

다층퍼셉트론의 잡음 강건성

On the Noise Robustness of Multilayer Perceptrons

오상훈

목원대학교 정보통신전파학부

Oh Sang-Hoon

Div. Information Communication and Radio
Engineering, Mokwon University

요약

이 논문에서는 MLP(Multilayer Perceptron)가 지닌 잡음 강건성에 대한 통계학적 분석을 하였다. 또한, MLP의 잡음 강건성을 향상시키기 위한 선형적 전처리 단계로써, ICA(independent component analysis)와 PCA(principle component analysis)를 고려하여, 이들이 지닌 잡음처리 효과를 분석한 후, MLP와 접목시 나타나는 잡음 강건성의 향상 여부를 필기체 숫자 인식의 시뮬레이션으로 확인하였다.

Abstract

In this paper, we analyze the noise robustness of MLPs(Multilayer perceptrons). Also, as a preprocessing stage of MLPs to improve noise robustness, we consider the ICA(independent component analysis) and PCA(principle component analysis). After analyzing the noise reduction effect using PCA or ICA, we verify the noise robustness of MLPs through handwritten-digit recognition simulations.

I. 서론

다층퍼셉트론(MLP: Multilayer Perceptron)은 충분한 수의 중간층 노드들이 있으면 임의의 함수를 제한된 오차 내에서 근사화 할 수 있으며[1], 이를 근거로 패턴 인식, 음성인식, 시계열 예측, 비선형 제어, 이동통신, ATM 교환 등 다양한 분야에 응용되고 있다[2][3][4][5][6][7]. 한편, 이러한 응용분야에서 MLP의 입력은 많은 잡음을 지닐 수 있으며, 이러한 경우의 성능저하에 대한 분석을 MLP의 실질적 문제 응용을 위해 행해져야 한다.

이제까지 행해졌던 MLP의 잡음에 의한 성능저하 분석은 먼저 중간층 노드들이 문턱함수(threshold function)인 경우에 행해졌다[8][9]. 중간층 노드들이 시그모이드 비선형 변환을 지닌 경우는, 출력을 입력에 대한 미분의 형태로 구해진 것이 대부분이었다[10].

이 논문에서는, 입력에 가우시안 잡음이 섞일 경우 출

력에 나타나는 성능을 인식률의 관점에서 살펴본다. 그리고, 잡음에 의한 인식률 저하 영향을 감소시키기 위한 선형적 전처리 단계로 PCA(principle component analysis)와 ICA(independent component analysis)를 고려하여, 이들이 지닌 잡음 감소 효과를 분석한 후, 이를 MLP와 접목 시 나타나는 잡음 강건성을 필기체 숫자의 시뮬레이션으로 확인하고자 한다.

II. MLP의 잡음 강건성

MLP가 잡음에 대한 강건성을 지니는 것을 정성적으로 설명하면 크게 두 가지이다. 첫째로는, 중간층 노드들 간의 orthogonal 성질 때문이다. 중간층 뉴런들은 학습 과정 중에 orthogonal 하도록 학습되는 경향이 있다. 따라서, 비록 입력에 잡음이 같이 들어오더라도 중간층 노드의 출력들은 서로 orthogonal 하므로 출력층 노드 계산 시 서로 상쇄되어 잡음량이 감소되는 효과가 발생한다. 두 번째는 정보이론적 관점에서 그 원인을 찾

* 이 논문은 2003년도 한국학술진흥재단의 지원에 의하여 연구되었음(KRF-2003-003-D00245)

을 수 있다. 즉, MLP의 학습 시 각 계층들은 입력 패턴의 인식에 유용한 정보는 간직하고, 유용하지 않은 정보는 제거하려고 한다. 이렇게 학습된 MLP에 입력과 함께 잡음이 들어오면, 입력 패턴의 분류에 유용하지 않은 잡음은 계층적 정보처리 과정에서 제거되는 효과를 지닌다.

이제 MLP의 잡음 강건성을 패턴인식으로 국한하여, 잡음이 패턴의 오인식률에 미치는 영향을 수식적으로 유도하는 형태로 분석하겠다. N-H-M (N 입력노드, H 중간층 노드, M 출력층 노드) 구조 MLP가 학습이 완료되었다고 하자. 이때, 입력패턴 $\mathbf{x}^{(p)} = [x_1^{(p)}, x_2^{(p)}, \dots, x_N^{(p)}]^T$ 가 $N(0, \sigma_p)$ 분포의 i.i.d. (independent identically distributed) 잡음 $\mathbf{x}^* = [x_1^*, x_2^*, \dots, x_N^*]^T$ 에 의해 변형되면, 입력은

$$\mathbf{x} = \mathbf{x}^{(p)} + \mathbf{x}^* \quad (1)$$

이 된다. 그리고, j 번째 중간층 노드에 입력되는 가중치 합은

$$a_j = a_j^{(p)} + \sum_{i=1}^N w_{ji} x_i^*, \quad j = 1, 2, \dots, H \quad (2)$$

와 같이 된다. 여기서, $\mathbf{W} = (w_{ji})$ 는 입력에서 중간층으로의 연결 가중치 행렬이고,

$$a_j^{(p)} = w_{j0} + \sum_{i=1}^n w_{ji} x_i^{(p)} \quad (3)$$

이다. 그러면, 가우시안 확률변수의 합은 가우시안이므로, $\mathbf{a} = [a_1, a_2, \dots, a_H]^T$ 는 가우시안 확률벡터가 되는 것을 쉽게 보일 수 있을 것이며, 그 p.d.f.(probability density function)는

$$f_a(\mathbf{a}) = \frac{1}{\sqrt{(2\pi)^H |\mathbf{C}|}} \exp\left[-\frac{1}{2}(\mathbf{a} - E\mathbf{a})^T \mathbf{C}^{-1}(\mathbf{a} - E\mathbf{a})\right] \quad (4)$$

이 된다. 여기서, $\mathbf{C} = (C_{ji})$ 는 \mathbf{a} 의 covariance 행렬

이며,

$$C_{ji} = \sigma_p^2 \sum_{k=1}^N w_{jk} w_{ik} \quad (5)$$

이다. 중간층 노드의 출력 벡터 $\mathbf{h} = [h_1, h_2, \dots, h_H]^T$ 는 시그모이드 비선형 변환

$$h_i = \frac{2}{1 + \exp(-a_i/T)} - 1 \quad (6)$$

에 의해 계산된다. 이 시그모이드 함수를 이용하여 \mathbf{h} 의 p.d.f.를 유도하면

$$f_h(\mathbf{h}) = \frac{f_a\left(-T \ln\left[\frac{2}{h_i+1} - 1\right], \dots, -T \ln\left[\frac{2}{h_H+1} - 1\right]\right)}{|J_h(a_1, a_2, \dots, a_H)|} \quad (7)$$

이고, 여기서

$$J_h(a_1, a_2, \dots, a_H) = \left(\frac{1}{2T}\right)^H \prod_{k=1}^H (1 - h_k)^2 \quad (8)$$

이다. \mathbf{h} 는 가우시안이 아니지만 분산은 유한함을 알 수 있다.

이제 k 번째 출력 노드에 입력되는 가중치 합은

$$b_k = v_{k0} + \sum_{j=1}^H v_{kj} h_j, \quad k = 1, 2, \dots, M \quad (9)$$

이고, 여기서 $\mathbf{V} = (v_{kj})$ 는 중간층과 출력층을 연결하는 가중치 행렬이다. \mathbf{b} 의 p.d.f.를 유도하기 위해서 $H-M$ 개의 보조 변수를

$$b_k = h_k, \quad k = M+1, M+2, \dots, H \quad (10)$$

와 같이 도입하자.

그리고, $\boldsymbol{\beta} = [b_1 - v_{10}, \dots, b_M - v_{M0}, b_{M+1}, \dots, b_H]^T$ 로 두면,

$$\beta = Zh \quad (11)$$

와 같이 표현할 수 있다. 여기서,

$$Z = \begin{bmatrix} v_{11} & v_{12} & \dots & v_{1,M+1} & \dots & v_{1H} \\ v_{21} & v_{22} & \dots & v_{2,M+1} & \dots & v_{2H} \\ \vdots & \vdots & & \vdots & & \vdots \\ v_{M1} & v_{M2} & \dots & v_{M,M+1} & \dots & v_{MH} \\ & & & 1 & & 0 \\ & & & & \ddots & \\ & & & & & 1 \end{bmatrix} \quad (12)$$

이다. 출력층 연결 가중치 행렬 V 가 full rank이면 Z 는 non-singular 이므로, 역행렬이 존재하고

$$h = Z^{-1}\beta \quad (13)$$

이다. 이 관계를 이용하면 β 의 p.d.f.는

$$f_{\beta}(\beta) = \frac{f_h(Z^{-1}\beta)}{|J(h_1, h_2, \dots, h_H)|} \quad (14)$$

로 구해지며, $J(h_1, h_2, \dots, h_H)$ 는 수식 (11)의 Jacobian 이다. 그러면, b 의 p.d.f.는 수식 (14)를 모든 보조변수에 대해 적분하여

$$f_b(b) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f_{\beta}(\beta) dh_{M+1} \dots dh_H \quad (15)$$

로 구할 수 있다.

만약, 입력 패턴 $x^{(p)}$ 가 첫 번째 class에 속하는 패턴이라면, 잡음의 영향 때문에 오인식될 확률은

$$\Pr(E_p) = 1 - \int_{-\infty}^{\infty} \int_{-\infty}^{b_1} \dots \int_{-\infty}^{b_1} f_b(b) db_M \dots db_2 db_1 \quad (16)$$

로 구해진다. 이 수식을 실제 수치해석적으로 계산하려면 H 차원의 중적분에 의한 계산량이 아주 크다.

III. 선형 변환에 의한 잡음 강건성 향상

1. PCA(Principle Component Analysis)

PCA는 일련의 벡터들로 주어진 데이터들의 평균을 0으로 처리하고 나서 상관행렬을 구한 후 이를 근거로 eigenvalue와 eigenvector들을 구하고, 이를 이용하여 데이터 벡터에 선형 변환을 가하는 것이다. 즉, PCA에 의해 구한 eigenvector들로 행렬을

$$E = [e_1, e_2, \dots, e_N] \quad (17)$$

와 같이 구성하자. 이때 각 eigenvector e_i 는 eigenvalue λ_i 를 지닌다고 할 때, 이들이 정렬되어 $\lambda_1 > \lambda_2 > \dots > \lambda_N$ 이 되도록 한다. 그러면, λ_i 는 데이터들이 e_i 방향으로 가지는 분산에 해당 된다. 행렬 E 를 이용하여 데이터를

$$u = Ex \quad (18)$$

와 같이 선형 변환을 시키면 PCA에서 구한 eigenvector들을 이용하여 데이터의 좌표축을 변형한 것이다. 물론, e_1 방향 데이터의 분산이 가장 크며 e_N 방향의 분산은 아주 작게 나타난다. 대부분의 데이터는 eigenvalue들이 큰 값에서 감소하다가 갑자기 아주 작은 값을 지니게 된다. 즉, $\lambda_1 \dots \lambda_L$ 는 어느 정도 큰 값을 지니고, $\lambda_{L+1}, \dots, \lambda_N$ 는 아주 작은 값을 지닌다고 하면, u 는 L 차원으로 줄여도 데이터에서 소실되는 정보가 아주 미미하다. 이와 같은 관점에서

$$\hat{u} = [e_1 e_2 \dots e_L] x = \sum_{j=1}^L (e_j \cdot x) e_j \quad (19)$$

와 같이 차원이 축소된 근사 데이터를 만든다.

식 (1)과 같이 신호원 벡터에 잡음이 섞인 경우 PCA에 의한 잡음 감소 효과를 살펴보자. 먼저, \hat{u} 에 있는 신호 성분의 power는

$$P_s = \sum_{j=1}^L (e_j \cdot x^{(p)})^2 \approx \sum_{j=1}^L (x_j^{(p)})^2 \quad (20)$$

로 주어지는데, 이것이 위의 수식과 같이 $x^{(n)}$ 의 power와 거의 같은 이유는 PCA의 성질 때문이다. 다음으로 \hat{u} 의 잡음성분 power는

$$P_n = \sum_{j=1}^L (e_j \cdot x^n)^2 = \sum_{j=1}^L (x_j^n)^2 = L\sigma_j^2 \quad (21)$$

이다. 여기서, 식 (21)은 x^n 이 i.i.d.이기 때문이다.

따라서, x 의 PCA 변환 후 \hat{u} 을 얻으면 잡음의 power는 $N\sigma_j^2$ 에서 $L\sigma_j^2$ 으로 줄어들므로 $10\log(N/L)$ 만큼 SNR이 향상된다.

2. ICA(Independent Component Analysis)

ICA 알고리즘[11]에 따라 얻은 unmixing 가중치 행렬 $M \times N$ 을 이용한 선형 변환 결과는 식(18)와 같이 표현된다. 이때, 신호 성분의 power는

$$P_s = \sum_{j=1}^N (e_j \cdot x^{(n)})^2 \quad (22)$$

이고, 잡음 성분의 power는

$$P_n = \sum_{j=1}^N (e_j \cdot x^n)^2 \quad (23)$$

로 구해진다. ICA에 의해 학습된 unmixing 가중치 행렬을 이루는 각 vector가 $\|e_j\|^2=1$ 이라고 하면 식 (22)와 (23)은 각각 x 의 신호 및 잡음의 power와 같다. 즉, $M \times N$ 형태의 ICA에서는 잡음의 감소가 이루어지지 않는다. 이는 PCA의 경우도 마찬가지일 것이다.

다음으로, ICA의 unmixing 행렬의 크기가 $P \times N$ 인 경우를 고려해보자. 이 경우, 신호 및 잡음의 power는 각각

$$P_s = \sum_{j=1}^P (e_j \cdot x^{(n)})^2 \quad (24)$$

와

$$P_n = \sum_{j=1}^N (e_j \cdot x^n)^2 \quad (25)$$

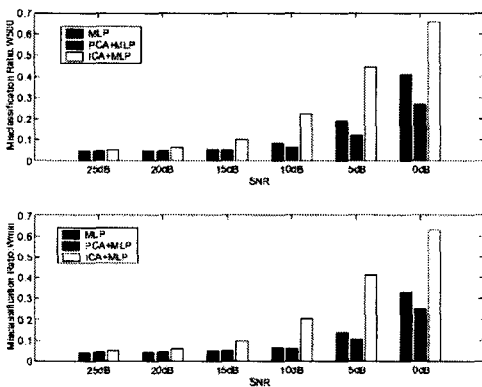
로 구해진다. 만약 PCA 경우에서처럼 ICA에 의해 얻은 unmixing 행렬에 의한 선형변환이 신호의 power를 그대로 간직한다고 하면 잡음의 power는 줄어들어 SNR의 개선이 이루어진다. 그렇지만, ICA의 변환 후 각 성분들은 통계적 독립성을 지니게 되므로, 이는 MLP가 학습에 의해 데이터로부터 규칙을 찾아내고자 하는 특성에 반한 형태의 데이터를 만든 것이므로 MLP가 학습하기에 어려운 형태의 변환이 이루어진 것이다.

즉, ICA를 전처리 단계로 한 후 MLP와 접합하여 인식기를 구성하는 것은 SNR 관점에서 아무런 이득을 기대할 수 없으며 오히려 MLP의 학습이 어려운 형태의 데이터 변환 때문에 학습 성능이 아주 저하되므로 적합하지 않은 형태이다.

IV. 시뮬레이션

MLP의 잡음 강건성을 확인하기 위한 시뮬레이션 대상으로 CEDAR 데이터베이스의 필기체 숫자 인식을 다룬다. 학습대상은 18,468개의 12×12 필기체 숫자 영상이며, 각 픽셀은 16 레벨의 그레이 값을 지닌다. 이를 144-30-10 구조의 MLP에 학습시킨 후, 2214개의 학습시키지 않은 시험 영상에 가우시안 잡음이 섞인 경우의 인식을 저하를 조사하였다. 다음으로, 학습대상인 144차원의 18,468개 필기체 숫자 영상에 대해 PCA를 수행하여 eigenvalue를 조사한 후, 40차원으로 데이터 차원을 축소한 데이터로 40-30-10 구조의 MLP를 학습시킨 후, 역시 마찬가지로 가우시안 잡음이 섞인 시험 영상 숫자의 인식을 저하를 조사하였다. 마지막 비교 방법은 144차원의 학습 데이터들로 ICA를 학습시켜 얻은 144차원의 데이터를 144-30-10 구조의 MLP에 학습시킨 후, 마찬가지로 방법으로 잡음의 강건성을 조사하였다. 이렇게 조사한 결과를 그림 1에 그렸다. 이 결과는 학습 횟수 500 sweep 후 얻은 가중치를 지닌 MLP를 사용한 결과와 학습 과정 중 가장 일반화 성능이 좋은 가중치를 지닌 MLP를 사용한 결과이다.

먼저, ICA로 전처리 후 MLP를 학습시킨 경우에는 예상하였던 바와 같이 잡음의 강건성이 제일 나쁘다. 이는 ICA에 의한 데이터 변환이 잡음 제거 효과는 지니지 못하며, 오히려 MLP의 학습이 어려운 상태로 데이터 변환을 수행하기 때문이다. 그리고, PCA로 전처리 후 MLP에 학습시킨 경우가 MLP 경우보다 잡음 강건성이 많이 향상되었음을 볼 수 있는 데, 이는 앞에서 주장하였던 바와 같이 PCA에서 얻은 eigenvector들을 이용한 차원 축소가 신호의 특성은 최대한 유지하면서 잡음을 제거하여 주었기에 얻어진 결과이다.



▶▶ 그림 1. 필기체 숫자 영상의 SNR에 따른 오인식률

PCA가 차원 축소 시 잡음 제거 효과를 지니므로, ICA로 차원을 축소한 경우에도 잡음 제거 효과가 있는지를 확인하기 위하여 시뮬레이션을 하였지만, ICA에 의한 차원 축소를 얻어진 데이터들은 서로 독립적인 성향 때문에 MLP에서 제대로 학습이 되지 않았다.

V. 결론

이 논문에서는 MLP가 지닌 잡음의 강건성을 통계학적으로 유도하였으며, 그 강건성을 향상시키기 위한 선형적 변환의 방법으로 PCA와 ICA를 고려하여 보았다. ICA는 잡음 제거 효과를 지니지 못하는 반면에 오히려 데이터의 특성들이 독립적이 되도록 하여, 데이터들로부터 학습을 통해 법칙을 찾아내는 MLP의 학습능력이 오히려 저하됨을 확인 하였다. PCA는 차원 축소 시 신호는 유지하면서 잡음을 제거하는 효과가 있어서, MLP

의 잡음 강건성이 향상됨을 시뮬레이션으로 보였다. 실제의 환경에서 MLP를 이용한 응용 문제를 다룰 때, PCA에 의한 전처리 과정은 잡음의 강건성을 향상시켜 줄 것이다.

■ 참고문헌 ■

- [1] K. Hornik, M. Stincombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks*, vol. 2, pp.359-366, 1989.
- [2] R. P. Lippmann, "Pattern classification using neural networks," *IEEE Communication Magazine*, pp.47-64, Nov. 1989.
- [3] J. B. Hampshire II and A. H. Waibel, "A novel objective function for improved phoneme recognition using time-delayed neural networks," *IEEE Trans. Neural Networks*, vol. 1, pp.216-228, June 1990.
- [4] A. S. Weigend and N. A. Gershenfeld, *Time Series Prediction: Forecasting the Future and Understanding the Past*, Addison-Wesley Publishing Co., 1994.
- [5] K. S. Narendra and K. Parthasarathy, "Identification and control of dynamic system using neural networks," *IEEE Trans. Neural Networks*, vol. 1, pp.4-27, 1990.
- [6] B. Aazhang, B.-P. Paris, and G. C. Orsak, "Neural networks for multiuser detection in code-division multiple-access communications," *IEEE Trans. Communications*, vol. 40, pp.1212-1222, 1992.
- [7] I. W. Habib, A. A. Tarraf, and T. N. Saadawi, "Intelligent traffic control for ATM broadband networks," *IEEE Communication Magazine*, vol. 33, pp.76-85, 1995.
- [8] M. Stevenson, R. Winter, and B. Widrow, "Sensitivity of feedforward neural networks to weight errors," *IEEE Trans. Neural Networks*, vol. 1, pp.71-90, Mar. 1990.
- [9] S.-H. Oh and Y. Lee, "Sensitivity analysis of single hidden-layer neural networks with threshold functions," *IEEE Trans. Neural Networks*, vol. 6, pp. 1005-1007, July 1995.
- [10] J. Y. Choi and C.-H. Choi, "Sensitivity analysis of multilayer perceptron with differentiable activation transformations," *IEEE Trans. Neural Networks*, vol. 3, pp.101-107, Jan. 1992.
- [11] T.-W. Lee, et al., "A unifying information-theoretic framework for independent component analysis," *Computers & Mathematics with Applications*, vol. 31, no. 11, pp.1-21, March 2000.