# 신경망 앙상블의 편기와 분산을 이용한 분류 패턴 선택

신현정° 조성준
서울대학교 산업공학과
{hjshin72, zoon}@snu.ac.kr

# Pattern Selection for Classification Using the Bias and Variance of Ensemble Network

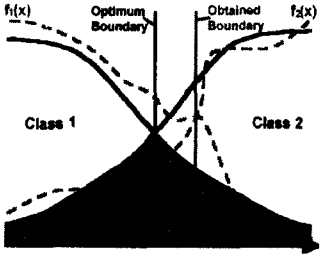
Hyunjung Shin° Sungzoon Cho
Dept. of Industrial Engineering, Seoul National University

## 요 약

분류문제에서 유용한 학습패턴은 클래스들간의 분류경계에 근접한 정상패턴들을 말한다. 본 연구에서는 다양한 구조와 학습 파라미터를 가진 신경망 앙상블을 구성하고 그 출력값의 편기와 분산에 기초한 패턴점수를 정의한다. 전체 학습패턴 중 일정한 임계값 이상의 패턴점수를 가진 패턴들만이 학습패턴으로 선정된다. 제안한 방법은 두 개의 인공문제와 두 개의 실제문제 (UCI Repository)에 적용, 검증되었다. 그 결과 선택된 패턴만으로 학습한 경우, 메모리 공간 절약 및 계산시간 단축의 효과 뿐만 아니라 복잡도가 큰 모델이라도 과적합을 하지 않았고 실험적으로 안정된 결과를 산출했으며, 적은 수의 학습패턴만으로도 일반화 성능을 향상시키거나 적어도 저하시키지 않았다는 것을 보였다.

## 1. 서 론

분류문제에 있어서 여러 분류기(classifier)간의 일반화 성능의 차이는 분류경계(decision boundary) 부근에서 발생한다[1]. 예를 들어 (그림1)과 같이 최적 분류경계가 B*인 2분류 문제인 경우, 두 개의 출력노드 $f_1(x)$, $f_2(x)$를 가진 신경망 분류기에 의해 얻어진 분류경계 B는 그림의 짙은 부분 중 한 곳에 위치하게 된다.




<그림1> 분류경계와 오분류 발생위치

|B*-B|의 값이 얼마나 되느냐는 분류기의 일반화 성능이 어떠한가를 의미한다. 이는 신경망 분류기의 경우에는 초기 네트워크 가중치, 네트워크의 구조, 학습알고리즘, 학습데이터 셋 등에 의해 달라진다[2]. 서로 다른 분류기들의 분류경계{$B_j$}들을 선형, 비선형적으로 결합하면 총론(consensus)적 분류경계 $B_{con}$을 얻을 수 있다. $B_{con}$은 B*에 보다 근사하게 됨으로써 개선된 분류 성능을 산출한다. 이러한 접근방법이 앙상블 분류 알고리즘의 기본 아이디어이다[2,3]. 동일 사실을 주어진 패턴 측면에서 생각해 보면 앙상블 구성 네트워크들의 결과값 분산이 큰 패턴들은 분류경계 부근에 위치한다는 것을 알 수 있다. 이들 중 클래스 레이블이 정상적인 패턴들이 분류문제에 유용한 패턴들이다. 본 연구에서는 앙상블 네트워크 결과값들의 분산(variance)과 편기(bias)를 이용하여 "혼잡도"와 "정확도"를 정의하고 이를 분류 패턴 선정에 이용하는 방법을 소개한다. 즉, 하나의 패턴에 대하여 앙상블 결과값들의 분산이 크면 혼잡도가 커지게 되고 이를 통하여 해당 패턴이 분류경계에 근접해 있음을 알 수 있다. 이들 중 앙상블 결과값의 편기가 작은 패턴들은 큰 정확도값을 갖게 되므로 정상패턴 여부를 판단할 수 있게 된다. 패턴 선정으로 인해 일반적으로 거둘 수 있는 일차적 성과는 학습 패턴 수 감소로 인한 계산 시간 단축 및 메모리 공간 절약의 효과이다[4,5]. 본 연구에서 제안하는 패턴 선정방법은 분류 경계부근의 정상패턴들만을 정련하여 선정하는 방법이므로 분류기로 쓰일 모델의 종류나 복잡도에 상관없이 과적합(overfitting)에 대한 대응 필요성이 없게 되는 부가적 효과가 있다. 예측 모델에 무관한 패턴 선택방법으로 [6]에서는 배이지안 네트워크 출력값들의 분포를 이용하여 예측(regression)문제에 대한 패턴 선택방법을 제안하였다. 베이지안 네트워크의 에러 바(error bar) 값이 큰 패턴은 예측문제에 있어서는 이상패턴일 가능성이 높으므로 이들을 제거함으로써 효율적 예측패턴만을 선택하는 방법이다[6,7]. MLP 신경망 분류기에 한정된 분류 패턴 선택방법으로서 최근 제안된 연구로는 nearest neighbor를 이용하는 방법과 cross-validation을 이용하는 방법이 있다[4,5]. 본 논문의 2절에서는 각 패턴의 패턴점수를 산출하는 방법과 이를 이용하여 유용한 패턴들을 분리하는 방법을 소개한다. 3절에서는 두 개의 인공문제와 두 개의 실제문제에 대한 실험 방법 및 결과에 대한 검증을 기술하였다. 본 연구의 실험에서는 분류기로서 single MLP와 bagging MLP를 사용하였다.

## 2. 패턴 선택 방법

### 2.1 패턴 점수 (Pattern Scoring)

각 패턴에 대하여 패턴점수를 산출하는 방법은 다음과 같다.

① $J$클래스 분류문제에 대하여 $L$개의 1-of-J MLP 네트워크를 전체학습 패턴 $M$에 대하여 학습시킨다. 각 네트워크들은 네트워크 구조 및 학습 파라미터에 교란(perturbation)을 주어 구성한다. 본 연구에서는 은닉층 및 은닉노드의 수, 활성화 함수, 학습 횟수 등의 설정을 랜덤화하였다.

② 패턴 $x_i$에 대하여 $l$번째 네트워크의 $j$개의 출력노드 값들로부터 네트워크 결과값 $F_l(x_i)$을 산출한다.

$$F_l(x_i) = \arg\max_j f_j(x_i),\quad j\in J,\ l=1...L,\ i=1...M. \qquad ...식(1)$$

③ 클래스 $j$별로 $L$개 네트워크의 다수 투표(majority voting) 결과확률을 계산한다.

$$P_j(x_i) \approx \frac{\sum_{l=1}^{L} 1(\text{if } F_l(x_i)=j)}{L},\quad j\in J. \qquad ...식(2)$$

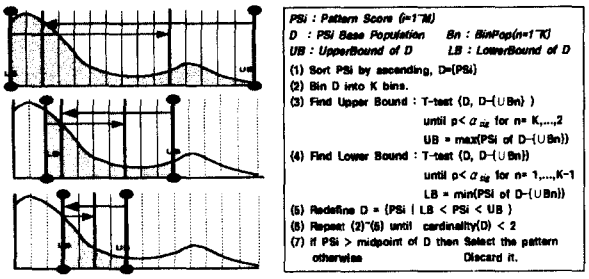
④ 각 패턴의 패턴점수(pattern score) $PS(x_i)$를 계산한다.

$$PS(x_i) = \alpha \times P_{j^*}(x_i) + (1-\alpha)\times \sum_{j\in J} P_j(x_i)\log_{|J|}\frac{1}{P_j(x_i)} \qquad ...식(3)$$

식(3)의 첫 번째 항 $P_{j^*}(x_i)$는 정확도로서, 패턴 $x_i$의 정답 클래스 $j^*$에 대한 네트워크들의 투표율을 나타낸다. 앙상블 결과값의 편기는 $\{1-P_{j^*}(x_i)\}$이므로 편기가 작을수록 정확도는 증가한다. 두 번째 항은 혼잡도로 정의되며 네트워크 투표결과의 분산정도를 나타낸다. 혼잡도는 투표 결과값의 정답여부와는 무관하며 단지 해당 패턴에 대한 네트워크 의견의 불일치성을 나타낸다. 따라서 분류경계 인접패턴들은(정상패턴인지 이상패턴인지에 상관없이) 식(3)의 혼잡도에 의하여 우선적으로 높은 패턴점수를 갖게되고, 이 들 중 정상패턴들은 정확도에 의하여 이상패턴들보다 더 높은 패턴점수를 갖게된다.

### 2.2 패턴 분리 (Pattern Separation)

$M$개의 패턴에 대하여 $PS(x_i)(i=1,...,M)$가 계산되면 이를 오름차순으로 정렬한 후, 패턴점수의 분포가 확연히 달라지는 임계값을 찾아 패턴을 분리한다. 임계값을 찾는 방법은 다음과 같다. 우선 전체 패턴점수의 분포를 기본(base)분포로 하고 이를 $K$개의 등간격구간(equal spaced bin)으로 분할한다. 분할이 끝나면 기본분포와 상,하한 1구간씩을 제외한 분포와의 평균 비교를 반복적으로 시행한다. 이는 두 분포평균에 대한 T-test결과가 유의하다는 판정이 나올 때까지 반복되는데, 유의판정이 나면 해당 구간이 다음 회 기본분포의 새로운 상,하한으로 설정된다. (그림2)에서는 이 과정을 묘사하였다.

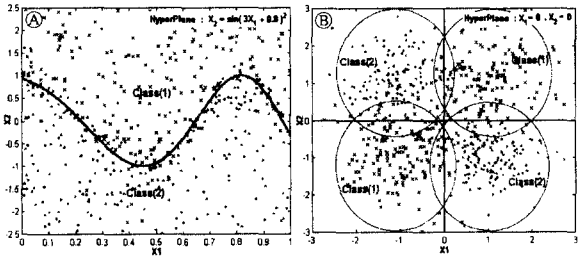



<그림2> 유의한 패턴분리 임계값을 찾는 과정

## 3. 실험 방법 및 결과

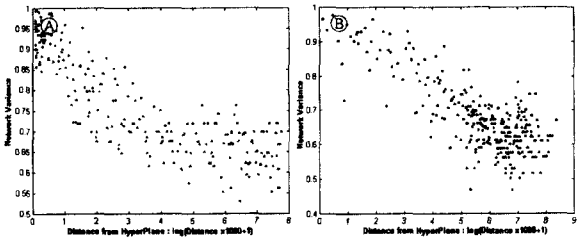
### 3.1 인공 분류 문제에 대한 실험

제안된 알고리즘은 우선 두 개의 인공문제 Ⓐ, Ⓑ에 적용되었다. Ⓐ의 경우, 각각 일양분포 U[0,1]를 따르는 $(x_1,x_2)$에 대하여 class1 = $\{(x_1,x_2)|x_2 > \sin(3x_1+0.8)^2\}$, class2 = $\{(x_1,x_2)|x_2 \le \sin(3x_1+0.8)^2\}$가 정의되었다. 전체 500개의 패턴이 생성되었으며 이 중 19%의 패턴들은 클래스 레이블이 바뀐 이상패턴들이다. Ⓑ에서는, 4개의 gaussian 분포로부터 class1= $\{(x_1,x_2)|N(C,0.5^2I),\ C=(1,1)\text{or}(-1,-1)\}$과 class2 = $\{(x_1,x_2)|N(C,0.5^2I),\ C=(-1,1)\text{or}(1,-1)\}$가 정의되었다. Ⓑ문제의 경우, 각 분포로부터 총 600개의 패턴이 생성되었다. 편의상 Ⓑ문제의 분류경계를 $x_1=0, x_2=0$으로 본다면 생성된 패턴들 중 10%가 이상패턴들이다. 두 문제의 차이점은, Ⓐ는 분류경계가 명확하고 이 부분에 패턴들이 밀집해 있는 반면, Ⓑ는 분류경계가 모호하고 경계부분에 가까울수록 패턴 수가 희박해 진다는 데에 있다(그림3).



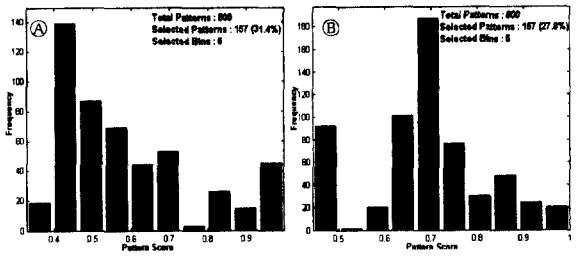

<그림3> 인공문제 Ⓐ와 Ⓑ

다음의 (그림4)는 네트워크 구조 및 학습 파라미터 랜덤화에 의해 30개의 네트워크로 구성된 앙상블 네트워크를 학습시킨 후, 각 패턴에 대한 네트워크 결과값의 분산 즉, 혼잡도를 분류경계로부터의 거리를 기준으로 나타낸 것이다. 두 문제의 경우 공통적으로 분류경계에 가까울수록 혼잡도가 커짐을 알 수 있다.



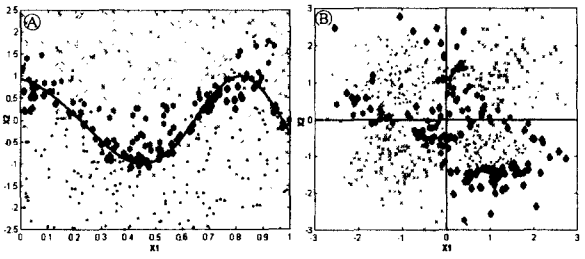

<그림4> 클래스 경계로부터의 거리(x)와 혼잡도(y)와의 관계

패턴분리 방법에 의하여 패턴점수가 높은 일부 패턴들이 선택되었다. Ⓐ의 경우에는 31.4% 인 157개가 Ⓑ의 경우에는 27.8%인 167개의 패턴이 선택되었으며 (그림5)의 짙은 부분이 이에 해당된다.



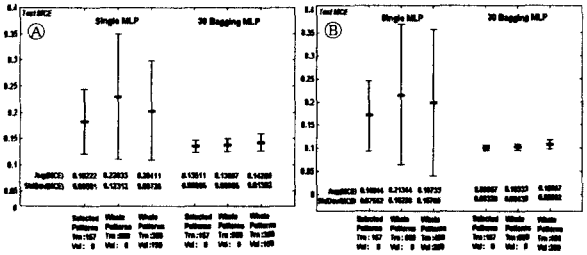

<그림5> 패턴분리 방법에 의해 선택된 패턴들의 분포

(그림6)에서는 혼잡도에 의하여 분류경계 부분에 위치한 패턴들이 선택되었으며 정확도에 의하여 이들 중 이상패턴들이 제거되었음을 보여준다.



<그림6> 패턴점수에 의하여 선택된 패턴들

패턴 선정 전,후의 효과를 검증하기 위하여 2-5-2 single-MLP와 30 2-5-2 bagging-MLP를 분류기로 하여 실험결과를 비교하였다. 두 분류기 모두 과적합(overfitting)을 유도하기 위하여 은닉노드의 수를 크게 설정하였으며, 마찬가지 이유에서 epoch수를 30으로 설정하였다. 학습 알고리즘으로는 Levenberg-Maquardt를 사용하였다. 패턴 선정을 하면 패턴 수 감소로 인하여 학습 데이터 셋과 검증 데이터 셋으로의 분할이 부적절할 수 있다. 그러나, 선택된 학습 데이터 셋은 이미 정련된 패턴들로 구성되었으므로 분류기가 과적합을 해도 무방하다. 즉, 검증 데이터 셋이 필요 없음을 시사하기 위하여 전체 패턴으로 학습하는 경우에 한해서만 검증 데이터 셋(전체 패턴의 30%)이 있는 경우와 없는 경우로 구분하여 실험을 실시하였다. 테스트 패턴들은 Ⓐ, Ⓑ 각각의 동일분포로부터 300개씩 생성되었으며 총 30회의 실험이 반복되었다. (그림7)은 실험결과를 오분류(MCE: miss-classification error)율로 나타낸 것이다. Sinlgle-MLP의 경우에는 선택된 패턴만으로 학습한 경우, 실험의 평균과 특히 안정성(실험편차) 측면에서 월등히 우수하였다. Bagging-MLP의 경우에는 선택된 패턴만으로 학습한 경우의 분류성과가 약간 우수하였으나, 통계적으로 유의하다고 볼 수는 없었다. 주목할만한 사항은 sinlge-MLP에 있어서는 과적합



<그림7> 패턴 선택 전후에 대한 예측모델 (Single 및 Bagging MLP) 성능비교

방지를 위한 검증 데이터 셋의 설정이 효과적이었으나, bagging-MLP의 경우에는 오히려 성능을 감소시키는 결과를 가져온다는 것이다. 이는 Bagging과 같은 앙상블 알고리듬인 경우에는 각 구성 네트워크들이 과적합을 할수록 결과값의 분산이 커져 결합(aggregation)의 효과가 크게 발생하게 되는데, early stopping을 하게 되면 이러한 효과가 감쇄되기 때문이다.

### 3.2 실제 분류 문제에 대한 실험

제안된 알고리즘은 실제 문제에도 적용되었다 : BreastCancer, PimaIndian[9]. 3.1절과 마찬가지로 분류기는 9-15-2 single MLP와 30 9-15-2 bagging-MLP를 사용하였다(PimaIndian은 8-15-2 MLP). 인공문제의 결과를 참조하여, 학습에 전체 패턴을 모두 사용하는 경우, single-MLP 에서는 검증 데이터 셋을 설정하였고 bagging-MLP 에서는 모든 패턴을 학습 데이터 셋으로 사용하였다. 실험은 각각의 테스트 패턴들(BreastCancer: 205개, PimaIndian :230개)에 대하여 30회씩 반복되었다. 패턴선택 결과, BreastCancer의 경우에는 302개(전체 패턴의 63.2%)가, PimaIndian의 경우에는 308개(전체 패턴의 57.2%)가 선택되었다. (표1)은 실험결과를 정리한 것이다.

<표1> 실제 분류문제에 대한 실험결과

| Misc-Classification Error Rate | | Breast Cancer | | Pima Indian | |
|---|---|---|---|---|---|
| # of Features | | 9 | 9 | 8 | 8 |
| | Avg. (MCE) | 9.188 | 6.147 | 30.217 | 28.971 |
| Single MLP | Std. Dev. | 11.183 | 1.841 | 10.247 | 5.327 |
| | P-value | (T-test) 0.1519<br>(F-test) 0.0000 | | (T-test) 0.5576<br>(F-test) 0.0007 | |
| | Avg. (MCE) | 4.976 | 4.163 | 22.256 | 22.957 |
| Bagging MLP | Std. Dev. | 0.620 | 0.611 | 2.535 | 2.523 |
| | P-value | (T-test) 0.0001<br>(F-test) 0.9374 | | (T-test) 0.2882<br>(F-test) 0.9797 | |

Single-MLP 결과를 살펴보면, 오분류율의 실험평균에 있어서는 선택된 패턴만으로 학습한 경우와 전체패턴으로 학습한 경우간에 유의한 성능 차이는 보이지 않았다(T-test P-value: 0.1519, 0.5576). 그러나, 실험편차를 비교하면 선택된 패턴만으로 학습한 경우가 확연히 안정적인 결과를 산출함을 알 수 있었다(F-test P-value: 0.0000, 0.0007). Bagging-MLP의 결과에서는 패턴 선택 전,후의 성능 차이가 거의 없었으므로, 패턴 수에 따른 앙상블 네트워크의 메모리 및 계산시간 측면에서 일반화 성능이 개선되었다고 볼 수 있다.

## 4. 결 론

본 연구에서는 신경망 앙상블 네트워크 결과값의 분산과 편기를 이용하여 분류패턴을 선정하는 방법을 제안하였다. 분류문제에서 유용한 학습 패턴이란, 클래스들간의 분류경계에 위치하고 클래스 레이블이 정상인 패턴들을 말한다. 본 연구에서는 패턴점수와 패턴분리방법에 의하여 이들을 선정하였으며, 두 개의 인공문제와 두 개의 실제문제에 대하여 적용하였다. 실험결과 선정된 학습 패턴들은 패턴 수 감소로 인하여 학습 시간 단축 및 필요 메모리 공간을 감소시키는 효과뿐 아니라 실험적으로 불안정한 예측모델 출력값을 안정화시키는 효과도 얻을 수 있었다. 특히, 과적합의 우려가 없음이 실험적으로 검증되었으므로 예측모델의 복잡도에 무관한 분류패턴들이 선정되었음을 보였다. 추후 연구과제로는 패턴점수 선정시의 파라미터 $\alpha$값은 앙상블 네트워크의 학습 오분류율과 연관성이 있으므로 이를 정량적으로 결정하는 방법에 대한 연구가 이루어져야 할 것이다.

## 5. 참고문헌


[1] Tumer, K. and Ghosh, J., "Error Correlation and Error Reduction in Ensemble Classifiers," *Connection Science, Vol.8, 385-404* (1996)

[2] Sharkey, A.J.C., "On Combining Artificial Neural Nets," *Connection Sci-ence, Vol.8, 299-313* (1996)

[3] Perrone, M.P. and Cooper, L. N., "When networks disagree: Ensemble methods for hybrid neural networks," *Artificial Neural Networks for Speech and Vision*, (1993)

[4] Hara, K. and Nakayama, K., "A Training Method with Small Computation for Classification," *Proceedings of the IEEE-INNS-ENNS International Joint Conference*, Vol 3, 543-548 (2000)

[5] Leisch, F., Jain, L.C. and Hornik, K., "Cross-Validation with Active Pattern Selection for Neural-network Classifiers," *IEEE Transactions on Neural Networks*, Vol 9, 35-41 (1998)

[6] Bishop, C.M., "Neural Networks for Pattern Recognition," *Oxford Unive.-rsity Press, NewYork*, 386-450 (1995)

[7] Cho, S. and Wong, P.M., "Data Selection based on Bayesian Error Bar," *The Six International Conference on Neural Information Processing Vol.1, 418-422* (1999)

[8] http://www.ics.uci.edu/~mlearn