적응적 정규화, 프루닝 및 BIC를 이용한 신경망 최적화 방법

이현진[†] · 박혜영[††]

요 약

주어진 문제에 대하여 최적의 성능을 가지는 신경회로망을 얻기 위해서는 학습을 통한 매개변수의 최적화(parameter optimization)와 모델 선택을 통한 구조 최적화(structure optimization)의 통합적인 과정이 필요하다. 본 논문에서는, 각 세부 방법들의 특성을 고려하여, 공통의 특성을 갖는 방법들을 결합함으로써 효율적이면서도 일반화 성능을 높이는 총체적인 신경회로망 최적화 방법을 제안한다. 먼저 다양한 오차 함수를 사용할 수 있는 자연 기울기 강하 학습에 적응적 정규화 방법을 도입함으로써 가중치 매개변수(weight parameter)들을 최적화 한다. 그리고 이렇게 최적화된 매개변수(parameter)들에 자연 프루닝(natural pruning)을 적용하여 불필요한 요소들을 제거하여 최적화 된 구조를 생성한다. 반복적인 과정에 의하여 후보 모델들을 구성하고 베이시안 정보 기준(Bayesian Information Criterion: BIC)을 이용하여 최적의 모델을 평가하여 선택하는 방법을 제안하였다. 벤치마크 데이터에 대한 실험을 통하여 제안하는 방법의 구조 최적화 능력과 일반화 성능의 우수성을 보였다.

An Optimization Method of Neural Networks using Adaptive Regulraization, Pruning, and BIC

Hyunjin Lee[†] and Hyeyoung Park[††]

ABSTRACT

To achieve an optimal performance for a given problem, we need an integrative process of the parameter optimization via learning and the structure optimization via model selection. In this paper, we propose an efficient optimization method for improving generalization performance by considering the property of each sub-method and by combining them with common theoretical properties. First, weight parameters are optimized by natural gradient learning with adaptive regularization, which uses a diverse error function. Second, the network structure is optimized by eliminating unnecessary parameters with natural pruning. Through iterating these processes, candidate models are constructed and evaluated based on the Bayesian Information Criterion so that an optimal one is finally selected. Through computational experiments on benchmark problems, we confirm the weight parameter and structure optimization performance of the proposed method.

Key words: 일반화, 최적화, 적응적 정규화, 자연 프루닝, 베이시안 정보 기준

1. 서 론

신경회로망은 학습을 통하여 스스로 지식을 획득하는 특성을 갖는다. 이러한 신경회로망의 특성으로 인하여 패턴인식, 문자인식, 음성인식, 로보틱스, 의사결정 시스템등 다양한 응용 분야에서 유용하게 사용되고 있다[1]. 신경회로망은 입·출력쌍으로 구성되어 있는 학습 데이터가 주어졌을 때, 학습을 통하여 학습 되지 않은 데이터에 대해서도 적절한 출력을

접수일 : 2002년 7월 2일, 완료일 : 2002년 10월 11일
[†] 정회원, 한국싸이버대학교 컴퓨터정보통신학부
[††] 일본 이화학연구소 뇌과학연구센터 뇌수리연구팀

주는 일반화 성능이 있으며, 이러한 성능으로 인하여 다양한 응용분야에서 적용되고 있다. 데이터가 무한히 많다면 신경회로망으로 원하는 오차범위까지 모델링이 가능하다고 알려져 있다.

하지만 실제로 학습 데이터는 입·출력 함수를 추정하기에는 부족하며, 그 안에는 노이즈(noise)가 존재한다. 따라서 학습 데이터에 대해 잘 학습을 시키면, 학습데이터의 특징만 발견하고 원래의 데이터를 생성하는 함수를 발견하지 못하기 때문에, 미지의 데이터에 대한 예측 능력이 저하되는 과다 적합(over-fitting) 현상이 발생한다[1-3]. 이러한 현상은 학습 데이터가 적을수록 , 신경회로망의 복잡도가 증가 할수록 노이즈의 영향은 더욱 커지기 때문에 심해진다. 이러한 과다적합에 의한 일반화 성능 저하를 개선하기 위한 방법에는 검증(validation) 방법, 모델 선택 방법, 가중치 매개변수의 수 조절 방법, 정규화 방법 등이 있다.

검증 방법은 신경회로망 학습시에 학습 데이터를 학습 집합과 검증 집합으로 나누어서 일반화 오차를 추정하는 방법이다. 이 방법은 학습데이터의 표집 방법에 따라 성능 차이가 발생하고 표집(sampling)에 따른 시간이 문제가 되고 있다[4].

모델 선택 방법은 과다학습이 모델의 복잡도와 관련된다는 사실에 기반 한다. 모델의 복잡도가 증가할수록 학습되지 않은 데이터에 대해 성능이 떨어지는 과다적합이 발생하고, 반대로 복잡도가 감소할수록 학습데이터에 대해 적합되지 못하는 과소적합(underfitting)이 발생 한다. 모델 선택 척도는 검증 집합을 따로 나눌 필요 없이 학습된 모델의 일반화 성능을 평가 할 수 있으며, 이러한 방법들로는 Akaike의 최종 예측오차 (Final Prediction Error: FPE)와 Akaike의 정보 기준(Akaike's Information Criterion: AIC), Akaike의 정보기준을 더 일반화 시켜서 비선형 모델과 정규화항이 존재하는 경우를 다룰 수 있는 일반화된 예측 오차 (Generalized Prediction Error: GPE)방법, 네트워크 정보 기준 (Network Information Criterion: NIC), 베이즈(Bayes)의 추정에 의해 유도된 베이시안 정보 기준(Baysian Information Criterion: BIC)등이 있다[5].

가중치 매개변수의 수를 조절하는 방법에는 프루닝(pruning) 방법과 그로잉(growing) 방법이 있다. 신경회로망의 매개변수의 수를 줄이는 프루닝 방법은 중요한 매개변수들 만을 남기고 점점 단순한 구조를 생성하는 방법이다. 전방향(feed-forward) 신경회로망의 프루닝에 가장 널리 쓰이는 방법에는 OBD (Optimal Brain Damage)와 OBS(Optimal Brain Surgeon)가 있다[1,3,6]. 이 방법은 어떤 가중치 매개변수가 제거될 때 발생하는 오차의 변화를 바탕으로 하여 매개변수를 제거하는 방법이다. 프루닝 방법과 반대로 그로잉(growing) 방법은 매우 간단한 모델로부터 시작해서 하나의 노드를 추가시키고 추가된 노드들과의 연결을 학습시켜서 점점 복잡한 구조를 형성하는 방법이다. 이러한 방법들에는 캐스케이드 상관관계(cascade correlation), 익스텐트론(extentron), 노드 분할 방법 등이 있다[4]. 프루닝과 그로잉과 같은 가중치 매개변수의 수 조절 방법은 가중치 매개변수의 수를 어느 정도까지 조절 해야 한다는 종료 조건이 없는 단점이 있다.

정규화 방법은 오차함수에 벌칙항(penalty term)을 추가 시켜, 신경회로망의 복잡도를 통제하는 방법이다. 이 방법의 목적함수는 학습오차에 관련된 항과 매개변수의 복잡도를 제어하는 벌칙항으로 구성된다. 이렇게 함으로써 최소한의 복잡도를 가지고 학습오차를 최소화하는 신경회로망을 얻을 수 있다. 가장 많이 사용되는 벌칙항으로는 가중치 감소(weight decay)항이 있다[3]. 이 방법은 벌칙항의 영향력을 조정하는 정규화 매개변수에 따라 성능의 큰 영향을 받으며, 따라서 이를 조절하는 것이 적용하는데 있어서 문제가 되고 있다.

최근 연구에서는 각각의 일반화 성능 향상 방법들의 단점을 극복하기 위하여 여러 방법들을 결합시킨 연구가 시도되고 있다. Hansen은 정규화 방법의 일종인 가중치 감소항 방법과 프루닝 방법의 일종인 OBS를 결합하여 일반화 성능을 향상시키는 방법을 제안하였다[6]. Larsen등은 크로스 검증 오차 또는 간단한 홀드 아웃(hold-out) 검증 오차의 최소화에 의한 적응적 정규화 방법을 제안하여 정규화의 영향력을 조절함으로써 일반화 성능을 높이는 방법을 제안하였다[7]. Andersen과 Hintz-Madsen은 정규화 방법과 프루닝 방법 그리고 일반화 오차 추정 방법을 결합시켜 신경 분류기를 설계하는 방법을 제안하였다[4,8]. 이현진 등은 베이시안 적응적 정규화 방법과 OBS 프루닝 방법을 적용하여 일반화 성능을 향상시키는 방법을 제안하였다[9].

본 논문에서는 통계적인 관점에서 신경회로망을 고찰하고, 이를 바탕으로 신경회로망의 일반화 성능 향상을 위한 최적화 방법을 제안하고자 한다. 본 논문의 구성은 다음과 같다. 2장에서는 신경회로의 최적화에 대한 문제점과 이를 해결하기 위한 제안하는 방법의 개념을 살펴본다. 3장에서는 2장에서 살펴본 개념을 바탕으로 제안하는 방법의 구체적인 구성방법에 대해 살펴본다. 4장에서는 실험결과를 살펴보고 분석한다. 그리고 마지막으로 5장에서는 결론을 내린다.

2. 신경회로망 최적화

실제적인 문제에서 신경회로망을 적용하여 효과적인 결과를 얻기 위해서는 두가지의 최적화가 필요하다. 하나는 가중치 매개변수 값의 최적화이다. 가중치 매개변수들이 최적의 값을 가질때 최적의 결과를 기대할 수 있으며, 신경회로망에서는 학습에 의해서 이러한 최적의 매개변수 값을 결정한다. 다른 하나는 신경회로망의 구조 최적화이다. 신경회로망의 은닉노드의 수와 가중치 매개변수의 수는 신경회로망의 일반화 성능에 중요한 영향을 미치며 따라서 최적의 성능을 얻기 위해서는 구조 최적화를 고려해 주어야 한다.

신경회로망의 가중치 매개변수값의 최적화에 쓰이는 학습방법 중에서 널리 쓰이는 방법에는 오류역전파 학습법, 뉴턴(Newton), 컨쥬게이트(conjugate) 기울기 학습, LM (Levenberg-Marquardt) 학습법등이 있다. 이러한 방법은 학습이 계속 반복됨에도 불구하고 한동안 오차가 줄지 않는 플라토(plateau)라 불리는 기간이 존재한다. 이러한 기간이 극소가 아님에도 불구하고 극소로 오인되어 중간에 학습을 중단하게 되고, 따라서 최적의 시스템을 찾기 못하게 되는 원인을 제공한다. 이러한 플라토 문제의 원인을 밝히고, 기존의 학습방법에 대한 문제점을 해결한 정보 기하 이론에 기반한 자연기울기 학습법이 제안 되었다[10-13].

실제적인 응용문제에 신경회로망의 구조에 따라 그 성능의 차이가 크게 발생하지만, 구조를 결정하는 노드수의 최적값이 어떤것인지 결정하는 것은 쉬운 일이 아니다. 따라서 이를 위하여 신경회로망의 구조 최적화 방법이 사용된다. 이러한 구조 최적화를 위하여 충분한 수의 노드로부터 시작하여, 노드수를 줄여가며 최적의 구조를 결정하는 프루닝 방법이 널리 쓰인다. 대표적인 프루닝 방법으로는 유클리디안 거리(Euclidean distance)에 기반한 OBD와 OBS가 있다. 하지만 최근 학습과 마찬가지로 정보 기하 이론에 기반한 리만 거리(*Riemannian distance*)를 사용하는 자연 프루닝 방법이 제안되었고, 이러한 방법이 유클리디안 거리에 기반한 방법들에 비해 성능이 우수하다는 것이 연구 되었다[14,15].

프루닝에 의한 구조 최적화 방법은 프루닝 전에 신경회로망의 가중치를 최적의 상태로 만드는 학습과정이 필요하고, 프루닝 후 오차의 증가가 심한 경우 감소시키기 위한 추가적인 학습 방법이 필요하다. 이러한 방법의 문제는 프루닝 방법이 가장 큰 신경회로망의 학습된 해가 최적의 해라는 가정 하에서 그 해를 중심으로 해서 신경회로망의 구조를 최적화하기 때문에, 프루닝 전의 학습 상태에 따라서 그 성능이 크게 영향을 받게 된다. 또한 프루닝 후 발생하는 오차 증가를 방지하기 위하여 사용되는 학습은 보통의 학습 방법으로는 최적해에 도달하도록 할 수 없다. 왜냐하면 보통 학습 방법은 목표해를 알 수 없기 때문에, 학습데이터를 잘 추정하도록 학습을 하기 때문이다. 하지만 학습데이터에는 노이즈가 존재하고, 원하는 목표 시스템을 추정할 만큼 충분한 수가 아니기 때문에 학습데이터를 잘 적합하도록 신경회로망을 학습시키면 학습데이터에 대해 잘 적합할 뿐, 실제 목표 시스템을 추정할 수 없게 된다. 따라서 지역극소에 도달하기 쉽게 되는 문제점이 발생한다. 이러한 문제점을 해결하기 위하여 학습에 정규화를 도입한다.

학습에 정규화를 도입하면 학습과정에서 지역 극소에 빠지는 것을 방지하여 가중치 매개변수의 값을 최적화 할 수 있을 뿐만 아니라, 목표해에 가깝게 해를 추정함으로써 프루닝에 의한 구조 최적화 성능의 향상을 기할 수 있다. 학습에 정규화를 도입하는데 있어서 정규화의 성능은 벌칙항의 영향력을 조절하는 정규화 매개변수에 따라 크게 성능 차이를 발생하며, 따라서 이러한 정규화 매개변수의 값을 정해주는 방법이 매우 중요하다. 따라서 정규화 매개변수의 값을 정해주는 방법이 필요하며 이를 위하여 적응적인 방법을 적용하여 정규화의 성능을 향상시키는 것이 필요하다.

프루닝에 의한 구조 최적화를 함에 있어서 필요한 것은 어떠한 구조가 최적의 구조인가를 결정하는 것이다. 따라서 이를 판단하기 위한 기준이 필요하며, 이러한 기준으로써 모델 선택 방법이 쓰인다.

본 논문에서는 앞에서 제시한 이러한 문제점과 해결 방법을 바탕으로 하여 각각의 방법을 유기적으로 결합한 총체적인 신경회로망 최적화 방법을 제안한다. 다음 장에서는 제안하는 방법의 자세한 구성에 대해 살펴보겠다.

3. 제안하는 방법의 구성

본 논문에서 2장에서 살펴본 신경회로망 최적화 방법의 문제점과 해결 방안을 바탕으로 신경회로망의 가중치 매개 변수 최적화와 구조 최적화를 통합한 총체적인 신경회로망 최적화 방법을 제안한다. 먼저 신경회로망에 주어진 문제에 적합한 오차함수를 사용할 수 있는 자연 기울기 강하 학습 방법에, 보다 일반화 성능이 우수한 최적화된 매개변수를 얻기 위하여 적응적 정규화를 도입한다. 그리고 이렇게 최적화된 매개변수에 구조 최적화를 위한 자연 프루닝을 도입함으로써 최적의 후보 모델들을 생성한다. 이렇게 생성된 후보모델들 중 최적의 모델을 베이시안 정보 기준에 의해 평가함으로써 일반화 성능이 우수한 최적의 모델을 얻는다.

2장에서 신경회로망 학습이 프루닝에 의한 구조 최적화 성능에 큰 영향을 미치는 것을 살펴보았다. 본 논문에서는 학습에 의해 지역 극소에 빠지는 것을 방지하고, 이를 통해 구조 최적화 성능을 향상시키기 위하여 학습에 적응적 정규화를 도입한다. 학습 방법과 프루닝 방법은 신경회로망 학습에 효과적인 자연 기울기 학습 방법과 자연 프루닝 방법을 적용하였다. 널리 쓰이는 프루닝 방법의 일종인 OBS 방법은 학습에 정규화를 도입하면, 프루닝 과정에서 정규화 항을 고려해 주어야 한다. 하지만 정보 기하 이론에 기반한 자연 프루닝은 프루닝의 척도가 오차함수가 아닌 신경회로망의 통계적인 특성에 영향을 받기 때문에 프루닝 과정에 정규화항을 고려해 주지 않아도 되는 장점을 지닌다. 또한 본 논문에서는 정규화 파라메터를 적응적으로 조절하는 베이시안 방법을 도입하였고, 이렇게 구성한 후보 모델들 중 최적의 모델을 선택하는데 있어서 베이시안 정보기준을 적용하여 최적의 모델을 찾는 방법을 제안한다.

제안하는 방법에 대한 전체적인 과정을 나타내면 그림 1과 같다. 제안하는 방법은 적응적 정규화 방법과 자연 프루닝을 결합하여 일반화 성능이 단순한 후보 모델을 구성하고 베이시안 정보기준에 의해 최적의 모델을 선택하며 이를 절차적으로 나타내면 그림 2와 같다.

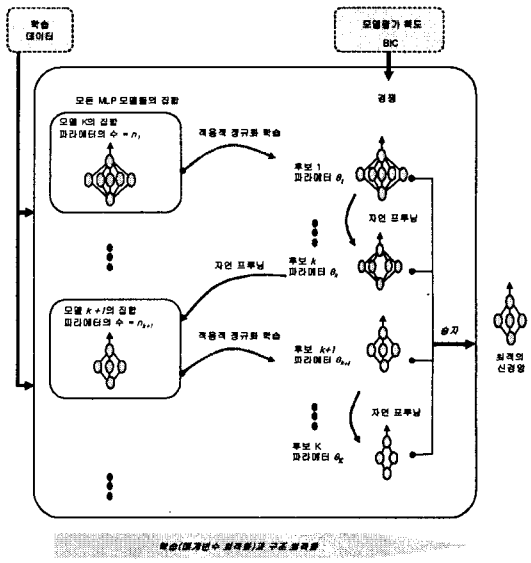


그림 1. 전체적인 신경회로망 최적화 방법

1단계:	큰 네트워크 구조에서 시작한다.
2단계:	**(적응적 정규화 자연기울기 강하 학습)** 학습에 의하여 정규화 항이 있는 오차함수를 최소화하는 최적의 가중치 매개변수를 구성한다.
3단계:	**(자연 프루닝)** 각각의 매개변수에 대한 중요도를 계산하여, 최소의 중요도를 갖는 매개변수를 찾아서 제거하고 남아 있는 매개변수들을 갱신한다.
4단계:	만약 거리의 변화가 임계치 보다 낮으면 3단계로 가고 그렇지 않으면 2단계로 간다.
5단계:	**(베이시안 정보 기준)** 마지막으로 후보 모델들 중에서 최적의 모델을 베이시안 정보 기준에 의하여 선택한다.

그림 2. 제안하는 방법의 단계적인 구성

다음은 제안하는 방법에 적용된 각각의 방법에 대해 자세히 살펴보도록 하겠다.

3.1 적응적 정규화 자연기울기 강하 학습

신경회로망은 N개의 학습쌍 $\boldsymbol{D}=\{(\boldsymbol{x}^{(v)},\boldsymbol{t}^{(v)})\}_{v=1}^{N}$에 의해서 학습된다. 적응적 정규화 방법을 도입한 방법의 오차함수는 식(1)과 같이 주어진다.

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

여기서 $E(x, y, \theta)$는 네트워크 모델과 입력 데이터에 의존하는 표준적인 성능 측정에 관계된 항이다. 이는 회귀문제에서 사용되는 가우시안 노이즈 모델의 경우 제곱합 오차가 되며, 분류문제에 사용되는 동전 던지기 모델의 경우 크로스 엔트로피 오차 함수가 된다[11]. $R(\theta)$는 모델의 복잡도에 의존하는 벌칙항이다. 이 항은 가중치 값이 커지는 것을 억제함으로써 매끄럽고 간단한 사상을 수행하도록 하는 역할을 한다. 이렇게 추가되는 벌칙항으로는 가중치 감소항이 널리 쓰이며 이는 신경회로망의 가중치의 제곱합으로 정의 된다.

식(1)과 같은 목적함수가 주어졌을 때 정보기하 이론에 기반한 자연 기울기는 식(2)와 같이 주어진다.

$$\begin{aligned}\tilde{\nabla}C(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta}_t) &= \boldsymbol{G}^{-1}(\boldsymbol{\theta}_t)\nabla C(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta}_t)\\ &= \boldsymbol{G}^{-1}(\boldsymbol{\theta}_t)(\nabla E(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta}_t)+\alpha\nabla R(\boldsymbol{\theta}_t))\end{aligned} \quad (2)$$

이때 G^{-1}는 피셔 정보행렬 G의 역행렬로 다층 퍼셉트론 공간에서의 G는 식(3)과 같이 주어진다.

$$\boldsymbol{G}_{ij}(\boldsymbol{\theta}) = E_{x,y}\left[\frac{\partial \log p(\boldsymbol{y}\mid\boldsymbol{x};\boldsymbol{\theta})}{\partial\theta_i}\frac{\partial \log p(\boldsymbol{y}\mid\boldsymbol{x};\boldsymbol{\theta})}{\partial\theta_j}\right] \quad (3)$$

여기서 $p(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta})$는 신경회로망의 확률 분포 함수이고, $E_{x,y}$는 데이터 $p(x,y)$의 진 분포에 대한 기대값을 나타낸다. 이러한 자연 기울기를 이용한 강하 학습 방법은 식(4)와 같다.

$$\begin{aligned}\boldsymbol{\theta}_{t+1} &= \boldsymbol{\theta}_t-\eta_t\tilde{\nabla}C(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta}_t)=\boldsymbol{\theta}_t-\eta_t\boldsymbol{G}^{-1}\nabla C(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta}_t)\\ &= \boldsymbol{\theta}_t-\eta_t\boldsymbol{G}^{-1}(\nabla E(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta}_t)+\alpha\nabla R(\boldsymbol{\theta}_t))\end{aligned} \quad (4)$$

학습을 통하여 궁극적으로 얻고자 하는 신경회로망은 학습데이터에 좋은 결과 만을 내는 신경회로망이 아니라 미지의 데이터에 대해 좋을 성능을 보이는 일반화 능력이 우수한 신경회로망을 얻는 것이다. 하지만 보통의 학습방법에는 일반화 성능을 향상시키는 기재가 존재하지 않는다. 따라서 학습과정에 일반화 성능 향상을 위한 기재가 필요하며 본 논문에서는 이를 위해 정규화 항을 도입하였다. 정규화는 정규화 매개변수에 따라 그 성능이 좌우된다. 본 논문에서는 학습하는 동안에 온라인으로 정규화 매개변수를 조정할 수 있고, 베이시안 유도과정에서 자연스럽게 정규화를 설명할 수 있는 베이시안 적응적 정규화를 도입한다. 이러한 베이시안 정규화 방법은 학습 집합과 검증 집합을 나눌 필요가 없이 모든 학습데이터를 학습에 사용하면서 일반화 오차를 추정할 수 있다는 장점을 지닌다. 식 (1)의 오차 함수에서 정규화 성능에 중요한 영향을 미치는 정규화 매개변수 는 베이시안 에비던스(evidence)에 의해 유도된 식(5)에 의해 결정된다(자세한 사항은[1] 참조).

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

여기서 N은 매개변수의 수이고 n은 데이터의 수이다.

3.2 자연 프루닝

신경회로망의 구조가 너무 단순하면 학습데이터나 테스트 데이터에 대해 만족할 만한 결과를 얻을 수 없고, 너무 복잡하면 학습데이터 내의 노이즈에 적합 되어서 학습데이터에 대해서는 좋은 결과를 보이나 테스트 데이터에 대해선 좋은 결과를 보이지 못하는 현상이 발생한다. 불행하게도 대부분의 실제 문제에 적합한 최적의 구조를 알아내는 것은 쉽지 않다. 따라서 구조 최적화 알고리즘을 통하여 최적의 구조를 찾아야 한다. 구조 최적화 방법의 하나인 프루닝 방법은 반복적인 절차로 완전 연결된 신경회로망으로부터 불필요한 가중치들을 제거하면서 최적의 구조를 찾아 내는 방법이다.

자연 프루닝은 학습에 의하여 추정된 가중치 $\hat{\boldsymbol{\theta}}$중에서 각각의 요소에 대한 중요도를 구하여 중요도가 가장 낮은 요소를 제거하는 방법이다. i 번째 요소 $\hat{\theta}_i$에 대한 중요도 $\delta F_i(\hat{\boldsymbol{\theta}})$는 식 (6)과 같이 계산 될 수 있다.

$$\delta F_i\left(\hat{\boldsymbol{\theta}}\right)=\left(\hat{\boldsymbol{\theta}}-\hat{\theta}^i\right)'\boldsymbol{G}\left(\hat{\boldsymbol{\theta}}\right)\left(\hat{\boldsymbol{\theta}}-\hat{\theta}^i\right) \quad (6)$$

여기서 $\boldsymbol{G}(\hat{\boldsymbol{\theta}})$는 $\hat{\boldsymbol{\theta}}$에서의 피셔 정보 행렬이다.

자연 프루닝은 널리 사용되고 있는 프루닝 방법인 OBD와 OBS보다 노이즈에 덜 민감하다. 또한 자연 프루닝의 중요도 계산에 사용되는 피셔정보 행렬은

오차함수에 의존하지 않고 단지 신경회로망 모델의 통계적인 특성에 의존하기 때문에 OBD와 OBS와 달리 정규화 방법의 도입시 프루닝 과정에서 정규화에 대해 고려하지 않아도 된다는 장점을 지닌다. Heskes는 제곱오차와 가우시안 근사를 할 경우 OBS의 헤시안 행렬과 자연 프루닝의 피셔정보 행렬이 같아져 동일한 방법이 된다는 것을 보였다[14]. 식(7)은 의 프루닝후 남아있는 가중치들의 변화를 최소화하기 위하여 가중치를 갱신하는 식이다(자세한 사항은 [14] 참조).

$$\theta_i^{new} = \hat{\theta}_i - \frac{G^{mi}(\hat{\theta})}{G^{mm}(\hat{\theta})}\hat{\theta}_m \tag{7}$$

여기서, G^{mi}는 피셔 정보 행렬의 m번째 행 i 번째 열의 요소이다.

3.3 베이시안 정보 기준(Bayesian Information Criterion)

신경회로망의 과다학습문제를 해결하기 위하여 두 가지 형태의 모델선택 방법이 자주 쓰인다. 하나는 크로스 검증 방법처럼 데이터를 나누어서 수행하는 표본외(out-of-sample) 모델 선택 방법이다. 다른 하나는 Akaike의 정보기준, 베이시안 정보 기준과 같이 데이터를 분할하지 않는 표본내(in-sample) 모델선택이다. 본 논문에서는 최적의 모델선택을 위하여 자연 프루닝에 의해 생성된 모델간의 패널티가 가해진 성능 측정도구로 베이시안 정보 기준을 적용한다. 이는 식(8)과 같으며 비선형 모델의 경우 $d>1$ 이고 d는 실험에 의하여 결정한다[16].

$$BIC = \log(E(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta})) + \frac{W^d \log(N)}{N} \tag{8}$$

여기서 N은 데이터의 수이고 W는 매개변수의 수이다. 본 논문에서는 적응적 자연기울기 강하 학습과 자연 프루닝의 반복적인 프로세스를 통해 얻어진 후보 모델들 중에서 일반화 성능이 우수한 최적의 모델을 선택하는 척도로서 베이시안 정보 기준을 사용하였다.

4. 실험 결과 및 분석

제안하는 NGARNP 방법은 자연기울기(Natural Gradient) 학습에 적응적 정규화(Adaptive Regularization)와 자연 프루닝(Natural Pruning)을 통합 시켜 신경회로망 구조를 최적화 시키고, 최적화된 모델들 중 최적의 모델을 베이시안 정보기준에 의하여 선택하는 방법이다. 이러한 제안하는 방법의 우수성을 검증하기 위하여 2가지 방법과의 비교를 통하여 우수성을 검증 하였다.

먼저 제안하는 방법은 정보 기하 이론에 기반한 방법들을 효과적으로 구성하였으며, 이러한 방법의 도입의 우수성을 검증하기 위하여 LMAROP 방법과 비교하였다. LMAROP방법은 LM (Levenberg-Marquardt) 학습 방법에 적응적 정규화(Adaptive Regularization)를 적용하고, OBS 프루닝 (OBS Pruning)을 결합한 방법이다. 이렇게 최적화 시킨 모델들 중 최적의 모델은 베이시안 정보 기준에 의하여 선택하였다. LMAROP 방법은 제곱합 오차(Sum of Squared Error)함수 밖에 사용할 수 없지만 제안하는 NGARNP 방법은 다양한 오차 함수를 사용할 수 있기 때문에 각각의 문제에 적합한 오차 함수를 적용할 수 있다. 따라서 회귀문제에는 제곱합 오차 함수를 분류문제에는 크로스 엔트로피(Cross Entropy) 오차함수를 사용할 수 있어 보다 좋은 성능을 보일 수 있다. 따라서 LMAROP 방법과의 비교를 통하여 제안하는 정보 기하 이론에 기반한 통합 프로세스의 우수성을 검증한다.

둘째, 제안하는 NGARNP방법은 적응적 정규화와 프루닝의 결합을 통하여 최적화의 성능을 높인 방법이다. 이를 확인하기 위하여 자연기울기(Natural Gradient) 학습에 자연 프루닝(Natural Pruning)으로 구성된 NGNP방법과의 비교를 하였다. 제안하는 NGARNP방법과 NGNP방법은 적응적 정규화의 유무의 차이가 있다. 이러한 적응적 정규화를 도입한 제안하는 방법의 우수성을 실험을 통하여 확인하였다.

LMAROP와 제안하는 NGARNP방법은 베이시안 정보 기준으로 최적의 모델을 선택하였다. 하지만 베이시안 정보 기준은 정규화등의 방법을 도입하여 최적해에 가깝게 근사를 해주어야 사용할 수 있는 척도이기 때문에, 이러한 방법을 도입하지 않은 NGNP 방법에는 사용할 수 없다. 따라서 NGNP 방법은 성능 비교를 위해 테스트 데이터에 대한 분류 오차가 최소인 모델 중 연결선 수가 최소인 모델을 최적의

모델로 선택하여 제안하는 방법과 비교하였다.

실험에 사용된 데이터와 신경회로망의 구조는 표 1과 같다. 본 논문에서는 벤치마크 데이터로 UCI 데이터를 사용하였다[17]. 회귀 문제로는 Boston Hosing과 Building 데이터를 사용하였다. 분류문제에 대해서는 2개의 클래스 분류 문제로 MONK3 데이터와 Diabetes 데이터를 사용하였고, 다중 클래스 분류 문제에 대해서 Glass 데이터와 Horse 데이터에 대해 실험을 하였다. 초기에 사용된 가중치 매개변수의 수는 바이어스를 포함하여 계산하였다. 실험은 가중치 초기화를 10번 다르게 수행하여 실험한 평균을 비교 하였다.

4.1 회귀(Regression)

회귀문제의 경우 제곱합 오차함수를 사용하며, 이러한 경우 학습과 프루닝 과정에서 NGARNP에 사용되는 피셔 정보 행렬과 LMARNP방법에 사용되는 헤시안 행렬이 동일해 진다[14]. 따라서 NGARNP방법과 LMAROP방법은 동일한 방법이 되기 때문에 NGNP방법과의 비교를 통하여 적응적 정규화를 도입한 통합 프로세스의 우수성을 살펴 보았다. 표 2의 실험 결과를 통하여 Boston Hosing 데이터와 Building 데이터에 대한 실험을 통하여 회귀문제들에 대하여, 제안하는 NGARNP 방법이 일반화 성능면에서 우수한 성능을 보이는 것을 확인 할 수 있었다.

다음은 10번 초기화 하여 실험한 결과중 하나의 예를 통하여 구조 최적화 과정과 일반화 성능의 변화를 살펴본 것이다. 그림 3은 Boston Housing 문제에 대한 NGNP와 NGARNP의 학습 결과의 예를 보인 것이다. (a)의 NGNP의 경우 테스트 데이터에 대해서 0.2886의 오차율을 보이면서 33개까지 구조를 최적화 시킨 그래프이고, (b)의 NGARNP의 경우 테스트 데이터에 대하여 0.149의 테스트 데이터 오차율을 보이면서 33개까지 구조 최적화를 수행한 그래프이다.

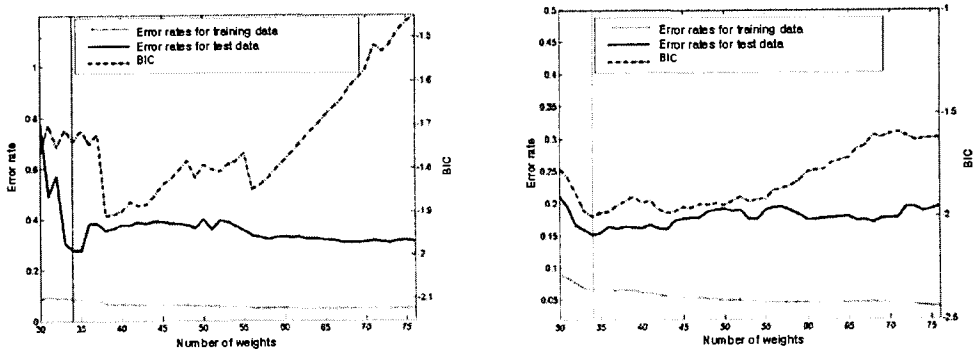
그림 4는 Building 문제에 대한 NGNP와 NGARNP의 결과의 한 예를 보인 것이다. (a)의 NGNP의 경우 테스트 데이터에 대해서 0.0213의 오차율을 보이면서 17개까지 구조를 최적화 시킨 그래프이고, (b)의 NGARNP의 경우 테스트 데이터에 대하여 0.0082의 테스트 데이터 오차율을 보이면서 19개까지 구조 최적화를 수행한 그래프이다. NGNP방법은 프루닝의 대상이 되는 모델이 최적해에 근사한다는 가정하에 수행되는 정규화를 적용하지 않으면 최적해를 제공할 수 없기 때문에 NGARNP 방법보다 일반화 성능이 낮았다. 구조최적화 성능은 같은 일반화 성능이나 일반화 성능이 우수 할때, 비교 할 수 있는 척도이기

표 1. 실험에 사용된 데이터 및 신경회로망 구조

		데이터 이름	입력수	은닉 노드수	출력수	매개변수의 수	총 데이터의 수 (학습/테스트)
회귀문제		Boston Housing	13	5	1	76	506(256/250)
		Building	14	5	3	93	4208(3156/1052)
분류문제	2개의 클래스	MONK3	17	5	1	96	432(122.432)
		Diabetes	8	5	1	57	768(576/192)
	다중 클래스	Glass	9	6	6	102	214(161/53)
		Horse	58	5	3	313	364(272/91)

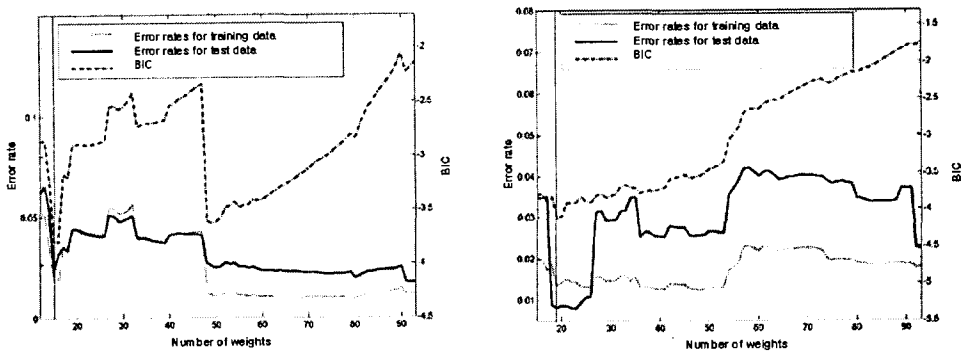
표 2. 회귀문제에 대한 가중치의 수 및 오차

		NGNP	NGARNP
Boston Housing	가중치의 수 (개)	36.3(1.702)	33.6(0.699)
	학습 데이터 오차	0.059(0.001)	0.052(0.011)
	테스트 데이터 오차	0.295(0.012)	0.151(0.002)
Building	가중치의 수 (개)	19.6(2.797)	22.2(5.712)
	학습 데이터 오차	0.012538(0.00522)	0.01334(0.00080)
	테스트 데이터 오차	0.026432(0.00775)	0.00826(0.00035)



(a) NGNP 방법 (b) NGARNP 방법

그림 3. Boston Hosing문제에 대한 학습 및 테스트 분류 오차와 일반화 오차



(a) NGNP 방법 (b) NGARNP 방법

그림 4. Building문제에 대한 학습 및 테스트 분류 오차와 일반화 오차

때문에 NGARNP 방법의 일반화 성능이 더 우수한 상태에서 NGNP의 가중치 수가 적다고 해서 구조최적화가 우수하다고 평가 할 수는 없다.

4.2 분 류 (Classification)

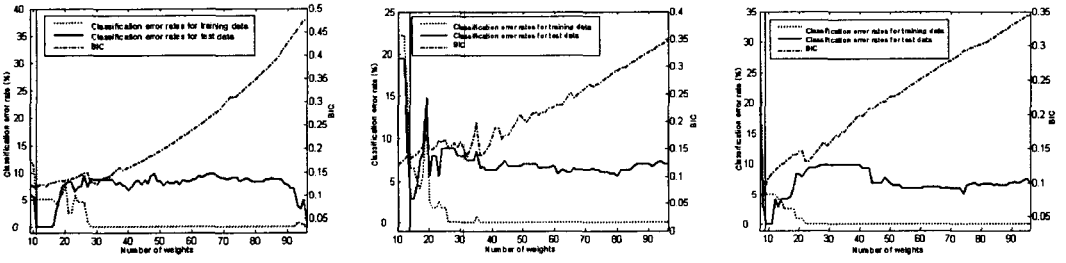
본 논문에서는 2개의 클래스 분류 문제로 MONK3와 Diabetes 데이터와 다중 클래스분류 문제로 Glass와 Horse 데이터에 대하여 실험을 하였다. 실험에 대한 결과는 표 3과 같다. 실험을 통하여 적응적 정규화를 도입한 제안하는 NGARNP 방법과 LMAROP 방법 의 비교를 통하여 적응적 정규화가 최적화 성능에 큰 영향을 미치는 것을 확인 할 수 있었다. 또한 제안하는 NGARNP 방법은 정보 기하 이론에 기반을 둔 방법들을 통합 구성함으로써 분류문제에 적합한 크로스엔트로피 오차함수 적용이 가능해졌고, 그것으로 인하여 LMAROP방법보다 우수한 성능을 보이는 것을 확인 할 수 있었다.

다음은 10번 초기화 하여 실험한 결과중 하나의 예를 통하여 구조최적화 과정과 일반화 성능의 변화를 살펴보도록 하겠다. 그림 5는 MONK3의 데이터에 대한 실험 결과의 예이다. MONK3의 경우 학습데이터에 5%의 노이즈가 포함되어 있는 경우이다. (a)의 LMAROP는 100%의 테스트 데이터에 대한 분류율을 보이면서 11개의 가중치 수를 가진 모델을 구성하였으며, (b)의 NGNP는 97.55%의 테스트 데이터에 대한 분류율을 보이는 가중치의 수가 15개인 모델을, (c)의 NGARNP는 100%의 테스트 데이터에 대한 분류율을 보이면서 9개의 가중치 수를 갖는 일반화 성능이 우수한 모델을 구성하였다.

그림 6은 Diabetes 문제에 대한 LMAROP, NGNP와 NGARNP의 학습 결과의 예를 보인 것이다. (a)의 LMAROP의 경우 테스트 데이터에 대해서 76.56%

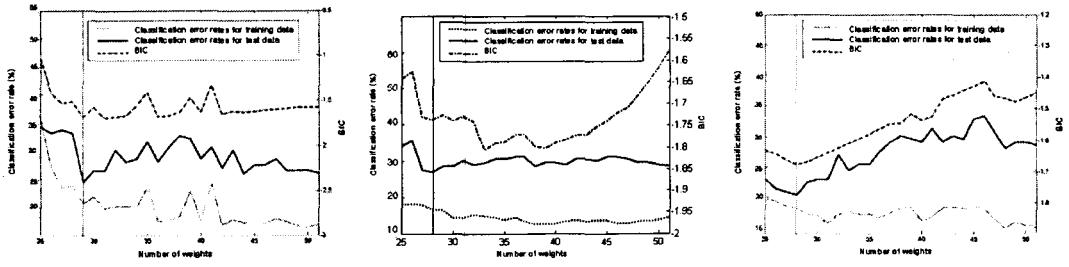
표 3. 분류 문제에 대한 가중치의 수 및 분류율

			LMAROP	NGNP	NGARNP
2개의 클래스	MONK3	가중치의 수 (개)	11.6(0.70)	17.0(1.05)	10.6(0.84)
		학습 데이터 분류율(%)	97.54(0.95)	95.9(1.96)	96.48(0.95)
		테스트 데이터 분류율(%)	100(0)	97.22(1.60)	100(0)
	Diabetes	가중치의 수 (개)	28.4(3.57)	30(4.16)	26.8(1.23)
		학습 데이터 분류율(%)	81.86(2.178)	83.85(2.639)	80.75(0.569)
		테스트 데이터 분류율(%)	77.40(0.858)	76.25(0.703)	77.60(0.491)
다중 클래스	Glass	가중치의 수 (개)	49.8(1.87)	62.5(2.27)	47.8(1.22)
		학습데이터 분류율 (%)	72.07(3.05)	78.88(2.12)	77.10(2.93)
		테스트 데이터 분류율(%)	70.56(1.32)	69.05(0.97)	73.77(1.39)
	Horse	가중치의 수 (개)	58.8(11.56)	53.4(3.17)	44.4(1.96)
		학습 데이터 분류율(%)	74.03(3.128)	74.66(2.500)	76.67(1.973)
		테스트 데이터 분류율(%)	73.74(0.959)	68.90(1.275)	75.17(1.182)



(a) LMAROP 방법 (b) NGNP 방법 (c) NGARNP 방법

그림 5. Monk-3문제에 대한 학습 및 테스트 분류 오차와 일반화 오차

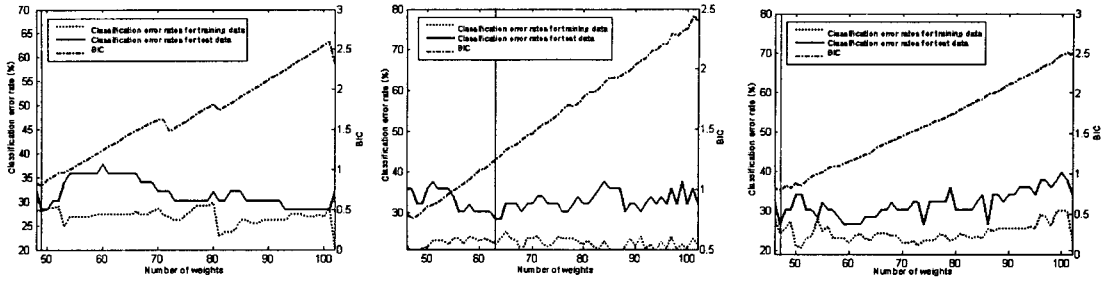


(a) LMAROP 방법 (b) NGNP 방법 (c) NGARNP 방법

그림 6. Diabetes문제에 대한 학습 및 테스트 분류 오차와 일반화 오차

의 분류율을 보이면서 29개 까지 구조를 최적화 시킨 그래프이고, (b)의 NGNP의 경우 테스트 데이터에 대해서 75.52%의 분류율을 보이면서 28개까지 구조를 최적화 시킨 그래프이고, (c)의 NGARNP의 경우 테스트 데이터에 대하여 78.13%의 테스트 데이터 분류율을 보이면서 28개까지 구조를 최적화 시킨 그래프이다. 전체 평균과 마찬가지로 NGARNP 방법과 NGNP 의 구조최적화 성능은 유사하지만, 일반화 성능은 NGARNP 방법이 우수하였다.

그림 7은 Glass 데이터에 대한 LMAROP, NGNP와 NGARNP의 학습 결과의 예를 보인 것이다. (a)의 LMAROP의 경우 71.7%의 테스트 데이터에 대한 분



(a) LMAROP 방법 (b) NGNP 방법 (c) NGARNP 방법

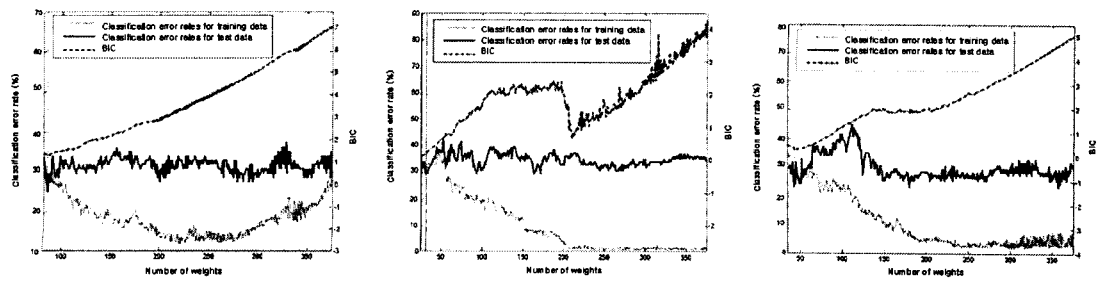
그림 7. Glass문제에 대한 학습 및 테스트분류 오차와 일반화 오차

류율을 보이면서 49개의 가중치를 가진 모델을 구성하였으며, (b)의 NGNP는 68.05%의 테스트 데이터에 대한 분류율을 보이는 가중치의 수가 63개인 모델을 구성하였고, (c)의 NGARNP는 73.59%의 테스트 데이터 분류율 보이면서 47개의 가중치 수를 갖는 모델을 구성하였다. 전체 평균과 마찬가지로 NGARNP 방법이 LMAROP와 NGNP 보다 우수한 구조최적화 성능과 일반화 성능을 보였다.

그림 8은 Horse 문제에 대한 LMAROP, NGNP와 NGARNP의 학습 결과의 한 예를 보인 것이다. (a)의 LMAROP의 경우 72.53%의 테스트 데이터에 대한 분류율을 보이면서 58개의 가중치를 가진 모델을 구성하였으며, (b)의 NGNP는 70.33%의 테스트 데이터에 대한 분류율을 보이는 가중치의 수가 55개인 모델을 구성하였고, (c)의 NGARNP는 74.36%의 테스트 데이터 분류율 보이면서 43개의 가중치 수를 갖는 모델을 구성하였다. 전체 평균과 마찬가지로 NGARNP 방법이 LMAROP와 NGNP 보다 구조 최적화 성능을 보이면서, 일반화 성능이 우수한 것을 확인할 수 있었다.

5. 결 론

본 논문에서는 신경회로망을 실제 문제에 적용하는데 있어서 일반화 성능을 향상시키는 통합적인 신경회로망 최적화 방법을 제안하였다. 이를 위하여 적응적 정규화를 도입한 학습에 의해서 가중치 매개변수를 최적화 하였으며, 이렇게 최적화된 가중치 매개변수들을 대상으로 프루닝에 의해 구조 최적화를 수행하였다. 또한 이렇게 생성된 후보 모델들 중 최적의 모델을 선택하는 기준으로 베이시안 정보 기준을 도입한 총체적인 신경회로망 최적화 과정을 제안하였다. 각각의 단계에 대하여 최적의 방법들로 구성하고, 일관된 방법들의 통합 구성을 통하여 효율성을 높였다. 제안하는 방법에 적용된 자연기울기 강하 학습법은 기존의 헤시안 행렬을 이용한 2차 근사 학습법이 오차 함수로 오차 제곱합만을 사용할 수 있는 것과 달리, 패턴 분류에 더 좋은 성능을 보이는 크로



(a) LMAROP 방법 (b) NGNP 방법 (c) NGARNP 방법

그림 8. Horse문제에 대한 학습 및 테스트 분류 오차와 일반화 오차

스 엔트로피 오차 함수도 사용할 수 있으며, 이로 인하여 좋은 일반화 성능을 기대 할 수 있다. 여기에 적응적 정규화를 도입하여 최적해에 가까운 근사를 보일 수 있도록 가중치 매개변수들을 최적화 함으로써, 일반화 성능 뿐만 아니라 구조 최적화 성능을 향상 시켰다. OBD와 OBS 프루닝에서 사용되는 헤시안 행렬과는 달리, 자연 프루닝의 피셔 정보 행렬은 오차 함수에 무관 하기 때문에 학습시 고려했던 정규화 항에 대한 고려를 프루닝 단계에서는 하지 않아도 되는 장점을 지닌다. 일관된 방법을 유기적으로 구성하였기 때문에 학습에 사용하였던 피셔 정보 행렬을 프루닝 단계에서 그대로 사용 할 수 있는 효율성을 가진다.

향후 연구과제는 계산 시간의 단축을 위하여 프루닝 단계에서 여러 가중치 매개변수들을 동시에 제거하고, 남아 있는 요소들을 갱신하는 방법의 도입이 필요하겠다.

참 고 문 헌

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

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

[3] Reed, R. D., Marks, R. J., Neural Smithing: Supervised Learning in Feedforward Artificial Neural Networks, MIT Press, 1999.

[4] Andersen, T., Rimer, M., Martinez, T., "Optimal Artificial Neural Network Architecture Selection for Baggin," *Proceedings of International Joint Conference on Neural Networks*, 2, 790-795, 2001.

[5] Ripley, B., *Pattern Recognition and Neural Networks*, Cambridge: Cambridge University Press, 1996.

[6] Hansen, L. K., Pedersen, M. W., "Controlled Growth of Cascade Correlation Nets," *Proceedings of International Conference on Neural Networks*, 797-800, 1994.

[7] Larsen, J., Svarer, C., Andersen, L. N., Hansen, L. K., "Adaptive Regularization in Neural Network Modeling, Neural Networks: Tricks of the Trade," *Lecture Notes in Computer Science*, 1524, Germany: Springer-Verlag, 113-132, 1998.

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

[9] Lee, H., Jee, T., Park, H., Lee, Y., "A Hybrid Approach to Complexity Optimization of Neutral Networks," *Proceedings of International Conference on Neural Information Processing*, 3, 1455-1460, 2001.

[10] Amari, S., "Natural gradient works efficiently in learning," *Neural Computation*, 10(2), 251-276, 1998.

[11] 박혜영, Efficient On-line Learning Algorithms Based on Information Geometry for Stochastic Neural Networks, *연세대학교 박사학위 청구 논문*, 2000.

[12] Amari, S., Park, H., Fukumizu, K., "Adaptive method of realizing natural gradient learning for multilayer perceptrons," *Neural Computation*, 12(6), 1399-1409, 2000.

[13] Park, H., "Practical Consideration on Generalization of Natural Gradient Learning," *Lecture Notes in Computer Science*, 2084, 402-409, 2001.

[14] Heskes, T., "On Natural Learning and Pruning in Multilayered Perceptrons," *Neural Computation*, 12, 1037-1057, 2000.

[15] Laar, P. V. D., Heskes, T., "Pruning Using Parameter and Neuronal Metrics," *Neural Computation*, 11, 977-993, 1999.

[16] Qi, M., Zhang, G. P., "An investigation of model selection criteria for neural network time series forecasting," *European Journal of Operational Research*, 132, 666-680, 2001.

[17] Murphy, P. M., Aha, D. W., "UCI Repository of Machine Learning Databases[Machine Readable Data Repository]," Univ. of California, Dept of Information and Computer Science, 1996.

이 현 진

1996년 순천향대학교 전산학과 (학사).
1998년 연세대학교 대학원 컴퓨터과학과 (석사).
2002년 연세대학교 대학원 컴퓨터과학과 (박사).
현재 한국 싸이버 대학교 컴퓨터 정보통신학부 전임강사.
관심분야 : 패턴인식, 신경회로망, 영상처리, 사이버교육

박 혜 영

1994년 연세대학교 전산과학과 (학사).
1996년 연세대학교 대학원 컴퓨터과학과 (석사).
2000년 연세대학교 대학원 컴퓨터·산업시스템 공학과 (박사).
현재 일본 이화학 연구소 뇌과학연구센터 뇌수리연구팀 연구원.
관심분야 : 계산학습이론, 통계적 정보처리 이론, 신경회로망, 패턴인식, 영상처리등

교신저자

이 현 진 (135-244) 서울특별시 강남구 개포4동 현대2차 apt 208동 702호