \Delta\sigma_N^{(t)}(x) \quad (5)$$

여기서 분산학습 신경망의 학습 오차는  $\Delta\sigma_N^{(t)}(x)$ 으로 정의되므로 이값의 적절한 추정이 매우 중요하다. 학습 규칙의 유

도를 단순화하기 위하여  $\Delta\sigma_N^{(t)}(x)$ 의 크기는 충분히 작은 일정한 값이며 부호만 바뀌는 것으로 가정한다. 이때 부호를 결정하는 학습 규칙은 아래와 같이 임의의 확률함수로 표현할 수 있다고 가정한다.

$$\begin{aligned} |\Delta\sigma^{(t)}| &\ll \sigma_N^{(t)} \\ \text{sign}(\Delta\sigma^{(t)}) &= \begin{cases} +1 & P[\Delta\sigma > 0] \\ -1 & P[\Delta\sigma < 0] \end{cases} \end{aligned} \quad (6)$$

여기서 확률함수는 다음과 같은 일반적인 3가지 제한 조건을 만족하여야 한다.

Constraint 1 :

$$E[\Delta\sigma | \sigma_N = \sigma_T] = 0 \quad (7.a)$$

Constraint 2 :

$$\begin{aligned} P[\Delta\sigma < 0 | \sigma_N > \sigma_A] &> P[\Delta\sigma < 0 | \sigma_N < \sigma_A] \\ P[\Delta\sigma > 0 | \sigma_N > \sigma_A] &< P[\Delta\sigma > 0 | \sigma_N < \sigma_A] \end{aligned} \quad (7.b)$$

Constraint 3 :

$$\begin{aligned} P[\Delta\sigma < 0 | y_s] &\leq 1, P[\Delta\sigma > 0 | y_s] \leq 1 \\ P[\Delta\sigma < 0 | y_s] + P[\Delta\sigma > 0 | y_s] &= 1 \end{aligned} \quad (7.c)$$

where  $\sigma_T^2$  : the true value of variance

$\sigma_N^2$  : the estimated variance by VALEAN

$\sigma_A^2$  : an arbitrary value of variance

$y_s$  : a stochastic relationship between input and output

첫번째 제한조건은 학습 결과가 바이어스되지 않음을 요구하는 것이며 두번째 제한조건은 참값에 수렴함을 요구하는 것이고 마지막 제한조건은 확률함수의 일반적인 요구조건이 된다. 여기서 두번째 제한 조건은 다음과 같이 서술될 수 있다.

“추정치  $\sigma_N^{(t)}(x)$ 가 임의의 값  $\sigma_A$ 보다 크면 그렇지 않은 경우보다 감소할 확률이 커진다”

“추정치  $\sigma_N^{(t)}(x)$ 가 임의의 값  $\sigma_A$ 보다 작으면 그렇지 않은 경우보다 증가할 확률이 커진다”

따라서 충분한 학습후에는 반드시 임의의 값에 수렴하게 된다.

위의 제한 조건을 모두 만족하는 학습 규칙을 찾기 위해 식 (7.a)를 다음과 같이 정리한다.

$$\begin{aligned} E[\Delta\sigma | \sigma_N = \sigma_T] &= \int P_{\sigma_T}(y_s) \cdot \{P[\Delta\sigma > 0] - P[\Delta\sigma < 0]\} |\Delta\sigma| dy_s \\ &= \int P_{\sigma_T}(y_s) \cdot \{1 - 2P[\Delta\sigma < 0]\} |\Delta\sigma| dy_s \\ &= 0 \end{aligned} \quad (8)$$

위식에서  $|\Delta\sigma|$ 가 일정하다고 가정하였으므로 다음과 같이

간단히 정리된다.

$$\int P_{\sigma_T}(y_s) \cdot P[\Delta\sigma < 0] \phi_s = \frac{1}{2} \quad (9)$$

정규화된 가우시안 분포 함수를 다음과 같이 정의 한다.

$$\bar{P}_{\sigma}(y_s) = \exp\left(-\frac{y_s^2}{2\sigma^2}\right) \quad (10)$$

정규화된 가우시안 분포함수와 가우시안 분포함수 사이에는 다음과 같은 관계가 성립한다.

$$\int P_{\sigma}(y_s) \bar{P}_{\sigma}^n(y_s) \phi_s = \frac{1}{\sqrt{n+1}} \quad (11)$$

만일 학습을 위한 확률 함수를 아래와 같이 선택한다면 앞에서 설명한 3 가지 제한 조건을 모두 만족하게 된다.

$$\begin{aligned} P[\Delta\sigma < 0] &= \bar{P}_{\sigma_N}^{-3}(y_s) \\ P[\Delta\sigma > 0] &= 1 - \bar{P}_{\sigma_N}^{-3}(y_s) \end{aligned} \quad (12)$$

From NNI

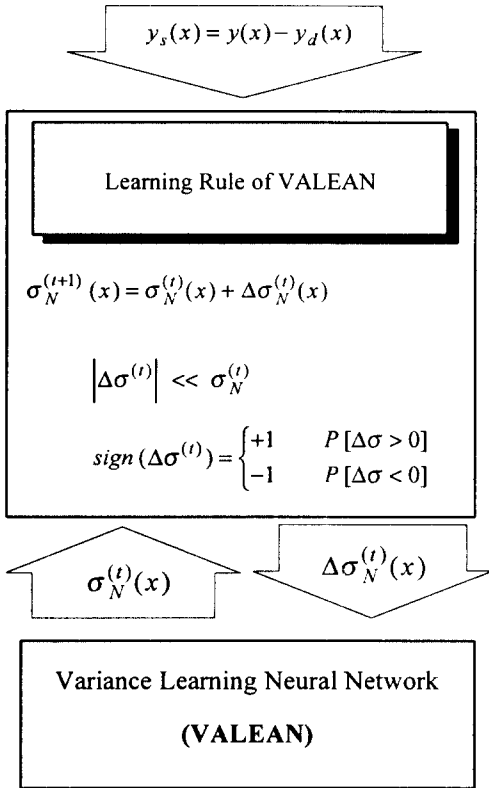


Figure 2. Learning Rule of VALEAN

분산 학습 알고리즘의 동작은 Figure 2의 모식도에서 보느냐와 같이 정리될 수 있다.

#### 4. 모의 실험

지금까지 설명한 학습 알고리즘의 모의 실험 결과를 살펴보

기 위하여 입력이 0.5 0.6 그리고 0.7에 대하여 각각 25개의 오류분포를 가진 75개의 출력 데이터를 만들었다. 3가지 종류의 입력패턴에 대하여 생성된 25개 출력 패턴의 표준편차는 각각 0.027304, 0.051190, 0.085624였다. 신경회로망은 1개의 입력노드, 3개의 은노드, 1개의 출력노드를 가진 다층 신경회로망으로 구성되었다. 알고리즘의 개발과정에서 설명한 바와 마찬가지로 수정항은 전체 학습 과정 동안 일정한 값을 가지도록 하는 방법을 사용하였으며 추가적으로 현재의 추정값에 비례한 값을 가지도록 하는 방법도 실험하였다. 추정치가 참값에 충분히 가까와 지고 학습이 충분히 천천히 진행된다면 수정항의 크기가 일정하다는 가정을 그대로 적용할 수 있다.

$$|\Delta\sigma| = k \quad (13)$$

$$|\Delta\sigma| = 0.01 \times \gamma \times (\sigma + \epsilon \cdot \gamma^2) \quad (14)$$

식 (14)에서 0.01이라는 수치가 사용된 이유는 주어진 모든 패턴에 대하여 오류 역전 과정을 행한후 신경망의 가중치를 수정하기 때문에 패턴의 갯수 75개의 평균을 취하기 위한 것이다.

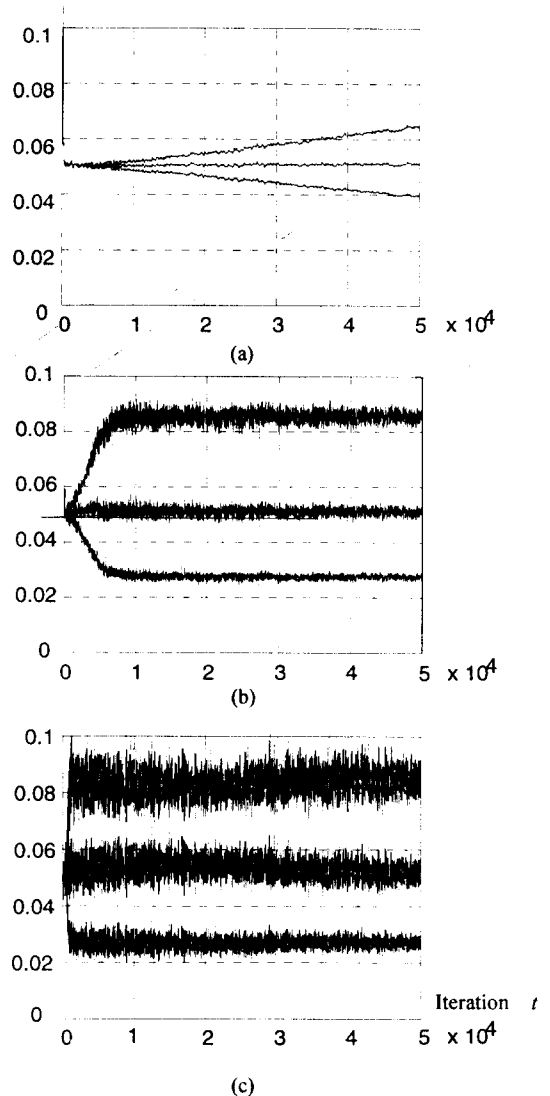


Figure 3. Learning history of VALEAN for constant correction

Figure 3 은  $k$  값이 각각 0.0001, 0.001 그리고 0.01 일 때의 학습 과정을 도시한 것이다. 여기서 보는 바와 같이 학습과정 전반에 걸쳐 일정한 수정값을 유지할 경우는 수정값의 크기에 민감하게 변하는 것을 알 수 있다. 첫번째의 경우는 수정값이 너무 작아서 50,000 번의 학습에도 수렴하지 못하고 너무 느린 학습을 보여 주었다. 이에 비해 세번째의 경우는 너무 큰 진동양상을 보여 일정한 값에 수렴 시킬 수 없었다.

Figure 4 에는 식 (14)를 이용한 학습 결과를 보여 준다. 여기서  $\gamma$  는 초기값은 1 이고 서서히 0 으로 수렴하는 수치로서 충분한 학습후는 진동을 일으키지 못하도록 하기 위한 값이다. 또한  $\epsilon$  은 초기에 추정값이 매우 작아 학습을 하지 못하는 현상을 피하기 위하여 모의 실험에서는 0.02 가 사용되었고 학습 초기에만 필요한 값이므로 빨리 소멸되도록 하였다. 이렇게 수정치의 크기를 정할 경우 그림에서 보는 바와 같이 빠르고도 안정적인 학습결과를 얻어낼 수 있었으며 학습 성능이 민감하게 변하지 않는 바람직한 결과를 얻을 수 있었다.

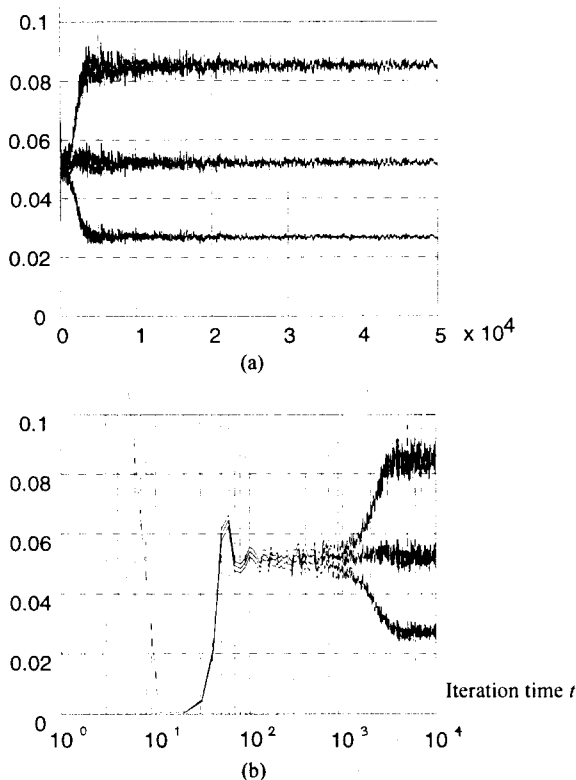


Figure 4. Learning history of VALEAN for flexible correction

## 5. 결론

높은 신뢰성이 요구되는 검사 작업과 같은 응용분야에 있어서 입출력 변수사이에 존재하는 불확실한 관계, 즉 확률적 관계의 규명은 매우 중요하다. 이런 경우 분산값의 추정은 검사

결과의 신뢰성을 보증하기 위하여 유용하게 사용될 수 있다. 본 논문에서는 분산값을 추정하기 위한 새로운 학습 알고리즘을 개발하였고 컴퓨터 모의실험을 통하여 검증을 하였다. 알고리즘의 유도를 단순화 하기 위하여 수정항이 학습과정 전반에 걸쳐 항상 일정한 값을 가지는 것으로 가정하였으나 2 가지 추가적인 파라미터를 사용하여 학습과정이 보다 안정적이고 빠른 수렴성능을 가지도록 향상시킬수 있었다. 1 차원 입력에 대한 예제만을 보이고 있으나 학습방법은 다차원 입력공간에서도 역시 동일하므로 입력변수가 다차원인 경우도 마찬가지로 적용할 수 있다.

## 참고문헌

- [1] Hecht-Nielsen, R. (1987). "Kolmogorov mapping neural network existence theorem." *IEEE International Conference on Neural Networks*, 3, 11-13
- [2] Hopfield, J. (1984). "Neurons with graded response have collective computational properties like those of two-state neurons." *Proceedings of National Academy of Sciences (U.S.A.)*, 81, 3088-3092.
- [3] Hornik, K., Stinchcombe, M., and White, H. (1989). "Multilayer feedforward networks are universal approximators." *Neural Networks*, 2, 359-366.
- [4] Irie, B., and Miyake, S. (1988). "Capabilities of three-layered perceptrons." *IEEE International Conference on Neural Networks*, 1, 641-648.
- [5] Parzen, E. (1962). "On the estimation of a probability density function and mode." *Annals of Mathematical Statistics*, 33, 1065-1076.
- [6] Rumelhart, D., Hinton, G., and Williams, R. (1986). "Learning internal representations by error propagation." In D. Rumelhart, & J. McClelland(Eds.), *Parallel Distributed Processing: Exploration in the Microstructure of cognition*. Cambridge, MA: MIT press.
- [7] Weigend, A. S., and Nix, D. A. (1994). "Predictions with confidence intervals (local error bars)." *Proc. Int. Conf. Neural Inform. Process.* Seoul, Korea, 847-852