제26회 한국정보처리학회 추계학술발표대회 논문집 제13권 제2호 (2006. 11)

2개층 전방향 인공신경망에서의 이원적인 기울기 하강 알고리즘

최범기, 이주홍, 박태수
인하대학교 컴퓨터정보공학과
e-mail:{neural, juhong}@inha.ac.kr, taesu@datamining.inha.ac.kr

Dual Gradient Descent Algorithm On Two-Layered Feed-Forward Artificial Neural Networks

BumGhi Choi, Ju-Hong Lee, Tae-Su Park
Dept. of Computer Science & Information Engineering, Inha University

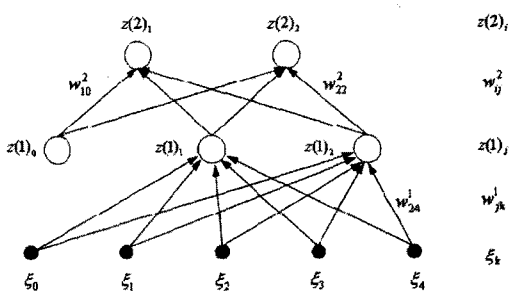
요 약

멀티레벨의 feed-forward 네트워크에 대한 학습 방법은 기울기 방법과 전역 최적화방법으로 나눌 수 있다. 역전파 또는 그 변형적인 방법들과 같은 기울기 하강 방법은 편리하기 때문에 여러 분야에서 다양하게 사용되고 있다. 하지만, 역전파와 관련된 가장 큰 문제는 지역 최소점에 빠진다는 것이다. 따라서 본 논문에서 기울기 하강 방법의 단순성을 침범하지 않고 지역 최소점을 극복할 수 있는 개선된 기울기 하강 방법을 제안한다. 제안하는 방법은 상위 연결과 하위연결을 분리하여 훈련하고 평가하기 때문에 이원적인 기울기 하강 방법이라 칭한다. 그렇기 때문에, 은닉층 유닛의 목표 값들은 하위 연결의 평가 툴로써 사용한다. 논문에서 제안하는 방법의 성능은 다양한 실험을 통해서 검증된다.

1. 서론

역전파에 대해 논하기 전에, 역전파와 같은 기울기 하강 방법이 더 좋다는 것을 밝히기 위해 전역 최적화에 대한 기술 조사가 선행되어야 할 것이다. 전역 최적화는 결정적인 방법과 확률적인 방법으로 분류할 수 있다. 결정적인 방법은 단지 정확한 함수의 클래스들에 대해서만 전역 최적점으로 수렴을 보장한다[1,2,3,4,5,6]. 그리고 추가적인 계산과정이 필요하다. 확률적인 방법은 탐색하는 수가 거의 무한대로 주어질 경우에만, 단지 확률적으로 전역 최적점으로 수렴을 보장한다[9,10]. Simulated annealing과 진화 알고리즘이 대표적인 방법들이다.

역전파는 많은 응용분야에서 감독 신경망 학습에서 주로 사용되는 모든 기울기 하강 방법 중에서 대표적으로 잘 알려진 방법이다. 기울기 하강 방법은 목표 값과 훈련하여 얻은 출력 값의 차이를 나타내는 오차 함수의 값을 최소화하기 위한 전방향 연결 가중치를 찾는데 사용된다. 본 논문에서는 (그림 1)과 같은 2개층 네트워크를 기반으로 설명한다.



(그림 1) 2개층 네트워크와 기호

여기서 $z(2)_i$는 i번째 출력층 유닛의 출력 값이고, $z(1)_j$는 j번째 은닉층의 출력 값이다. w_{ij}^2는 j번째 은닉 유닛에서 i번째 출력 유닛 사이의 가중치이고,

w_{ik}^1은 k번째 입력 유닛에서 j번째 은닉 유닛사이의 가중치이다. ξ_k는 k번째 입력의 출력 값이다.

본 논문에서 사용하는 오차 측정도구 혹은 비용함수는 다음과 같다.

$$E[w] = \frac{1}{2}\sum_{\mu i}\left[\zeta_i^\mu - g\left(\sum_j w_{ij}^2 g\left(\sum_k w_{jk}^1 \xi_k^\mu\right)\right)\right]^2$$

여기서 g는 시그모이드 또는 하이퍼블릭 탄젠트 함수이고, ζ_i^μ는 입력 패턴 μ에 대한 i번째 출력 유닛의 목표 값을 나타낸다.

상위 연결에 대한 기울기 하강 규칙은 다음과 같다.

$$\Delta w_{ij}^2 = -\eta\frac{\partial E}{\partial w_{ij}^2} = \eta\sum_\mu\left[\zeta_i^\mu - g\left(h(2)_i^\mu\right)\right]g'\left(h(2)_i^\mu\right)z(1)_i^\mu$$
$$= \eta\sum_\mu \delta(2)_i^\mu z(1)_i^\mu$$
$$\delta(2)_i^\mu = \sum_i \left\lfloor \zeta_i^\mu - z(2)_i^\mu \right\rfloor g'(h(2)_i^\mu)$$

여기서 η는 0과 1사이의 값을 가지는 학습율을 나타내고, $h(2)_i^\mu$는 입력 패턴μ에 대한 i번째 출력 유닛의 입력 값을 나타낸다.

하위 연결에 대한 오차 함수는 w_{jk}에 대한 미분 값이다.

$$\Delta w_{jk}^1 = -\eta\frac{\partial E}{\partial w_{jk}^1} = \eta\sum_\mu \frac{\partial E}{\partial z(1)_j^\mu}\frac{\partial z(1)_j^\mu}{\partial w_{jk}^1}$$
$$= -\eta\sum_\mu \frac{\frac{1}{2}\sum_i\left[\zeta_i^\mu - g\left(\sum_j w_{ij}^2 z(1)_j^\mu\right)\right]^2}{\partial z(1)_j^\mu}\frac{g\left(\sum_k w_{jk}^1\xi_k^\mu\right)}{\partial w_{jk}^1}$$
$$= \eta\sum_{\mu i i}\left[\zeta_i^\mu - z(2)_i^\mu\right]g'\left(h(1)_j^\mu\right)w_{ij}^2 g'\left(h(1)_j^\mu\right)\xi_k^\mu$$
$$= \eta\sum_\mu\sum_i \delta(2)_i^\mu w_{ij}^2 g'\left(h(1)_j^\mu\right)\xi_k$$
$$= \eta\sum_\mu \delta(1)_j^\mu \xi_k^\mu$$
$$\delta(1)_j^\mu = \sum_i \delta(2)_j^\mu w_{ij}^2 g'\left(h(1)_j^\mu\right)$$

여기서 $h(1)_j^\mu$는 입력 패턴μ에 대한 j번째 은닉 유닛의 입력 값이다.

역전파 방법은 구현이 쉽고 계산이 단순하다는 장점이 있다. 역전파의 가장 큰 문제점은 전역 최소점으로의 수렴을 보장하지 못하고 지역 최소점이나 포화점에 빠진다는 것이다. 지역 최소점은 최적이 수평면이 아닌 차선의 수평면으로 여기서 시스템 오차는 0이 아니며, 은닉 공간 행렬은 단수이다.

지역 최소점의 문제를 해결하기 위해서는 은닉 공간 행렬은 임의적인 혼동이나 진화 알고리즘을 사용하여 비특이(non-singular) 행렬이 되도록 바꾸어야한다. 하지만, 탐색 방향의 임의적인 혼동과 현재 가중치 집합을 조정하는 다양한 확률적인 방법들은 사용가능한 네트워크에서 지역최소점으로부터 탈출하는데 효과적이지 않고, 주어진 반복횟수 동안 전역 최소점으로의 수렴을 실패하게 만든다.

본 논문에서는 기울기 하강 방법의 단순성을 유지하면서 부분적으로 지역 최소점을 해결할 수 있는 새로운 방법을 제안한다. 제안하는 방법은 상위 연결과 하위 연결을 분리하여 그들의 오차 함수를 평가하기 때문에 이원적인 기울기 하강 방법이라고 한다. 또한, 은닉 유닛의 목표 값들은 하위 연결의 오차 함수를 생성하기 위해 새롭게 소개할 것이다. 본 논문에서 제안하는 방법의 정당성은 다음 장에서 설명할 것이다.

2. Dual Gradient 학습방법

2.1 네트워크 분리를 통한 지역최소점 문제 해소

제안하는 방법에서 네트워크는 2부분으로 나눠진다. 첫 번째 부분에서, 하위 연결을 고정하고, 상위 연결은 평가되고 갱신된다. 그런 후에, 하위 연결에서는 상위 연결을 고정하고, 상위 연결과는 다르게 평가하고, 갱신된다. 이 과정을 허용 오차내의 기대 값보다 작거나 같을 때 까지 반복한다.

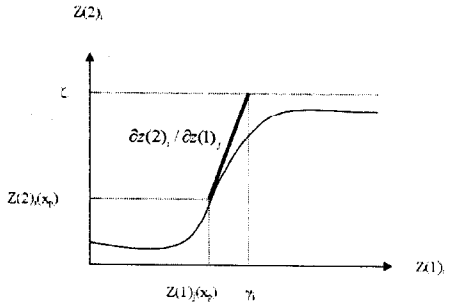
2개층 전방향 네트워크에서 지역최소점문제는 비특이 은닉 공간 행렬에 의해 발생한다. 이런 상황을 피하기 위해서, 은닉 유닛이 은닉 공간 행렬이 비특이 행렬을 갖도록 유도해야하고, 그래서 출력층 유닛들이 목표 값에 도달할 수 있도록 해야한다. 하지만, 모든 가중치들이 동일한 오차 함수로 갱신되면 은닉 유닛은 의미 있는 값을 가지지 못하게 된다. 하위 연결의 분리는 결정적으로 값을 가리키도록 하기 위해 은닉 유닛의 지도에 도움을 준다. 따라서 지역 최소점으로부터 자유로운 상태를 기대할 수 있다.

2.2 은닉 유닛의 목표 값으로 근사화

이원적인 기울기 학습방법의 성공은 은닉 유닛의 목표 값을 얼마나 정확하게 유도할 수 있는지에 달려있다. 그래서 하위 연결을 목표로 향하도록 한다. 여기서, 은닉 유닛의 목표 값은 가능한 한 기대 값에 가까운 출력을 선택하도록 만드는 은닉 유닛의 값을 의미한다.

은닉층과 출력층의 매핑에서 입력은 은닉 유닛의 출력 값이고, 출력은 출력층의 목표 값이다. 여기서 우

리가 원하는 것은 출력의 목표 값에 대한 상위 연결의 주어진 가중치 벡터에 사상된 행렬과 같은 은닉 공간의 솔루션 행렬을 찾는데 있다. 즉, 출력층에서 은닉층으로의 역으로 사상함으로써 문제를 해결할 수 있다. 해결책의 존재를 확실하게 보장하지는 못하지만, 은닉에서 출력으로의 사상을 통해 생성된 출력층의 목표 값과 같은 은닉 공간의 목표 값들은 뉴튼의 방법을 이용하여 근사시킬 수 있다. 이 근사는 (그림 2)를 통해 설명된다. 은닉 유닛의 목표 값 γ_j은 근사적으로 유도될 수 있다.



(그림 2) 은닉노드의 목표 값에 대한 근사

근사는 미분을 통한 출력오차로부터 은닉 유닛의 오차를 계산함으로써 할 수 있다. $z(2)_i$는 다변수 함수이지만, 은닉 유닛 $z(1)_j$의 목표 값을 유도하기 위해서, 그 함수는 다른 은닉 유닛들을 상수로 가정하는 단일 변수 함수로써 고려되어야한다. 이 방법에서, $z(1)_j$의 목표 값γ_j는 모든 출력 유닛들 i에 의해서 합해지고, 평균을 구할 수 있다. 이 과정을 수학 표기를 통해 설명하도록 하겠다. 단일 출력 유닛에 대하여, 은닉 유닛 j의 오차는 다음과 같다.

$$\gamma_j - z(1)_j = \frac{(\zeta_i - z(2)_i)}{\partial z(2)_i / \partial z(1)_j}$$

은닉 유닛 j의 목표 값은 모든 출력 유닛의 전체 수 N에 대하여 합하고, 평균 낼 수 있다.

$$\gamma_j = \frac{1}{N}\sum_i \frac{(\zeta_i - z(2)_i)}{g'(h(2)_i)w_{ij}} + z(1)_j$$

하위 연결의 은닉 유닛 j에 대한 평가 함수는 위의 식에서 목표 값을 통하여 유도될 수 있다.

$$E_j[w] = \frac{1}{2}\sum_\mu \left[\gamma_j^\mu - g\left(\sum_k w_{jk}^1 \xi_k^\mu\right)\right]^2$$

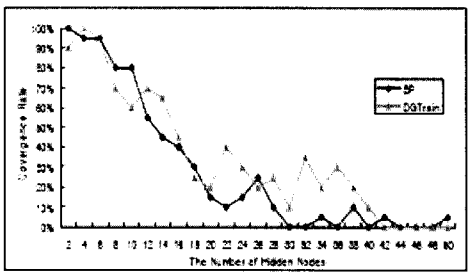
그래서, 하위 연결에 대한 갱신 규칙은 다음과 같이 변환된다.

$$\begin{aligned}\Delta w_{jk}^1 &= -\eta\frac{\partial E}{\partial w_{jk}^1} = \eta\sum_\mu \frac{\partial E}{\partial z(1)_j^\mu}\frac{\partial z(1)_j^\mu}{\partial w_{jk}^1} \\ &= -\eta\frac{1}{2}\sum_\mu \frac{\left[\gamma_j^\mu - z(1)_j^\mu\right]^2}{\partial z(1)_j^\mu}\frac{g\left(\sum_k w_{jk}^1 \xi_k^\mu\right)}{\partial w_{jk}^1} \\ &= \eta\sum_{\mu ii}\left[\gamma_j^\mu - z(1)_j^\mu\right]^2 g'\left(h(1)_j^\mu\right)\xi_k^\mu \\ &= \eta\sum_\mu \delta(1)_i^\mu \xi_k\end{aligned}$$

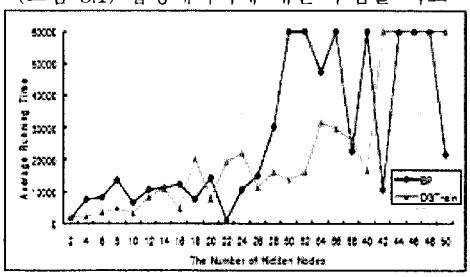
만약 정확한 은닉 유닛의 목표 값을 찾는다면, 은닉 유닛의 비특이 행렬을 만들 수 있다. 왜냐하면, 은닉 유닛의 목표 값으로부터 출력 유닛의 기대 값 전파를 보장하기 때문이다. 하지만 은닉 유닛의 정확한 기대 값을 찾을 수 없다. 단지 근사만이 가능하다. 은닉 유닛의 목표 값의 결정적인 계산 때문에 여전히 어려움으로 여겨지고 있다. 따라서 임의로 혼동을 하는 것보단 단일 은닉 공간 행렬을 비특이 행렬로 변환하는 것이 기대된다.

3. 실험 결과

합성데이터와 Iris 데이터를 이용하여 DGTrain과 역전파에 대하여 은닉 유닛이 증가함에 따른 성공한 수렴율과, 전체 반복 횟수, 평균 수행시간, 총 합계 평균 오차 등을 비교하였다. 또한 약 60초의 제한시간을 두었고, 수렴오차는 0.01로 설정하였다.



(그림 3.1) 합성데이터에 대한 수렴율 비교



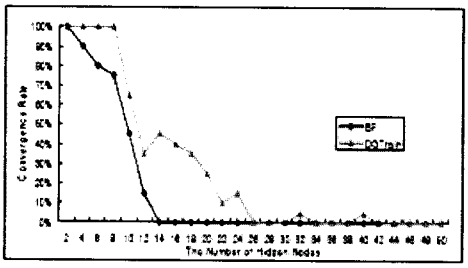
(그림 3.2) 합성데이터에 대한 수행시간 비교

합성데이터를 이용한 실험은 은닉 유닛을 2에서 50개까지 증가시키면서 역전파와 DGTrain방법을 비교

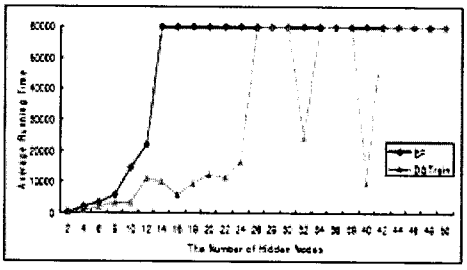
하였다. 수렴율의 경우, 은닉 유닛의 수가 적을 때 역전파가 제안하는 방법보다 좋은 성능을 보인다. 하지만, 은닉 유닛이 증가할수록 제안하는 방법의 수렴율이 더 높은 것을 알 수 있다. 또한, 수행시간의 경우 본 논문에서 제안하는 방법이 은닉 유닛의 증가에 관계없이 월등한 성능을 보였다.

Iris Data는 꽃받침의 길이(sepal length), 꽃받침의 두께(sepal width), 꽃잎의 길이(petal length), 꽃잎의 두께(petal width)의 4개 변수로 구성되며, 클래스는 붓꽃의 3가지 종류(Setosa, Versilcolor, Vignica)로 데이터의 수는 50개씩 150개이다.

(그림 4.1,2)는 iris 데이터에 대한 실험결과를 나타낸다. 수렴율의 경우 제안하는 방법이 역전파보다 높은 수렴율을 나타내고, 특히 역전파는 은닉유닛의 수가 14개 이상일 경우에는 수렴하지 않은 것을 볼 수 있다. 그러나 본 논문에서 제안하는 방법은 역전파보다 좀더 많이 수렴한 것을 볼 수 있다. 또한, 수행시간의 경우 은닉 유닛의 증가에 상관없이 더 빠른 것을 볼 수 있다.



(그림 4.1) Iris데이터에 대한 수렴율 비교



(그림 4.2) Iris데이터에 대한 수행시간 비교

4. 결 론

본 논문에서 제안하는 방법은 결정적인 방법과 휴리스틱 방법을 병합하였다. 이것은 전역최소점으로의 수렴을 보장하지 못하지만, 비교적으로 빠르고, 부분적으로 지역최소점 문제를 해결할 수 있다. 또한 본 논문에서 제안하는 방법은 매우 단순하여 이해하기 쉽고 다른 알고리즘들과 병합하여 사용할 수 있다. 네트워크의 분리는 입력 패턴수가 많은 매우 복잡한 네트워크를 단순화하는 새로운 방법이다. 지역최소점과 같은 문제를 해결하기 위해서는 좀더 많은 은닉 유닛이 필요하다. 따라서 이런 복잡한 문제를 해결하기 위해서는 본 논문에서 제안하는 방법이 매우 효과적이다. 위의 두 실험에서도 보였듯이, 본 논문에서 제안하는 방법은 은닉 유닛의 증가에 매우 유연한다.

참고문헌

[1] Cetin, 1. Burdick, and J. Barhen "Global descent replaces gradient descent to avoid local minima problem in learning with ANN," Proc. of lEEE Conf. on NN, vol. 2, 1993, pp 836-842.

[2] Handbook of Global Optimization, R. Hont and P. Padalos, Eds. Dordrecht: Kluwer, 1995.

[3] P. Plagianakos. G. D. Magoulas, M. Vrahatis, "Deterministic non-monotone strategies for the effecive training of multilayer perceptrons," IEEE Trans. on N. Networks, vol. 13, no. 6, pp. 1268-1284, 2002.

[4] D. Jones, C. Perttunen, and B. Stuckman, "Lipschitzian Optimization without the Lipschitz Constant," J. of Optimization Theory and Applications, vol. 79, 1993,pp. 157-181.

[5] M. Vrahatis, G. Androulakis, J. Lambrinos, and G. Magoulas, "A class of gradient unconstrained minimization stepsize," *J. of Computational and Applied Mathematics* 114, 2000, pp. 367-386.

[6] Ivan N. Jordanov, and Tahseen A. Rafik, "Local Minima Free Network Learning" *Second lEEE International Conference On Intelligent Systems,* June 2004, pp. 34-39

[7] G. Bilbro, "Fast stochastic global optimisation," *IEEE Trans. On Systems, Man, and Cybernetics*, vol. 24, pp. 684-689, 1994.

[8] W. Huyer and A. Neumaier, "Global Optimization by Multilevel Coordinate Search," J. *of Gl. Optimizalion*, vol. 14, 1999, pp. 331-335.

[9] A. Tom and S. Vitanen, "Topographical Global Optimization Using Pre-Sampled Points," *J. of Global Optimization* 5, 1994, pp.267-276.

[10] S. C. Ng, S. H. Leung and A. Luk, "A Hybrid Algorithm of Weight Evolution and Generalized Back-propagation for finding Global Minimum", *Proceedings of IEEE International Joint Conference on Neural Networks (IJCNN'99),* 1999.