수렴성 구조를 이용한 강인한 선행 신경망 구현

김 준석° 서 진헌
서울대학교 전기공학과

Implementation of Robust Feedforward Neural Network Using Classifier Structure

Joonsuk Kim° Jin-Heon Seo
Dept. of Electrical Eng., Seoul National University

Abstract

In this paper, we improve feedforward neural network performance by eliminating the effect of gross error using classifier structure. At first, we prove the output of classifier converges to the posteriori probability of each pattern given input x, $f_\theta(\theta_i \mid x)$. And we apply filtering approach based on the robust statistics before reconstructing continuous output. The data distorted with noise can be rejected by this process. Finally, we suggest neurofilter structure.

Simulation result shows that our structure yields consistent estimates even in the presence of noise.

1. 서론

신경회로망의 특징은 병렬연산과 적응학습으로 요약할 수 있다. 즉, 뉴론간의 상호연결에 의한 분산기억과 병렬처리에 의해 단시간내에 오차보정능력이 높은 신호처리를 할 수 있을 뿐 아니라 학습법칙에 의해 스스로 새로운 것을 배워 나갈 수 있다. 특히 신경회로망은 학습능력에 따라 명시화된 식이 주어지지 않은 경우에도 비선형 함수를 표현할 수 있는 장점이 있다.

신경회로망은 크게 수렴성 회로망과 비수렴성 회로망으로 나눌 수 있다. 대표적인 비수렴성 신경회로망에는 선행 신경망 등이 있는데, 이것은 몇가지 표본값들로 학습한 후 표본 사이값들은 내삽을 통해 적절하게 출력할 수 있도록 학습하는 회로망이다. 하지만 학습 데이타에 잡음이 실린 경우나 잘못 입력되었을 경우, 결과에 대한 신뢰도가 떨어지고, 데이타가 충분치 못하면 학습된 신경회로망은 의미가 없어지게 된다. Mukhopadhyay[1]는 입력 차수를 증가시켜 많은 데이타의 평균치를 계산하는 방법으로 잡음문제를 해결하려 하였으나 근본적인 잡음 제거는 하지 못하였다.

이에 비해 Classifier로 대표되는 수렴성 신경회로망은 어느 정도의 잡음에 강인하다는 장점이 있다. Huang[2]은 BP(Back Propagation)를 사용한 신경망이 데이타의 확률분포에 상관없이 오차를 최소화시키므로 기존의 Classifier보다 outlier에 대해 더 강인함을 보였다. 하지만 수렴 신경망은 내삽이 불가능하고 설정된 패턴이외의 정보는 알 수 없다. 이를 보완하기 위해 Kosko[3]와 Traven[4]은 신경망의 결과값을 패턴에 대한 확률분포로 설정하여 조정하였고, Specht[5]는 Bayes 규칙을 이용한 결정함수를 정의하여 PNN(Probability Neural Network)을 제안하였다. 이밖에도 Baysian 추정자를 이용한 확률추정 연구가 진행되었다[6,7]. 하지만 이경우에도, 설정된 표준 데이타가 왜곡된 경우에는 문제가 발생한다.

이러한 문제를 해결하기 위해 수렴성 신경망의 내삽특성과 비수렴성 신경망의 강인성을 취합하여, 보다 잡음에 강인한 선행 신경망을 구성하였다. 이를 위해 확률분포를 표현하는 신경망구조를 구현하였고 보다 타당한 결과값을 내기 위해 각 패턴으로 분석된 결과를 로버스트 통계학관점[12][13]에서 필터링하는 기법을 도입하여 outlier를 잘라내는 신경망필터(NeuroFilter)를 구성하였다.

본 논문에서는 먼저 Bayesian 추정자의 개념을 설명하고 수렴성 신경망의 확률값 수렴도를 증명한 후, 신경망필터(NeuroFilter)의 구조와 특성에 대해 전개하였다. 그리고 간단한 예제를 통해 기존의 신경망에 비해 우수함을 보였다.

2. 본론

1) Bayesian 추정자와 결정규칙

$\theta = \{\theta\}$를 패턴의 범주집합이라 하고, $X = \{x\}$를 분류될 입력벡터들의 집합이라고 하자. 일단 θ와 X는 유한집합이라고 가정한다. 그리고, $P(\theta)$를 θ상에서의 사전확률분포라 하고 $W(x \mid \theta)$는 패턴 θ에 대한 관측된 x의 조건확률값이라고 한다.

결정규칙(decision rule) $\psi : X \rightarrow \theta$에 따른 손실함수 $\rho : \theta \times \theta \rightarrow R^+ = [0, \infty)$를 정의하면 위험함수(expected risk function)는 다음과 같이 표현된다.

$$r(P, W, \psi) = \sum_{\theta \in \theta} \sum_{x \in X} P(\theta) W(x \mid \theta) \rho(\psi(x), \theta) \tag{1}$$

이제 모든 결정규칙들의 집합을 Δ라고 할 때 위험함수를 최소로 하는 결정함수 ϕ를 설정하면,

$$r(P, W, \phi) = \min_{\psi \in \Delta} r(P, W, \psi) \tag{2}$$

이때 결정된 r을 Bayes 최적 위험함수라 하고, 최소화시킨 결정규칙 ϕ를 조건확률 $P(\theta)W(x \mid \theta)$에 대한 Bayes 최적규칙이라 한다. Bayes 결정규칙은 항상 존재하며 다음 식으로 계산이 가능하다[9].

$$\phi(x) = \arg \min_{a \in \theta} \sum_{\theta \in \theta} P(\theta) W(x \mid \theta) \rho(a, \theta) \tag{3}$$

실제로는 사전확률분포 P나 조건확률분포 W가 사전에 주어지지 않기 때문에, 위와 같은 수식으로 실제문제를 풀기는 불가능하다. 하지만 상당히 많은 표본 데이타로 학습시킨 수렴성 신경망은, 수학적 표현은 불가능하지만 실제의 Bayes 규칙에 근사하는 결정함수를 만들어내어 사전확률 P와 조건부확률 W에 대한 정보를 이끌어 낼 수 있다[10,11,12].

BP같은 지도학습으로 신경망을 학습시킨다고 할 때 에너지함수는 다음과 같이 설정할 수 있다.

$$E = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{2} \| z(x_i, w) - t(\theta_i) \|^2 \tag{4}$$

여기서 학습표본 σ가 m개의 데이타로 이루어져 있다고 하자.

$$\sigma = \{ (\theta_1, x_1), (\theta_2, x_2), \cdots, \cdots, (\theta_m, x_m) \} \quad \theta_i \in \theta, \ x_i \in X, \quad i = 1, 2, \cdots, m \tag{5}$$

이때 $t(\theta)$는 패턴 θ에 해당하는 원하는 출력값이고 $z(x, w)$는 신경망의 가중치 w와 입력벡터 x에 따른 실제 출력값이다.

이제 연관확률분포(Joint Probability Distribution) $q_\sigma(\theta, x)$을 다음과 같이 정의한다.

$$q_\sigma(\theta, x) = \frac{1}{m} \sum_{i=1}^{m} \delta((\theta, x), (\theta_i, x_i)) \quad \forall (\theta, x) \in \theta \times X \tag{6}$$

위의 확률함수를 사용하여 (4)식을 다시 쓰면 다음과 같다.

$$E = \sum_{x \in X} \sum_{\theta \in \theta} q_\sigma(\theta, x) \frac{1}{2} \| z(x, w) - t(\theta) \|^2 = \sum_{i=1}^{|\theta|} \left[\sum_{x \in X} q_\sigma(\theta_i, x) \frac{1}{2} \| z(x, w) - t(\theta_i) \|^2 \right] \tag{7}$$

Bayes 추정규칙과 (7)식을 비교해 보면 에너지 함수 E는 Bayes 위험함수에 상응하고 출력 $z(x,w)$는 결정함수 $\psi(x)$에 상응한다는 것을 알 수 있다. 이제 최적의 E와 $z(x,w)$를 구하면 E^*는 Bayes 최적 위험함수 $r(P,W,\phi)$에, $z(x,w^*)$는 Bayes 최적 결정함수 ϕ에 해당한다.

계산상의 편의를 위해 A와 $t(\theta)$를 다음과 같이 정의하자.

$$A = \sum_{x\in X} q_0(\theta_i,x)\frac{1}{2}\| z(x,w)-t(\theta_i)\|^2 \quad (8)$$

$$t(\theta_i)=(0,0,0,\cdots,1,\cdots,0), \quad (i\text{번째 요소가 } 1) \quad (9)$$

이때, A는 (10)식이 된다.

$$\begin{aligned} A &= \frac{1}{2}\sum_{x\in X}[\sum_{k\neq i}^{|\theta|} q_0(\theta_k,x)\| z_i(x,w)\|^2 \\ &\quad + q_0(\theta_i,x)\| z_i(x,w)-1\|^2] \\ &= \frac{1}{2}\sum_{x\in X}[\sum_{k\neq i}^{|\theta|} q_0(\theta_k,x)\| z_i(x,w)\|^2 \\ &\quad + q_0(\theta_i,x)\| z_i(x,w)\|^2 + q_0(\theta_i,x)(-2z_i(x,w)+1)] \\ &= \frac{1}{2}\sum_{x\in X}[\sum_{k=1}^{|\theta|} q_0(\theta_k,x)\| z_i(x,w)\|^2 \\ &\quad + q_0(\theta_i,x)(1-2z_i(x,w))] \end{aligned} \quad (10)$$

다음과 같이 확률함수 $f_0(\theta_i,x)$를 정의하자.

$$f_0(\theta_i\mid x) = \frac{q_0(\theta_i,x)}{Q(x)}, \qquad Q(x)=\sum_{\theta\in\theta} q_0(\theta,x) \quad (11)$$

이때, A는 다음과 같다.

$$\begin{aligned} A &= \frac{1}{2}\sum_{x\in X}[\, Q(x)\sum_{k=1}^{|\theta|} f_0(\theta_k\mid x)\| z_i(x,w)\|^2 \\ &\quad + Q(x)f_0(\theta_i\mid x)(1-2z_i(x,w))] \end{aligned} \quad (12)$$

$$\begin{aligned} &= \frac{1}{2}\sum_{x\in X}[Q(x)\| z_i(x,w)-f_0(\theta_i\mid x)\|^2 \\ &\quad + Q(x)f_0(\theta_i\mid x)(1-f_0(\theta_i\mid x))] \end{aligned} \quad (13)$$

그러므로,

$$\begin{aligned} E &= \frac{1}{2}\sum_{i=1}^{|\theta|}\sum_{x\in X} Q(x)\| z_i(x,w)-f_0(\theta_i\mid x)\|^2 \\ &\quad + \frac{1}{2}\sum_{i=1}^{|\theta|}\sum_{x\in X} Q(x)f_0(\theta_i\mid x)(1-f_0(\theta_i\mid x)) \end{aligned} \quad (14)$$

따라서, E는 w함수의 항과 w에 대한 상수항으로 나눌 수 있다. 여기에 지도학습을 통하여 에너지 함수 E를 최소화하는 최적의 w^*와 $z_i^*(x,w^*)$를 구하면 첫째 항의 값이 최소가 된다. 이때 결과식 $z_i(x,w)$은 첫째 항에만 포함되며 합의 각 요소값이 모두 비음수이므로 $z_i(x,w)$는 $f_0(\theta_i,x)$에 수렴한다.

이때 $z(x,w^*)$는 다음과 같은 의미를 가진다.

$$z(x,w^*) = (\, f(1\mid x),\cdots,f(\theta\mid x),\cdots,f(\mid\theta\mid x)) \quad (15)$$

이것은 최적의 w^*를 구했을때, 손실함수로서 Hamming Distance d_H를 설정하면 신경망은 주어진 입력 x의 패턴 θ에 대한 사후확률값을 출력하게 된다는 의미가 된다. 즉, 지도학습으로 학습된 수렴성 신경망은 주어진 연관확률분포 $q_0(\theta,x)$에 대하여 에너지함수 E를 최소화시키는 최적의 Bayes 결정함수 ϕ를 결정한다.

2) 신경망필터(NeuroFilter)

먼저 출력값에 따른 패턴을 정의하기 위해 수렴성 신경망을 구성한다. 이때 패턴은 출력에 따라 적당한 크기로 나누며 그 간격은 복원의 일관성을 위해 등간격으로 일정하게 순서대로 양자화한다. 신경망의 입력갯수와 출력갯수는 각각 입력좌표 데이타의 차수와 양자화된 패턴갯수로 설정된다. 마지막의 WTA(Winner Take All)단으로는 홉필드 신경회로망이 많이 쓰인다.(그림 1 참조)

$$\begin{cases} u_i^{(l+1)} = \sum_{j=1}^{N_l} w_{ij}^{(l+1)} v_j^{(l)} + w_i^{(l+1)} \\ v_i^{(l+1)} = g(u_i^{(l+1)}), \qquad 1 \le i \le N_{l+1} \end{cases}$$

$$g(x) = \frac{1}{1+\exp(-x)}$$

$$E = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}\| z(x_i,w) - t(\theta_i)\|^2$$

지도학습 알고리즘으로는 에너지 함수 E를 최소화시키도록 학습하는 어떠한 알고리즘이라도 쓰일 수 있다.

기존의 수렴성 신경망처럼 $t(\theta_i)$를 직교좌표로 설정한다. 표준 데이타로 취한 샘플링 집합이 잡음으로 왜곡된 경우를 대비하여 하나의 표준입력 데이타에 대한 출력집합을 여러번 취하여 학습시킨다. 이때, 최종 WTA(Winner Take All)단 직전의 각 패턴에서 나오는 결과값은, 앞에서 유도한 바와 같이, 각각 해당하는 확률값이 됨을 알 수 있다. 출력되는 확률값은 표준 데이타로 취한 샘플링 집합에서 각 패턴에 해당되는 빈도수의 비율이며, 표본을 많이 취할 수록 Law of Large Number에 의해 실제 확률값으로 수렴한다. 이 확률값을 가중치로 하여 각 패턴에 해당하는 값을 모두 합하면, 모든 표본 데이타들의 평균치로서 기존의 선행 신경망과 같은 결과값이 출력된다.

그리고 이 구조는 각 패턴에 대한 확률값 분석이 이루어 지므로 그 분포도 해석이 가능하다. 이로써 로버스트 통계학[13]에서의 outlier판정이 가능하게 된다. 즉, 모든 표준 데이타로 학습시킨 신경망 구조에서 평균값을 중심으로 분포된 패턴경향 분석이 가능하여 신뢰성이 없는 데이타를 제거하는 필터링을 수행할 수 있다. 필터링에는 여러가지 방법이 있겠는데 본 논문에서는 Minimax Descending M-Estimator[13][14]를 사용하였다. 이러한 필터링을 거친 보정된 값들로 다시 정규화시켜 평균값을 취하면 보다 타당한 결과값을 구할 수 있다. 이것은 평균값에서 상당히 멀어진 출력값을 outlier로 간주하고 제거하여 선행 신경망에서 문제가 되는 잡음문제를 해결할 수 있는 구조적 방법이다.

$$\overline{x^*} = \frac{1}{n}\sum_{i=1}^{i=n}\psi(x_i)$$

$$\psi(x) = -\psi(-x) = \begin{cases} x, & \text{for } 0\le x\le a \\ b\tanh[\frac{1}{2}b(c-x)], & \text{for } a\le x\le c \\ 0, & \text{for } c\le x \end{cases}$$

$$(\,\overline{x}=0,\ \sigma=1 \text{ 이라고 가정})$$

제안된 신경망 구조는 신경망을 이용한 패턴 확률값 변환(Pattern Probability Transform)구조라 할 수 있는데, 마치 주파수 영역에서 고주파를 잘라내는 저주파 필터(Low Pass Filter)처럼 샘플 데이타를 각 패턴에 해당하는 확률값으로 변환(Transform)시킨 후 (LPF의 주파수 분석에 해당 : 확률값 분석), 평균값과의 차이가 설정된 값 이상인 패턴의 확률값은 가중치를 주거나 영으로 두어 제거하고 (c : Cutoff frequency에 해당), 다시 나머지 값들을 정규화시켜 최종 결과로 출력한다.

역변환 (Inverse Transform: 이산처리된 데이타의 연속값 복원)은 출력되는 필터링된 확률값을 가중치로 하여 각 패턴에 대한 값을 합하는 방법으로 수행한다.(그림 3 참조)

신경망 필터는 작은 잡음은 각 패턴값으로 수용하고 계산한다. 주로 검출오차에서 오는 작은 잡음은 데이타를 여러번 취하면 자동적으로 평균에 의해 실제 데이타에 근접하게 된다. 문제가 되는 것은 전체 평균값을 크게 왜곡할 수 있는 큰 잡음이다. 신경망 필터는 이러한 큰 잡음을 제거하는 필터링기법을 구현한 것이다. 즉, 작은 잡음은 양자화시켜 대표값으로 수용하고 그렇게 대표된 값에서 잘못된 데이타를 제거하는 것이다.

이것은 기존의 수렴성 신경망구조를 이용하되 결과를 확률적인 값으로 나오게 하여 보다 정확한 결과값을 추정해내는 방법이다. 이는 각 패턴에 해당하는 판정도를 분석하고 타당하지 않은 데이타를 잘라내는 필터링을 통해 큰 잡음를 제거하여 기존의 선행 신경망의 한계를 극복한 것이다. 뿐만 아니라 신경망의 내삽성질을 이용하여 표본과 표본사이의 값들도 적절하게 추정해 낼 수 있다. 특히, 출력 패턴의 갯수를 하나가 아닌 여러개로 늘려 학습함으로써, 표본값에 연관성을 가진 표본사이의 내삽복원범위가 넓어져, 표본 데이타의 변화가 거의 없는 부분에서는 취하는 표본갯수를 현저히 줄일 수 있게 된다. 이는 데이타전송이나 압축효율면에서도 좋은 결과를 나타낸다. 특히 이 구조는 주어진 모든 표본 데이타를 학습한 신경망으로, 사용자의 목적과 시스템의 특성에 따라 다양하게 필터링하여 보다 타당한 결과를 이끌어 낼 수 있다는 점에서 응용도가 매우 높다. 경우에 따라서는 경계선 구분을 확실히 하기 위해 부분적으로 WTA을 하여 불연속적인 출력값이 나오도록 사용할 수 있다.

3. 시뮬레이션

신경망의 기본적 특징인 비선형성을 이용하여 임의의 비선형 함수를 추정하는데 사용할 수 있다.

$$y = \Psi(t) + w$$

여기서 Ψ는 비선형함수이고, t는 입력, w는 잡음이다.

본 논문에서는 간단한 비선형 함수를 잡아 모의실험을 수행하였다.

$$y = \cos(t) + w \qquad t\in[0,\pi]$$

입력벡터로서 1차변수인 t를 택하고 w는 정규분포의 잡음을 발생시켰다. 기존의 선행 신경망으로는 1-5-1구조를 사용하였고, 신경망 필터는 출력을 51단계로 양자화하여 1-50-51구조를

사용하였다. 학습시 표본 데이타는 25%를 취하였고, 잡음발생시 표본 데이타는 6번씩 취하였다. 지도학습 알고리즘은 BP를 사용하였고, 학습률 η는 0.8, 모멘텀 α는 0.7로 하였다. 필터링하기위한 추정자로는 Minimax M-Estimator를 택하였고 오염도 ε는 0.1로 가정하여 a,b,c를 설정하였다. 다만 출력 패턴갯수가 많이 설정된 데 따른 평균값 편중을 보정하기 위해 확률값의 임계치를 정하고 그 이상의 값이 나오는 패턴에 대해서만 고려하였다.(그림 4, 그림 5 참조)

잡음이 없을 때 기존의 선행 신경망보다 크게 성능이 떨어지지 않았고, 잡음이 있는 경우 현저한 성능향상을 보였다.

마지막 부분의 오차는 BP알고리즘의 학습시 국부최소(Local Minimum)에 걸린것으로 생각된다. 시뮬레이티드 어닐링(simulated Anealing)등의 방법을 써서 보완한다면 이러한 문제는 피할 수 있을 것이다.

4. 결론

본 논문이 제시한 방법은, 잡음에 약한 선행 신경망의 단점을 보완하기 위해 상대적으로 잡음에 강한 수렴성 신경망의 구조를 도입하여 각 패턴에 대한 분석을 통해 잡음에 해당하는 패턴 출력값을 잘라내는 필터링을 거침으로써, 보다 타당하고 신뢰성있는 결과값을 내는 신경망을 구성하는 방법이다. 이것은 제한된 데이타로 전체 데이타를 추정해 내는 방법으로 신경회로망을 사용하되, 왜곡된 데이타의 영향을 최소화시킬 수 있으며 표본사이의 값들에 대해 보다 많은 연관성을 줄 수 있어 데이타압축과 전송에도 효과적인 구조이다. 그리고 필요에 따라 필터링을 다르게 적용하여 목적과 상황에 맞게 타당한 결과를 이끌어 낼 수 있다.

이 방법은 잡음이 심한 데이타의 복원이나, 불완전한 검출기가 취한 데이타의 복원 등에 특히 효과가 있다. 다만 패턴은 항상 순차적인 순서를 가져야 하며 그 크기 간격도 일정해야 한다.

5. 참고문헌

[1] S.Mukhopadhyay and K.S.Narendra, "Disturbance Rejection in Nonlinear Systems Using Neural Networks", IEEE Trans. Neural Networks, vol.4, No.1, pp.63-72, Jan.,1993.

[2] W.Y.Huang and R.P.Lippmann, "Comparisons Between Neural Net and Conventional Classifiers", in Proc. IEEE Int. Conf. Neural Networks, vol.4, pp.485-493, 1987.

[3] B.Kosko, "Stochastic Competitive Learning", IEEE Trans. Neural Networks, vol.2, No.5, pp.522-529, Sep., 1991.

[4] Hans G.C.Traven, " A Neural Network Approach to Statistical Pattern Classification by 'Semiparametric' Estimation of Probability Density Functions", IEEE Trans. Neural Networks, vol.2, No.3, pp.366-377, May., 1991.

[5] D.F.Specht, "Probabilistic Neural Networks and the Polynomial Adaline as Complementary Techniques for Classification", IEEE Trans. Neural Networks, vol.1, No.1,pp.111-121, Mar., 1990.

[6] T.Kohonen, G.Barna and R.Chrisley, "Statistical Pattern Recognition with Neural Networks : Benchmarking studies", in Proc. IEEE 2nd. Int. Conf. Neural Networks, vol.1, pp.61-68, 1988.

[7] S.Geman and D.Geman, "Stochastic Relaxation, Gibbs Distributions and Bayesian Restoration of Images", IEEE Trans. Patt. Anal. Machine Intell., vol.PAMI-6, pp.721-741, Nov., 1984.

[8] L.L.Scharf, *Statistical Signal Processing*, Addison - Wesley Publishing Co., 1991.

[9] F.Kanaya and S.Miyake, "Bayes Statistical Behavior and Valid Generalization of Pattern Classifying Neural Network.", IEEE Trans. Neural Networks, vol.2, No.4, pp.471-475, Jul., 1991.

[10] S.Miyake and F.Kanaya, "A Neural Nework Approach to a Bayesian Statistical Decision Problem", IEEE Trans. Neural Networks, vol.2, No.5, pp.538-540, Sep., 1991.

[11] E.Wan, "Neural Network Classification : A Bayesian Interpretation", IEEE Trans. Neural Networks, vol.1, pp.303-305, Dec., 1990.

[12] F.R.Hampel *et al.*, *Robust Statistics*, John Wiley & Sons Inc., 1986.

[13] P.J.Huber, *Robust Statistics*, Johm Wiley & Sons Inc., 1981.

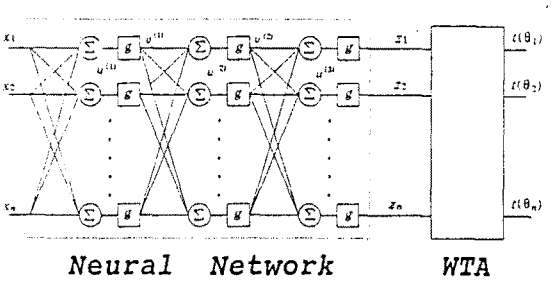

그림 1. 수렴성 신경 회로망

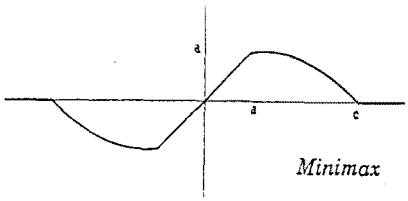

그림 2. Minimax Descending M-Estimator

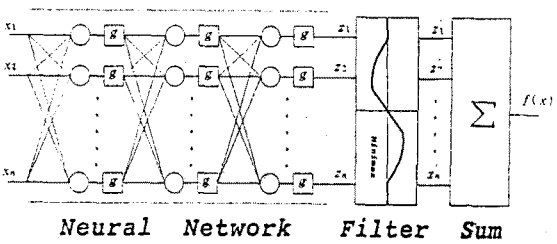

그림 3. 신경망 필터 구성

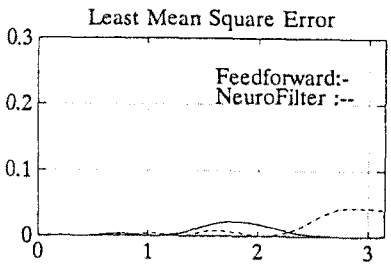

그림 4. 잡음이 없는 경우

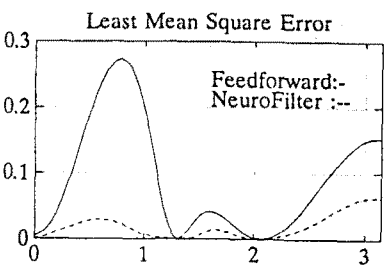

그림 5. 잡음이 있는 경우