적응적 가중치 감소항을 적용한 Optimal Brain Surgeon

이현진+, 지태창++, 박혜영+++, 이일병+
+ 연세대학교 컴퓨터과학과 ++ LG-EDS 시스템 기술 연구 부문 +++ Brain Science Institute RIKEN

Optimal Brain Surgeon with Adaptive Weight Decay Term

Hyunjin Lee+, Taechang Jee++, Hyeyoung Park+++, Yillbyung Lee+
+ Dept. of Computer Science, Yonsei University ++Dept. of R&D Group, LG-EDS System
+++ Brain Science Institute, RIKEN, JAPAN

요 약

본 논문에서는 다층 퍼셉트론 신경망에서 연결선 수를 최소로 하면서 일반화 성능을 향상시키기 위해 가장 널리 쓰여지고 있는 Optimal Brain Surgeon 을 이용한 프루닝(pruning)을 기반으로 하여 오차 함수의 가중치 감소항을 추가 시키는 방법을 사용한다. 이때 학습 및 프루닝의 성능에 많은 영향을 미치는 가중치 감소항의 반영정도를 베이시안 테크닉에 기반하여 적응적으로 최적화 하는 방법을 제안한다. 제안하는 방법의 성능을 검증하기 위해 벤치마크 데이터를 이용하여 실험을 수행하였다. 순수한 OBS 방법과 고정된 반영정도를 가진 가중치 감소항을 추가시킨 OBS, 그리고 제안하는 적응적 가중치 감소항을 적용한 OBS 방법을 비교하여 제한하는 방법이 기존의 두 방법에 비해 신경망 구조의 최적화 능력이 뛰어남을 확인 할 수 있었다.

1. 서론

프루닝 알고리즘은 신경망에서 불필요한 가중치를 제거 시키면서 학습하기 위한 방법이다. 프루닝에 의한 신경망의 최적화는 많은 실용적인 응용분야에 유용하게 사용될 수 있다. 프루닝으로 신경망 구조를 단순화하면 일반화 능력을 향상 시킬 수 있으며, 계산량을 감소시킬 수 있을 뿐만 아니라 신경망에 설명능력을 부여할 수 있도록 하는 장점이 있다[6][7].

전방향(feed-forward) 신경망의 프루닝에 가장 널리 쓰이는 방법에는 OBD(Optimal Brain Damage) 와 OBS(Optimal Brain Surgeon) 가 있다. 이 두 방법은 어떤 연결선을 제거할 때 발생하는 오차의 변화를 바탕으로하여 프루닝 하는 방법이다. OBD 와 OBS 의 차이점은 다음과 같다. OBD 는 중요도가 최소인 가중치를 제거만 시키는데 반하여, OBS 는 그 가중치를 제거할 때 생기는 오차의 증가를 최소로 하기 위해 남아 있는 가중치들을 재조정한다[1][4][6].

Hansen은 정규화 항(regularization term) 방법의 일종인 가중치 감소항(weight decay term) 을 갖는 OBS 를 제안하여 일반화 성능을 높였다[3]. 하지만 이 방법은 가중치 감소항의 반영을 어느 정도로 하는가에 따라 성능차이가 발생한다. 가중치 감소항의 영향력이 작으면 과다학습이 될 수 있고, 너무 크면 학습이 되지 않는 문제가 발생한다[8]. 이러한 문제점을 극복하기 위하여 본 논문에서는 베이시안 최적화로 가중치 감소항의 반영정도를 적응적으로 조정 하므로서 프루닝 성능을 향상 시키는 방법을 제안한다.

본 논문의 구성은 다음과 같다. 2 장에서는 Hansen 이 제안한 가중치 감소항을 가진 OBS 에 대해 설명한다. 3 장에서는 가중치 감소항의 반영정도의 적응적 최적화에 대해 설명을 한다. 4 장에서는 제안하는 시스템의 구성에 대해 설명 하며 5 장에서는 실험 및 결과를 다룬다. 마지막으로 6 장에서는 결론을 내린다.

2. 가중치 감소항을 가진 OBS

학습 비용함수(cost function)는 다음과 같이 정의 된다.

$$C(\mathbf{w}) = E(\mathbf{w}) + \frac{1}{2}\mathbf{w}^T\mathbf{D}\mathbf{w} \qquad (1)$$

여기서 E(w)는 학습데이터 집합에 대한 평균 제곱오차이고 D는 정방 행렬이다. D는 $\mathbf{D}_{ij} = \alpha\delta_{ij}$ 인 단위 벡터이다. OBS 에서 가중치의 중요도는 어떤 가중치가 제거될 때 남아 있는 가중치들에 대해서 수식(1)을 최소로 하도록 재조정 하였을 때 학습 오차의 변화로 정의된다.

중요도는 다음과 같은 단계로 계산된다[3].

1. 비용함수를 2차 극한으로 확장 시킨다.
2. j째 가중치 제거후 남아 있는 가중치의 변화를 찾고, 2차 근사로 남아 있는 가중치를 재조정 시킨다.
3. 가중치 조정후 학습오차의 변화를 계산한다.

그림 1 중요도의 계산 단계

제약이 없는 극소를 $\mathbf{w}_0$ 라 하자. $\mathbf{w} = \mathbf{w}_0 + \delta\mathbf{w}$ 에

서 이점을 중심으로 비용함수를 전개하면 다음과 같다.

$$C(\mathbf{w}) = C(\mathbf{w}_0) + \frac{1}{2}\delta\ \mathbf{w}^T(\mathbf{A}+\mathbf{D})\delta\ \mathbf{w} \qquad (2)$$

$\mathbf{w}_0$이 최적화 되었다고 가정하면 첫번째 항은 0 이 된다. A는 학습오차의 2 차 미분 행렬이다. 비용함수의 헤시안은 $\mathbf{H} = \mathbf{A} + \mathbf{D}$ 이다.

제약이 있는 극소를 유도하면 다음과 같다. j 번째 가중치의 제거는 $\mathbf{e}_j$가 j 번째 단위 벡터일 때 $\delta\ \mathbf{w}^R\mathbf{e}_j = -\mathbf{w}_0^T\mathbf{e}_j$로 나타낼 수 있다. 제약이 있는 극소는 라그랑지 승수 (Lagrange multiplier) 를 이용하여 찾는다.

$$\tilde{C}_j(\mathbf{w}) = C(\mathbf{w}) + \lambda\,(\delta\ \mathbf{w} + \mathbf{w}_0)^T\mathbf{e}_j \qquad (3)$$

극소는$\delta\ \mathbf{w}_j = -\lambda_j\mathbf{H}^{-1}\mathbf{e}_j$이고 λ_j는 다음과 같이 표현된다.

$$\lambda_j = \frac{\mathbf{w}_0^T\mathbf{e}_j}{\mathbf{e}_j^T\mathbf{H}^{-1}\mathbf{e}_j} \qquad (4)$$

j번째 가중치의 중요도는 다음과 같이 계산된다.

$$\delta\ E_j(\mathbf{w}) = \lambda_j\mathbf{w}_0^T\mathbf{D}\mathbf{H}^{-1}\mathbf{e}_j + \frac{1}{2}\lambda_j^2\mathbf{e}_j^T\mathbf{H}^{-1}\mathbf{A}\mathbf{H}^{-1}\mathbf{e}_j \qquad (5)$$

3. 적응적 파라메터 최적화

목적함수 C(w)는 신경망 학습에 널리 쓰이는 오차의 제곱의 합 E(w)와 정규화 항인 가중치의 제곱의 합 $\mathbf{w}^T\mathbf{D}\mathbf{w}$에 각각의 파라메터를 곱한 합으로 계산된다.

$$C(\mathbf{w}) = \beta\ E(\mathbf{w}) + \alpha\ \mathbf{w}^T\mathbf{D}\mathbf{w} \qquad (6)$$

목적함수의 파라메터 α와 β의 상대적인 크기는 학습이 어떠한 것에 중점을 두고 있는지를 나타낸다. 만약 $\alpha \ll \beta$이면 학습은 오차를 최소로 하는 방향으로 진행되는 것이다. 만약 $\alpha \gg \beta$라면 학습은 오차를 줄이는 것은 희생하고 가중치의 값을 감소시키는 것에 중점을 두고 있는 것이다. 이 결과 유연하게 반응하는 신경망이 생성된다 [2].

MacKay는 정규화를 최적화 시키는 베이시안 테크닉을 제안하였다[5]. 정규화 파라메터를 최적화하기 위해서는 헤시안 행렬의 계산이 필요하다. Foresees는 계산량을 줄이기 위하여 베이시안 정규화를 위한 Gauss-Newton 근사 알고리즘을 개발하였으며 그 알고리즘은 다음과 같다[2].

1. α=0 와 β=1 로 가중치 초기화한다.
2. 목적함수 (6) 을 최소화 시키기 위하여 Levenberg-Marquardt 알고리즘을 수행한다[1].
3. Levenberg-Marquardt 알고리즘에서 헤시안에 대해서는 Gauss-Newton 근사를 한다. 효과적인 파라메터의 수 γ는 식(7)로 구한다.

$$\gamma = N - 2\alpha\ \mathrm{tr}(\mathbf{H})^{-1} \qquad 7)$$

4. 새로운 파라메터 α,β를 식(8)로 추정한다.

$$\alpha = \frac{\tilde{a}}{2\ \mathbf{w}^T\mathbf{D}\mathbf{w}} \qquad \beta = \frac{n-\gamma}{2E(\mathbf{w})} \qquad (8)$$

5. 2-4 까지 수렴 할 때 까지 반복한다.

그림 2 Gauss-Newton 근사에 기반한 Bayesian 학습

4. 제안하는 시스템

제안하는 시스템의 구성은 그림 3과 같다. 신경망의 학습은 Gauss-Newton 근사에 기반한 베이시안 학습 알고리즘인 그림 2를 적용하여 정규화 파라메터를 최적화 하였다. 이를 통해 얻어진 최적화된 파라메터를 기반으로 하여 OBS 을 유도하여 프루닝을 수행하였다.

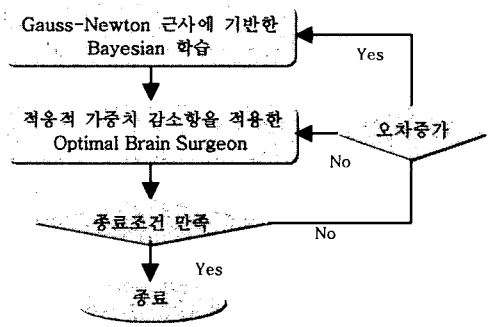

그림 3 제안하는 방법의 구성

5. 실험 및 결과

실험에 사용한 데이터는 MONK 문제1 데이터를 사용하였다. 432개의 데이터중 124개를 학습에 사용하였다. 신경망은 한층의 은닉층을 가진 다층 퍼셉트론을 대상으로 하였다. 최초의 신경망의 구성은 입력노드는 17개 은닉노드는 5개 출력노드는 1개이며 은닉노드와 출력노드에 바이어스가 있어서 총 96개의 연결선으로 구성되어 있다. 신경망의 가중치를 20번 다르게 초기화 하여 실험하였다.

5.1 가중치 감소항이 없는 OBS와 가중치 감소항이 있는 OBS와 제안하는 적응적 가중치 감소항을 적용한 OBS 성능 비교

연결선이 12 개 일 때의 학습데이터의 인식률과 테스트 데이터의 인식률을 살펴보았다. 가중치 감소항이 없는 OBS 의 경우 학습시간은 1 분58초였다. 가중치 감소항이 있는 OBS 의 경우 가중치 감소항의 반영정도를 0.0001로 하여 실험하였고 이 경우 학습시간은 2분 8초였다. 제안하는 적응적 가중치 감소항을 적용한 OBS의 경우 학습 시간은 2분 12초였다. <표 1>을 보면 가중치 감소항이 없는 OBS 보다 가중치 감소항이 있는 OBS 가 인식률이 우수하였다. 3 가지 방법 중에서 제안하는 적응적 가중치 감소항을 적용한 OBS 는 일반화 능력이 가장 우수하였다.

<표 1>학습과 테스트 데이터에 대한 인식률비교

	학습	테스트
가중치 감소항이 없는 OBS	85.69%	80.90%
가중치 감소항이 있는 OBS	96.45%	93.33%
적응적 가중치 감소항을 적용한 OBS	99.56%	99.17%

5.2 연결선 분석

그림 4,5,6은 가중치를 다르게 하여 20번 실험한 것 중 한 예이다. 위에서 비교한 3 가지 방법을 적용하여 연결선을 12 개까지 줄였을 때 남아 있는 연결선은 다음과 같다. 가중치 감소항이 없는 OBS 의 연결선은 그림 4, 가중치 감소항이 있는 OBS 방법의 연결선은 그림 5, 제안하는 적응적 가중치 감소항을 적용한 OBS 방법의 연결선은 그림 6 과 같다.

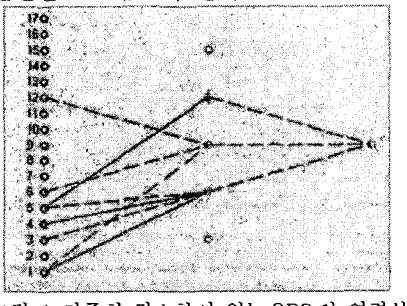

그림 4 가중치 감소항이 없는 OBS 의 연결선

그림 4 의 가중치 감소항이 없는 OBS 방법은 은닉층의 바이어스가 제거 되지 않았다. 또한 중요한 연결선인 1,2,3,4,5,6,12 번 중 2 번 연결선이 제거가 되었기 때문에 학습 데이터에 대한 인식률이 89.52%에 테스트 데이터에 대해 83.33%의 인식률에 그쳤다.

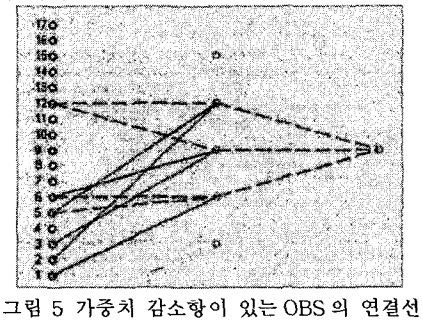

그림 5 가중치 감소항이 있는 OBS 의 연결선

그림 5 의 가중치 감소항이 있는 OBS 는 중요한 연결 중 하나인 4 번 연결선이 끊어졌다. 하지만 가중치 감소항에 의해 정규화가 되었기 때문에 학습과 테스트 데이터에 대해 100% 의 인식률을 유지할 수 있었다.

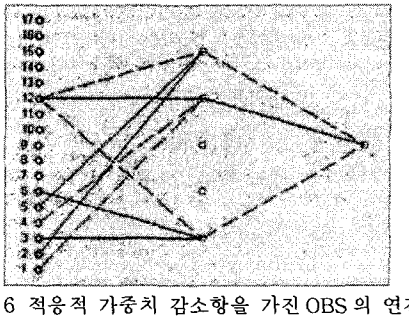

그림 6 적응적 가중치 감소항을 가진 OBS 의 연결선

그림 6 의 제안하는 적응적 가중치 감소항을 적용한 OBS를 살펴보면 중요한 연결선인 1,2,3,4,5,6,12 번을 유지하면서 학습과 테스트 데이터에 대해 100% 의 인식률을 보였다.

5.3 인식률이 100%일때 최소한의 연결선의 수

학습데이터와 테스트 데이터의 인식률이 100% 일때 연결선의 수를 분석하면 <표 2>와 같다. 가중치 감소항이 없는 OBS는 평균 14개의 연결선 까지 줄일 수 있었다. 가중치 감소항이 있는 OBS는 평균 13개까지 줄일 수 있었다. 제안하는 베이시안 최적화 OBS는 평균 11.95개 까지 연결 선을 줄여서 3 가지 방법 중 가장 적은 수의 연결선으로 우수한 인식률을 보였다.

<표 2> 인식률 100% 일때 최소한의 연결선의 수

	가중치 감소항이 없는 OBS	가중치 감소항이 있는 OBS	적응적 가중치 감소항을 적용한 OBS
평균	14개	13개	11.95개

6. 결론

본 연구에서는 다층 퍼셉트론 신경망의 연결선의 수는 최소로 하면서 일반화 성능을 향상시키기 위한 효과적인 프루닝 방법을 제안 하였다.

제안하는 적응적 가중치 감소항을 적용한 OBS는 오차 제곱합과 가중치 감소항의 반영정도를 적응적으로 조정하므로써 신경망이 과다 학습 되는 것을 방지 하기 때문에 성능 향상이 되었다.

앞으로의 과제는 제안하는 적응적 가중치 감소항을 적용한 OBS 프루닝에서 연결선을 몇 개 까지 줄일 것인가를 자동적으로 정하는 연구가 필요하겠다.

7. 참고 문헌

[1]C.M.Bishop, "Neural Networks for Pattern Recognition,"Oxford University Press, 1995.

[2]F.D.Foresee,M.T.Hagan,"Gauss-Newton approximation to Bayesian learning,"ICNN, Vol.3, pp.1930-1935, 1997.

[3]L.K.Hansen, M.W.Pedersen,"Controlled Growth of Cascade Correlation Nets,"ICANN, pp.797-800,1994.

[4]B. Hassibi, D.G.Stork, G.J.Wolff, "Optimal Brain Surgeon and General Network Pruning,"ICNN, pp.293-299, 1993.

[5]D.J.C.MacKay, "Bayesian interpolation,"Neural Computation, Vol.4, No.3, pp.415-447,1992.

[6]M.W.Pedersen, L.K.Hansen, J.Larsen, "Pruning with generalization based weight saliencies: γOBD, γOBS,"NIPS8, pp.521-527,1996.

[7]R. Reed, "Pruning algorithms-a survey," IEEE Transactions on Neural Networks, Vol. 4, pp. 740 -747, 1993.

[8]R. Setiono, "A penalty-function approach for pruning feedforward neural networks,"Neural Computation, Vol.9, No.1, pp. 185-204, 1997.