# 자연 프루닝과 베이시안 선택에 의한 신경회로망 일반화 성능 향상

## (Improving Generalization Performance of Neural Networks using Natural Pruning and Bayesian Selection)

이 현 진 † 박 혜 영 †† 이 일 병 †††

(Hyun Jin Lee) (Hye Young Park) (Yill Byung Lee)

**요 약** 신경회로망 설계 및 모델선택의 목표는 최적의 구조를 가지는 일반화 성능이 우수한 네트워크를 구성 하는 것이다. 하지만 학습데이타에는 노이즈(noise)가 존재하고, 그 수도 충분하지 않기 때문에 최종적으로 표현하고자 하는 진확률 분포와 학습데이타에 의해 표현되는 경험확률분포(empirical probability density) 사이에는 차이가 발생한다. 이러한 차이 때문에 신경회로망을 학습데이타에 대하여 과다하게 적합(fitting)시키면, 학습데이타만의 확률분포를 잘 추정하도록 매개변수들이 조정되어 버리고, 진 확률 분포로부터 멀어지게 된다. 이러한 현상을 과다학습이라고 하며, 과다 학습된 신경회로망은 학습데이타에 대한 근사는 우수하지만, 새로운 데이타에 대한 예측은 떨어지게 된다. 또한 신경회로망의 복잡도가 증가 할수록 더 많은 매개변수들이 노이즈에 쉽게 적합 되어 과다학습 현상은 더욱 심화된다.

본 논문에서는 통계적인 관점을 바탕으로 신경회로망의 일반화 성능을 향상시키는 신경회로망의 설계 및 모델 선택의 통합적인 프로세스를 제안하고자 한다. 먼저 학습의 과정에서 적응적 정규화가 있는 자연기울기 학습을 통해 수렴속도의 향상과 동시에 과다학습을 방지하여 진확률 분포에 가까운 신경회로망을 얻는다. 이렇게 얻어진 신경회로망에 자연 프루닝(natural pruning) 방법을 적용하여 서로 다른 크기의 후보 신경회로망 모델을 얻는다. 이러한 학습과 복잡도 최적화의 통합 프로세스를 통하여 얻은 후보 모델들 중에서 최적의 모델을 베이시안 정보기준에 의해 선택함으로써 일반화 성능이 우수한 최적의 모델을 구성하는 방법을 제안한다. 또한 벤치마크 문제를 이용한 컴퓨터 시뮬레이션을 통하여, 제안하는 학습 및 모델 선택의 통합 프로세스의 일반화 성능과 구조 최적화 성능의 우수성을 검증한다.

**키워드** : 적응적 정규화, 자연 프루닝, 일반화, 신경회로망 설계, 모델 선택

***Abstract*** The objective of a neural network design and model selection is to construct an optimal network with a good generalization performance. However, training data include noises, and the number of training data is not sufficient, which results in the difference between the true probability distribution and the empirical one. The difference makes the learning parameters to over-fit only to training data and to deviate from the true distribution of data, which is called the overfitting phenomenon. The overfitted neural network shows good approximations for the training data, but gives bad predictions to untrained new data. As the complexity of the neural network increases, this overfitting phenomenon also becomes more severe.

In this paper, by taking statistical viewpoint, we proposed an integrative process for neural network design and model selection method in order to improve generalization performance. At first, by using the natural gradient learning with adaptive regularization, we try to obtain optimal parameters that are not overfitted to training data with fast convergence. By adopting the natural pruning to the obtained optimal parameters, we generate several candidates of network model with different sizes. Finally, we

† 정 회 원 : 한국싸이버대학교 컴퓨터정보통신학부
hjlee@mail.kcu.ac
†† 학 생 회원 : 일본이화학연구소 뇌수리연구팀
hypark@brain.riken.go.jp
††† 종신회원 : 연세대학교 컴퓨터과학과 교수
yblee@csai.yonse.ac.kr
논문접수 : 2002년 6월 24일
심사완료 : 2003년 2월 4일

select an optimal model among candidate models based on the Bayesian Information Criteria. Through the computer simulation on benchmark problems, we confirm the generalization and structure optimization performance of the proposed integrative process of learning and model selection.

**Key words** : Adaptive Regularization, Natural Pruning, Generalization, Neural Networks Design, Model Selection

## 1. 서 론

신경회로망 설계 및 모델 선택의 목표는 학습데이타를 통해서 가능한 작은 예측오차를 갖는 입출력 관계를 찾아내는 것이다[1]. 하지만 학습 데이타만으로는 이러한 입출력 관계를 구성하기엔 부족하다. 따라서, 학습데이타에 대해서만 잘 학습하게 되면 미지의 데이타에 대한 예측 능력이 저하되는 현상이 발생한다. 이러한 현상을 과다학습(overfitting)이라고 하며, 이는 학습데이타의 특징만 발견하여 원래의 데이타를 생성하는 함수를 발견하지 못하여 발생한다[2]. 또한, 학습데이타가 적을수록 노이즈(noise)의 영향력은 더 커지게 되고, 신경회로망의 복잡도가 증가 할수록 이러한 노이즈의 영향이 커지게 된다. 이러한 노이즈는 신경회로망의 가중치에 반영되어 일반화 성능을 저하시킨다[3]. 따라서 이러한 현상을 극복하여, 일반화 성능 향상 시키기 위한 다양한 방법들이 제안되었다.

첫째, 검증방법은 신경회로망 학습시에 학습데이타를 학습집합과 검증집합으로 나누어 일반화 오차를 추정하는 방법이다. 이 방법은 학습데이타를 학습집합과 검증집합으로 나눌 때의 샘플링 방법에 따라 성능 차이가 발생하고 샘플링 시간이 오래 걸린다는 단점이 있다[4].

둘째, 통계학에 기반을 둔 모델 선택 방법들이 있으며, 이는 과다학습이 모델의 복잡도와 관계가 있다는 사실에 기반한다. 모델의 복잡도가 증가하면 학습오차의 최소값은 0에 가까워지나, 과다학습으로 인하여 일반화 성능이 저하된다. 복잡도가 감소하게 되면 학습오차의 최소값은 증가하게 되나, 과다학습은 발생하지 않게 된다. 하지만 지나치게 복잡도를 감소시키면, 원하는 목표에 도달하지 못하는 현상이 발생한다. 따라서 복잡도를 적절이 조절하는 것이 중요하다.

셋째, 검증 집합을 따로 나눌 필요 없이 학습된 모델의 일반화 성능을 평가 할 수 있는 모델선택 척도에 의한 방법들이 있다. 이러한 방법들로는 선형모델에 대해서 일반화 성능을 평가하는 Akaike의 최종 예측오차(Final Prediction Error: FPE)와 Akaike의 정보 기준(Akaike's Information Criterion: AIC)등이 있다[1]. Akaike의 정보기준을 더 일반화 시켜서 비선형 모델과 정규화항이 존재하는 경우를 다룰 수 있는 일반화된 예측 오차(Generalized Prediction Error: GPE)방법과 네트워크 정보 기준(Network Information Criterion: NIC)을 이용하는 방법이 있다[5]. 또한 베이즈(Bayes)의 추정에 의해 유도된 베이시안 정보 기준(Baysian Information Criterion: BIC)과 부호장 최소화의 원리로부터 나온 최소 기술 거리(Minimum Description Length: MDL)등이 있다[5].

넷째, 가중치 매개변수의 수를 조절하는 방법에는 프루닝(pruning) 방법과 그로잉(growing) 방법이 있다. 신경회로망의 가중치 매개변수의 수를 줄이는 프루닝 방법은 중요한 가중치 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 두 방법은 어떤 가중치 매개변수를 제거했을 때 발생하는 오차의 변화를 추정하여 이를 바탕으로 가중치 매개변수를 제거하는 방법이다. 프루닝 방법과 반대로 그로잉 방법은 매우 간단한 모델로부터 시작해서 하나의 노드를 추가시키고, 연관된 가중치들을 추가시키고, 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.

마지막으로 신경회로망의 복잡도를 통제하기 위하여 오차함수에 패널티항을 추가 시켜 수행하는 정규화 방법이 있다. 이 방법의 목적함수는 학습오차에 관련된 항과 가중치 매개변수의 복잡도를 제어하는 패널티항(penalty term)으로 구성된다. 이렇게 함으로써 최소한의 복잡도를 가지고 학습오차를 최소화하는 신경회로망을 얻을 수 있다. 가장 많이 사용되는 패널티항으로는 가중치 감소(weight decay)항이 있다[3]. 이 방법은 패널티항의 영향력을 조정하는 패널티항 매개변수에 따라 성능이 좌우되며, 이를 조절하는 것이 성능향상에 관건이 되고 있다.

최근 연구에서는 각각의 일반화 성능 향상 방법들의 단점을 극복하기 위하여, 다양한 방법들을 결합시키는 연구가 시도되고 있다. Hansen은 정규화 방법의 일종인

가중치 감소항 방법과 프루닝 방법의 일종인 OBS를 결합하여 일반화 성능을 향상시킨 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 신경회로망 설계 및 모델선택의 통합 방법을 제안하고자 한다. 먼저 주어진 문제에 적합한 오차함수를 사용함으로써 일반화 성능을 향상 시킨다. 제안하는 방법에서 사용하는, 자연 기울기 학습 방법은 2차 근사 학습 방법에 비해 다양한 오차함수를 적용하는 것이 가능하다. 따라서 회귀문제에 대해서는 우수한 성능을 보이는 제곱 오차 합(sum of squared error) 오차함수를 적용할 수 있고, 분류문제에 대해서는 이에 적합한 크로스엔트로피(cross entropy) 오차함수를 적용하는 것이 가능하다[1]. 이러한 자연기울기 강하 학습 방법에 베이시안 적응적 정규화 방법을 도입함으로써 신경회로망의 복잡도를 간접적으로 통제하여 일반화 성능이 우수한 모델을 얻을 수 있다. 하지만 이러한 신경회로망 내에는 불필요한 가중치들이 존재하며 이들은 신경회로망의 복잡도를 증가시키고, 노이즈에 쉽게 적합 되도록 하여 일반화 성능의 저하를 발생 시킨다. 따라서 신경회로망을 단순화 시키기 위하여, 자연 프루닝을 수행한다. 무조건 모델을 단순화 시키면, 과소적합(underfitting)이 발생하기 때문에, 베이시안 정보 기준에 의해서 프루닝에 의해 생성된 단순한 후보모델들의 일반화 성능을 비교하여 최적의 모델을 선택한다.

본 논문의 구성은 다음과 같다. 2장에서는 적용 대상이 되는 통계적 신경회로망에 대해 살펴본다. 3장에서는 제안하는 방법의 구성 및 기반 이론에 대해 살펴본다. 4장에서는 실험결과를 살펴보고 분석한다. 그리고 마지막으로 5장에서는 결론을 내린다.

## 2. 통계적 신경회로망

먼저 본 논문의 적용 대상이 되는 통계적인 신경회로망에 대해 살펴본다. 먼저 매개변수 $\boldsymbol{\theta}$와 신경회로망의 구조를 나타내는 입출력간의 매핑을 결정하는 함수 $f(\boldsymbol{x};\boldsymbol{\theta})$를 갖는 식(1)과 같은 전방향 신경회로망 모델을 가정하자.

$$\boldsymbol{y}=f(\boldsymbol{x};\boldsymbol{\theta}) \tag{1}$$

이때 $\boldsymbol{x}$는 입력벡터이고 $\boldsymbol{y}$는 출력벡터이며, $\boldsymbol{\theta}$는 가중치 벡터이다. 입력 $\boldsymbol{x}$가 주어졌을 때 결정되는 출력 $\boldsymbol{y}$는 에 $f(\boldsymbol{x};\boldsymbol{\theta})$의해 결정되고, 이 신경회로망의 확률적인 프로세스는 식(2)와 같은 조건부 확률 밀도에 의해 표현될 수 있다.

$$p(\boldsymbol{y}\mid\boldsymbol{x};\boldsymbol{\theta})=r(\boldsymbol{y}\mid f(\boldsymbol{x};\boldsymbol{\theta})) \tag{2}$$

$$f(\boldsymbol{x};\boldsymbol{\theta})=(f_1(\boldsymbol{x};\boldsymbol{\theta}),\cdots,f_L(\boldsymbol{x};\boldsymbol{\theta})) \tag{3}$$

각각의 출력 노드 $y_i$는 결정함수 $f_i(\boldsymbol{x};\boldsymbol{\theta})$와 신경회로망의 통계적인 특성을 나타내는 함수 $r$에 의해 결정된다. 결정함수 $f_i(\boldsymbol{x};\boldsymbol{\theta})$는 식(4)와 같이 정의된다.

$$f_1(\boldsymbol{x};\boldsymbol{\theta})=\varphi_0\Big(\sum_j v_{ij}\varphi_h(\boldsymbol{w}_j^T\boldsymbol{x}+b_j)+b_{o_i}\Big) \tag{4}$$

여기서 $v_{ij}$, $w_j$, $b_j$, $b_{o_i}$는 신경회로망의 매개변수이다. 함수 $\varphi_o$, $\varphi_h$는 출력노드와 은닉노드의 활성화 함수이다. 통계적 특성을 나타내는 함수 $r$을 어떻게 결정하는 가에 따라 오차 함수가 달라진다[10]. 이러한 통계적인 측면에서 신경회로망은 확률 밀도 함수 $p(\boldsymbol{y}\mid\boldsymbol{x},\boldsymbol{\theta})$를 갖는 네트워크의 작용으로 나타낼 수 있다.

입력 $\boldsymbol{x}$와 출력 $\boldsymbol{y}$와 매개변수 $\boldsymbol{\theta}$를 갖는 통계적 신경회로망의 확률밀도 함수 공간 $\{p(\boldsymbol{y}\mid\boldsymbol{x};\boldsymbol{\theta})\mid\boldsymbol{\theta}\in R^M\}$을 생각해 보자. 학습의 목표는 주어진 오차함수 (5)를 최소화 하는 최적의 매개변수 $\boldsymbol{\theta}^*$를 찾는 것이다. 오차함수는 우도 함수(likelihood function)의 음의 로그로 나타내진다. 즉 오차함수를 최소화 하는 것은 주어진 데이타에 대한 우도를 최대화 하는 것을 의미한다.

$$E(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta})=-\log p(\boldsymbol{y}\mid\boldsymbol{x};\boldsymbol{\theta}) \tag{5}$$

$p(\boldsymbol{y}\mid\boldsymbol{x},\boldsymbol{\theta})$의 확률공간은 리마니안 공간의 특성을 가지고, 이때 매트릭 텐서(metric tenser)는 식(6)과 같이 피셔 정보 행렬(Fisher information matrix) 로 주어진다[10].

$$G_{ij}(\boldsymbol{\theta})=E_{\boldsymbol{x},\boldsymbol{y}}\left[\frac{\partial\log p(\boldsymbol{y}\mid\boldsymbol{x};\boldsymbol{\theta})}{\partial\theta_i}\,\frac{\partial\log p(\boldsymbol{y}\mid\boldsymbol{x};\boldsymbol{\theta})}{\partial\theta_j}\right] \tag{6}$$

$E_{\boldsymbol{x},\boldsymbol{y}}$는 데이타 $p(\boldsymbol{x},\boldsymbol{y})$의 진 분포에 대한 기대값을 나타낸다. 따라서 통계적인 신경회로망에 대한 자연기울기는 식 (7)과 같고, 이때의 학습은 식 (8)과 같다.

$$\tilde{\nabla}E(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta})=\boldsymbol{G}^{-1}(\boldsymbol{\theta})\nabla E(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta}) \tag{7}$$

$$\boldsymbol{\theta}_{t+1}=\boldsymbol{\theta}_t-\eta_t\tilde{\nabla}E(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta}_t)=\boldsymbol{\theta}_t-\eta_t\boldsymbol{G}^{-1}\nabla E(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta}_t) \tag{8}$$

Amari는 자연기울기 학습의 개념을 제안하고 자연기

울기가 피셔 효율성을 획득할 수 있다는 것을 증명하였다[11]. 자연기울기는 학습이 계속 반복됨에도 불구하고 한동안 오차가 줄지 않는 플라토(plateau)를 피하거나 완화 시킬 수 있기 때문에, 학습 속도를 향상시킬 수 있다[12,13]. 전방향 신경회로망의 학습에 자연기울기의 개념을 적용하는데 있어서 피셔 정보 행렬과 그 역행렬의 계산상의 문제를 해결하기 위해 자연기울기의 추정치를 계산하는 적응적 자연기울기 방법이 제안되었고 다양한 통계적인 모델로 확장하는 연구들이 진행되었다[11,12,13]. 또한 이러한 리마니안 공간상의 거리척도를 이용한 일반화 성능 향상 방법들이 제안되고 있다. 신경회로망 공간에서 리마니안 거리척도를 이용하는 프루닝 방법이 제안되었고, 이는 기존의 유클리디안 거리에 기반한 방법에 비해 효과적이라는 것이 밝혀졌다[14,15]. 또한 자연기울기 강하 학습법에 정규화를 도입하여 일반화 성능을 향상시키는 연구가 제안되었다[13].

## 3. 제안하는 방법의 구성

목표가 되는 입력벡터 $\boldsymbol{x}$를 받고 출력벡터 $\boldsymbol{y}$를 내놓는 통계적인 시스템을 생각해보면, 입력벡터 $\boldsymbol{x}$는 확률 $q(\boldsymbol{x})$라는 조건하에 이루어지고, 출력벡터 $\boldsymbol{y}$는 $\boldsymbol{x}$에 의해 구체화 되는 조건부 확률 분포 $q(\boldsymbol{y}\mid\boldsymbol{x})$에 따라 출력을 내놓는다. 같은 방법으로 신경회로망은 조건부 확률 분포 $p(\boldsymbol{y}\mid\boldsymbol{x};\boldsymbol{\theta})$를 가진다. 여기서 $\boldsymbol{\theta}$는 $W$ 차원을 갖는 가중치 매개변수의 집합이다. 신경회로망 학습과 모델선택의 목표는 경험적인 분포 $p(\boldsymbol{y}\mid\boldsymbol{x};\hat{\boldsymbol{\theta}})$가 목표 시스템 $q(\boldsymbol{y}\mid\boldsymbol{x})$에 도달하도록 하는 가중치 $\hat{\boldsymbol{\theta}}$를 찾는 것이다. 하지만 목표 시스템 $q(\boldsymbol{y}\mid\boldsymbol{x})$를 추정할 수 있을 만큼 학습데이타가 충분하지 않고, 학습데이타에는 노이즈가 존재하기 때문에 $p(\boldsymbol{y}\mid\boldsymbol{x};\hat{\boldsymbol{\theta}})$와 $q(\boldsymbol{y}\mid\boldsymbol{x})$ 사이에는 차이가 발생한다. 이러한 차이를 일반화 오차라 하며 이는 식(9)와 같이 나타낼 수 있다. 이러한 일반화 오차는 식(10)과 같이 목표 시스템 $q(\boldsymbol{y}\mid\boldsymbol{x})$와 경험적인 확률 분포 $p(\boldsymbol{y}\mid\boldsymbol{x};\hat{\boldsymbol{\theta}})$사이의 쿨백-라이블러 발산으로 정의된다.

$$E_{gn}(\hat{\boldsymbol{\theta}}) = d(q(\boldsymbol{y}\mid\boldsymbol{x}):p(\boldsymbol{y}\mid\boldsymbol{x},\hat{\boldsymbol{\theta}}) \tag{9}$$

$$= E_{x,y}\left[\log\frac{q(\boldsymbol{y}\mid\boldsymbol{x})}{p(\boldsymbol{y}\mid\boldsymbol{x},\hat{\boldsymbol{\theta}})}\right] \tag{10}$$

$$= E_{x,y}[\log q(\boldsymbol{y}\mid\boldsymbol{x}) - \log p(\boldsymbol{y}\mid\boldsymbol{x},\hat{\boldsymbol{\theta}})] \tag{11}$$

$$= E_{x,y}[\log q(\boldsymbol{y}\mid\boldsymbol{x}) + \log p(\boldsymbol{y}\mid\boldsymbol{x},\boldsymbol{\theta}^*) - \log p(\boldsymbol{y}\mid\boldsymbol{x},\boldsymbol{\theta}^*) - \log p(\boldsymbol{y}\mid\boldsymbol{x},\hat{\boldsymbol{\theta}})] \tag{12}$$

$$= E_{x,y}\left[\log\frac{q(\boldsymbol{y}\mid\boldsymbol{x})}{p(\boldsymbol{y}\mid\boldsymbol{x},\boldsymbol{\theta}^*)}\right] + E_{x,y}\left[\log\frac{\log q(\boldsymbol{y}\mid\boldsymbol{x},\boldsymbol{\theta}^*)}{\log p(\boldsymbol{y}\mid\boldsymbol{x},\hat{\boldsymbol{\theta}})}\right] \tag{13}$$

$$= d(p(\boldsymbol{y}\mid\boldsymbol{x},\boldsymbol{\theta}^*):q(\boldsymbol{y}\mid\boldsymbol{x})) + d(p(\boldsymbol{y}\mid\boldsymbol{x},\hat{\boldsymbol{\theta}}):p(\boldsymbol{y}\mid\boldsymbol{x},\boldsymbol{\theta}^*)) \tag{14}$$

식(11)부터 식(13)까지의 유도 과정을 거쳐서 식 (14)를 얻을 수 있으며, 이때 $p(\boldsymbol{y}\mid\boldsymbol{x},\boldsymbol{\theta}^*)$는 신경회로망으로 구현할 수 있는 최적의 시스템이다. 식(14)에서 $d(p(\boldsymbol{y}\mid\boldsymbol{x},\boldsymbol{\theta}^*):q(\boldsymbol{y}\mid\boldsymbol{x}))$는 실제 시스템이 신경회로망에 의해 완벽하게 구현될 수 없기 때문에 발생하는 차이이며 이는 고정적인 값이다. 따라서 일반화 오차는 학습데이타에 의한 경험 분포 $p(\boldsymbol{y}\mid\boldsymbol{x},\hat{\boldsymbol{\theta}})$와 최적의 시스템 $p(\boldsymbol{y}\mid\boldsymbol{x},\boldsymbol{\theta}^*)$ 사이의 차이인 $d(p(\boldsymbol{y}\mid\boldsymbol{x},\hat{\boldsymbol{\theta}}):p(\boldsymbol{y}\mid\boldsymbol{x},\boldsymbol{\theta}^*)$로 정의된다. 즉 일반화 성능 향상이라고 하는 것은 $d(p(\boldsymbol{y}\mid\boldsymbol{x},\hat{\boldsymbol{\theta}}):p(\boldsymbol{y}\mid\boldsymbol{x},\boldsymbol{\theta}^*)$의 최소화를 의미한다.

자연 프루닝은 $d(p(\boldsymbol{y}\mid\boldsymbol{x},\hat{\boldsymbol{\theta}}):p(\boldsymbol{y}\mid\boldsymbol{x},\boldsymbol{\theta}^*)$를 척도로 하여 이를 최소화 할 수 있는 가중치 $\boldsymbol{\theta}^*$ 매개변수를 제거한다. 하지만 실제 최적의 가중치 $\hat{\boldsymbol{\theta}}$를 알 수 없기 때문에 경험적인 분포를 사용할 수 밖에 없다. 따라서 자연 프루닝에서는 큰 모델의 가중치 $\hat{\boldsymbol{\theta}}$가 최적의 가중치라 간주하고 프루닝을 수행한다. 그렇기 때문에 자연 프루닝에서는 큰 모델의 가중치 $\hat{\boldsymbol{\theta}}$가 얼마나 $\boldsymbol{\theta}^*$에 유사한가에 따라서 그 성능이 좌우된다. 정규화에 방법이 일반화 성능을 향상시킨다는 것은 이론적으로 밝혀져 있으며[16], 이는 정규화 되지 않은 가중치 매개변수에 비해 정규화된 가중치 매개변수가 최적의 가중치 매개변수와 유사함을 의미한다.

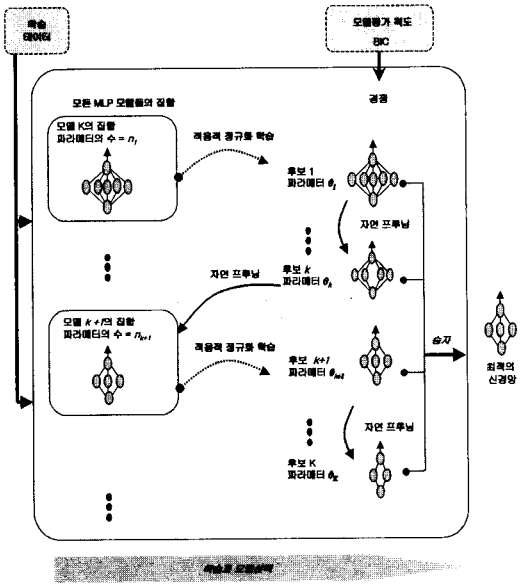



그림 1 제안하는 방법의 구성

본 논문에서는 적응적 정규화 방법을 도입하므로써 $p(y \mid x, \theta^*)$에 유사한 경험 분포를 추정 할 수 있는 최적화된 $\hat{\theta}$를 생성하고, 이들을 프루닝에 전달하여 후보 생성한다. 이러한 적응적 정규화 학습과 프루닝의 반복적인 프로세스를 통하여 후보 모델들을 생성하고, 이들 중 최적의 모델을 선택하는 방법을 제안한다. 제안하는 방법의 구성은 그림 1과 같다. 본 논문에서는 정규화 매개변수를 온라인으로 적절하게 조절할 수 있고, 다른 세부 방법들과 일관된 관점에서 설명할 수 있는 베이시안 정규화 방법을 자연 기울기 강하 학습에 도입하여, 가중치 매개변수를 최적화 시킨다. 그리고 자연프루닝을 통하여 일반화 성능이 우수한 단순한 후보 모델을 구성하고 베이시안 정보기준에 의해 최적의 모델을 선택하는 방법을 제안하였다. 전체적인 과정을 기술하면 그림 2와 같다.

1 단계: 큰 네트워크의 구조에서 시작한다.
2 단계: $E(x, y; \theta) + \alpha \|\theta\|^2$를 최소화 시키는 추정치 $\hat{\theta}$를 찾는다.
3 단계: 각각의 요소 $\hat{\theta}_i$에 대해서 중요도 $\delta F_i(\hat{\theta})$를 계산하고, 최소의 중요도 $\delta F_m(\hat{\theta})$를 갖는 $\hat{\theta}_m$을 찾는다.
4 단계: $\hat{\theta}_m$을 제거하고 남아 있는 요소들을 갱신한다.
5 단계: 학습오차와 베이시안 정보 기준이 계속적으로 증가하면6단계로 간다. 그렇지 않을 때는 거리의 변화가 임계치 보다 낮으면 3단계로 가고, 그렇지 않으면 2단계로 간다.
6 단계: 이렇게 형성된 단순화된 후보 모델중 일반화 성능이 우수한 최적의 모델은 베이시안 정보 기준에 의해서 선택한다.

그림 2 제안하는 방법의 과정

2 단계에서 정규화 매개변수 $\alpha$는 $\hat{\theta}$를 잘 추정 하기 위해서 매우 중요하며, 본 논문에서는 적응적인 방법을 사용한다. 자세한 사항은 3.2 절에서 다룬다. 3단계에서는 제거될 요소를 선택하기 위해서 중요도 계산이 필요하며, 4단계에서는 제거된 뒤 남아 있는 매개변수에 대한 갱신이 필요하다. 이에 대한 자세한 사항은 3.3 절에서 다룬다.

다음은 제안하는 방법을 구성하고 있는 정규화항이 있는 자연기울기 강하 학습, 정규화 매개변수 최적화, 자연프루닝 그리고 베이시안 정보 기준에 대해 좀더 자세히 살펴보도록 한다.

### 3.1 패널티 항이 있는 자연기울기 강하 학습

신경회로망은 N개의 학습쌍 $D = \{(x^{(u)}, t^{(u)})\}_{u=1}^{N}$에 의해서 학습된다. 적응적 정규화 방법을 도입한 방법의 오차함수는 식와 (15)같이 주어진다.

$$C(x, y, \theta) = E(x, y, \theta) + \alpha R(\theta) \quad (15)$$

여기서 $E(x, y, \theta)$는 네트워크 모델과 입력 데이타에 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 패널티 항이다. 이 항은 가중치 값이 커지는 것을 억제함으로써 매끄럽고 간단한 사상을 수행하도록 돕는다. 이렇게 추가되는 패널티 항으로는 가중치 감소항이 널리 쓰이며 이는 네트워크의 적응적 매개변수의 제곱합으로 정의 된다.

식 (15)와 같은 목적함수가 주어졌을 때 정보기하 이론에 기반한 자연기울기는 식 (16)과 같이 주어진다.

$$\tilde{\nabla} C(x, y, \theta_t) = G^{-1}(\theta_t) \nabla C(x, y, \theta_t)$$
$$= G^{-1}(\theta_t)(\nabla E(x, y, \theta_t) + \alpha \nabla R(\theta_t)) \quad (16)$$

이때 $G^{-1}$는 식 (6)과 같이 주어지는 피셔 정보 행렬 $G$에 대한 역행렬이다. 이러한 자연 기울기를 바탕으로 한 학습은 식(17)과 같다.

$$\theta_{t+1} = \theta_t - \eta_t \tilde{\nabla} C(x, y, \theta_t) = \theta_t - \eta_t G^{-1} \nabla C(x, y, \theta_t)$$
$$= \theta_t - \eta_t G^{-1}(\nabla E(x, y, \theta_t) + \alpha \nabla R(\theta_t)) \quad (17)$$

실제 구현에서 피셔 정보행렬 $G$의 역행렬 $G^{-1}$를 매번 구하는 것은 거의 불가능하기 때문에 근사 방법이 필요하며 이를 적응적 자연기울기라 한다(자세한 사항은 [10,12] 참조).

### 3.2 적응적 정규화 매개변수 최적화

우리가 학습을 통하여 궁극적으로 얻고자 하는 신경회로망은 학습데이타에 좋은 결과 만을 내는 신경회로망이 아니라 미지의 데이타에 대해 좋을 성능을 보이는 일반화 능력이 우수한 신경회로망을 얻는 것이다. 하지만 보통의 학습방법에는 일반화 성능을 향상시키는 메커니즘이 존재하지 않는다. 따라서 학습과정에 일반화 성능을 향상 시키는 메커니즘이 필요하며 본 논문에서는 이를 위해 정규화 항을 도입한다. 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. 이

러한 베이시안 최적화 방법은 학습집합과 검증 집합을 나눌 필요가 없이 모든 학습데이타를 학습에 사용하면서 일반화 오차를 추정할 수 있다는 장점을 지닌다.

식 (15)의 정규화 매개변수 (를 결정하기 위하여 베이시안 에비던스(evidence) 개념을 적용한다. 통계적인 관점으로 볼 때 표준 오차 함수는 조건부 확률 밀도 함수 $p(\boldsymbol{y} \mid \boldsymbol{x}; \boldsymbol{\theta})$의 음의 로그우도로 간주할 수 있다. 정규화의 경우 매개변수 $\boldsymbol{\theta}$의 확률밀도 함수도 고려해야 하며, 따라서 학습의 최종 목표는 사후 확률 밀도 함수 $p(\boldsymbol{y} \mid \boldsymbol{x}; \boldsymbol{\theta})p(\boldsymbol{\theta})$를 최대화 하도록 하는 것이다. 이에 상응하는 오차함수는 식(18)과 같이 나타낼 수 있다.

$$C(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}) = -\log p(\boldsymbol{y} \mid \boldsymbol{x}; \boldsymbol{\theta})p(\boldsymbol{\theta}) \quad (18)$$

$$= E(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}) - \log p(\boldsymbol{\theta}) \quad (19)$$

식 (19)와 (15)를 비교하면 사전 가정 $p(\boldsymbol{\theta})$와 정규화항 사이의 관계를 알 수 있다. 이를 위해 학습전의 매개변수가 식 (2)와 같은 가우시안 분포라고 가정한다.

$$p(\boldsymbol{\theta}) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left\{-\frac{1}{2}\|\boldsymbol{\theta}\|^2\right\} \quad (20)$$

여기서 $n$은 매개변수의 수이다. 여기서 상수에 관련된 항을 무시하면 식 과 같이 이의 관계를 쉽게 알 수 있다.

$$R(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|^2, \alpha = \frac{1}{\sigma^2} \quad (21)$$

정규화 항에 대한 이러한 통계적인 관점은 Mackay에 의해 소개 되었고 이를 적용하여 $\alpha$를 적응적으로 결정하는 방법을 제시하였다[1]. 주어진 학습 집합 $\boldsymbol{D} = (\boldsymbol{x}^{(n)}, \boldsymbol{t}^{(n)})_{n=1}^{N}$에 대해서 $\alpha$는 베이시안 에비던스를 최대화 하는 것에 의해 최적화되고 이는 다음의 식 (22)와 같이 정의될 수 있다.

$$p(\boldsymbol{\theta} \mid \boldsymbol{D}) \propto p(\boldsymbol{D} \mid \boldsymbol{\theta})p(\boldsymbol{\theta}) \quad (22)$$

$$= \frac{1}{p(\boldsymbol{D})} \int p(\boldsymbol{D} \mid \boldsymbol{\theta})p(\boldsymbol{\theta} \mid \alpha)p(\alpha)d\alpha \quad (23)$$

여기서 $p(\alpha)$는 하이퍼 매개변수 $\alpha$의 사전확률이다. 적절한 $\alpha$의 값을 알 수 없기 때문에, $\alpha$의 사전확률은 무정보적 분포라 가정한다. 척도 모수에 대한 무정보적 분포는 일반적으로 균등한 로그 스케일 $p(\log(\alpha)) = 1$로 표현된다[1]. 이러한 사전확률과 식 (23)을 통하여 다음의 식 (24)를 유도해 낼 수 있다.

$$-\ln p(\boldsymbol{\theta} \mid \boldsymbol{D}) = \sum_{i=1}^{N} E(\boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{\theta}) + \frac{n}{2}\log R(\boldsymbol{\theta}) + const \quad (24)$$

여기서 $n$은 매개변수의 수이다. 한편 식 (15)로 부터 식 (25)를 얻을 수 있다.

$$\sum_{i=1}^{N} C(\boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{\theta}) = \sum_{i=1}^{N} E(\boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{\theta}) + N\alpha R(\boldsymbol{\theta}) + const \quad (25)$$

식 (25)에 대한 편미분은 식 (26)과 같다.

$$\sum_{i=1}^{N} \nabla C(\boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{\theta}) = \sum_{i=1}^{N} \nabla E(\boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{\theta}) + N\alpha \nabla R(\boldsymbol{\theta}) \quad (26)$$

또한 식 (24)에 대한 편미분은 식 (27)과 같다.

$$-\nabla \ln p(\boldsymbol{\theta} \mid \boldsymbol{D}) = \sum_{i=1}^{N} \nabla E(\boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{\theta}) + \frac{n}{2R(\boldsymbol{\theta})} \nabla R(\boldsymbol{\theta}) \quad (27)$$

그러므로 $p(\boldsymbol{\theta} \mid \boldsymbol{D})$를 최대화 하는 것은 비용함수의 합 $\sum_{i=1}^{N} C(\boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{\theta})$를 최소화 하므로써 구현될 수 있으며, 이때 $\alpha$는 식 (28)에 의해 주기적으로 갱신된다.

$$\alpha = \frac{n}{2NR(\boldsymbol{\theta})} \quad (28)$$

이 방법은 베이시안 에비던스를 사용하여 정규화 매개변수 $\alpha$를 최적화 시키는 단순한 방법이다. 보다 정확한 공식이 있기는 하지만 본 논문에서는 실제적인 응용에 적용할 수 있는 계산적으로 효율적인 이러한 단순한 형태를 적용한다 (자세한 사항은[1]을 참조).

### 3.3 자연 프루닝

신경회로망의 구조가 너무 단순하면 학습데이타나 테스트 데이타에 대해 만족할 만한 결과를 얻을 수 없고, 너무 복잡하면 학습데이타 내의 노이즈에 적합되어서 학습데이타에 대해서는 좋은 결과를 보이나 테스트 데이타에 대해선 좋은 결과를 보이지 못하는 현상이 발생한다. 불행하게도 대부분의 실제 문제에 있어서는 최적의 구조를 알 수 없다. 구조 선택 알고리즘의 목표는 이러한 최적의 구조를 찾아 내는 것이다. 구조 선택방법의 하나인 프루닝 방법은 반복적인 절차로 완전 연결된 신경회로망으로부터 불필요한 매개변수들을 제거하면서 최적의 구조를 찾아 내는 방법이다.

프루닝이라고 하는 것은 학습에 의해 얻은 매개변수 $\hat{\boldsymbol{\theta}}$가 최적의 매개변수 $\boldsymbol{\theta}^*$의 요소 보다 많다는 가정하에서 출발한다. 프루닝에서는 $\hat{\boldsymbol{\theta}}$가 주어졌을 때 제거될 요소를 평가하기 위한 척도가 필요하다. 단순한 척도로는 각각의 요소의 크기 $|\hat{\theta}_i|$를 이용하는 것이다. 하지만 매개변수들이 네트워크의 행위와 무관하지 않기 때문에 이러한 단순한 척도만으로는 네트워크 구조를 최적으로 만들 수 없다. 이러한 문제를 해결하기 위해서, 학습오차의 변화에 기반한 OBD와 OBS의 중요도 (saliency) 계산 방법이 있다. 이는 학습 오차의 변화를 기반으로 하며, 2차 근사 방법으로 각 요소 $\hat{\theta}_i$에 대한 유사도 $\delta E_i(\hat{\boldsymbol{\theta}})$를 계산한다.

$$\delta E_i(\hat{\boldsymbol{\theta}}) = (\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}^i)' \boldsymbol{H}(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}^i) \quad (29)$$

여기서 $\boldsymbol{H}(\hat{\boldsymbol{\theta}})$는 학습오차의 헤시안(Hessian) 행렬이다. $\hat{\boldsymbol{\theta}}^i$는 $\hat{\theta}_i$를 영으로 만들어서 얻어진 매개변수 벡터이다. 이는 유사도가 단순한 척도를 이용한 방법에 비해서

는 보다 좋은 성능을 나타낸다. 하지만 유사도는 학습데이타의 오차에 기반하기 때문에 현재 추정된 $\hat{\theta}$에 편중되어서, 이 유사도가 언제나 일반화 성능에 효과적이라고는 할 수 없다. 이러한 문제를 해결하기 위해 Pedersen 등이 $\gamma$OBD와 $\gamma$OBS를 제안하였으며, 이는 중요도를 구할 때 학습 오차 대신 추정한 일반화 오차를 사용하는 방법이다[17]. 그러나 이 일반화 된 오차는 모델에 대하여 제약을 가함으로써 얻어진 것이며, 그 결과가 일반적인 모델에 적용하는데 성공적이라고 말하기는 어렵다.

반면, 자연 프루닝은 신경회로망 공간의 기하학적인 고려에 의해 발전되었다[14,15]. 자연 프루닝은 피셔 척도를 이용하여 $\hat{\theta}$와 $\theta^*$ 사이의 거리를 측정한다. $\hat{\theta}_i$의 유사도 $\delta F_i(\hat{\theta})$는 식 (30)과 같이 계산 될 수 있다.

$$\delta F_i(\hat{\theta}) = (\hat{\theta} - \hat{\theta}^i)' G(\hat{\theta})(\hat{\theta} - \hat{\theta}^i) \tag{30}$$

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

자연 프루닝은 여러면에서 OBD와 OBS보다 우수하다. 첫째 자연 프루닝의 중요도 $\delta F_i(\hat{\theta})$는 신경회로망의 확률밀도함수 $p(\boldsymbol{y} \mid \boldsymbol{x}; \theta)$의 차이에 기반하나, OBD와 OBS의 중요도 $\delta E_i(\hat{\theta})$는 학습오차의 변화에 의존한다. 중요도 $\delta F_i(\hat{\theta})$는 $\delta E_i(\hat{\theta})$에 비하여 노이즈에 덜 민감하다. 둘째, 피셔정보 행렬은 오차함수에 의존하지 않고 단지 네트워크 모델의 통계적인 특성에 의존한다. 따라서 다양한 오차함수에 적용할 수 있다. Heskes는 제곱오차와 헤시안 행렬의 가우시안 근사를 할 경우 $\delta E_i(\hat{\theta})$가 $\delta F_i(\hat{\theta})$와 같아 진다는 것을 보였다[14]. 프루닝 후 남아 있는 매개변수에 대한 갱신은 식 (31)에 의해서 수행된다.

$$\theta_i^{new} = \hat{\theta}_i - \frac{G^{mi}(\hat{\theta})}{G^{mm} - (\hat{\theta})} \hat{\theta}_m \tag{31}$$

이는 $\hat{\theta}_m$의 제거에 의한 거리의 증가를 최소화 하도록 하는 것이다. 여기서는 $G^{mi}$ 피셔 정보 행렬의 $m$번째 행 $i$번째 열의 요소이다.

### 3.4 베이시안 정보 기준(Bayesian Information Crite

신경회로망의 과다학습문제를 해결하기 위하여 두 가지 형태의 모델선택 방법이 자주 쓰인다. 하나는 크로스 검증 방법처럼 데이타를 나누어서 수행하는 아웃 오브 샘플(out-of-sample) 모델선택 방법이다. 다른 하나는 Akaike의 정보기준, 베이시안 정보 기준과 같이 데이타를 분할하지 않는 인 샘플(in-sample) 모델선택이다. 본 논문에서는 최적의 모델선택을 위하여 자연 프루닝에 의해 생성된 모델간의 패널티가 가해진 성능 측정도구로 베이시안 정보 기준을 적용한다. 이는 식와 같으며 비선형 모델의 경우 $d > 1$ 이고 $d$는 실험에 의하여 결정한다[18].

$$BIC = \log(E(\boldsymbol{x}, \boldsymbol{y}, \theta)) + \frac{W^d \log(N)}{N} \tag{32}$$

여기서 $N$은 데이타의 수이고 $W$는 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 사용한다.

## 4. 실험 결과 및 분석

제안하는 방법은 크게 회귀문제와 분류문제에 대한 벤치마크데이타[19]에 적용하였으며 이는 표 1과 같다.

표 1 실험에 사용된 데이타

| | 데이터 이름 | 출처 | 항목수 | 출력수 | 총 데이터수 (학습/테스트) |
|---|---|---|---|---|---|
| 회귀문제 | Boston Housing | UCI | 13 | 1 | 506(256/250) |
| | Building | UCI | 14 | 3 | 4208(3156/1052) |
| 분류문제 | Diabetes | UCI | 8 | 1 | 768(576/192) |
| | Glass | UCI | 9 | 6 | 214(161/53) |
| | Horse | UCI | 58 | 3 | 364(272/91) |

제안하는 NGARP 방법은 자연기울기 학습에 적응적 정규화와 자연 프루닝을 적용시킨 방법이다. SSEARP는 제곱합 오차 함수에 적응적 정규화와 프루닝을 적용시킨 방법이다. NGNRP방법은 자연기울기 학습에 정규화없이 프루닝을 적용한 방법이다. 실험 결과는 가중치 초기화를 다르게 하여 10번 실험한 평균을 비교 하였다. 회귀문제의 경우 제안하는 NGARP방법과 SSEARP방법은 동일하므로, 회귀 문제에 대해서는 적응적 정규화에 의한 매개변수 최적화의 유무의 차이가 있는 NGARP 방법과 NGNRP의 일반화 성능과 구조 최적화의 성능을 비교하였다. 분류 문제는 2개의 클래스의 문제와 여러개의 클래스 문제의 데이타에 대하여 적용하였다. 제안하는 방법은 분류 문제에 있어서 오차 제곱 합 보다 좋은 성능을 보이는 크로스엔트로피 오차함수를 적용할 수 있기 때문에, 분류 문제에 대해서 좋은 성능을 기대할 수 있다. 따라서 이를 검증하기 위하여 제안하는 NGARP방법과 오차함수의 차이가 있는 정규화 방법을 적용한 SSEARP 방법과 비교한다. 또한 오차함수는 같으나 거기에 적응적 정규화에 의한 차이를 보이기 위하여 NGNRP 방법과 비교를 하였다. SSEARP와 제안하는 NGARP방법은 베이시안 정보 기준으로 최적의 모델을 선택하였으나 NGNRP는 베이시안 정규화에 기반하는 방법이 아니기 때문에 베이시안 정보 기

준으로 모델을 선택하기가 힘들다. 따라서 NGNRP는 성능 비교를 위해 테스트 분류 오차가 최소인 모델 중 연결선 수가 최소인 모델을 선택하여 제안하는 방법과 비교하였다.

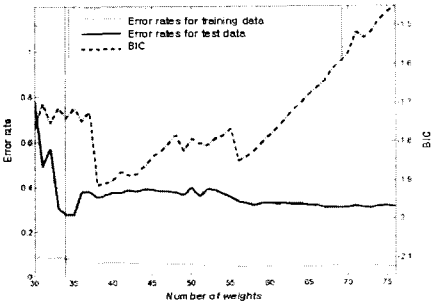
### 4.1 회귀(Regression)

Boston Housing 문제는 통계학 논문에서 벤치마크 데이타로 광범위하게 사용되고 있다. 전체 506개의 데이타로 구성되어 있고, 학습 집합으로는 256개를 사용하였다. 신경회로망은 14개의 입력 노드, 5개의 은닉 노드, 1개의 출력 노드로 구성 되어 있으며 바이어스를 포함하여 실험에 사용된 초기 모델의 가중치 수는 총 76개였다. 학습률은 0.01로 하여 실험하였다. 표 2는 NGNRP 방법과 NGARP 방법을 가중치 수 및 오차율을 이용해서 비교한 것이다. 가중치의 수를 보면 NGARP 방법이 33.6개로 NGNRP 보다 평균 3개 이상 더 구조 최적화를 하였다. 또한 NGARP의 표준 편차도 0.699로 NGNRP의 1.702 보다 더 안정적이었다. 학습 데이타 오차와 테스트 데이타 오차에서도 NGARP 방법이 NGNRP 방법보다 더 우수한 성능을 보였다.

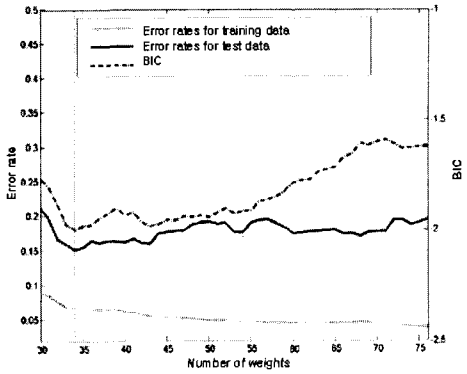
표 2 회귀문제에 대한 가중치의 수 및 오차

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

그림 3은 Boston Housing 문제에 대한 NGNRP와 NGARP의 학습 결과의 예를 보인 것이다. (a)의 NGNRP의 경우 테스트 데이타에 대해서 0.2886의 오차율을 보이면서 33개까지 구조를 최적화 시킨 그래프이고, (b)는 테스트 데이타에 대하여 0.149의 테스트 데이타 오차율을 보이면서 33개까지 구조 최적화를 수행한 그래프이다.




(a) NGNRP 방법




(b) NGARP 방법

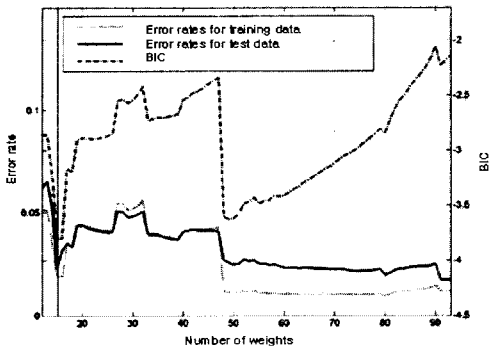
그림 3 Boston Hosing문제에 대한 학습 및 테스트 분류 오차와 일반화 오차

Building 문제는 건물에서 에너지 소비량을 예측하기 위한 데이타이다. 날짜, 시간, 외부 온도, 외부 습도, 태양 열, 풍속에 기반하여 전기에너지, 온수, 냉수의 시간당 소비량을 예측하는 것이다. 전체 4208개의 데이타로 구성되어 있고, 학습 집합으로 3156개를 사용하였다. 신경회로망은 14개의 입력 노드, 5개의 은닉 노드, 3개의 출력 노드로 구성 되어 있으며 바이어스를 포함하여 실험에 사용된 초기 모델의 가중치 수는 총 93개였다. 학습률은 0.003으로 하였다. 표2는 NGNRP 방법과 NGARP 방법을 가중치 수 및 오차율을 이용해서 비교한 것이다. 일반화 성능을 확인할 수 있는 테스트 데이타 오차에서는 NGARP 방법이 NGNRP 방법보다 우수성을 보였다.

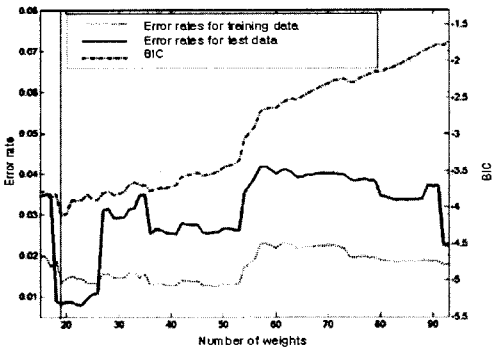
그림 4는 Building 문제에 대한 NGNRP와 NGARP의 학습 결과의 예를 보인 것이다. (a)의 NGNRP의 경우 테스트 데이타에 대해서 0.0213의 오차율을 보이면서 17개까지 구조를 최적화 시킨 그래프이고, (b)는 테스트 데이타에 대하여 0.0082의 테스트 데이타 오차율을 보이면서 19개까지 구조 최적화를 수행한 그래프이다. 전체 평균과 마찬가지로 NGARP 방법이 NGNRP 보다 우수한 일반화 성능을 보였다.

제안하는 NGARP방법은 학습과정에 베이시안 적응적 정규화를 도입하여 일반화가 우수한 모델을 프루닝에 전달한다. 하지만 NGNRP는 학습과정에서 과소적합과 과다적합에 의한 통제를 할 수 없기 때문에, 현재의 모델이 최적의 모델이라는 보장을 할 수가 없다. 따라서 표 2에서와 같이 현재의 모델이 최적의 모델이라는 가정하에서 모델을 축소 시키는 프루닝 방법을 적용하는 데 있어서, 성능의 차이가 발생하였다. 그림 3과 4에서

제안하는 NGARP 방법은 베이시안 적응적 정규화에 의해서 보다 최적화된 가중치를 구조 최적화 과정에 제공하기 때문에, 정규화가 없는 NGNRP 방법의 그래프 (a)에 비해 제안하는 방법의 그래프 (b)가 프루닝에 의해 발생하는 시스템의 불안정성이 감소한 것을 확인할 수 있었다.




(a) NGNRP 방법




(b) NGARP 방법

그림 4 Building문제에 대한 학습 및 테스트 분류 오차와 일반화 오차

### 4. 2 분류(Classification)

Diabetes 문제는 성별, 나이, 임신 회수와 같은 개인 정보에 혈압, 신체 질량 지수(Body Mass Index), 당부하 검사(Glucose Tolerance Test) 등의 의료 검사정보를 이용해서 인디언 개인이 당뇨병에 걸려 있는지 아닌지를 분류 하는 문제이다.

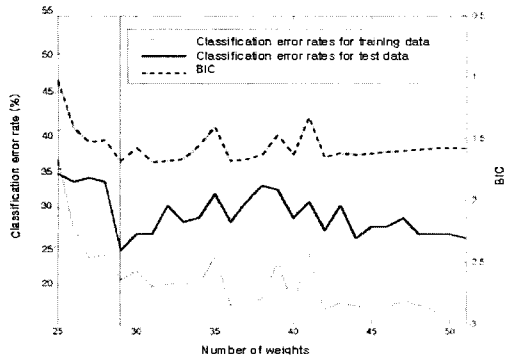
신경회로망은 8개의 입력 노드, 5개의 은닉 노드, 1개의 출력 노드로 구성 되어 있으며 바이어스를 포함하여 실험에 사용된 초기 모델의 가중치 수는 총 57개였다. 학습률은 0.015로 하여 실험하였다. 표3의 결과에서 가중치의 수를 보면 NGARP 방법이 26.8개로 가장 최적화 된 구조를 보였다. 또한 NGARP의 표준 편차는

표 3 분류 문제에 대한 가중치의 수 및 테스트 분류율

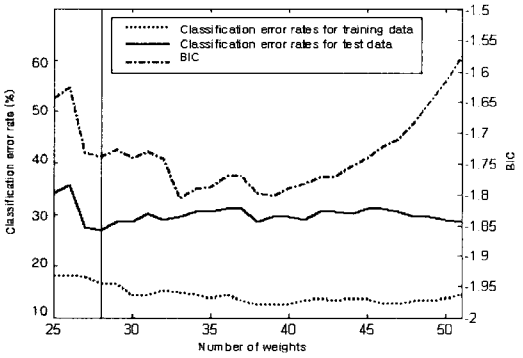
| | | SSEARP | NGNRP | NGARP |
|---|---|---|---|---|
| Diabetes | 가중치의 수 (개) | 28.4(3.57) | 30(4.16) | 26.8(1.23) |
| | 학습 데이터 분류율(%) | 81.86(2.178) | 83.85(2.639) | 80.75(0.569) |
| | 테스트 데이터 분류율(%) | 77.40(0.858) | 76.25(0.703) | 77.60(0.491) |
| Glass | 가중치의 수 (개) | 49.8(1.87) | 62.5(2.27) | 47.8(1.22) |
| | 학습 데이터 분류율(%) | 72.07(3.05) | 78.88(2.12) | 77.1(2.93) |
| | 테스트 데이타 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 데이타 분류율(%) | 73.74(0.959) | 68.90(1.275) | 75.17(1.182) |

1.23으로 SSEARP의 3.57이나, NGNRP의 4.16 보다 훨씬 안정된 모습을 보인다. 학습 데이타 분류율에서는 NGNRP 방법이 83.85%로 NGARP 방법의 80.75%보다 3% 이상의 우수한 성능을 보였다. 하지만, 테스트 데이타 분류율에서는 NGARP 방법이 77.6%로 SSEARP 방법이나 NGNRP 방법보다 더 우수한 성능을 보였다. NGARP 방법의 표준오차는 학습 데이타 분류율이나 테스트 데이타 분류율 모두 0.5 내외로 안정성을 보였다. 표3의 결과에서 정보 기하 이론에 기반한 제안하는 방법은 분류문제에 적합한 크로스엔트로피 오차함수를 적용할 수 있기 때문에, SSEARP 방법에 비하여 일반화 성능이 우수하였다. 또한 같은 정보 기하 이론에 기반한 방법 하에서 적응적 정규화를 적용한 제안하는 방법은 정규화를 적용하지 않은 NGNRP방법보다 최적의 모델을 프루닝에 전달 할 수 있기 때문에 구조최적화와 일반화 성능면에서 우수성을 보였다. 그림 5는 Diabetes 문제에 대한 SSEARP, NGNRP와 NGARP의 학습 결과의 예를 보인 것이다. (a)의 SSEARP의 경우 테스트 데이타에 대해서 76.56%의 분류율을 보이면서 29개 까지 구조를 최적화 시킨 그래프이고, (b)의 NGNRP의 경우 테스트 데이타에 대해서 75.52%의 분류율을 보이면서 28개까지 구조를 최적화 시킨 그래프이고, (c)는 테스트 데이타에 대하여 78.13%의 테스트 데이타 분류율을 보이면서 28개까지 구조를 최적화 시킨 그래프이다. 전체 평균과 마찬가지로 NGARP 방법과 NGNRP의 구조최적화 성능은 유사하지만, 일반화 성능은 NGARP 방법이 우수하였다.

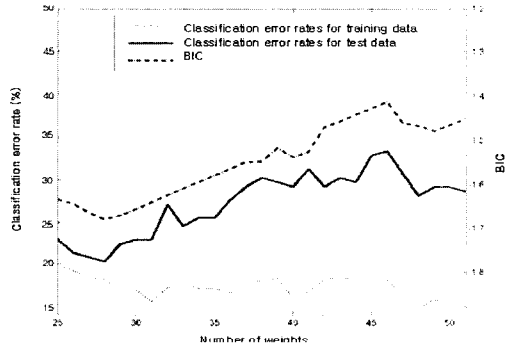
Glass 문제는 UCI 데이타 집합 중의 하나이다. 유리의 파편에 대한 화학적인 분석과, 굴절률을 이용해서 유리를 건물 유리, 차 유리, 램프 등의 6개로 분류하는 문




(a) SSEARP 방법




(b) NGNRP 방법



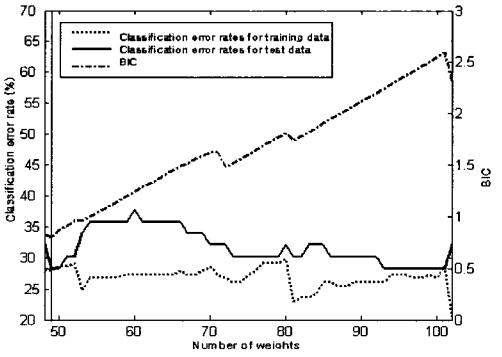

(c) NGARP 방법

그림 5 Diabetes문제에 대한 학습 및 테스트 분류 오차와 일반화 오차

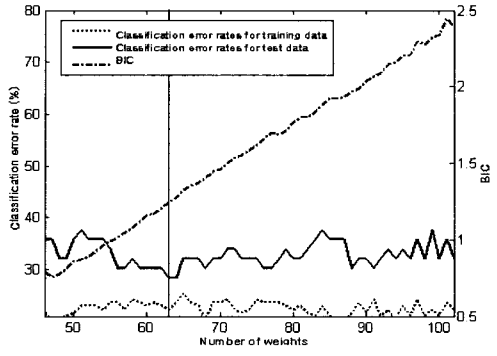
제이다. 실험에 사용된 네트워크의 구조는 입력 노드는 9개, 은닉 노드는 6개, 출력 노드는 6개이고 초기 네트워크의 매개변수의 수는 바이어스를 포함하여 총 102개이다. 가중치 초기화를 10번 다르게 하여 실험하였고 학습률은 0.03으로 하였다.

표 3의 결과에서 가중치의 수를 보면 NGARP 방법이 47.8개로 가장 최적화 된 구조를 보였다. 또한 NGARP의 표준 편차는 1.22로 SSEARP의 1.87이나, NGNRP의 2.27 보다 안정성을 보였다. 학습 데이타 분류율은 NGNRP 방법이 78.88%로 NGARP 방법의 77.1%보다 우수한 성능을 보였다. 하지만, 테스트 데이타 분류율에서는 NGARP 방법이 73.77%로 SSEARP 방법이나 NGNRP 방법보다 더 우수한 성능을 보였다. 이를 통해 NGARP 방법의 일반화 성능이 우수하다는 것을 확인할 수 있다.

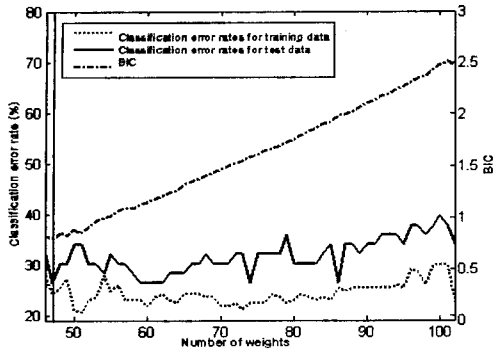
그림 6은 Glass 데이타에 대한 SSEARP, NGNRP와 NGARP의 학습 결과의 예를 보인 것이다. (a)의 SSEARP의 경우 71.7%의 테스트 데이타에 대한 분류율을 보이면서 49개의 가중치를 가진 모델을 구성하였으며, (b)의 NGNRP는 68.05%의 테스트 데이타에 대한 분류율을 보이는 가중치의 수가 63개인 모델을 구성하였고, (c)의 NGARP는 73.59%의 데스트 데이타 분류율 보이면서 47개의 가중치 수를 갖는 모델을 구성하였다. 전체 평균과 마찬가지로 NGARP 방법이 SSEARP와 NGNRP 보다 구조최적화 성능과 일반화 성능이 우수하였다.




(a) SSEARP 방법




(b) NGNRP 방법



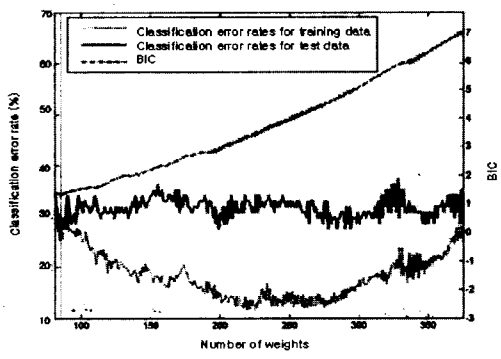

(c) NGARP 방법

그림 6 Glass문제에 대한 학습 및 테스트 분류 오차와 일반화 오차

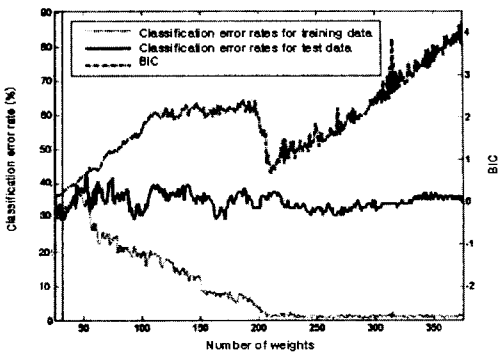
Horse 문제는 산통이 있는 말의 운명을 예측하는 문제이다. 산통이 있는 말의 수의학적인 진단으로 그 말이 살아날 것인지, 죽을 것인지, 안락사 시킬 것인지를 판별해야 하는 문제이다. 전체 데이타는 364개의 데이타로 이루어져 있고, 272개를 임의로 선택하여 학습에 사용하였다. 신경회로망은 58개의 입력 노드, 5개의 은닉 노드, 3개의 출력 노드로 구성 되어 있으며 바이어스를 포함하여 실험에 사용된 초기 모델의 가중치 수는 총 313개였다. 학습률은 0.07로 하여 실험하였다.

표3의 결과에서 가중치의 수를 보면 NGARP 방법은 44.4개로 다른 두 방법에 비하여 보다 최적화 된 구조를 보였다. 학습 데이타 분류율에서는 NGARP 방법이 76.67%로 NGNRP 방법의 74.66%보다 우수한 성능을 보였다. 또한, 테스트 데이타 분류율에서도 NGARP 방법이 75.17%로 SSEARP 방법의 73.74%보다 우수한 성능을 보였다.

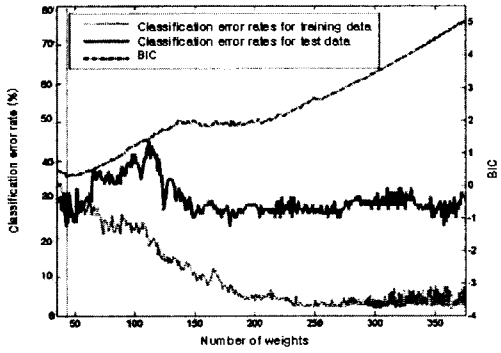
그림 7은 Horse 문제에 대한 SSEARP, NGNRP와 NGARP의 학습 결과의 예를 보인 것이다. (a)의 SSEARP




(a) SSEARP 방법




(b) NGNRP 방법




(c) NGARP 방법

그림 7 Horse문제에 대한 학습 및 테스트 분류 오차와 일반화 오차

의 경우 72.53%의 테스트 데이타에 대한 분류율을 보이면서 58개의 가중치를 가진 모델을 구성하였으며, (b)의 NGNRP는 70.33%의 테스트 데이타에 대한 분류율을 보이는 가중치의 수가 55개인 모델을 구성하였고, (c)의 NGARP는 74.36%의 테스트 데이타 분류율 보이면서 43개의 가중치 수를 갖는 모델을 구성하였다. 전체 평균과 마찬가지로 NGARP 방법은 SSEARP와 NGNRP에 비해 구조최적화와 일반화 성능에서 우수성을 보였다.

## 5. 결 론

본 논문에서는 신경회로망의 일반화 성능 향상 문제를 통계적인 관점에서 살펴보고, 이를 바탕으로 학습과 모델선택의 통합 프로세스를 통하여 신경회로망의 일반화 성능을 향상 시키는 설계 및 모델 선택 방법을 제안하였다. 벤치마크 데이타에 대한 실험을 통하여 제안하는 방법의 우수성을 보였다. 프루닝에 의한 구조 최적화 방법과 제안하는 학습과 모델선택의 통합 프로세스의 비교를 통하여 제안하는 학습

및 모델선택 통합 프로세스가 우수하다는 것을 보였다. 또한 자연기울기 학습법은 기존의 헤시안 행렬을 이용한2차 근사 학습법이 오차 제곱합 함수만을 사용할 수 있는 것과 달리, 평균제곱오차뿐만 아니라 패턴분류에 더 좋은 성능을 보이는 크로스 엔트로피 오차 함수도 사용할 수 있다. 따라서 실험을 통하여 오차 제곱합 함수를 사용하는 방법과 크로스 엔트로피 오차함수까지 사용 할 수 있는 제안하는 방법과의 비교를 통하여 제안하는 방법의 우수성을 보였다.

향후 연구과제는 다음과 같다. 계산 시간의 단축을 위하여 프루닝 단계에서 다중 연결선 제거 방법 및 매 개변수 갱신 방법에 대한 연구가 필요하겠다. 또한 제안하는 방법은 기호적 인공지능 방법에 비해 신경회로망 방법이 지니고 있는 일반화 성능 향상을 극대화함과 동시에 구조 최적화를 수행하고 있다. 신경회로망의 단점으로 알려진 설명력 부재의 단점을 극복하기 위한 규칙 추출 연구에 제안하는 방법을 도입함으로써 장점인 일반화 성능은 극대화 하면서, 설명력 부여에 유용하게 적용 될 수 있을 것이다.

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

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

[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 Property 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.

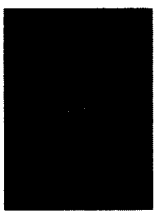
[15] Laar, P. V. D., Heskes, T., " Pruning Using Parameter and Neuronal Metrics," *Neural Computation*, 11, 977-993, 1999.

[16] Krogh, A., Hertz, J. A., "A Simple Weight Decay Can Improve Generalization," *Advances in Neural Information Processing Systems*, 4, 950-957, 1992.

[17] Pedersen, M. W., Hansen, L. K., Larsen, J., " Pruning with generalization based weight saliencies: (OBD, ( OBS," *Advances in Neural Information Processing Systems*, 8, 521-527, 1996.

[18] 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

[19] 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 컴퓨터과학과(석사).
2002년 연세대학교 컴퓨터과학·산업시스템공학과(박사). 현재 한국 싸이버 대학교 컴퓨터정보통신학부 전임강사. 관심분야는 패턴인식, 신경회로망, 영상처리, 데이타마이닝, 사이버교육

박 혜 영
1994년 연세대학교 전산과학과 (학사).
1996년 연세대학교 컴퓨터과학과(석사).
2000년 연세대학교 컴퓨터과학·산업시스템공학과(박사). 현재 일본 이화학 연구소 뇌과학연구센터 뇌수리연구팀 연구원. 관심분야는 계산학습이론, 통계적 정보처리이론, 신경회로망, 패턴인식

이 일 병
1976년 연세대학교 전자공학과(학사).
1980년 Illinois대학교 컴퓨터과학(석사).
1985년 Massachusetts대학교 전산정보과학(박사). 현재 연세대학교 컴퓨터과학·산업시스템공학과 교수. 관심분야는 생체인식, 데이타마이닝, 신경회로망, 패턴인식