제22회 한국정보처리학회 추계학술발표대회 논문집 제11권 제2호 (2004. 11)

신경망 학습의 일반화 성능향상을 위한 초기 가중값과 학습률 그리고 계수조정의 효과

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

The Effect of Initial Weight, Learning Rate and Regularized Coefficient on Generalization Performance

YeoChang Yoon
Dept. of e-Information Engineering, Woosuk University

요 약

본 연구에서는 신경망 학습의 중요한 평가 척도로써 고려될 수 있는 일반화 성능과 학습속도를 개선시키기 위한 방안으로써 초기 가중값과 학습률과 같은 주요 인자들을 이용한 신경망 학습 영향을 살펴본다. 특히 초기 가중값과 학습률을 고정시킨 후 새롭게 조정된 계수들을 점차적으로 변화시키는 새로운 인자 결합방법을 이용하여 신경망 학습량과 학습속도를 비교해 보고 계수조정을 통한 개선된 학습 영향을 살펴본다. 그리고 단순한 예제를 이용한 실증분석을 통하여 신경망 모형의 일반화 성능과 학습 속도 개선을 위한 각 인자들의 개별 효과와 결합 효과를 살펴보고 그 개선 방안을 제시한다.

1. 서 론

신경망 연구에서 모형의 일반화 문제와 그 효용성 판단 평가는 가장 중요한 과제중의 하나다. 신호처리, 패턴인식, 제어 그리고 통신 등과 같은 응용분야에서 비선형 문제의 해결을 위해 적용되는 관련 시스템들은, 주어진 문제의 학습을 통하여 그 문제에 내포되어 있는 비선형 함수로 적절하게 적합되어야 한다. 추정된 모형이 충분한 정확도를 가지면서 비선형 함수로 적합되기 위한 학습을 위해서는 좋은 일반화 성능을 보이는 비선형의 적응적 학습시스템(nonlinear adaptive learning system)이 필요하다. 또한 최근의 무선통신이나 네트워킹과 같은 많은 신경망 응용분야들은 급격히 변화되는 확률과정적 환경에 적응하기 위하여 실시간으로 실행될 수 있는 학습시스템이 요구되고 있다. 그러므로 좋은 일반화 성능을 보일 수 있고 실시간에 적용될 수 있는 적응적 학습시스템을 어떻게 개발하느냐가 중요한 문제다.

이 연구에서는 비선형의 적응적 학습시스템을 위한 신경망 모형의 일반화 성능과 실시간 학습에 영향을 주는 몇 가지 중요한 인자(factor)들을 살펴본다. 특히 Zhang, et al[1]의 개선된 알고리즘을 이용한 학습을 통하여 신경망의 일반화 성능과 학습속도에 크게 영향을 주는 있는 초기 가중값의 범위, 학습률 그리고 계수조정 등을 통한 학습의 효과를 살펴본다. 이들 인자들의 영향 비교를 토대로 세가지 인자들을 동시에 고려하는 Wu 와 Zhang[2]의 학습방법을 적용하면서 인자 결합방법을 새롭게 적용한 신경망 학습량과 학습오차를 살펴본다. 학습률과 계수조정의 효과는 실증분석을 통하여 살펴보고, 제시된 방법들을 결합한 학습에서의 신경망 학습량과 학습오차들의 비교 결과를 제시한다. 마지막으로 결론과 향후 연구방향을 제시하고자 한다.

2. 초기 가중값의 효과

이 장에서는 일반화 성능과 효율에 영향을 주는 초기 가중값이 신경망 학습과정에 어떻게 영향을 주는가를 살펴본다. 신경망 학습 알고리즘에 일반적으로 많이 이용되고 있는 격자감소(gradient descent) 알고리즘은 학습 초기화를 위해 발생시킨 초기 가중값에 대한 해 평면의 초기 조건에 가장 밀접하게 위치한 목

적함수의 지역 최소값에 수렴시키는 경향이 있다. 과다 적합된 네트워크인 경우에도 목적함수의 최소값은 오차 평면을 따라가며 학습해 나간다. 격자감소 학습 알고리즘은 일반적으로 초기 조건에 근접한 해 평면상의 한 점으로 가중값이 찾아가게 한다. 만약 선택된 초기 가중값이 아주 작으면 최종적으로 얻게 되는 학습된 가중값도 작아지는 경향이 있다. 따라서 모형의 일반화 성능과 효율성의 측면에서 무엇이 초기 가중값의 가장 좋은 분포인가를 찾는 것이 중요하다. 그러므로 주어진 문제의 복잡도에 최적으로 일치하며 가장 좋은 일반화 문제를 구할 수 있고 가장 빠른 학습속도를 나타내는 초기 가중값의 범위를 구하는 문제가 이 연구에서 중요한 과제다.

가중값 초기화의 가장 보편적인 방법은 구간 $[-c, c]$에서 작은 확률난수로 가중값을 설정하는 것이다. 그러나 이와 같은 초기값은 최종 해의 정확도를 구하는데 있어서 다소 정형적인 효과만을 갖기 때문에 아주 좋은 작은 초기값이라는 개념은 다소 모호한 정의일 수 있다.

Cherkassky 와 Shepherd[3]는 역전파 모형에서 가중값 초기화에 대한 정형적인 영향을 연구하였다. 이들은 초기화의 정형화된 효과를 보여주는 예제를 통한 결과에서 가장 좋은 예측모형은 c 의 범위가 0.0001 ~ 0.001 에서 얻을 수 있다고 하였다.

3. 학습률의 효과

오차 역전파 알고리즘을 이용한 신경망 학습에서 적용되는 학습률은 가중값들이 학습자료에 대하여 관측된 오차에 따라 얼마나 크게 변화될 수 있는지를 결정한다. 이와 같은 학습 알고리즘을 적용하는 대부분의 연구자들은 최적 학습률의 선택문제에 직면하고 있지만, 어떤 값을 사용해야 하는지에 대한 일반적인 제시가 없다. 이는 가장 좋은 학습률을 사용한다고 하여도 이 학습률이 특정 학습자료에만 크게 영향을 주기 때문이다. 일부 알고리즘은 학습률을 자동으로 학습과정 중에 보정하기도 하지만[4,5], 이 방법은 전형적으로 수렴속도를 개선시키는데 초점이 있고 일반화 성능에는 알맞지 않다. 신경망 학습률의 선택은 학습속도뿐만 아니라 일반화 성능에 크게 영향을 줄 수 있다.

현재 많은 신경망 분석자들은 학습속도를 높이고 수렴이 잘 되도록 하기 위하여 가능하면 가장 큰 학습률을 사용하고 있다. 그러나 학습자료가 복잡한 문제일 경우에 학습률을 너무 크게 하면 추정된 모형의 일반화 성능을 떨어뜨리고 또한 학습속도도 저하시키게 된다. 다시말하여 학습률이 적절하지 않으면 일반화 성능과 학습속도의 개선이 없이 컴퓨터 자원만을 낭비하게 된다.

격자감소 학습 알고리즘을 이용한 학습에서 오차에 대한 격자감소는 가중값 공간의 현재 점에서 계산하기 시작한다. 오차를 줄여주기 위해서는 격자감소 방향이 반대 방향으로 가중값들을 변화시킨다. 그러나 감소 방향을 어느 방향으로 움직이게 할 수는 있어도 가중값을 어느정도 멀리 안전하게 변화시켜야 하는지에 관한 연구는 활발하지 못하다. 이러한 학습률 설정의 모호성으로 인하여 한편으로는 정확한 방향에서 너무 많이 변화하도록 하므로써 오차 평면 위를 과도하게 넘나들게 하고, 따라서 학습의 정확성을 떨어뜨리게 한다. 이러한 영향 때문에 학습률을 너무 크게 설정하면 더 많은 학습시간이 요구되며, 오차 평면 위를 발산하는 형태로 학습하게 됨으로써 수렴이 용이하지 않고 학습시간만 늘어나게 된다. 학습과정의 불안정성은 가중값들이 발산하지 않고 최소값으로 충분히 수렴되는 결과를 보여주지 않음으로써 일반화 성능을 떨어뜨리게 한다.

Wilson 과 Martinez[6]는 학습률이 학습시간과 일반화 정도에 어떻게 영향을 미치는지를 실험하였고, 일반화 정도를 최대화 시켜주는 학습률을 어떻게 효과적으로 선택해야 하는지를 연구하였다. 그들의 결론은 학습률이 작으면 일반화 정도에 좋은 결과를 줄 수 있다고 하였다

4. 계수조정의 효과

계수조정의 일반적인 방법은 신경망에 영향을 주는 복잡도를 수정된 학습 알고리즘이나 전략을 이용하여 제어하여 결과적으로 모형의 일반화 정도를 개선시키는 것이다[7,8].

계수조정의 방법은 다음과 같다. 여기서 기존 알고리즘과의 가장 큰 차이점은 개선된 오차함수에 있다. 오차함수는 다음과 같이 정의한다.

$$\hat{E} = E + \gamma \sum \quad , \tag{1}$$

$$E = \frac{1}{M}\sum_{k=1}^{M}(t_k - y_k)^2 . \tag{2}$$

식 1 은 식 2 와 같은 일반적인 오차 항 E 와 계수조정 항 $\gamma\Sigma$ 의 합으로 표현된다. 여기서 Σ 는 계수조정 함수다. γ 는 계수조정값이다. M 은 출력층 노드의 개수다. t_k 는 k 번째 노드의 목표값이고, y_k 는 k 번째 노드의 실제 출력값이다.

계수조정의 간단한 방법중의 하나는 가중값을 감소시키는 것이다. 여기서 가중값을 감소시키는 간단하고 일반적인 형식 중의 하나는 신경망의 모든 적응적 모수들에 대한 가중값들의 제곱합 형태로 나타낼 수 있다. 이를 이용한 계수조정함수는 식 3 과 같다. 여기서 w_i 는 각 학습주기마다 나타나는 가중값들이다.

$$\sum = \sum_i w_i^2 . \tag{3}$$

일부 연구결과에서 이 방법은 일반화 능력을 크게 개선시킬 수 있다고 알려졌지만 다음과 같은 두 가지의 단점이 있다. 그 하나는 가장 최적의 모수를 구하기가 어렵고 다른 하나는 각 학습주기마다 새로운 계수조정 식을 계산해야 하기 때문에 학습속도가 매우 느리다는 것이다. 더욱이 이 방법은 계수조정 결과에 매우 민감하게 가중값이 반응한다고 알려졌다. 만약 조정된 계수가 매우 작으면 이 방법은 효과가 없으며, 계수가 매우 크면 모든 가중값은 제로로 가게 되므로

계수를 신중하게 선택해야 한다. 그러나 계수를 어떻게 정확하게 선택하고 선택된 계수를 어떻게 조정해야 하는지는 여전히 어려운 문제로 남아있다.

5. 일반화 성능과 학습속도의 개선을 위한 신경망 학습의 결합방법

신경망의 일반화 성능과 효용에 대하여 초기 가중값의 범위, 학습률 그리고 계수조정의 효과를 분석하면서 그 방법들을 토대로 결합분석을 제시한다. 이 연구에서 우리는 식 (4)와 같이 Wu 와 Zhang[2]의 결합방법을 위한 또 다른 계수조정함수를 사용한다.

$$\sum = \sum_i \frac{w_i^2}{1+w_i^2}. \qquad (4)$$

일반화 성능과 학습속도 개선을 위한 결합방법의 알고리즘을 의사코드화 하면 다음과 같다.

단계 1: 가장 작은 학습오차와 빠른 학습속도를 얻기 위한 가중값 c 를 찾기 위하여 서로 다른 초기 가중값의 범위를 이용한 몇 번의 주기로 신경망을 학습한다.

단계 2: 가장 작은 오차와 빠른 학습속도를 얻기 위한 충분히 큰 학습률을 찾기 위하여 서로 다른 학습률을 이용한 몇 번의 주기로 신경망을 학습한다.

단계 3: 가장 작은 오차와 빠른 학습속도를 얻기 위한 적절한 값의 범위를 찾기 위하여 서로 다른 계수조정을 이용한 몇 번의 주기로 신경망을 학습한다.

단계 4: 학습오차에 따라 최적의 초기가중값 범위 c 와 학습률의 조정 그리고 계수조정을 이용하여 신경망 학습을 계속한다.

단계 5: 만약 허용오차를 만족하면 학습을 정지한다

6. 실증분석

적은 개수의 학습자료를 이용하여 많은 개수의 모수들로 이루어진 신경망을 의도적으로 학습하기 위하여 다음과 같은 함수를 추정한다고 하자. 또한 신경망 학습과정에는 격자감소법을 이용한다.

$$y = e^{-(x-1)^2} + e^{-(x+1)^2}, \quad x \in [-2.6, 2.6]. \qquad (5)$$

분석을 위한 모형으로는 한 개의 입력노드 x 와 출력노드 y 그리고 16 개의 은닉노드들로 이루어진 네트워크를 이용한다. 은닉층과 출력층에 대해서 시그모이드형 변환함수를 이용하고 학습오차는 0.005 로 설정한다. 실험을 단순화 하기 위하여 10 개의 모의 학습자료와 검정자료를 구간 [-2.6, 2.6]의 균등분포에서 각각 발생시키며, 발생된 표본은 평균이 0 이고 분산이 0.005 인 정규분포를 따르는 오차를 포함한다고 하자. 이때 실험은 세가지 경우로 나뉘어 실행한다. 첫째, 학습률을 0.5 로 고정한다. 초기 가중값은 구간 $[-c, c]$ 의 균등분포로부터 랜덤하게 생성시킨다. 여기서 c 의 범위는 c=0.001~5 이다. 둘째, 초기 가중값을 구간 [-1, 1]의 균등분포로부터 랜덤하게 생성시킨다. 학습률은 구간 [0.01, 20] 사이에서 변경시킨다. 셋째, 일반화 성능과 학습속도에 대하여 초기 가중값과 학습률을 고정시킨 후 조정된 계수들의 효과를 조사한다. 조정된 계수들은 구간 [0.000001, 0.00005]에 위치한다.

서로 다른 초기 가중값, 학습률 그리고 계수조정에 대한 학습결과는 표 1,2,3 과 같다. Zhang, et al[1]의 개선된 알고리즘을 Wu 와 Zhang[2]의 학습방법에 적용했을 때, 표와 같이 이 연구에서 결합방법을 이용한 학습량과 오차가 많이 개선되고 있음을 알 수 있다. 표 1 은 네트워크를 초기화시키기 위한 가중값의 주어진 범위에 따른 학습량과 학습 오차를 보여주고 있다. 여기서 가장 좋은 학습량과 오차 결과는 c=0.01 인 경우이며 이는 Atiya 와 Ji[9]의 결과와 유사하다. 표 2 는 다양한 학습률을 이용한 학습결과이다. 여기서 가장 좋은 학습량과 오차 결과는 학습률이 10 으로 설정된 경우이다. 표 3 은 초기 가중값과 학습률을 고정시킨 후 조정된 계수들의 효과를 조사한 결과이다. 두가지 인자의 범위가 제시된 상태에서 계수조정에 따른 학습결과는 전 영역에서 유사하게 수렴되고 있음을 알 수 있다. 여기서 학습오차를 중심으로 살펴보면 가장 좋은 결과가 계수조정 값이 0.000015 인 경우이다. 학습량은 계수조정 값을 미세하게 줄여나감에 따라 개선되고 있다.

표 1. 초기 가중값의 범위에 따른 학습결과

초기가중값의 범위	학습량	오차
(-0.00.,0.001)	658	0.01247579
(-0.01,0.01)	212	0.00920548
(-0.1,0.1)	936	0.01011324
(-1,1)	284	0.01073516
(-2,2)	348	0.01116621
(-5,5)	발산	

표 2. 학습률에 따른 학습결과

학습률	학습량	오차
0.01	13958	0.01070594
0.05	2783	0.01070959
0.5	293	0.01073516
1.0	186	0.01043836
1.5	135	0.01019269
2.0	178	0.00841369
10.0	80	0.00650959
20.0	발산	

표 3. 계수조정에 따른 학습결과

계수조정	학습량	오차
0.000050	530	0.01208128
0.000025	401	0.01025114
0.000015	318	0.01018539
0.000010	317	0.01041553
0.000005	309	0.01050411
0.000001	295	0.01067854

7. 결 론

이 연구에서는 비선형의 적응적 학습시스템을 위한

일반화 성능과 실시간 학습에 영향을 주는 중요한 인자들인 초기 가중값의 범위, 학습률 그리고 계수조정 등을 통한 신경망 학습의 효과를 살펴보았다. 그리고 이들 인자들의 영향 비교를 토대로 세가지 인자들을 동시에 고려하는 결합방법을 이용하였다. 실증분석을 통하여 결합방법을 이용한 학습효과를 살펴보고, 제시된 방법들을 이용한 학습에서의 신경망 학습량과 학습오차들의 비교 결과를 제시하였다. 그 결과로써 학습률, 초기 가중값의 범위 그리고 계수조정 등이 일반화 성능과 학습속도에 중요한 영향을 준다는 것을 살펴 보았다. 이들 인자들의 값이 너무 크거나 작은 경우는 좋은 학습 결과를 나타내지 않고 있다. 즉 너무 큰 가중값과 학습률 또는 너무 큰 계수조정값을 이용한 신경망 학습 알고리즘은 수렴되지 않을 수 있다. 일반적으로 말하여, 비교적 큰 학습률은 더 좋게 일반화할 수 있고 학습속도를 더 높일 수 있다. 너무 작은 학습률은 수렴속도를 저하시키는 경향을 보일 수 있다. 세가지 인자들의 결합방법을 이용하면 신경망 학습의 일반화 성능과 효율이 제시된 영역에서 가장 좋았다. 향후 연구로는 좀더 이론적인 접근이 필요하며 더 좋은 일반화 성능과 효율을 얻기 위한 학습 알고리즘을 제안할 필요가 있다고 본다.

참고문헌

[1] Y. Zhang, D. Liu and T.S. Chang, " A New Learning Algorithm for Feedforwad Neural Networks," Proceedings of the IEEE International Symposium on Intelligent Control, pp.39-44, 2001.

[2] Y. Wu and L. Zhang, "The Effect of Initial Weight, Learning Rate and Regularization on Generalization Performance and Efficiency," ICSP Proceedings, pp1191-1194, 2002.

[3] V. Cherkassky, R. Shepherd, "Regularization effect of weight initialization in back propagation networks," 1998 World Congress on Computational Intelligence, pp.2258-2261, 1998

[4] R.A. Jacobs, "Increased Rates of convergence Through Learning Rate Adaptation," Neural Networks, Vol.1, No.4, pp.295-307, 1988.

[5] T. Tollenaere, "SuperSAB: Fast Adaptive back-propagation with Good Scaling Properties," Neural Networks, Vol.3, No.5, pp.561-573, 1990.

[6] D.R. Wilson and T.R. Martinez, "The Need for Small Learning Rates on Large Problem," International Joint Conference on Neural Network, pp.115-119, 2001.

[7] A.S Weigend, D.E. Rumelhart and B.A. Huberman, "Generalization by Weight-Elimination with Application to Forecasting," in Advances in Neural Information Processing Systems, San Mateo, CA: Morgan Kaufmann, pp.875-882, 1991.

[8] S. Bo, "Optimal Weight Decay in Perceptron," Proceedings of the International Conference on Neural Networks, pp.551-556, 1996.

[9] A. Atiya and C.Y. Ji, "How Initial Conditions affect Generalization Performance in Large Networks," IEEE Transaction on Neural Networks, Vol.8, No.2, pp.448-451, 1997.