제 19 회 한국정보처리학회 춘계학술발표대회 논문집 제 10 권 제 1 호 (2003. 5)

다층 퍼셉트론에서 구조인자 제어의 영향

윤여창
우석대학교 전산정보학부
e-mail : yoonyc@woosuk.ac.kr

On the factors controlling effects at MLP Networks

YeoChang Yoon
Dept. of Computer Science and Statistics, Woosuk University

요 약

다층 퍼셉트론(Multi-Layer Perceptron, MLP) 구조를 이용한 비선형 적합은 실제문제에 매우 다양하게 적용되고 있다. 이때 일반화된 MLP 구조의 적합을 위해서는 은닉노드의 개수, 초기 가중값 그리고 학습 회수와 같은 구조인자들을 함께 고려해야 한다. 만약 구조인자들이 부적절하게 선택되었다면 일반화된 MLP 구조의 적합효율이 매우 저하될 수 있다. 그러므로 MLP 구조에 영향을 주는 인자들의 영향을 살펴보는 것은 중요한 문제다. 이 논문에서는 제어상자(controller box)를 통한 학습결과와 더불어 MLP 구조를 일반화할 때 영향을 줄 수 있는 구조인자(factor)들의 실증분석과 이들의 상대효과를 살펴본다.

1. 서 론

MLP 구조에 대한 학습은 일반적으로 유한개수의 표본을 이용한다. 학습표본은 보다 많은 입력과 출력의 쌍으로 구성되며, 전체 모집단을 일반화시키기 위한 네트워크로 학습시키기 위하여 매우 중요하다. 제한된 표본으로부터 추정된 모형을 이용하여 일반화된 네트워크의 적합능력을 제어하기 위해서는 복잡도 제어(complexity control)를 이용한다[1,2].

본 논문에서는 Zhong 과 Cherkassky[3]의 MLP 구조의 복잡도 제어에 관하여 논의하고 Yoon[4]의 제어상자를 통한 학습과정을 이용하여 그 결합 영향을 비교한다. 일반적으로 모형의 복잡도를 제어하기 위한 방법은 은닉노드의 개수 조절이다. 그러나 MLP 구조에 일반적으로 사용되는 역전파(backpropagation) 알고리즘은 은닉노드 개수나 가중값을 쉽게 정량화시킬 수 없다.

학습 알고리즘의 일반화된 영향을 이해하기 위해서 다음과 같은 논의가 필요하다. 먼저 가중값 공간의 경로(path)를 따라 MLP 학습을 위한 최적의 비선형 처리절차를 규정해야 하며, 역전파 알고리즘에서 경사하강추적(gradient descent)방법상의 경로를 따라 경험적인 오류를 감소시켜야 한다. 이를 통한 가능한 학습 결과가 적절한 경로상에 위치하여야 한다. 학습 과정의 경로는 학습 자료, 비선형 변환함수, 학습 경로상의 초기 가중값 그리고 최종 학습결과를 보여주는 학습회수등에 주로 영향을 받게 된다.

MLP 의 오차평면은 많은 지역 최소값들을 갖기 때문에 특별한 지역최소값은 학습 경로상의 초기 가중값 그리고 최종 학습결과를 보여주는 학습회수를 변화시키면서 찾는다. 예를들어 초기 가중값을 확률난수로부터 발생시킨 작은 값으로 설정할 때, 역전파 알고리즘은 매우 작은 가중값으로 인한 지역최소값에 수렴하는 경향이 있다. 그러나 매우 큰값을 경사하강추적법 또는 허용오차로 부여하여 학습회수를 제어하면 초기 가중값에서 시작된 경로상에서 적절한 해들을 찾지 못하는 경우도 발생한다. 따라서 초기 가중값들은 학습에 매우 큰 영향이 있으므로 학습과정중에 실시간으로 재설정함으로써 지역최소값을 벗어날 수 있게 하는 방법을 함께 고려할 수 있다.

역전파 알고리즘을 이용한 대부분의 연구들은 초기 가중값들의 조건과 변환함수들을 적절히 이용한다. 이 논문에서는 실제적인 학습의 결과로서 모형 복잡도에 대한 초기 가중값과 제어상자를 통한 초기 가중값 그리고 최종 학습결과를 보여주는 학습회수등의 구조인자들에 대한 결합영향을 살펴본다. 결합영향을 살펴보는 것은 MLP 구조를 과다적합하여 추정하는 것을 제어할 수 있다. 다음 장에서 우리는 SRM 구조하에서 서로 다른 구조인자들에 대한 MLP 복잡도 제어를 논의한다. 3 장에서는 MLP 구조의 일반화를 제어하기 위

하여 사용될 수 있는 MLP 복잡도 제어를 논의한다. 4 장은 이에 대한 결론을 다룬다.

2. SRM 을 이용한 MLP 복잡도 제어

본 논문에서는 SRM(Structural risk minimization) 구조와 Yoon 의 제어상자를 이용한 구조인자의 실시간 선택방법에 주로 기초하고 있다[2,4]. SRM 에서 학습은 주어진 MLP 구조 하에서 학습 오차의 최소화 그리고 가장 작은 예측오차를 제공하는 요소를 선택할 수 있는 모형선택등에 영향을 주로 받는다.

MLP 에서 그 구조는 초기 가중값, 은닉노드의 개수 그리고 허용오차에 따라 각각 다음과 같이 정의될 수 있다[2].

1)초기 가중값

먼저 다음과 같은 구조를 고려하자.

$$S_i = \{A : f(x,w), \| w_j^0 \| \le c_i\}, c_1 < c_2 < c_3 \cdots \quad (1)$$

여기서 $w = \{w_j^0\}$ 는 알고리즘 A 의 초기 가중값 벡터이며 i 는 그 구조의 첨자이다. 경사하강추적법은 초기 가중값 주변에서 지역최소값을 찾기 때문에, 전역최소값은 $\| w_j^0 \| \le c_i$ 을 만족하는 여러가지 초기조건에서 경험적 위험을 최소화시키며 이때 가장 좋은 가중값을 선택한다. 식 (1)의 구조식 S_i 는 초기 조건인 w 를 갖는 함수들에 적용되는 모수 추정을 위한 최적 알고리즘 A 에 대하여 각각 정의된다. 그리고 경험적 오류는 $\| w_j^0 \| \le c_i$ 를 만족하는 모든 초기 조건에 대하여 최소화 된다.

또한 학습의 시작단계에서 확률난수 대신에 학습자료의 특성에 따라 실시간으로 가장 잘 적합될 수 있는 초기값을 선택함으로써 지역 최소값을 쉽게 벗어나게 하는 학습과정을 고려할 수 있다[3]. 이러한 초기 가중값의 선택은 다음 장에서 설명하는 모형 선택에 이용될 수 있다.

2) 은닉노드의 개수

MLP 의 사전적 정의는 다음과 같다.

$$f(x,w,V) = \sum_{j=1}^{m} w_j g_j(x, v_j) + w_0 \quad (2)$$

여기서 $g_j(x, v_j)$ 는 시그모이드형 변환함수이며 모수 v_j 를 갖는다. 이 구조의 각 요소는 MLP 네트워크의 구조인자들이며, m 은 은닉노드의 개수다. 따라서 모형선택의 작업은 주어진 학습자료에 대한 최적의 은닉노드 개수를 갖는 MLP 네트워크를 선택하는 문제다.

3) 허용오차

학습의 허용오차를 크게 하는 것은 MLP 네트워크의 과다적합을 피하는 일반적인 방법이다. 여기서는 수렴과정의 정형화된 영향을 소개하고 과다적합되기 전에 학습을 정지시키는 방법으로 이용될 수 있는 허용오차를 논의한다. 다음과 같은 구조를 고려하자.

$$S_i = \{A : \ R_{emp} \le e_i\}, \ e_1 > e_2 > e_3 \cdots \quad (3)$$

여기서 R_{emp} 은 경험적인 오류를 줄이기 위한 학습정지 시점을 의미하고, e_i 는 최적 알고리즘 A 에 대한 최종 학습오차이다.

지금까지의 접근방법들은 MLP 모형의 복잡도를 제어하기 위하여 사용될 수 있다. 이 논문에서는 이들 요소들의 영향을 경험적으로 살펴본다. MLP 복잡도 제어에 대한 구조인자들은 경험적으로 특정 인자가 모형선택에 이용될 때 다른 인자는 고정시킨다. MLP 모형의 예측 효용은 상자그림(boxplot)을 이용한다.

3. 실증분석

제어상자를 통한 초기 가중값의 선택과 더불어 앞장에서 설명한 세가지 구조인자들이 모형의 일반화에 어떻게 영향을 주는지를 조사하기 위하여 적은 개수의 학습자료를 이용하여 많은 개수의 모수들로 이루어진 MLP 네트워크를 학습한다. 즉 다음과 같은 함수를 추정하기 위한 네트워크를 학습한다고 하자.

$$f(x) = e^{-(x-1)^2} + e^{-(x+1)^2}, \ x \in [-2.5, 2.5]. \quad (4)$$

학습자료: 12 개의 모의 학습자료 x 는 구간 [*-2.5, 2.5*]의 균등분포에서 발생되며, y 값은 평균 0, 분산 0.005 인 정규분포를 따르는 오차를 포함한다.

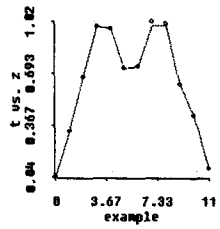
네트워크 구조: 한 개의 입력값 x 와 그에 대응하는 한 개의 출력값 y 그리고 n 개의 은닉노드로 이루어진 $1 \times n \times 1$ MLP 네트워크를 이용한다. 입력과 출력노드는 선형이며 은닉층은 로지스틱 변환함수를 이용한다.

학습알고리즘: 일반적인 LM(Levenberg-Marquardt) 알고리즘을 이용한다. LM 알고리즘은 비선형함수의 오차제곱합(MSE)을 최소화시키기 위하여 고안되었으며 뉴턴 방법의 일종이다. 이 방법은 효율성을 측정하는 척도가 오차제곱합인 신경망 학습에 잘 적용된다. 그리고 LM 알고리즘은 적정 개수의 네트워크 모수에 대한 가장 빠른 신경망학습 알고리즘으로 알려져 있다[5]. 최대 학습 회수는 200 회 까지로 한정하면서 오차 최소화를 검토한다.

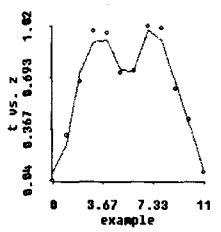
초기 가중값: 초기값의 범위는 *0<c<10* 이다. 각각의 c 값에 대하여 구간 [*-c, c*]로 부터 뽑은 확률난수를 이용하여 30 회 학습하고, 가장 작은 학습오차를 보이는 초기값을 네트워크의 구조인자로 선택한다. 제어상자를 이용한 실시간 초기 가중값의 선택은 같은 c 값 구간에서 초기값이 발생된 결과를 살펴보고 네트워크에 잘 적합되는 가중값들을 실시간으로 다시 초기 가중값으로 선택한 후 학습한다. 그러므로 선택된 네트워크는 확률난수를 이용한 초기화에 의한 가장 근접한 최종 예측모형으로 진행될 수 있다.

학습의 중단시점: 초기 가중값의 영향과 은닉노드의 개수를 조사할 때 0 으로 먼저 설정되며, 학습중단 규칙은 서로다른 충분히 작은 값을 설정한 최종 학습오차 e 에 의해서 제어된다.

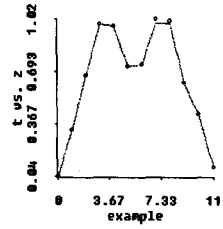
예측효율: 진실함수와 추정된 모형으로부터 구한 MSE 를 이용하여 측정한다. 그리고 시각적인 비교를 위하여 네트워크에 의해 추정한 실제 출력값을 그래픽으로 표현한다. 추정량들은 같은 크기의 모의추출된 학습표본에서 구한 네트워크에 의해 구해진다. 각 실험은 같은 크기의 모의추출된 학습표본을 이용하여 200 회 반복한다. 그리고 예측된 MSE 의 경험 분포는



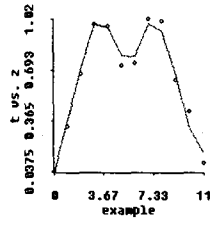
(a) c=10 인 경우



(b) c=0.1 인 경우

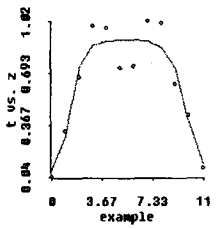


(c) c=10 에서 제어상자를 이용한 학습결과

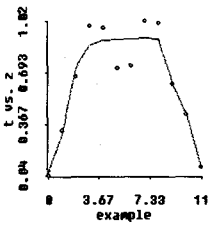


(d) c=0.1 에서 제어상자를 이용한 학습결과

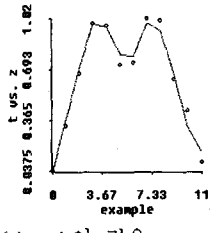
(그림 1) 초기 가중값의 영향: c 와 과다적합 관계.



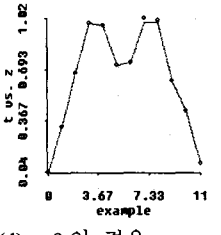
(a) n=2 인 경우



(b) n=3 인 경우

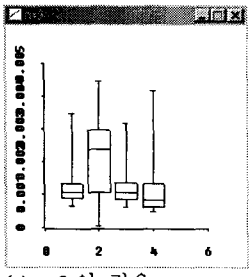


(c) n=4 인 경우

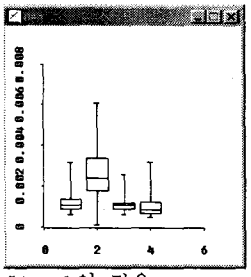


(d) n=9 인 경우

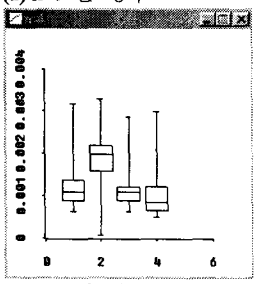
(그림 3) 은닉노드의 영향: n 과 과다적합 관계.



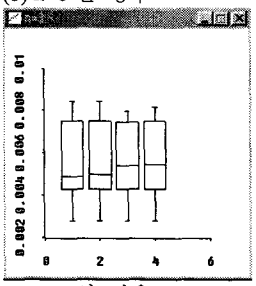
(a) n=9 인 경우



(b) n=6 인 경우

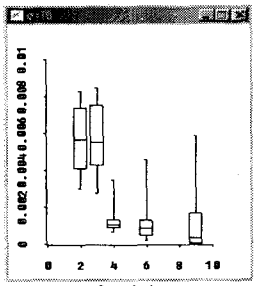


(c) n=4 인 경우

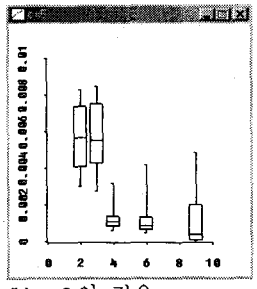


(d) n=3 인 경우

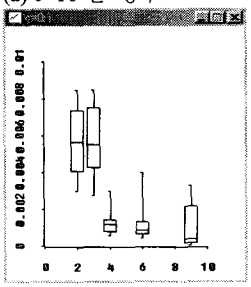
(그림 2) 은닉노드 개수에 따른 초기 가중값의 영향.



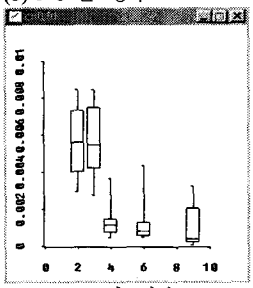
(a) c=10 인 경우



(b) c=5 인 경우



(c) c=0.1 인 경우



(d) c=0.01 인 경우

(그림 4) 초기 가중값에 따른 은닉노드 개수의 영향.

상자그림을 이용하여 살펴본다.

1) 초기 가중값의 효과[6,7]:

초기 가중값의 효과를 파악하기 위하여 먼저 $n=9$, $e=0$ 그리고 초기값의 범위를 $c = \{10, 0.1\}$로 설정한다. 최대 학습 회수는 200 회이다. 학습표본의 개수가 12 개이고 은닉노드의 개수가 9 개인 경우에 (그림 1)의 (a)와 (c)는 과다적합인 경우로서 가중값의 초기화가 MLP 네트워크의 적합능력에 영향을 주고 있음을 보여준다. 은닉노드의 개수를 9 로 고정하였을 때, 초기 가중값이 크면 복잡한 모형으로 인한 과다적합을 일으킨다. 그러나 초기 가중값이 작으면 더 좋은 MLP 모형을 유도하는 경향이 있다. 은닉노드의 개수가 각각 4, 6 인 경우에는 (그림 2)의 (b)와 (c)와 같이 주어진 c 의 결과간의 차이는 매우 작다. 그러나 (d)는 모형의 모수가 과소 적용되어 어떠한 초기 가중값에도 영향을 받지 않음을 알 수 있다

(그림 1)의 (c)와 (d)는 제어상자를 통한 초기값의 설정을 통한 학습 결과이다. 여기서 우리는 경험적으

로 유의한 차이는 아니지만 과다 적합을 피하고 있음을 확인할 수 있다.

학습 회수가 아주 충분한 경우에는 예측의 정도가 $c <= 5$ 인 주어진 c 에 민감하게 반응하지 않는 것을 알 수 있다. 이러한 결과로 좋은 추정 효율은 초기화 범위값 c 를 충분히 작게 함으로써 제어할 수 있다.

2) 은닉노드 개수의 효과:

은닉노드 개수의 효과를 논의하기 위하여 (그림 3)과 같이 $c=10, e=0$ 그리고 은닉노드의 개수변화는 $n = \{2, 3, 4, 6, 9\}$을 설정한다. 여기서 은닉노드의 개수가 $n=2,3$ 으로 작은 경우는 과소적합을 보이고 충분한 학습이 되지 않았음을 알 수 있다. 즉 앞에서 설명한 바와 같이 주어진 학습 표본자료가 $n \leq 3$ 인 은닉노드로는 충분히 적합시킬 수 없기 때문이다. 그러나 $n=9$ 와 같이 너무 많은 경우에는 과다적합을 보인다. (그림 4)의 상자그림은 은닉노드가 4 개인 경우에서 상대적으로 최적의 결과를 보인다. 여기서 은닉노드의 개수는 학습 결과에 민감한 반응을 보임을 알 수 있다.

3) 허용오차의 효과:

먼저 $n=9$, $c=10$ 그리고 학습중지 시점을 위한 허용오차의 변화는 $e = \{0, 1e\text{-}6, 0.001, 0.01\}$이라 하자. n 과 c 는 빠른 수렴결과를 얻기 위하여 큰 값으로 설정한다. 이 경우에 $e=0$ 은 다소 과다적합을 보이고 $e=0.01$ 는 과소적합을 보인다. 이는 허용오차가 모형 복잡도를 제어하는 구조인자로써 영향을 주고 있음을 보여준다. 또한 $e=0$ 에 대한 예측 효용은 e 의 다른 값에 대한 것들보다 유의적으로 더 나쁘지 않음을 보여준다. 그러므로 이 인자는 주어진 학습자료에 대한 초기 가중값과 은닉노드의 개수와 같은 모형 복잡도 제어에 유의하지는 않음을 알 수 있다.

4) 구조인자들의 상대효과:

(그림 2)는 $n=\{3,4,6,9\}$의 서로 다른 값에 대한 초기 가중값의 상자그림이다. 여기서 각 창의 세번째와 네번째 상자그림은 주어진 c 에 대하여 제어상자를 통한 초기 가중값으로 학습한 경우이다. 은닉노드의 개수가 $n=4,6$ 인 경우에, 각 c 에 대하여 거의 유사한 예측효용을 갖는 모형을 추정할 수 있다. $n=3$ 인 경우에는 과소 적합으로 인한 충분한 학습이 되지않고 있음을 알 수 있다. 다시말하면 예측효용은 최적의 은닉노드 개수를 선택하였을 경우에만 c 값의 영향을 줄일 수 있다. 즉 모형의 복잡도 제어는 c 값 보다는 은닉노드의 개수에 따라 더 영향을 받을 수 있음을 의미한다.

(그림 4)는 $c=\{10, 1, 0.1, 0.01\}$에 대한 은닉노드 개수의 영향을 보여주는 상자그림이다. 여기서 c 가 작은 경우에 서로다른 n 은 유사한 예측정도를 나타냄을 알 수 있다. 그러므로 충분히 작은 c 값을 선택하면 n 의 영향을 줄일 수 있다. 즉 복잡도 제어는 n 값보다는 초기 가중값이 더 중요한 인자임을 보여준다.

(그림 2)와 (그림 4)는 초기 가중값과 은닉노드 개수의 결합 영향을 나타내고 있다. 주어진 네트워크에서 한 개의 구조인자를 잘 선택하면 다른 구조인자에 관계없이 좋은 모형을 도출할 수 있음을 보았다. 이는 복잡도 제어가 어떤 이상적인 고정값으로 한가지를 정한 후 다른 두가지 구조인자중의 하나에 대하여 실행될 수 있다는 것을 의미한다.

예를들어 본 논문에서 고려한 학습표본의 분포특성을 고려하여 n 은 3 보다 큰 특정값인 4 이상의 값으로 설정될 수 있고 c 의 최적값은 교차타당성 검토에 의하여 선택될 수 있다. 다른 경우로서 c 를 10 으로 설정하고 최적의 n 값을 선택할 수 있다. 이는 MLP 네트워크의 모형 복잡도를 제어하기 위한 c 를 이용할 때 과다 적합의 n 값을 이용할 수 있게 된다. 즉 $n \leq 3$ 인 아주 작은 값으로 인하여 c 에 관계없이 과소 적합된 MLP 모형을 초래하는 경우가 있다. 이러한 값들은 실제 응용문제에서 일반적으로 알려지지 않고 있지만 충분히 큰 n 값에 의한 과다적합은 최적화된 c 값에 의하여 피할 수 있다.

4. 결 론

본 논문에서는 상자그림을 이용하여 MLP 구조인자들의 영향을 일반화시킬 수 있는 네트워크들의 적합능력을 살펴보았다. 상자그림과 같은 통계 그래픽스는 구조인자들의 영향에 대하여 직관적인 평가를 제시할 수 있었다. 허용오차는 다른 두 가지 인자인 초기 가중값과 은닉노드의 개수만큼 영향을 주지 않음을 확인하였다. 즉 허용오차의 적절한 설정은 MLP 네트워크의 일반화를 제어하는 더 좋은 척도가 됨을 보았다. 제어상자를 통한 초기 가중값의 실시간 적용 결과는 과다 적합을 피할 수 있는 대안이 될 수도 있음을 확인하였지만 그 차이가 유의하지는 않았다. 본 연구를 통하여 MLP 구조에 영향을 주는 요인들의 상호 작용이 네트워크의 복잡도 제어에 상대적으로 적은 영향만이 있음을 보았고, 이러한 논의는 실제적인 모형 제어전략의 한 방법으로 제시될 수 있겠다.

참고문헌

[1] V. Cherkassky and F. Mulier, "Learning from data - Concepts, Theory and Methods", Wiley, New York, 1998

[2] V. Vapnik, "The Nature of Statistical Learning Theory", Wiley, New York, 1995

[3] S. Zhong and V. Cherkassky, "Factors Controlling Generalization Ability od MLP Networks," IJCNN'99.

[4] Y. Yoon, "An improved Learning Process of Simple Neural Networks using the Controller box," Journal of KISS: Software and Applications, Vol.28, No.4, April, pp338-345, 2001.

[5] M.T. Hagan, H.B. Demuth and M. Beale, "Neural Network Design", PWS, Boston, 1995

[6] A. Atiya and C. Ji, "How Initial Conditions Affect Generalization Performance in Large Networks," IEEE Trans on Neural Networks, Vol.8, No.2, March, pp.448-451, 1997

[7] V. Cherkassky and R. Shepherd, "Regularization Effect of Weight Initialization in Back Propagation Networks," IJCNN'98, pp.2258-2261,1998