적응적 정규화 자연기울기 학습과 자연프루닝을 통한 신경망의 일반화 성능 향상

이현진[+], 박혜영[++], 지태창[+++], 이일병[+]
[+] 연세대학교 컴퓨터과학과 ++ 일본 이화학연구소 뇌과학 연구센터 +++ LGCNS 기술 내재화팀

Improving Generalization in Neural Networks using Natural Gradient Learning with Adaptive Regularization and Natural Pruning

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

요 약

본 논문에서는 적응적 정규화 자연기울기 학습법과 자연 프루닝(pruning) 방법의 결합을 통하여 일반화 성능이 우수한 신경망을 구성하고자 한다. 먼저 적응적 정규화 자연기울기 학습을 통하여 신경망의 가중치를 최적화 시키고, 자연 프루닝에 의하여 신경망의 구조를 단순화 시킨다. 이러한 모델들 중 최적의 모델은 베이시안 정보 기준에 의해 선택함으로써 일반화 성능이 우수한 신경망을 구성하는 방법을 제안한다. 벤치마크(benchmark) 데이터로 제안하는 방법과 유클리디안(Euclidean) 거리에 기반한 결합 방법과 자연 프루닝만을 적용한 방법을 비교함으로써 우수성을 검증한다.

1 서론

신경회로망 학습의 목표는 학습데이터를 통해서 가능한 작은 예측오차를 갖는 입출력 관계를 찾아내는 것이다[1]. 하지만 학습데이터는 이러한 입출력 관계를 구성하기엔 부족하다. 따라서, 학습데이터에 대해서 잘 학습하게 되면 미지의 데이터에 대한 예측능력이 저하되는 과다학습(overfitting)이 발생하며, 이는 학습데이터의 특징만 발견하고 원래의 데이터를 생성하는 함수를 발견하지 못하여 발생한다[1][2].

일반화 성능 향상을 위한 방법에는 다음과 같은 것들이 있다. 첫째, 크로스 벨리데이션(cross-validation) 방법은 신경망 학습시에 학습 데이터를 학습 집합과 평가집합으로 나누어서 일반화 오차를 추정하여 최적의 구조를 찾는 방법이다[3]. 하나의 학습집합과 평가집합만을 사용하는 방법의 단점을 극복하기 위하여 여러 리샘플링(*resampling*) 방법들이 연구되었다. 크로스 벨리데이션 방법은 학습데이터를 학습집합과 평가집합으로 나눌 때의 샘플링 방법에 따라 성능차이가 발생하고 샘플링 시간이 오래 걸린다는 단점이 있다. 둘째, 수학적인 모델선택 척도에 의한 방법은 벨리데이션 집합을 따로 나눌 필요 없이 학습된 모델의 일반화 성능을 평가 할 수 있으며, 이러한 방법들로는 선형모델에 대해서 일반화 성능을 평가하는 Akaike의 최종 예측오차(FPE)와 Akaike의 정보기준(AIC)등이 있다. Akaike의 정보기준을 더 일반화 시켜서 비선형 모델과 정규화항이 존재하는 경우를 다룰 수 있는 Moody의 일반화된 예측 오차 (GPE)방법과 Murata의 네트워크 정보 기준(NIC)을 이용하는 방법 등이 있다[2]. 이 방법은 비교대상이 될 후보모델을 생성하는 기준이 없다는 문제점이 있다. 셋째, 가중치 파라메터의 수를 조절하는 프루닝(pruning)방법과 그로잉(growing)방법이 있다. 신경망의 파라메터의 수를 줄이는 프루닝 방법은 중요한 파라메터들 만을 남기고 점점 단순한 구조를 생성하는 방법이고, 그로잉 방법은 단순한 구조에서 복잡한 구조를 생성해 가는 방법이다. 이 방법은 파라메터 수가 언제일 때까지 조절해야 한다는 종료조건이 없다는 단점이 있다. 넷째, 정규화는 신경망의 복잡도를 통제하기 위하여 오차함수에 패널티항(penalty term)을 추가 시켜 수행하는 방법이다. 이 방법은 간접적으로 신경망의 복잡도만 최적화 할 뿐 직접적으로 신경망의 파라메터 수는 최적화 시키지 못한다.

최근 연구에서는 각각의 일반화 성능 향상 방법들의 단점을 극복하기 위하여 여러 방법들을 결합시킨 연구가 시도되고 있다. Hansen은 정규화 항 방법의 일종인 가중치 감소항 방법과 프루닝 방법의 일종인 OBS(Optimal Brain Surgeon)를 결합하여 일반화 성능을 향상시키는 방법을 제안하였다[4]. Larsen 등은 크로스 벨리데이션 오차 또는 간단한 홀드 아웃(hold-out) 벨리데이션 오차의 최소화에 의한 적응적 정규화 항 방법을 제안하여 정규화의 영향력을 조절함으로써 일반화 성능을 높이는 방법을 제안하였다[5]. Hintz-Madsen 은 정규화항 방법과 프루닝 방법 그리고 일반화 오차 추정 방법을 결합시켜 신경 분류기를 설계하는 방법을 제안하였다[6]. 이현진

등은 베이시안 적응적 정규화 방법과 OBS 프루닝 방법을 적용하여 일반화 성능을 향상시키는 방법을 제안하였다[7].

기존의 결합방법들은 다들 유클리디안(Euclidean) 거리에 기반한 방법들이다. 본 연구에서는 신경회로망의 일반화 성능 향상을 위하여 리마니안(Riemannian) 거리에 기반한 적응적 정규화 자연기울기 학습과 자연 프루닝의 결합을 통하여 일반화 성능을 향상시키는 방법을 제안하고자 한다.

2 제안하는 방법의 구성

제안하는 방법의 구성은 그림 1과 같다. 먼저 신경망 구조가 들어오면 최적의 가중치를 갖도록 적응적 정규화가 있는 자연기울기 학습을 시킨다. 학습이 끝나면 자연 프루닝을 통하여 단순한 모델을 생성한다. 생성한 모델이 최적의 가중치 이면 다시 자연 프루닝으로 구조를 단순화 시키고 최적의 모델이 아니면 다시 적응적 정규화가 있는 자연기울기 학습으로 최적의 가중치를 갖도록 최적화 시킨다. 최적의 모델은 베이시안 정보 기준(BIC)를 통하여 가장 최적의 모델을 선택한다.

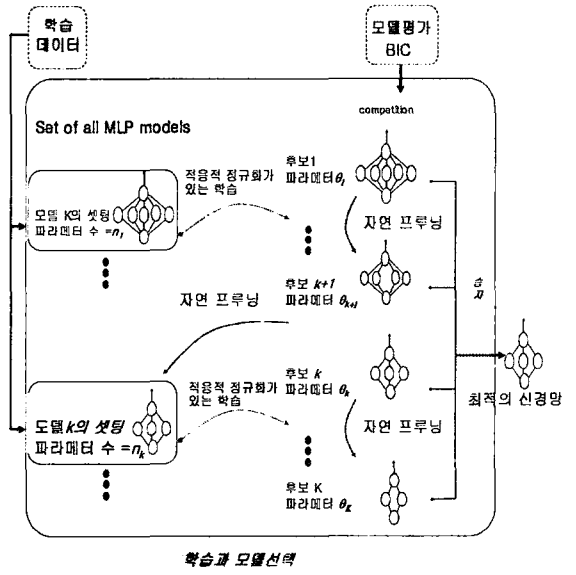

그림 1 제안하는 방법의 구성

3 적응적 정규화가 있는 자연기울기 학습

정규화 항이 있는 오차 함수는 식(1)과 같이 정의 된다.

$$C(\boldsymbol{x},y,\boldsymbol{\theta}) = E(\boldsymbol{x},y,\boldsymbol{\theta}) + \alpha R(\boldsymbol{\theta}) \quad (1)$$

여기서 $E(\boldsymbol{x},y,\boldsymbol{\theta})$는 제곱 오차합, 크로스엔트로피 오차와 같은 표준적인 오차함수이다. 정규화 항은 식(2)와 같이 가중치 파라메타 $\boldsymbol{\theta}$의 식으로 표현된다[1].

$$R(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|^2 \quad (2)$$

이런 정규화 항이 있는 경우 자연기울기 학습은 식 (3)과 같이 주어진다[8].

$$\begin{aligned}\boldsymbol{\theta}_{t+1} &= \boldsymbol{\theta}_t - \eta_t \tilde{\nabla} C(\boldsymbol{x},y,\boldsymbol{\theta}_t) = \boldsymbol{\theta}_t - \eta_t \boldsymbol{G}^{-1} \nabla C(\boldsymbol{x},y,\boldsymbol{\theta}_t) \\ &= \boldsymbol{\theta}_t - \eta_t \boldsymbol{G}^{-1}(\nabla E(\boldsymbol{x},y,\boldsymbol{\theta}_t) + \alpha \nabla R(\boldsymbol{\theta}_t))\end{aligned} \quad (3)$$

η은 학습률 이고, 리마니안 거리 텐서(tensor)는 식(4)와 같이 계산된다[7].

$$\begin{aligned}\boldsymbol{G}(\boldsymbol{\theta}) &= \iint \frac{\partial \log p}{\partial \boldsymbol{\theta}} \left(\frac{\partial \log p}{\partial \boldsymbol{\theta}}\right)^T p(y|\boldsymbol{x},\boldsymbol{\theta}) q(\boldsymbol{x}) dy d\boldsymbol{x} \\ &= E_{\boldsymbol{x}}\left[E_{y|\boldsymbol{x},\boldsymbol{\theta}}\left[\frac{\partial \log p(y|\boldsymbol{x},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \left(\frac{\partial \log p(y|\boldsymbol{x},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^T \right] \right]\end{aligned} \quad (4)$$

식 (3)의 정규화항 파라메터 α에 의해 정규화의 성능차이가 발생하며 이를 베이시안 방법에 의해 적응적으로 조절할 수 있으며 이는 식 (5)와 같다(자세한 유도는 [1] 참조).

$$\alpha = \frac{n}{2\,NR(\boldsymbol{\theta})} \quad (5)$$

여기서 n은 가중피의 수이고, N은 데이터의 수이다.

4 자연기울기 프루닝

신경회로망의 학습에 의해 최적화된 파라메터 $\boldsymbol{\theta}^* = (\theta_1^*, \ldots, \theta_n^*)$ 라고 할 때 자연 프루닝은 i번째 가중치 파라메터의 값 0이 되었을 때 현재 모델에 가장 적은 변화를 주는 식(6)과 같은 파라메터를 찾아서 제거하는 방법이다.

$$\hat{\boldsymbol{\theta}}(i) = \arg\min_{\theta_i^*=0} F(\boldsymbol{\theta}_i^*) \quad (6)$$

이때 자연 프루닝은 제거 될 파라메터를 결정하는데 있어서 식(7) 과 같은 피셔 메트릭을 사용한다.

$$F(\boldsymbol{\theta}_i^*) = (\boldsymbol{\theta}^* - \boldsymbol{\theta}^i)^T \boldsymbol{G}(\boldsymbol{\theta}^*)(\boldsymbol{\theta}^* - \boldsymbol{\theta}^i) \quad (7)$$

5 베이시안 정보 기준

신경망의 과다학습문제를 해결하기 위하여 두 가지 형태의 모델선택이 자주 쓰인다. 첫째는 크로스 벨리데이션 방법처럼 데이터를 나누어서 수행하는 아웃오브샘플(out-of-sample) 모델선택 방법이다. 다른 하나는 AIC, 베이시안 정보기준 과 같이 데이터를 분할하지 않는 인샘플(in-sample) 모델 선택이다. 최적의 모델선택을 위하여 자연 프루닝에 의해 생성된 모델간의 패널티가 가해진 성능 측정도구로 베이시안 정보기준(BIC)을 적용하였다.

Granger(1993)는 식(8)과 같은 변형된 베이시안 정보기준을 제안 하였으며 비선형 모델의 경우 d>1 이고 d는 실험에 의하여 결정한다[9].

$$BIC = \log\left(\frac{1}{N}\sum_{i=1}^{N} E(\boldsymbol{x}_i, y_i, \boldsymbol{\theta})\right) + \frac{n^d \log(N)}{N} \quad (8)$$

6 실험 및 결과

제안하는 방법의 우수성을 검증하기 위하여 다음과 같은 실험을 하였다. SSEARP는 제곱합 오차(Sum of Squared Error) 함수에 적응적 정규화(Adaptive Regularization)와 프루닝(Pruning)을 결합한 방법이다. NGNRP는 자연기울기 학습에(Natural Gradient) 적응적 정규화(Adaptive Regularization)없이 프루닝(Pruning)을 적용시킨 예이다. NGARP는 자연기울기 학습(Natural

Gradient)에 적응적 정규화(Adaptive Regularization)와 자연 프루닝(Pruning)을 결합한 방법이다. 제안하는 방법은 제곱합 오차에 비해 분류에 있어서 더 좋은 성능을 나타낸다고 알려진 크로스엔트로피 오차함수를 적용 할 수 있고[1], 따라서 오차함수로 이를 사용하였다.

6.1 MONK3 문제

총 432개의 데이터중 122개를 학습에 사용하였다. 학습데이터에 5%의 노이즈가 포함되어 있다. 신경망은 한층의 은닉층을 가진 다층 퍼셉트론을 대상으로 하였다. 최초의 신경망의 구성은 입력노드 17개, 은닉노드 5개, 출력노드 1개 이며 은닉노드와 출력노드에 바이어스가 있어서 총 96개의 연결선으로 구성되어 있다. 신경망의 가중치를 10번 다르게 초기화 하여 실험 하였다. 실험결과는 <표 1>과 같다. 적응적 정규화와 프루닝의 결합방법인 SSERP와 NGARP의 경우는 100%테스트 데이터 분류율을 보였으며, 제안하는 방법은 구조 최적화 면에서 SSERP보다 좀 더 우수한 성능을 보였다. 적응적 정규화가 없이 자연 프루닝만을 수행한 NGNRP는 테스트 데이터에 대해 100%의 인식율을 보이지 못했다.

<표 1> Monk3 문제에 대한 가중치 수 및 분류율

	SSERP	NGNRP	NGARP
가중치 수 (개)	11.6	17.0	10.6
학습 분류율 (%)	97.6	95.9	96.5
테스트 분류율 (%)	100	97.2	100

6.2 Glass 문제

데이터 집합은 107개의 학습집합과 54개의 벨리 데이션 집합과 53개의 테스트 집합으로 총 214개로 구성되어 있다. 본 방법은 벨리데이션 집합이 필요 없기 때문에 원래의 학습데이터인 107개로만 학습하는 작은학습 집합(S)과 벨리데이션 집합까지 학습집합에 포함한 161개의 학습데이터로 실험하는 큰학습 집합 (L)로 실험 하였다. 네트워크의 구조는 입력노드는 9개, 은닉노드는 6개, 출력노드는 6개이고 총 초기 네트워크의 파라메터의 수는 바이어스를 포함하여 총 102개이다. 가중치 초기화를 10번 다르게 하여 실험하였다. <표 2>에서 보는 바와 같이 제안하는 방법이 다른 두 방법에 비해 구조최적화와 일반화 성능면에서 우수하였다. 본 방법은 벨리데이션을 사용하지 않기 때문에 벨리데이션 집합을 학습집합에 통합시키면 좀더 많은 데이터를 활용할 수 있으므로 좀 더 우수한 성능을 얻을 수 있었다.

7 결론

본 논문에서는 리마니안 거리에 기반한 적응적 정규화 자연기울기 학습방법과 자연 프루닝의 결합방법을 제안하고 베이시안 정보기준에 의해서 최적의 모델을 선택하는 방법을 제안하였다. 이 방법은 기존의 유클리디안 거리에 기반한 방법보다 구조 최적화와 일반화 성능 면에서 우수한 성능을 보였다. 제안하는 방법은 크로스 엔트로피 오차함수를 사용 할 수 있기 때문에 패턴분류문제에 더 우수한 성능을 보일 수 있으며, 다중 분류의 경우 소프트맥스 기법을 변형 없이 쓸 수 있다. 적응적 자연기울기 학습 방법과 자연프루닝은 정보기하이론에 기반한 방법이다. 본 연구는 정보기하이론에 기반한 일반화 성능향상 연구에 관한 기초연구로 현재 이러한 결합방법에 관한 이론적 해석 연구가 진행중이다.

<표 2> Glass 문제에 대한 가중치 수 및 분류율

		SSERP	NGNRP	NGARP
가중치 수 (개)	S	53.8	67.2	51.2
	L	49.8	62.5	47.8
학습 분류율 (%)	S	73.83	79.43	77.57
	L	72.05	78.88	77.02
테스트 분류율 (%)	S	70.10	62.26	71.70
	L	70.6	69.05	73.52

8 참고 문헌

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

[2] S. Haykin, Neural Networks; A Comprehensive Foundation, Prentice-Hall :Second Edition, Inc., 1999.

[3] T. Andersen , M. Rimer, T. Martinez, Optimal Artificial Neural Network Architecture Selection for Baggin, Proceedings. IJCNN '01. International Joint Conference on Neural Networks, vol. 2, 790 - 795, 2001

[4] L. K. Hansen, M. W. Pedersen, Controlled Growth of Cascade Correlation Nets, Proceedings of ICNN, 797-800, 1994.

[5] J. Larsen, C. Svarer, L. N. Andersen, L. K. Hansen, Adaptive Regularization in Neural Network Modeling, Neural Networks: Tricks of the Trade, L NCS 1524, Springer-Verlag, 113-132, 1998.

[6] M. Hintz-Madsen, L.K. Hansen, J. Larsen, M. W. Pedersen, M. Larsen, Neural classifier construction using regularization, pruning and test error estimation, Neural Networks, Vol. 11, 1659-1670, 1998.

[7]H. Lee, T. Jee, H. Park, Y. Lee, A Hybrid Approach to Complexity Optimization of Neutral Networks, 8th International Conference on Neural Information Processing, Vol. 3, 1455-1460, 2001.

[8]H. Park, Practical Consideration on Generalization Property of Natural Gradient Learning, LNCS, 2084, 402-409, 2001.

[9]C. Granger, Strategies for modeling nonlinear time series relationship, Neural Computation, 12, 881-901, 1993.