Journal of the Korean Institute of Industrial Engineers
Vol. 28, No. 1, pp. 112-127, March 2002.

앙상블의 편기와 분산을 이용한 패턴 선택

신현정 · 조성준†

서울대학교 산업공학과

Pattern Selection Using the Bias and Variance of Ensemble

Hyunjung Shin · Sungzoon Cho

Department of Industrial Engineering, Seoul National University, Seoul, 151-744

A useful pattern is a pattern that contributes much to learning. For a classification problem those patterns near the class boundary surfaces carry more information to the classifier. For a regression problem the ones near the estimated surface carry more information. In both cases, the usefulness is defined only for those patterns either without error or with negligible error. Using only the useful patterns gives several benefits. First, computational complexity in memory and time for learning is decreased. Second, overfitting is avoided even when the learner is over-sized. Third, learning results in more stable learners. In this paper, we propose a pattern "utility index" that measures the utility of an individual pattern. The utility index is based on the bias and variance of a pattern trained by a network ensemble. In classification, the pattern with a low bias and a high variance gets a high score. In regression, on the other hand, the one with a low bias and a low variance gets a high score. Based on the distribution of the utility index, the original training set is divided into a high-score group and a low-score group. Only the high-score group is then used for training. The proposed method is tested on synthetic and real-world benchmark datasets. The proposed approach gives a better or at least similar performance.

Keywords: pattern selection, ensemble network

1. 서 론

패턴 선택은 주어진 전체 학습 패턴으로부터 일부 유용한 패턴들을 추출하여 학습에 이용하는 방법이다. 유용한 패턴에 대한 척도는 패턴이 모델의 학습에 전달하는 정보의 양에 따라 결정된다. 분류문제에서는 클래스들간의 분류경계면에 근접한 패턴들이 클래스의 내부에 분포한 패턴들보다는 학습에 기여하는 바가 크다(Lee, 1997; Leich, 1998; Foody, 1999). 한편, 예측문제에 있어서는 추정면에 근접한 패턴들이 학습에 유용한 패턴들이다(Mackay, 1992; Plutowski, 1993; 1994; Röbel, 1994; Bishop, 1995; Cho, 1999). 두 경우 모두 측정 에러가 없거나 무시할 만한 정도의 에러를 가진 정확한 패턴들에 한해서만 유용성이 정의된다.

개별 패턴이 갖는 유용성을 측정하여 학습에 불필요하거나 위해한 패턴들을 사전 여과하면 다음과 같은 이점들이 있다. 우선, 학습에 소요되는 메모리 및 시간의 계산 복잡도를 감소시킬 수 있다(Leisch, 1998; Hara, 2000). 특히 근래 활발히 연구되고 있는 SVM(support vector machine)과 같은 경우에 있어서는 학습 패턴수가 QP(quadratic programming)의 제약식 수와 동일하므로 패턴의 다소 여부와 모델의 성능이 밀접하게 관련되어 있다(Burges, 1998; Gunn, 1998; Hearst, 1998; Kwok, 1999; Vincent, 2000). 또한 데이터마이닝의 모델 선택 절차에서처럼, 대량의 패턴 셋으로 여러 모델들간의 성능을 비교한다거나 반복 실험하는 경우에는 패턴 선정의 활용도가 증폭된다. 둘째, 문제에 비해 모델의 복잡도 및 학습 파라미터가 과도하게 설정되었을 경우에도 과적합(overfitting)을 방지할 수 있다. 대부분의 경우, 문제의 복잡도를 사전에 알기는 어렵기 때문에 이에 적합한 모델의 구조나 학습 파라미터를 설정하는 일은 시행착오를 통해서 이루어진다. 문제의 복잡도는 주어진 패턴 셋에 이상 패턴들이 포함된 경우에 과대평가된다. 이로 인하여 모델의 복잡도도 과대 설정되므로 과적합이 발생한다

† 연락저자 : 조성준 교수, 151-744 서울시 관악구 신림동 산 56-1 서울대학교 산업공학과, Fax : 02-883-4913, e-mail : zoon@snu.ac.kr
2001년 11월 접수, 1회 수정(2주 소요) 후, 2002년 1월 게재 확정.

(Mitchell, 1997). 그러나 사전 작업을 통하여 이상 패턴들을 제거하고 학습에 크게 기여하는 정상 패턴들만을 선정한다면, 과대설정된 파라미터와 구조를 가진 모델도 과적합되는 것을 방지할 수 있다. 셋째, 불안정한 모델(unstable learner)에 있어서는 출력값의 변동 편차를 안정화시킬 수 있다. 신경망, decision tree 등과 같은 모델들은 주어진 학습 패턴 셋의 변화에 민감하다(Breiman, 1996a; 1996b). 따라서 매 학습시마다 출력값의 변동이 크므로 모델의 신뢰도가 강하된다. 이러한 경우에 있어서도 이상 패턴 및 학습에 불필요한 패턴들이 제거되면 모델의 출력값 변동이 안정된다.

분류문제에 있어서 유용성이 큰 패턴들을 분류경계 부근의 패턴들로 정의하고 이를 검증한 유사연구들은 다음과 같다. Lee 등은 신경망을 위한 분류경계 특징추출(decision boundary feature extraction)에 사용할 분류경계 패턴들을 다음과 같은 방법으로 선정하였다. 예를 들어, 2분류 문제에 있어서 분류경계 $B(\vec{x})$는 1-of-2 신경망의 두 출력노드 값 $f_1(x), f_2(x)$의 크기가 같은 패턴(x)들이 연결된 면으로 정의되었다. 이 x 패턴들을 찾는 과정은 다음과 같다. 우선 학습된 신경망에 의해 정분류된 클래스1 패턴(y)과 이와 최단 거리에 있는 클래스2 패턴(z)을 찾는다. 이때 두 점을 잇는 선분은 분류경계를 통과하게 된다. 이 선분과 분류경계의 교점, 즉 이 선분 내에 있으면서 $B(\vec{x}) = f_1(\vec{x}) - f_2(\vec{x}) \cong 0$ 을 만족시키는 패턴 x가 분류경계 위의 점 또는 근접점이 되는 것이다(Lee, 1997). 한편, Leisch 등은 cross-validation 수행시 중요한 검증 패턴이 분류경계부근에 분포함을 시사하고, 이를 선정하기 위한 방법으로 에러의 크기를 이용하는 방법을 소개하였다. 실험결과, 학습에 사용되지 않은 패턴들 중 에러를 크게 유발시키는 패턴들은 주로 분류경계에 분포하고 있었으며, 이들은 클래스 내부에 있는 패턴들보다 분류학습에 기여도가 크다는 것을 보였다(Leisch, 1998). Foody는 분류 경계에 인접한 학습패턴들을 찾기 위하여 클래스 중심과의 마할라노비스(mahalanobis) 거리를 이용하였다. 양분된 두 개의 학습 셋 즉, 클래스 경계부근 패턴 셋과 클래스 중심부근 패턴 셋에 대한 성능 비교 실험 결과, 전자의 성능이 월등히 우수함을 보였다(Foody, 1999). Hara 등은 서로 다른 클래스에 속해 있는 정분류된 nearest neighbor를 학습 패턴으로 선정하였으며, 최종적으로 선정된 패턴들은 분류경계 부근의 정분류된 패턴들임을 실험적으로 제시하였다(Hara, 2000).

예측문제에 있어서 유용성이 큰 패턴들을 선정하는 방법에 대한 연구들은 다음과 같다. Röbel은 신경망 훈련 초기에는 패턴수가 적은 서브 셋으로 학습을 시작하고, 모델의 성능이 일정 수준 이상이 되면 원 학습 셋의 패턴들 중 에러를 가장 크게 유발하는 패턴을 선정하여 학습 셋에 추가하는 방법을 제안하였다(Röbel, 1994). Plutowski 등도 이와 유사한 방법으로서, 에러가 큰 패턴을 선정하면서 점진적으로 학습을 진행해 나가는 방법을 소개하였다(Plutowski, 1994). 학습된 모델이 새로이 학습할 패턴을 찾기 위해서는 새로운 정보를 많이 가진 패턴이 유용하다. 이러한 유용한 패턴은 에러가 크다는 근거로 제안된 방법들이다. 반대로 에러의 크기가 작은 패턴들만을 학습패턴으로 선정하는 방법들이 있다. Cho와 Qu 등의 연구에서는 베이지안 네트워크의 에러바(error bar)를 이용하여 이상 패턴들을 제거하였다. 베이지안 네트워크에서는 주어진 패턴에 대한 모델 출력값의 신뢰구간을 제공하는데, 이를 에러바라고 한다(Mackay, 1992; Bishop, 1995). 실험에서는 에러바가 큰 패턴들이 제거된 후, 학습 모델의 일반화 성능이 향상됨을 보였다(Cho, 1999; Qu, 2001). 추정면으로부터 멀리 떨어진 패턴, 즉 학습 에러가 큰 패턴일수록 베이지안 네트워크의 에러바 값이 증가한다는 상관관계를 이용한 성과이다. Röbel과 Plutowski의 연구에서의 전제 조건은 원 학습 패턴 셋의 노이즈가 0 또는 ε 이하의 깨끗한 패턴들이어야 한다는 데에 있다. 따라서 이들 연구는 Cho와 Qu 등이 제안한 방법처럼 에러가 큰 이상 패턴들을 제거하는 전처리 과정 후에 적용될 수 있는 방법들이다.

본 연구에서는 분류 및 예측 문제에 대하여 각 패턴의 유용성을 점수화하는 척도로서 패턴의 "효용지수(utility index)"를 제안한다. 효용지수는 각 패턴에 대하여 앙상블 네트워크의 출력값이 갖는 편기와 분산을 이용하여 계산된다. 앙상블 네트워크는 다양한 구조 및 학습 파라미터를 가진 여러 개의 네트워크들로 이루어졌으므로, 각 패턴마다의 출력값 분포를 얻을 수 있다(Perrone, 1993a; Tumer, 1996; Sharkey, 1997). 분류문제에서는 편기가 작고 분산이 큰 패턴이 높은 효용지수를 얻는다. 예측문제에서는 편기와 분산이 작은 패턴이 높은 효용지수를 얻는다. 이 효용지수의 분포를 근거로 하여, 원 학습 패턴 셋은 높은 효용지수 그룹과 낮은 효용지수 그룹으로 분리된다. 분리과정은 효용지수의 분포를 편중되지 않게 양분하되 최대한 상이하게 만든다는 원칙하에 이루어진다. 이 중 높은 효용지수 그룹에 속한 패턴들만이 학습에 사용된다.

방법론 측면에서 본 연구는, 앙상블 네트워크를 패턴 선택 방법으로 사용하였다는 점이 기존연구들과 비교하여 차별성을 갖는다. 특히 구성 네트워크들의 개별 결과값들로부터 얻을 수 있는 엔트로피가 활용되었다. 이는 앙상블의 특성상, 다수투표(majority voting)나 단순평균(simple averaging) 등의 방법에 의하여 사장되었던 개별 네트워크 정보들이 십분 활용되었다는 점에서도 다르다고 할 수 있다.

활용도 측면에서 제안하는 패턴 선정방법은 전처리 작업에 속한다. 즉, 후에 선정된 학습패턴을 사용하게 될 모델과는 독립된 절차이다. 이는 Leisch, Foody, Hara, Plutowski, Röbel의 연구에서처럼 학습중에 모델이 학습하고자 하는 일부 패턴만을 선택하는 능동 샘플링(active sampling)이나 능동 학습패턴 선택(active pattern selection)방법과는 다르다. 전자는 학습중에 모델이 패턴의 목표값과는 상관없이 입력 분포로부터 패턴을 선택해 나가는 방법이고, 후자는 모델이 학습하고자 하는 목표값을 참조하여 패턴을 선택하는 방법으로 두 방법 모두 모델의 학습 과정과 패턴선택 과정이 점진적으로 함께 이루어

진다(Cachin, 1994; Zhang, 1993; 1994; Plutowski, 1994; Röbel, 1994; Leisch, 1998; Food, 1999; Hara, 2000).

이에 반해 제안하는 패턴 선택방법은 모델의 학습 과정과는 분리된 별개의 과정이다. 따라서 학습 모델의 종류에 관계없는 범용적 학습 셋을 얻을 수 있다. 이는 여러 학습 모델간의 성과 비교나 반복 실험이 요구되는 상황에서 그 활용도가 크다. 제안하는 방법에 의한 실험 결과, 정련된 학습 패턴들은 원 학습 패턴의 수에 그 수가 현저히 감소되었음에도 불구하고, 모델의 종류나 복잡도에 상관없이 과적합에 대한 대응 필요성을 소진시켰으며, 부가적으로는 이상 패턴 제거로 인하여 실험 변동성을 안정시키는 효과를 보였다.

본 논문의 구성은 다음과 같다. 2절에서는 각 패턴의 효용지수를 계산하는 방법을 소개하고, 3절에서는 효용지수를 근거로 하여 전체 학습 셋으로부터 효용지수가 높은 패턴그룹을 분리하는 방법에 대하여 소개한다. 4절에서는 제안한 방법의 타당성 검증을 위하여 합성 데이터 실험 설정을 기술하고 이의 결과를 제시한다. 5절에서는 실제 데이터에 대한 실험을 통하여 제안한 방법의 일반화 성능을 검증한다. 6절에서는 결론 및 추후 연구과제에 대하여 기술한다.

2. 효용지수 계산

분류문제의 경우, 분류경계면(decision boundary surfaces)으로부터 멀리 떨어진 패턴들은 학습 성능을 저하시키지는 않으나 학습에 대한 기여도가 낮은 불필요한 패턴들이다. 또한 분류경계 부근의 패턴들이라도 클래스 레이블이 바뀐 패턴들은 이상 패턴들로서 학습을 어렵게 만들므로 제거되어야 할 패턴들이다(Lee, 1997; Leich, 1998; Foody, 1999; Hara, 2000). 예측문제에 있어서는, 추정면(estimated surface)으로부터 멀리 떨어진 패턴들은 이상 패턴일 가능성이 높으므로 학습에 방해가 되는 패턴들이다. 이러한 패턴들은 학습된 모델 출력값에 대한 신뢰도를 떨어뜨리므로 제거되면 학습성능이 향상된다(Cho, 1999; Qu, 2001).

본 연구에서는 각 문제에 있어서 패턴이 분류경계면 또는 추정면에 가까운 정도를 "근접도(proximity)"로 정의한다. 분류경계면 또는 추정면에 가까울수록 패턴의 근접도가 증가한다. 또한 패턴의 정상 여부에 대한 척도로서 "정확도(correctness)"를 정의한다. 정상 패턴은 정확도가 높다.

제안하는 방법에서는 근접도와 정확도를 구하기 위하여 앙상블 네트워크를 이용한다. 앙상블 네트워크는 여러 개의 다양한 네트워크들을 학습시킨 후, 이들의 출력값들을 결합하여 총론적 결과를 얻는 방법이다. 앙상블 구성 네트워크들 간에는 에러 상관(error correlation)이 작을수록 보다 나은 일반화 성능이 산출된다. 이를 도모하기 위한 방법으로는 대개 두 가지 범주, 즉 구성 네트워크 학습 셋에 교란을 주는 방법과 구성 네트워크의 구조 및 파라미터를 다양화시키는 방법이 있다(Perrone, 1993a; 1993b; Krogh, 1995; Breiman. 1996; Tumer. 1996; Parmanto, 1996; Sharkey, 1996; 1997). 제안하는 방법에서는 후자의 방법에 의하여 앙상블 네트워크를 구성한다. 전체 학습 셋에 대하여 학습된 앙상블 네트워크는, 각 패턴에 대한 출력값 분포를 제공한다. 예를 들어, 예측문제인 경우, 하나의 패턴 (x, y)에 대하여 L개의 네트워크들이 추정하는 출력값 $F_l(x)$의 분산과 편기는 다음과 같다.

$$\text{편기}: |y-\overline{F}|,$$

$$\text{분산}: \sum_{l=1}^{L}(F_l(x)-\overline{F})^2/(L-1)$$

$$\text{where } \overline{F}=\sum_{l=1}^{L}F_l(x)/L.$$

출력값 분포의 분산은 근접도를 유도하는 데 사용되며, 출력값 편기(평균과 목표값의 차이)는 정확도를 유도하는 데 사용된다. 구해진 각 패턴의 근접도와 정확도의 선형 결합은 "효용지수(utility index)"로 정의되며 패턴의 유용성이 클수록 효용지수는 높은 값을 갖는다. 다음의 각 절에서는 각 문제별로 패턴의 효용지수를 유도하는 방법에 대하여 설명한다.

2.1 분류문제

분류 패턴들에 대한 근접도(proximity, 해당 패턴이 분류경계면에 인접해 있는가)와 정확도(correctness, 정상 패턴인가)를 유도하기 위하여 이에 대한 앙상블 출력값의 편기 및 분산과의 관계를 살펴보면 다음과 같다. 예를 들어, <그림 1>과 같이 최적 분류경계가 B^*인 2분류 문제인 경우, 두 개의 출력노드 $f_1(x), f_2(x)$를 가진 신경망 모델의 분류경계 B는 그림의 짙은 부분 중 한 곳에 위치한다(Tumer, 1996). 따라서 모델의 분류성능은 이 $|B^*-B|$의 크기를 얼마나 작게 줄이느냐에 따라 평가된다. 앙상블 네트워크의 분류경계 B_{com}의 경우, 개별 모델의 분류경계 B보다는 B^*에 좀더 가깝게 근사할 수 있다. 이는 B_{com}이 l개의 구성 네트워크들의 분류경계 $\{B_l\}$들을 선형,비선형적으로 결합해서 얻어졌기 때문이다(Perrone, 1993b; Sharkey, 1996).

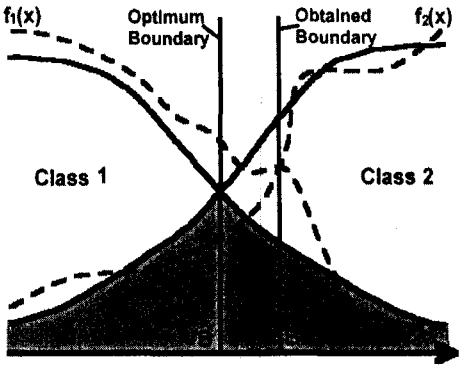


그림 1. 분류경계와 오분류 발생위치.

이러한 앙상블 효과-에러 분산 감소효과-는 주로 에러 발생 확률이 많은 분류경계 부근에서 작용한다(Tumer, 1996).

이를 패턴 측면에서 생각해 보면, 앙상블 네트워크의 출력값 분산이 큰 패턴들은 분류경계 부근에 위치할 확률이 크다는 것을 알 수 있다. 또한 이들 중 정상 패턴들은 앙상블 네트워크 출력값의 편기를 이용하여 이상 패턴들과 구별될 수 있다.

따라서 분류문제에서는 앙상블 네트워크의 출력값 분산이 클수록 근접도가 크며, 출력값 편기가 작을수록 정확도가 증가한다. 분류문제에 있어서 패턴의 효용지수를 계산하는 과정은 다음과 같다.

① J클래스 분류문제에 대하여 L개의 1-of-J MLP 네트워크를 전체 학습 패턴 M에 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 랜덤화 하였다.

② 패턴 x_i에 대하여 l번째 네트워크의 j개의 출력노드 값들로부터 네트워크 결과값 $F_l(x_i)$을 산출한다.

$$F_l(x_i) = \arg\max_{j\in J} f_j(x_i), \quad l=1\ldots L,\ i=1\ldots M \tag{1}$$

③ 클래스 j별로 L개 네트워크의 다수 투표(majority voting) 결과확률을 계산한다.

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

④ 각 패턴의 효용지수(utility index) $UI(x_i)$를 계산한다. 단, $0\log 0$은 0으로 정의한다.

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

식 (3)의 첫 번째 항 $P_{j^*}(x_i)$는 정확도로서, 패턴 x_i의 정답 클래스 j^*에 대한 네트워크들의 투표율을 나타낸다. 앙상블 출력값의 편기는 $\{1-P_{j^*}(x_i)\}$이므로 편기가 작을수록 정확도는 증가한다. 두 번째 항은 근접도로서 투표결과의 분산 정도를 나타낸다. 분산 정도가 클수록 근접도는 증가한다. 다음 <그림 2>는 위의 과정에 대한 앙상블 네트워크의 구조를 나타낸다.

<그림 3>은 3-클래스 분류문제인 경우, 6개의 1-of-3 MLP 네트워크에 의해 계산된 4개 패턴의 근접도와 정확도 계산과정을 예시한 것이다.

근접도는 투표결과 값의 정답 여부와는 무관하며, 단지 해당 패턴에 대한 네트워크 의견의 불일치성을 나타낸다. 분류경계 부근의 패턴들은 클래스 내부의 패턴들에 비하여 오분류

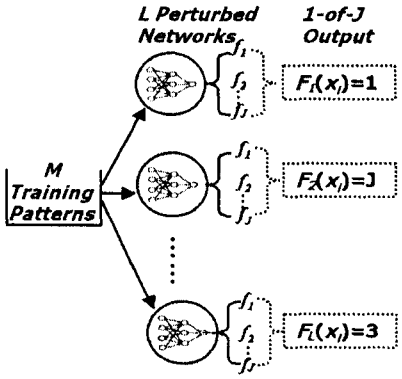


그림 2. 앙상블 네트워크의 구조.

	J*	F1	F2	F3	F4	F5	F6
X1	1	1	1	2	3	1	1
X2	1	1	1	1	1	1	1
X3	2	1	1	2	2	3	3
X4	3	3	3	2	2	3	1

ⓐ

	P1	P2	P3
X1	4/6	1/6	1/6
X2	6/6	0/6	0/6
X3	2/6	2/6	2/6
X4	1/6	2/6	3/6

ⓑ

Pj*	$\sum_{j\in J} P_j(x_i)\log_{\lvert J\rvert}\frac{1}{P_j(x_i)}$
4/6	0.7897
6/6	0
2/6	1
3/6	0.9227

ⓒ

그림 3. 근접도와 정확도 계산과정: ⓐ 4개의 학습 패턴에 대한 6개 앙상블 구성 네트워크들의 출력값, ⓑ 클래스별 투표확률, ⓒ 근접도와 정확도.

될 가능성이 높으므로 각 클래스에 대한 네트워크들의 투표확률 $P_j(x_i)$들이 분산된다.

반대로 클래스 내부 패턴들은 정답 클래스 j^*에 대한 투표확률 $P_j{}^*(x_i)$가 다른 클래스에 대한 투표확률 $P_j(x_i)$들에 비해 압도적으로 큰 값을 갖는다. 그러므로 투표결과의 엔트로피인 근접도 $\{-\sum_{j\in J} P_j(x_i)\log_{|J|}P_j(x_i)\}$는 분류경계 부근의 패턴들에게서 상대적으로 큰 값을 나타낸다. 따라서 분류경계 인접 패턴들은(정상 패턴인지 이상 패턴인지에 상관없이) 식 (3)의 근접도에 의하여 우선적으로 높은 효용지수를 갖게 되고, 이들 중 정상 패턴들은 정확도에 의하여 이상 패턴들보다 더 높은 효용지수를 갖게 된다.

다음의 <그림 4>는 패턴 선정에 있어서 근접도와 정확도가 미치는 영향을 도식화한 것이다. <그림 4>의 ⓐ와 같은 분류문제에 대하여 근접도만을 적용하면 ⓑ와 같이 분류 경계부근의 패턴들이 높은 점수를 얻게 되는데 이들 중 상당 부분이 이상 패턴에 해당된다. 따라서 근접도만으로는 노이즈 패턴들을 여과할 수 없음을 의미한다. ⓒ의 경우는 정확도만을 고려한

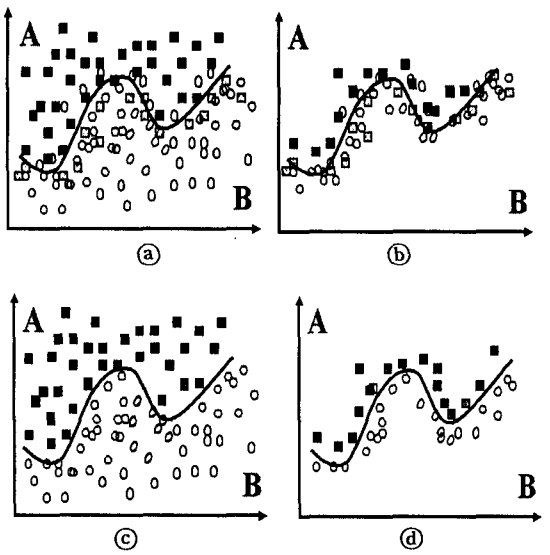


그림 4. 근접도와 정확도가 분류 패턴선정에 미치는 영향: ⓐ 원 학습패턴, ⓑ 근접도가 큰 패턴들, ⓒ 정확도가 큰 패턴들, ⓓ 근접도와 정확도를 함께 고려한 경우 선정된 패턴들(■: 클래스A, ●: 클래스B).

경우로, 이때에는 분류경계로부터 먼 패턴들도 모두 포함된다. 이 경우에 있어서는 정확도만으로는 학습에 기여도가 적은 불필요한 패턴들을 여과할 수 없음을 나타낸다. 따라서 근접도와 정확도를 모두 고려하면 ⓓ에서처럼 학습에 중요한 영향을 미치는 정상 패턴들만을 선정할 수 있다.

다음은 식 (3)에서 정확도와 근접도를 선형조합하는 파라미터 값 α를 결정하는 과정에 대하여 설명한다. α값의 결정은 제안한 효용지수가 최적의 변별력을 갖춘 척도가 되도록(분류경계 부근의 정상 패턴들에게 높은 점수를 부여하도록) 조정하는 과정이다. 주어진 x_i에 대하여 L개의 네트워크의 다수투표(majority voting)에 의해 결정된 클래스 $\bar{j}$를 다음과 같이 정의한다.

$$\bar{j} = \arg\max_j P_j(x_i)$$

다수 투표결과인 클래스 레이블 $\bar{j}$가 실제 클래스 레이블인 j^*와 일치하지 않는 패턴들은 사전분류과정(preliminary classification)에 의해 오분류된 패턴들로서 학습 셋에 내재되어 있는 이상 패턴들로 가정한다. 내재된 이상 패턴의 개수는 다음과 같다.

$$\sigma_M = \sum_{i=1}^{M} 1 \text{ if } (\bar{j} \neq j^*)$$

α 값은 근접도가 큰 패턴들을 최대한 지향하되, 이들 중 학습 셋에 내재된 이상 패턴들을 최대한 제거하도록 설정한다. 우선 임의의 α에 대하여 $UI(x_i)$를 정렬한 후, 가장 큰 값으로부터 $\overline{M} = 2 \cdot \sigma_M$개의 패턴들을 고려한다. $\overline{M}$을 이같이 설정하는 이유는 근접도만을 고려했을 경우($\alpha = 0$), σ_M개의 이상 패턴들은 대부분 이 범위 내에 포함되기 때문이다. 예를 들어, 100개의 학습 패턴 중 20%가 이상 패턴들이라면, 이들은 대부분 분류경계 부근에 분포하므로 높은 근접도 값을 갖는다. 따라서 근접도만으로 정렬했을 때, 최악의 경우에는 경계 부근의 정상 패턴들과 이상 패턴들의 효용지수 순위가 번갈아가며 바뀌게 된다. 그러므로 $40(= 2 \times 100 \times 20\%)$개의 상위 패턴들을 고려하면 20개의 이상 패턴들 중 대다수가 $\overline{M}$에 포함된다. 이러한 경우, $\alpha = 0$에 의하여 계산된 효용지수는 변별력이 없으므로 $\alpha = 0.1$로 값을 증가시켜 효용지수에 대한 정확도의 가중치를 높인다. 새로이 계산된 효용지수에 대하여도 이러한 과정을 반복한다. 이때 $\overline{M}$에 포함된 이상 패턴의 수가 현저히 감소하는 점을 α 값으로 결정한다. α 값과 이상 패턴 비율의 민감도는 다음의 두 가지 척도를 이용한다.

(i) 상대적 이상 패턴 비율
(relative noise ratio) $= (\sigma_{\overline{M}} / \sigma_M) \times 100\%$

(ii) 절대적 이상 패턴 비율
(absolute noise ratio) $= (\sigma_{\overline{M}} / M) \times 100\%$

상대적 이상 패턴 비율은 $\overline{M}$에 포함된 이상 패턴의 수($\sigma_{\overline{M}}$)를 전체 이상 패턴수(σ_M)에 대하여 대비시킨 것이고, 절대적 이상 패턴 비율은 전체 학습 패턴의 수(M)에 대비시킨 것이다. 다음의 <그림 5>는 α값의 변화에 따른 이상 패턴 비율의 민감도를 나타낸다.

상기한 예에서 $\alpha = 0$일 때, 40개의 패턴에 포함된 이상 패턴이 18개이면 상대적 이상 패턴 비율은 $(18/20) \times 100\%$로 90%이고, 절대적 이상 패턴 비율은 $(18/100) \times 100\%$로 18%이다. $\alpha = 0.1$로 값을 증가시키면 정상 패턴들에 대한 효용지수 값이 증가하므로 $\overline{M}$에 포함된 이상 패턴의 수가 감소하게 된다. 이때 포함된 이상 패턴의 수를 14라고 하면 상대적 이상 패턴 비율은 70%[$= (14/20) \times 100\%$]이고, 절대적 이상 패턴 비율은 14%이다. 최종적으로 $\alpha = 1$인 경우에 이르면 효용지수에는 정확도만이 고려되어 $\overline{M}$ 내에는 이상 패턴이 거의 포함되지 않는다. 그러나 이 경우에는 정상 패턴들 중에서도 분류경계면에서 멀리 떨어진 패턴들이 높은 효용지수를 얻는다.

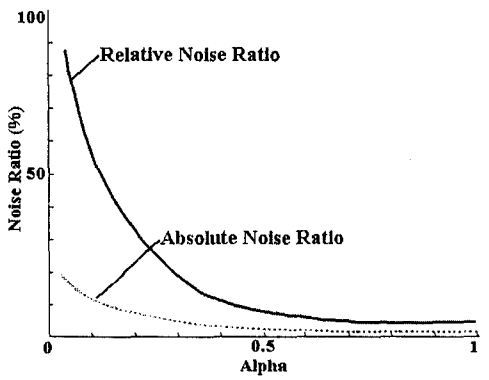


그림 5. α 값과 이상 패턴 비율의 민감도.

그러므로 분류경계면에 가까운 정상 패턴들이 보다 높은 효용지수를 갖도록 하기 위해서는 상대적 이상 패턴 비율 및 절대적 이상 패턴 비율의 감소가 안정화되는 시작점에서 α를 설정한다. <그림 5>에서는 $\alpha=0.5\sim0.6$ 사이에서 결정하는 것이 적절하다.

α 값의 결정으로 최적의 변별력을 갖춘 효용지수가 마련되면, 이를 척도로 하여 주어진 학습 패턴의 유용성을 정량화한다. 이를 기준으로 정렬된 패턴들은 3절에서 설명될 패턴 분리 작업에 의해 일부 학습 셋으로 축소된다.

2.2 예측문제

예측문제에 있어서 유용한 패턴은 추정면에 근접한 패턴들이다. 추정면으로부터 멀리 떨어진 패턴들은 학습에 있어서 비교적 중요도가 낮거나 방해가 되는 패턴들이다. 이러한 패턴들은 이상 패턴일 가능성이 높다.

예측 패턴들에 대한 근접도(해당 패턴이 추정면에 인접해 있는가)와 정확도(정상 패턴인가)를 유도하기 위하여 이에 대한 앙상블 출력값의 편기 및 분산과의 관계를 살펴보면 다음과 같다. 우선, 예측문제에 있어서 추정면은 전체 에러를 최소화하게끔 설정된다. 따라서 충분히 학습된 네트워크에 의해 형성된 추정면은 패턴 밀집 구간의 중심부근을 통과한다 (Mackay, 1992; Bishop, 1995; Freund, 1997; Drucker, 1997; 1999; Haykin, 1999). 예를 들어, <그림 6>의 ⓐ에서 A와 C구간은 패턴이 밀집된 구간으로, 앙상블 구성 네트워크 추정선 [$F_l(x)$]들이 이들 중심 부근을 통과한다. 이를 패턴 측면에서 살펴보면, <그림 6>의 ⓑ에서처럼 이들 구간에 속한 패턴들에 대해서는 네트워크 출력값들의 분산이 작음을 알 수 있다. 또한 이들 구간에서는 네트워크 출력값들의 평균과 목표값과의 차이가 작은 것을 알 수 있다. 그러나 <그림 6>의 B구간처럼 패턴 수가 희박한 경우에는 네트워크들의 학습이 잘 이루어지지 않았기 때문에 네트워크 출력값들의 분산이 크다. 그러나 반드시 출력값들의 평균과 목표값과의 차이도 큰 것은 아니다. B구간에서 보면 ①과 ③으로 표시된 패턴들은 이상 패턴들이므로 제거되어야 하고, ②패턴은 선정되어야 한다. 그러나 <그림 6>의 ⓑ에서처럼 세 패턴에 대한 출력값 분산은 동일하므로 이것만으로는 이들을 구별할 수 없다. 따라서 네트워크 출력값의 편기를 이용한다. 즉, ②패턴에 대한 네트워크 출력값들의 편기는 ①과 ③패턴들에 비해 상당히 작으므로, 이를 이용하여 패턴이 희박한 구간에서도 정상 패턴을 이상 패턴으로부터 선별할 수 있다.

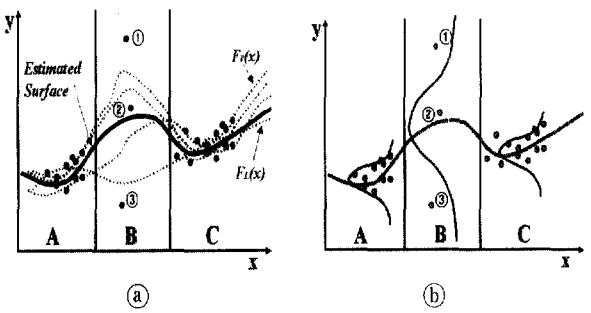


그림 6. 근접도와 정확도가 예측 패턴선정에 미치는 영향: ⓐ 학습 패턴 밀도에 따른 네트워크들의 추정선들 $F_L(x)$, ⓑ 학습 패턴 밀도와 추정값들의 분포.

제안하는 방법에서는 예측문제인 경우, 네트워크 출력값들의 분산이 작을수록 근접도가 증가하며 편기가 작을수록 정확도가 증가하도록 설정하였다. 이 근접도와 정확도를 이용하여 이상 패턴이 제거된 추정면 부근의 유용한 패턴들을 구별해 낼 수 있다. 이에 근거하여 예측 패턴의 효용지수를 계산하는 방법은 다음과 같다.

① L개의 네트워크들을 전체 학습 패턴 M에 대하여 학습시킨다. 각 네트워크들은 네트워크 구조 및 학습 파라미터에 교란(perturbation)을 주어 구성한다. 교란을 주는 방법은 2.1절과 마찬가지로 각 네트워크들의 은닉층 및 은닉노드의 수, 활성화 함수, 학습 횟수 등의 설정을 랜덤화한다.

② 패턴 x_i에 대하여, 각 네트워크의 출력값 $F_l(x_i)$의 평균 $\overline{F_i}$를 구한다.

$$\overline{F_i}=\frac{\sum_{l=1}^{L}F_l(x_i)}{L},\quad l=1,\cdots,L,\ i=1,\cdots,M \tag{4}$$

③ 마찬가지로 각 패턴 x_i에 대한 네트워크들의 출력값 분산 Σ_i를 구한다.

$$\Sigma_i=\frac{\sum_{l=1}^{L}(F_l(x_i)-\overline{F_i})^2}{L-1},\quad l=1,\cdots,L,\ i=1,\cdots,M \tag{5}$$

④ 각 패턴의 효용지수(utility index) $UI(x_i)$를 계산한다. y_i는 i 번째 패턴의 목표값이다.

$$UI(x_i)=\alpha\times e^{-|y_i-\overline{F_i}|}+(1-\alpha)\times e^{-\Sigma_i} \tag{6}$$

식 (6)의 첫 번째 항은 정확도로서, 지수 부분은 앙상블 출력값의 편기이다. 즉, 앙상블 네트워크의 평균 출력값 $\overline{F_i}$와 목표값 y_i의 차이가 작을수록 정확도는 증가한다. 식 (6)의 두 번째 항은 근접도로서 앙상블 출력값의 분산 Σ_i가 작을수록 값이 증가한다. 이때의 선형조합 파라미터는 다음과 같이 결정한다. 2.1절과 마찬가지로 α값은 효용지수가 최적의 변별력을 갖도록(추정면과 가까운 패턴이 높은 점수를 갖도록) 조정한다. 예측문제에서는 임의의 α에 대하여 $UI(x_i)$를 정렬한

후, 가장 큰 값으로부터 $\overline{M}=(1/2)\cdot M$개의 패턴들을 고려한다. 각 a에 대하여 다음과 같이 상대적 에러제곱합의 비율(relative sum of squared error ratio)을 구하고 민감도 분석을 실행한다.

- 상대적 에러제곱합의 비율(relative sum of squared error ratio)

$$= \frac{\sum_{i=1}^{\overline{M}}(y_i-\overline{F_i})^2}{\sum_{i=1}^{M}(y_i-\overline{F_i})^2}\times 100\%$$

효용지수가 추정면 부근의 패턴들에게 높은 점수를 부여하는 적절한 척도라면, 이에 의해 정렬된 상위 $\overline{M}$개의 에러제곱합은 전체 에러제곱합에서 작은 부분만을 차지한다. 따라서 상대적 에러제곱합의 비율이 안정화되는 점에서 a값을 결정한다.

3. 패턴 분리(Pattern Separation)

2절에서의 근접도와 정확도에 의하여 M개의 패턴에 대한 효용지수 $UI(x_i)(i=1,\ldots,M)$가 계산되면 이를 근거로 하여 유용성이 큰 패턴들을 적정 수준에서 분리하는 작업이 필요하다. 즉, $UI(x_i)$ 값의 분포를 UI로 두고 이를 효용지수가 낮은 패턴그룹 UI_L과 효용지수가 높은 그룹 UI_H로 양분한다. 이 중 UI_H는 후에 분류 및 예측 모델에 사용될 학습 데이터 셋이고, UI_L은 버려질 패턴집합이다. UI를 서로 상이한 두 그룹 UI_L, UI_H로 양분하는 원칙은 다음과 같다.

(1) UI의 분리 임계점은 두 그룹의 돗수를 가능한 한 균형되게 만든다.
(2) 임계점은 UI의 모양이 확연히 달라지는 저밀도 구간에 설정한다.

첫 번째 조건은 학습 데이터의 수가 지나치게 감소되거나 또는 패턴 선정의 효과가 무마되는 것을 방지한다는 의미이다. 두 번째 조건은, 되도록 편중되지 않게 UI를 분리하되, 양분된 그룹을 최대한 상이하게 만든다는 의미이다. 만약 밀도가 높은 구간에 임계점이 설정되면 양분된 UI_L과 UI_H가 비슷한 $UI(x_i)$값을 갖는 패턴들을 다수 포함하게 된다. 이는 두 그룹을 상이하게 만드는 데 반대효과를 낸다.

본 연구에서는 분포의 양분을 위하여 T-test의 아이디어를 활용하였다. T-test는 두 분포간 평균차이의 유의성을 검정하는 통계적 방법이다. 샘플의 크기가 각각 n_1, n_2이고, 평균이 $\overline{X_1}$, $\overline{X_2}$, 분산이 s_1^2, s_2^2인 두 분포가 있을 때 T-검정 통계량은 다음과 같다(유의 수준 : a_{sig}). 검정 통계량의 값이 $T>t(\nu, a_{sig})$이면,

T-검정 통계량:

$$T=\frac{(\overline{X_1}-\overline{X_2})}{\sqrt{s_1^2/n_1+s_2^2/n_2}}\sim t(\nu, a_{sig}),\ \text{where}$$

$$\nu=\frac{(s_1^2/n_1+s_2^2/n_2)^2}{[(s_1^2/n_1)^2/(n_1-1)]+[(s_2^2/n_2)^2/(n_2-1)]}$$

두 평균의 차가 유의하다는 의미로 두 분포가 동일하지 않음을 말한다. 이 때 T-검정통계량은 비교되는 두 분포의 분산이 등분산인 경우와 이분산인 경우에 따라 달라진다.

본 연구에서는 점차적으로 UI의 분포구간을 좁혀가면서 T-test를 반복 적용하였는데, 이를 이후 Drilldown T-test라 명명한다. UI의 분포구간은 매회 정해지는 상한과 하한에 따라 좁혀져 간다. 이는 T-test를 연속적으로 수행할 기본 분포 $D(k)$를 생성하기 위함이다. <그림 7>에서는 이러한 과정을 묘사한다. 우선, 기본 분포의 하한을 결정하는 방법은 다음과 같다. UI를 B개의 등간격구간(equal spaced bin)으로 분할하고, 각 구간의 패턴들의 집합을 $\mathrm{Bin}_p\,(p=1,\cdots,B)$로 정의한다. 그림에서는 $B=15$로 설정되었다. 초기에는 UI가 기본 분포 $D(1)=\{\cup_{p=1}^{B}\mathrm{Bin}_p\}$가 된다. $D(1)$으로부터 최하한 구간을 제외한 분포를 $D_L(1)=D(1)-\{\mathrm{Bin}_1\}$으로 둔다. 이

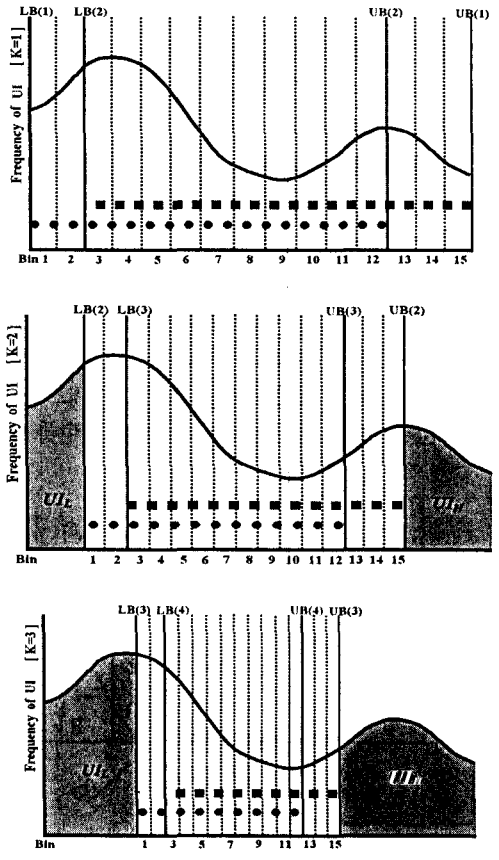


그림 7. Drilldown T-test에 의해 패턴 분리 임계점을 찾는 과정, $k=1\cdots 3$, ■: $DL(k)$, ●: $DU(k)$.

$D(1)$과 $D_L(1)$의 T-test 검정 통계량이 유의하지 않으면 $D_L(1)$을 $D_L(1)=D(1)-\{\mathrm{Bin}_1\cup\mathrm{Bin}_2\}$로 새롭게 정의하고 다시 $D(1)$과의 검정을 반복한다. 만일 이때의 검정 통계량이 유의하면, <그림 7>에서처럼 다음 회의 기본분포 $D(2)$의 하한 $LB(2)=\min\{UI(x_i)\in\mathrm{Bin}_3\}$가 결정된다. 이는 Bin_1 구간까지는 분포의 밀도가 크지 않았으며, Bin_2에서야 기본 분포에 영향을 미칠 만큼 누적 밀도가 급증했다는 의미이다. 두 번째 분리조건에 의하여 Bin_2는 밀도가 높은 구간이므로 이 구간 내에서는 임계점을 설정하지 않는다. 그러므로 다음 회의 기본 분포에서 제외되며, Bin_1과 Bin_2는 우선적으로 UI_L 그룹에 포함된다. 마찬가지 방법으로 상한은 $D(1)$과 $D_U(1)=D(1)-\{\mathrm{Bin}_{15}\}$으로부터 비교 검정이 시작된다. 그림에서처럼 유의한 검정 통계량 값을 내는 $D_U(1)=D(1)-\{\mathrm{Bin}_{13}\cup\cdots\cup\mathrm{B}_{15}\}$이 발견되면 상한은 $UB(2)\max\{UI(x_i)=\in\mathrm{Bin}_{12}\}$로 정해진다. 여기서는 $\mathrm{Bin}_{13},\cdots,\mathrm{Bin}_{15}$이 UI_H 그룹에 우선적으로 포함되는 것이다. 구해진 상한과 하한으로부터 다음 회($k=2$)의 기본 분포 $D(2)$가 $\{\mathrm{Bin}_3\cup\cdots\cup\mathrm{Bin}_{12}\}$로 재설정된다. 재설정된 기본 분포 $D(2)$는 다시 B개의 등간격구간으로 나누어지며 위의 과정이 반복된다. 이러한 과정은 기본 분포 내에 포함된 패턴의 수가 충분히 작을 때까지 반복되며 최종 K회에서는 $\{LB(K)+UB(K)\}/2$를 임계점으로 하여, 패턴들이 분리된다. 이 과정을 통하여 임계점은 (1), (2)의 조건과 부합되도록 UI 분포 중심 부근의 저밀도구간에 설정된다.

상기한 패턴 분리과정을 정리하면 다음과 같다.

⓪ 다음의 기호를 정의한다.

- UI_H: $UI(x_i)$가 높은 패턴들의 그룹
- UI_L: $UI(x_i)$가 낮은 패턴들의 그룹,
  $UI(x_h)>UI(x_l)$, $x_h\in UI_H$, $x_l\in UI_L$
  $UI_H\cap UI_L=\varnothing$,
  $UI_H\cup UI_L=\{x_i,\forall_i\ i=1,\cdots,M\}$
- $D(k)$: k회의 기본 분포
- $LB(k)$: $D(k)$의 상한값
- $UB(k)$: $D(k)$의 하한값

① $UI(x_i)$에 따라 패턴들을 오름차순으로 정렬한다.

($i=1,\cdots,M$)

② 다음과 같이 초기화한다.

$k=1$,
$LB(1)=\min\{UI(x_i)\}$, $UB(1)=\max\{UI(x_i)\}$,
$D(1)=\{UI(x_i),\forall_i\}$.

③ $D(k)$를 B개의 등간격구간(equal spaced bin)으로 분할한다.

④ 다음 회의 하한 $LB(k+1)$을 설정하기 위하여 T-test를 수행한다.

(1) $b=1$로 초기화하고 다음 과정을 반복한다.

(2) $D(k)$와 $D(k)-\{\cup_{p=1}^{b}\mathrm{Bin}_p\}$에 대한 T-test 검정통계량 T를 구한다:

$$T=\ T\text{-test}\,[D(k),D(k)-\{\cup_{p=1}^{b}\mathrm{Bin}_p\}]$$

(3) 만약 $T\le t(\nu,\alpha_{sig})$이면, $b=b+1$로 하고 (2)로 돌아간다($b<B$).

(4) 아니면 다음 회의 하한이 결정한다:

$$LB(k+1)=\min\{UI(x_i)\in B_{b+1}\}$$

⑤ 다음 회의 상한 $UB(k+1)$을 설정하기 위하여 T-test를 수행한다.

(1) $b=B$로 초기화하고 다음 과정을 반복한다.

(2) $D(k)$와 $D(k)-\{\cup_{p=b}^{B}\mathrm{Bin}_p\}$에 대한 T-test 검정통계량 T를 구한다:

$$T=T\text{-test}\,[D(k),D(k)-\{\cup_{p=b}^{B}\mathrm{Bin}_p\}]$$

(3) 만약 $T\le t(\nu,\alpha_{sig})$이면, $b=b-1$로 하고 (2)로 돌아간다($b>1$).

(4) 아니면 다음 회의 상한을 결정한다:

$$UB(k+1)=\max\{UI(x_i)\in B_{b-1}\}$$

⑥ 기본 분포를 재설정한다.

(1) $k=k+1$

(2) $D(k)=\{LB(k)\le UI(x_i)\le UB(k)\}$

⑦ 만약 기본 분포의 원소의 개수가 θ보다 작으면 ($|D(k)|<\theta$, $\theta>2$) 다음을 수행하고 끝낸다.

(1) 임계점 λ를 계산한다:

$$\lambda=\frac{LB(k)+UB(k)}{2}$$

(2) λ를 중심으로 패턴들을 UI_L과 UI_H로 양분한다:

$$UI_L=\{x_i|UI(x_i)\le\lambda\}$$
$$UI_H=\{x_i|UI(x_i)>\lambda\}$$

⑧ 아니면 ③으로 돌아간다.

다음의 <그림 8>은 임의의 UI 분포를 생성하고 제안한 Drilldown T-test 방법에 의해 설정된 임계점을 표시한 것이다. 분포에 상관없이 임계점은 양분된 두 그룹의 돗수가 균형을 이루게끔 분포 중심 부근에 설정되었으며, 동시에 두 그룹이 최대한 상이하도록 저밀도구간에서 설정되었음을 알 수 있다.

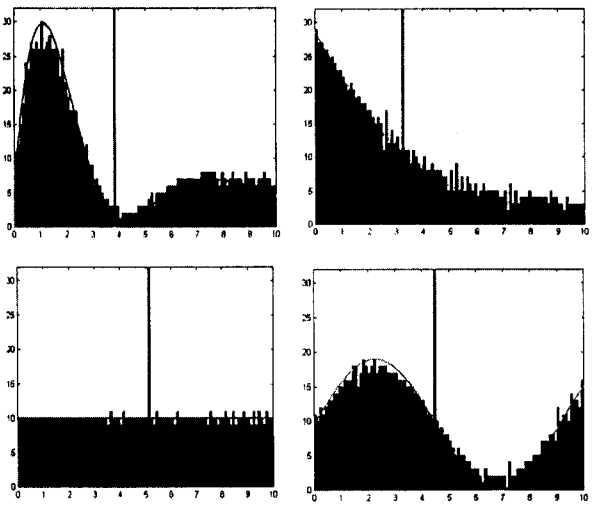


그림 8. Drilldown T-test에 의해 설정된 임계점.

4. 합성 데이터에 대한 실험 및 결과

제안된 효용지수 산정방법과 이에 근거한 패턴 분리방법의 타당성을 검증하기 위하여, 이를 합성 데이터 문제에 적용하였다. 4.1절에서는 분류문제에 대한 실험 및 결과를, 4.2절에서는 예측문제에 대한 실험 및 결과를 기술한다.

4.1 분류문제

제안된 알고리듬은 <그림 9>처럼 두 개의 합성문제 ⓐ, ⓑ에 적용되었다. ⓐ의 경우, 각각 일양분포 $U[0, 1]$을 따르는 (x_1, x_2)에 대하여 총 500개의 학습 패턴이 생성되었으며, 이 중 약 19%의 패턴들은 클래스 레이블이 바뀐 이상 패턴들이다. 클래스는 다음과 같이 정의되었다.

[문제 ⓐ]:

$$\text{class1} = \{(x_1, x_2) | x_2 > \sin(3x_1 + 0.8)^2\}$$

$$\text{class2} = \{(x_1, x_2) | x_2 \leq \sin(3x_1 + 0.8)^2\}$$

합성문제 ⓑ에서는 4개의 gaussian 분포로부터 총 600개의 패턴이 생성되었으며, 이들 중 약 10%가 이상 패턴들이다. 클래스 정의는 다음과 같다.

[문제 ⓑ]:

$$\text{class1} = \{(x_1, x_2) | N(C, 0.5^2 I), \; C = (1,1) \text{or} (-1,-1)\},$$

$$\text{class2} = \{(x_1, x_2) | N(C, 0.5^2 I), \; C = (-1,1) \text{or} (1,-1)\}.$$

두 문제의 차이점은 ⓐ는 분류경계 부근에 패턴들이 밀집해 있는 반면, ⓑ는 경계 부분에 가까울수록 패턴 밀도가 희박해진다는 데에 있다. 다음의 <그림 10>은 학습 패턴이 분류경

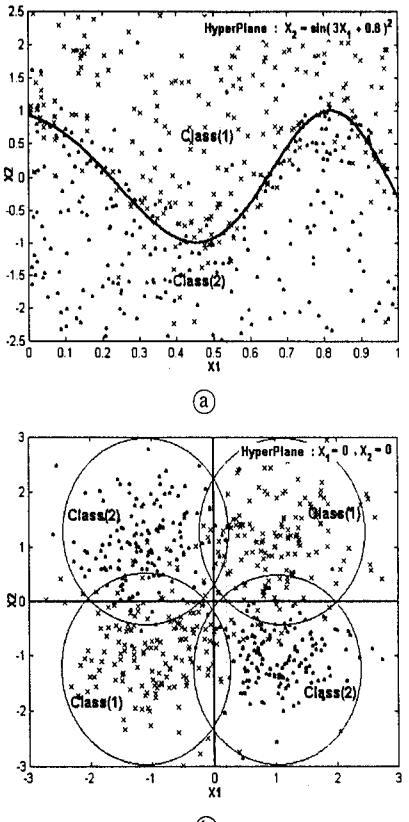


그림 9. 합성문제 ⓐ와 ⓑ, X : class(1), ● : class(2).

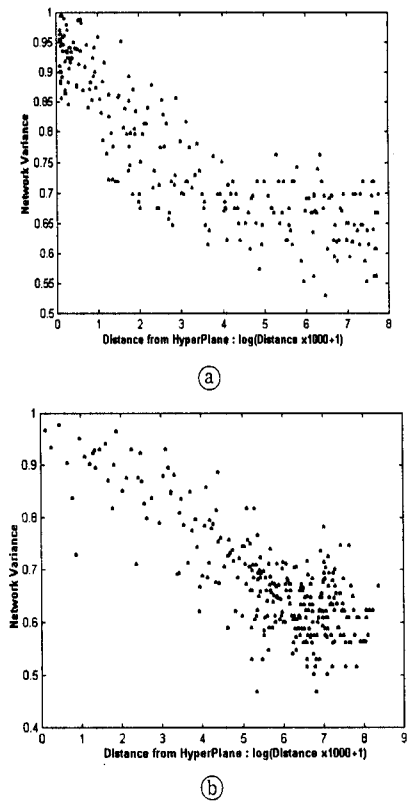


그림 10. 분류경계로부터의 거리(x)와 근접도(y)와의 관계.

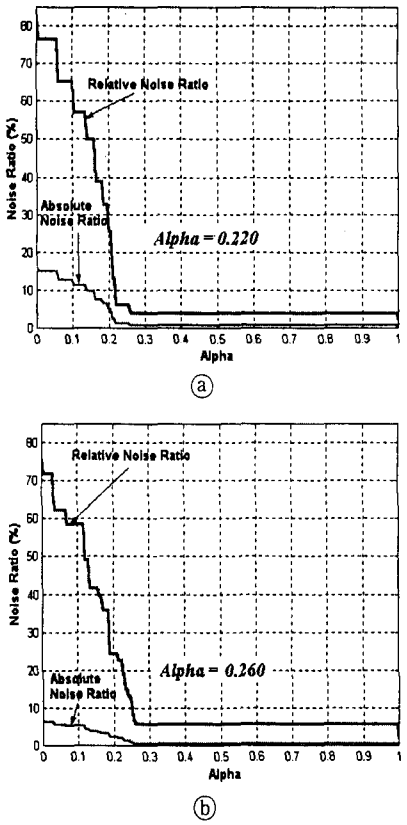


그림 11. 이상 패턴 비율변화에 따른 α 값의 결정.

계로부터 떨어진 거리와 앙상블의 근접도와의 관계를 나타낸 것이다. 분류경계에 가까울수록 근접도가 증가함을 알 수 있다.

다음의 <그림 11>은 효용지수 $UI(x_i)$의 선형조합 파라미터 α를 결정하기 위한 민감도 분석결과이다. α값은 상대적 이상 패턴 비율 및 절대적 이상 패턴 비율의 안정화가 시작되는 점에서 결정하였다. ⓐ에서는 $\alpha=0.220$으로 결정되었다. 이 경우, 효용지수 값 상위 196 ($=2\times19.6\%\times500$) 개의 패턴 내에 포함된 이상 패턴의 수는 5개 [상대적 이상 패턴 비율 : 5.1% ($=5/98\times100\%$), 절대적 이상 패턴 비율 : 1% ($=5/500\times100\%$)]였다. 마찬가지로 ⓑ에서는 $\alpha=0.260$에 의하여 정확도와 근접도를 선형조합한 결과, 효용지수 값 상위 120 ($=2\times10\%\times600$)개의 패턴 내에 포함된 이상 패턴의 수는 3개 [상대적 이상 패턴 비율 : 5.0% ($=3/60\times100\%$), 절대적 이상 패턴 비율 : 0.5% ($=3/600\times100\%$)]였다. 각 경우에 있어서 결정된 α 값은 분류경계 부근의 정상 패턴들이 높은 효용지수를 갖도록 하는 적절한 척도임을 알 수 있다.

다음의 <그림 12>는 패턴분리 방법(drilldown T-test)에 의하여 선택된 패턴그룹 UI_H를 나타낸 것으로 그림에서는 짙은 부분이 이에 해당된다. ⓐ의 경우에는 31.4%인 157개가 ⓑ의 경우에는 27.8%인 167개의 패턴이 선택되었다.

<그림 13>에서는 UI_H 그룹에 속한 패턴들을 원 패턴과 함께 대비시킨 것이다. 그림에서는 근접도와 정확도에 의하여 분류경계 부근의 패턴들 중 정상 패턴들이 선정되었음을 보여

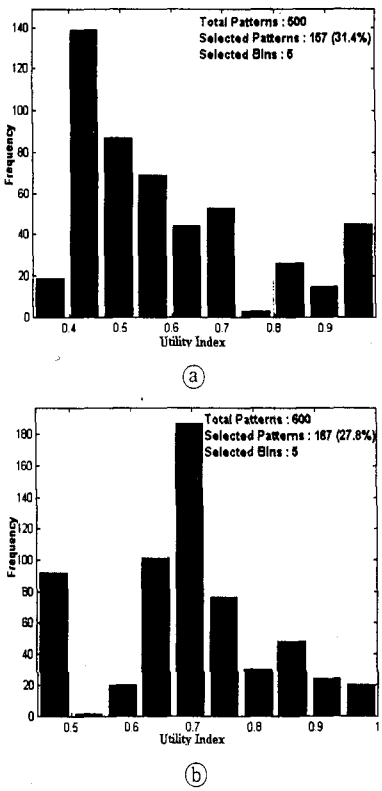


그림 12. 효용지수의 분포(UI)와 선택된 패턴그룹 (UI_H), 선택된 그룹은 진하게 표시됨.

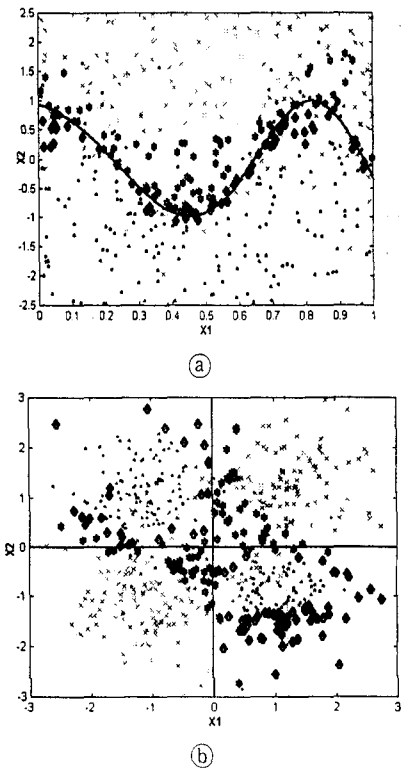


그림 13. 효용지수에 의하여 선택된 패턴들, 진하게 표시된 패턴들이 선택된 패턴들임.

준다.

패턴 선정 전, 후의 효과를 검증하기 위하여 대조실험을 실시하였다. 특히 다음과 같은 측면을 고려하였다.

(1) 선정된 패턴들은 분류모델의 성질과 상관없이 좋은 결과를 내는가.

(2) 선정된 패턴들로 학습시에도 과적합(over-fitting)에 대한 대응책이 필요한가.

(1)과 관련하여, 분류 모델로는 2-5-2 single-MLP와 30 2-5-2 bagging-MLP를 사용하였다. 이는 학습 셋의 변화에 민감하게 반응하는 불안정한(unstable) 학습 모델과 비교적 안정된(stable) 학습 모델에 대하여, 선정된 패턴들이 어떻게 영향을 미치는지를 알아보기 위한 설정이었다. (2)는 다음과 같은 이유에서 고려되었다. 패턴 선정을 하면 패턴 수 감소로 인하여 학습 셋과 검증 셋으로의 분할이 부적절할 수 있다. 그러나 선택된 학습 셋은 이미 정련된 패턴들로 구성되었으므로 분류모델이 과적합을 해도 무방하리라 기대되었다. 즉, 과적합을 방지하기 위한 검증 데이터 셋이 필요 없음을 시사하기 위한 설정이었다. 이를 위하여 두 분류모델 모두 과적합이 유도되도록 학습 파라미터를 조정하였다. 우선 은닉 노드의 수를 다소 많게 설정하였고, 많은 마찬가지 이유에서 epoch수도 30으로 하였다. 학습 알고리듬으로는 Levenberg-Maquardt를 사용하였는데, 주어진 문제와 학습 알고리듬을 고려해 보면, 설정된 epoch 수는 다소 과도한 설정이다. 분류 모델은 선정된 패턴으로 학습하는 경우에는 검증 셋 없이 실험을 실시하였으며, 전체 패턴으로 학습하는 경우에는 검증 셋(전체 패턴의 30%)이 있을 때와 없을 때를 구분하여 실험하였다. 테스트 패턴들은 ⓐ, ⓑ 모두 각각의 동일분포로부터 300개씩 생성되었으며 총 30회의 실험이 반복되었다.

<그림 14>는 두 문제 ⓐ, ⓑ에 대한 실험결과를 오분류(MCE : missclassification error)율의 평균과 편차로 나타낸 것이다. Single-MLP의 경우에는 선택된 패턴만으로 학습한 모델이 가장 좋은 성능을 보였으며, 특히 실험편차가 상당히 좁혀졌음을 알 수 있다.

Bagging-MLP의 경우에는 선택된 패턴만으로 학습한 모델의 분류성과가 약간 우수하였으나, 그 차이가 통계적으로도 유의하다고 볼 수는 없었다. 주목할 만한 사항은 원 학습 패턴을 모두 사용하는 경우, single-MLP에 있어서는 검증 셋의 설정이 효과적이었으나, bagging-MLP의 경우에는 오히려 성능을 감소시키는 결과를 가져온다는 것이다. 이는 bagging과 같은 알고리듬은 각 구성 네트워크들이 과적합을 할수록 이들간의 에러상관이 작아져서 결합(aggregation)의 효과가 크게 발생하게 되는데, 검증 셋을 설정하게 되면 이러한 효과가 감쇄되기 때문이다.

4.2 예측문제

마찬가지로 예측문제에 대해서도 합성함수 추정 문제에 대하여 실험하였다. 출력값 y의 노이즈는 정규분포를 따르되,

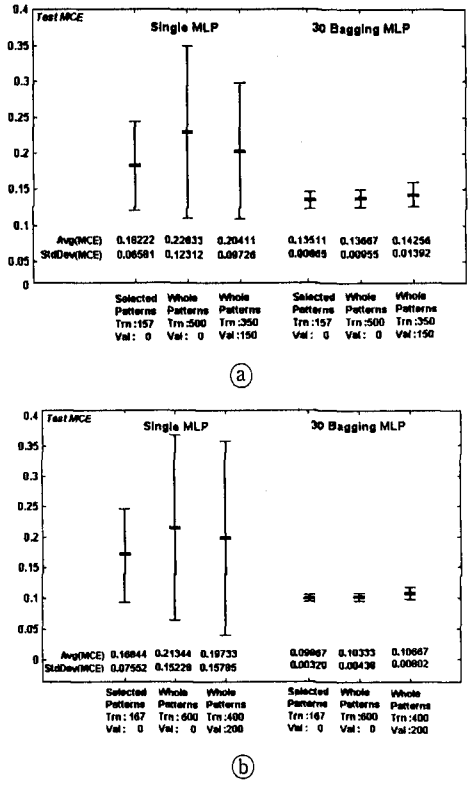


그림 14. 패턴 선택 전, 후에 대한 모델(Single MLP 및 Bagging MLP) 성능비교.

정상 패턴과 이상 패턴을 명확히 하기 위하여 분산의 크기가 다른 정규분포 변량 $\varepsilon_1, \varepsilon_2$를 생성한 후 이를 50:50으로 통합하였다.

$$\{(x, y_1) \cup (x, y_2) | x \sim U[0,1],$$
$$y_1 = \sin 3(x+0.8)^2 + \varepsilon_1,\ \ \varepsilon_1 \sim N(0.0, 0.3^2),$$
$$y_2 = \sin 3(x+0.8)^2 + \varepsilon_2,\ \ \varepsilon_2 \sim N(0.0, 0.8^2)\}$$

학습 패턴으로는 총 100개의 학습 패턴을 생성하였으며 검증 셋은 사용하지 않았다. 테스트 셋은 동일 분포를 따르는 80개의 패턴을 생성하였다. 앙상블 구성 네트워크의 수 $L=20$으로 각 네트워크의 네트워크 구조 및 학습 파라미터는 3.2절의 ①에 의해 랜덤화하였다. 학습 알고리듬으로는 Levenberg-Maquardt를 사용하였다. 다음의 <그림 15>는 원래의 학습 패턴들과 학습이 끝난 후 패턴들을 재대입한 L개의 추정선들을 점선으로 나타낸 것이다. 그림에서 $[x:0.1-0.3]$ 구간은 이상 패턴으로 인하여 과소추정(underestimation)이 나타나는 구간으로, 네트워크들의 추정선들이 실제 추정하고자 하는 실선보다 아래쪽으로 치우쳐져 있음을 알 수 있다.

다음의 <그림 16>의 ⓐ, ⓑ는 식 (6)에 포함된 두 항들의 적정성을 검사하기 위하여 단일 항에 의해서만 패턴 점수를 구한 후, 이 값이 큰 상위 40%의 패턴들을 묘사한 것이다. ⓐ는 정확도 ($\alpha=1$)에 의해서만 패턴 점수를 계산하였고. 선정된

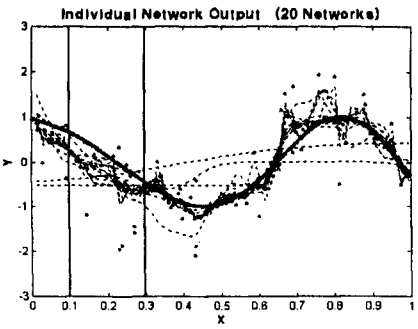
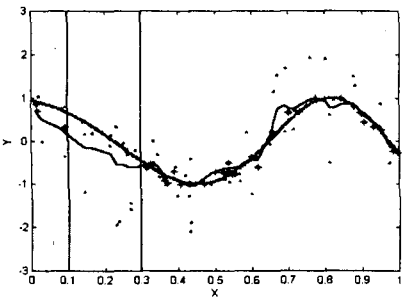
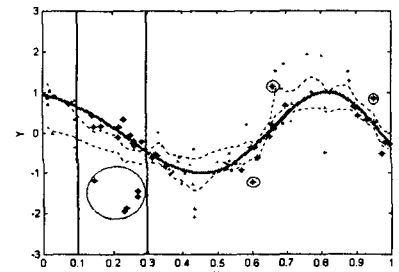


그림 15. 구성 네트워크별로 학습패턴(•)을 재대입한 추정선 $F_i(x_i)$ (---).



ⓐ



ⓑ

그림 16. ⓐ 정확도 $e^{-|y_i-\overline{F_i}|}$(—로 표시)에 의해 선정된 학습패턴들(*), ⓑ 근접도 $e^{-\Sigma_i}$ (--로 표시) 에 의해 선정된 학습패턴들(*).

패턴들은 (*)로, 네트워크들의 출력값 평균($\overline{F_i}$)은 실선으로 표시하였다. 정확도가 큰 패턴들은 전반적으로 추정선 부근의 정상 패턴들임을 알 수 있다. 그러나 [x:0.1−0.3] 구간에서는 정상 패턴이 많은데도 불구하고 정확도 값이 작아져서 이들이 선택되지 않았다. ⓑ는 근접도 ($\alpha=0$)에 의해서만 선정된 패턴들(*로 표시)을 나타낸다. 점선은 각 패턴 x_i에 대한 네트워크들의 출력값 분산 Σ_i를 나타낸다. 그림에서 보듯이 Σ_i에 의하여 선정된 패턴들도 정상 패턴일 가능성이 많음을 알 수 있다.

그러나 [x:0.1−0.3] 구간에서는 정상 패턴들뿐만 아니라 이상 패턴들도 다수 포함되었다(○로 표시). 이는 이 구간에서의 네트워크 출력값 분산이 작기 때문이다. 따라서 보다

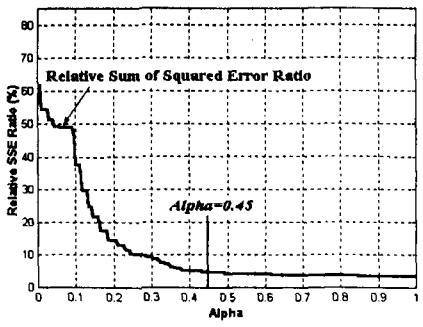


그림 17. α값과 상대적 에러제곱합의 비율.

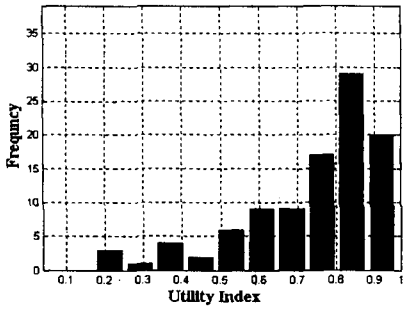


그림 18. 패턴 분리방법에 의해 선정된 패턴들(상위 3구간).

유용한 정상 패턴들을 선정하기 위해서는 정확도와 근접도를 동시에 고려해야 함을 알 수 있다.

이때의 선형조합 파라미터 α는 <그림 17>에서 나타내듯이 $\alpha=0.45$로 결정되었다. <그림 18>은 $\alpha=0.45$일 때의 효용지수를 이용하여 정렬한 후, 패턴 분리방법(drilldown T-test)에 의해 이들을 분리한 결과이다. 선정된 패턴의 수는 원 학습패턴의 57%이다.

<그림 19>는 선정된 학습 패턴들을 도시한 것이다. <그림 16>에서 문제시되었던 [x:0.1−0.3] 구간에서의 이상 패턴들은 제거되고, 정상 패턴들이 잘 선정되었음을 알 수 있다.

패턴 선정의 효과를 측정하기 위한 간단한 예비 검증단계로 1-7-1 single MLP를 예측모델로 설정하고 패턴 선정 전, 후의 실험 결과를 비교하였다. 분류문제의 실험에서와 마찬가지로 과적합(overfitting)을 유도하기 위하여 은닉노드 수와 학습 파

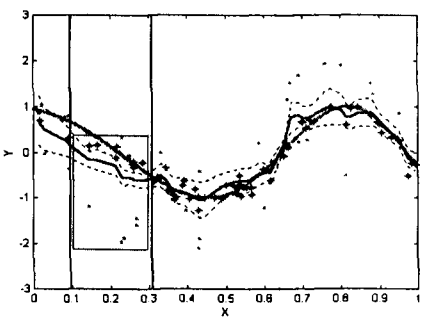


그림 19. 정확도(—) 및 근접도(--)를 함께 고려하여 선정된 학습패턴들(*).

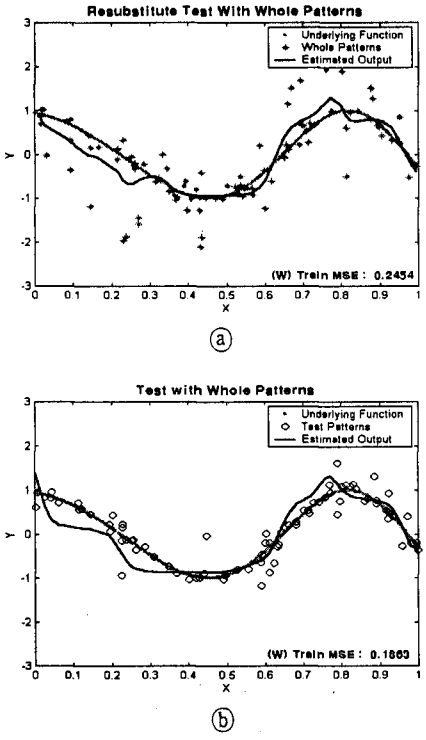


그림 20. 전체 학습 패턴(UI)으로 학습한 모델의 추정선 (진한 실선으로 표시) : ⓐ 재대입 추정선, ⓑ 테스트 추정선.

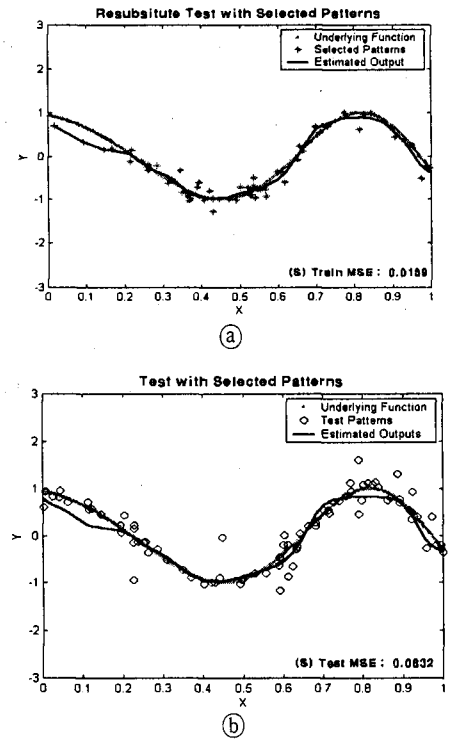


그림 21. 선택된 학습 패턴(UI_H)으로 학습한 모델의 추정선 (진한 실선으로 표시) : ⓐ 재대입 추정선, ⓑ 테스트 추정선.

표 1. 합성문제에 대한 패턴 선정 전, 후 비교

MSE($\times 10^{-2}$)	패턴 선정 전	패턴 선정 후
Train MSE	24.54	1.89
Test MSE	18.63	8.32

라미터의 값을 과도하게 설정하였다. <그림 20>은 전체 학습 패턴(100개)으로 학습한 예측 모델의 추정선을 나타낸다. 즉, 그림의 진한 실선이 예측모델의 추정선으로, ⓐ에서는 학습 패턴을 재대입한 추정선을, ⓑ에서는 테스트 패턴에 대한 추정선을 나타낸다. <그림 21>에서는 선택된 패턴(57개)으로 학습한 후 마찬가지 방법으로 각 추정선들을 나타낸 것이다.

<그림 20>과 <그림 21>을 비교해 볼 때, 과적합의 우려가 있는 상황에서도 패턴 선정이 잘 이루어지면 우수한 결과가 산출됨을 가시적으로 확인할 수 있다. 이는 제안한 패턴 선정방법이 유용한 패턴들을 미리 정련하는 전처리 과정이기 때문이다. 패턴 선정 전, 후의 성능을 mse(mean squared error)로 비교하면 <표 1>과 같다. 따라서 제안한 방법은 분류문제에서와 마찬가지로, 예측문제에서도 모델의 복잡도에 관계없이 보다 적은 수의 패턴으로 좋은 성능을 산출함을 알 수 있다.

5. 실제 데이터에 대한 실험 및 결과

제안된 패턴 선택방법은 각 문제별 실제 데이터에 대해서도 적용되었다.

5.1 분류문제

분류문제에 대한 실제 데이터 셋은 다음과 같다 : Breast Cancer, PimaIndian(UCI Repository Of Machine Learning Databases). 분류 모델로는 9-15-2 single MLP와 30 9-15-2 bagging-MLP를 사용하였다(PimaIndian은 8-15-2 MLP). 4.1절의 합성문제의 결과를 참조하여 실험설정을 다소 조정하였다. 즉, 전체 패턴을 모두 사용하는 경우, single-MLP에서는 검증 셋을 설정하였고 bagging-MLP에서는 검증 셋을 설정하지 않았다. 실험은 두 문제 각각의 테스트 패턴들(BreastCancer: 205개, PimaIndian :230개)에 대하여 30회씩 반복되었다. 효용지수의 선형조합 파라미터 α는 BreastCancer에서는 $\alpha=0.05$로, PimaIndian에서는 $\alpha=0.15$로 결정되었다. 패턴 선택 결과, BreastCancer의 경우에는 302개(전체 패턴의 63.2%)가, PimaIndian의 경우에는 308개(전체 패턴의 57.2%)가 선택되었다. <표 2>는 실험결과를 정리한 것이다.

Single-MLP 결과를 살펴보면, 합성데이터에 대한 실험결과와 마찬가지로 선택된 패턴만으로 학습한 모델의 실험 평균이 우수하였다. 특히 실험 편차 면에서는 선택된 패턴들로 학습

표 2. 실제 분류문제에 대한 실험결과

MissClassification Error Rate		Breast Cancer		Pima Indian	
# of Patterns		478	302	538	308
Single MLP	Avg ($x10^{-2}$)	9.188	6.147	30.217	28.971
	Std ($x10^{-2}$)	11.183	1.841	10.247	5.327
	P-value	(T-test) 0.1519 (F-test) 0.0000		(T-test) 0.5576 (F-test) 0.0007	
Bagging MLP	Avg ($x10^{-2}$)	4.976	4.163	22.256	22.957
	Std ($x10^{-2}$)	0.620	0.611	2.535	2.523
	P-value	(T-test) 0.0001 (F-test) 0.9374		(T-test) 0.2882 (F-test) 0.9797	

한 경우에 상당히 안정적인 결과를 얻을 수 있었다(F-test P-value: 0.0000, 0.0007). Bagging-MLP의 결과에서는 패턴 선택 전, 후의 성능 차이가 거의 없었다. 그러나 보다 적은 수의 학습 패턴으로 동등한 일반화 성능이 산출되었음을 알 수 있다.

5.2 예측문제

예측문제에 사용된 실제 데이터는 다음과 같다: Boston Housing, Ozone(UCI Repository Of Machine Learning Databases). 예측 모델로는 13-10-1 single MLP와 25 13-10-1 bagging-MLP를 사용하였다(Ozone은 8-7-1 MLP). 총 패턴 수는 각각 506개, 330개이고, 이 중 30%(152개, 100개)가 테스트 셋으로 남겨졌다. 효용지수의 선형조합 파라미터는 두 데이터 셋 모두 $\alpha=0.5$로 결정되었다. 이 후 패턴 분리방법에 의하여 학습 패턴의 81.07% (287개), 83.91%(199개)가 각각 선정되었다. 4.2절과 마찬가지로 전체 패턴 셋에 대하여 학습하는 경우에 한해서만 검증 셋을 설정하였다(학습 패턴수의 30%). <표 3>은 각 문제의 테스트 셋에 대한 30회 반복 실험결과를 정리한 것이다.

Single-MLP 결과를 살펴보면, mse의 실험평균에 있어서 패턴 선택 전보다 패턴 선정 후가 우수한 성능을 산출함을 알 수 있었고, 그 차이는 대체로 통계적으로도 유의하다(T-test P-value: 0.0017, 0.0283). 또한 분류문제의 결과에서와 같이 실험편차가 확연히 안정화되었다(F-test P-value: 0.0001, 0.0008).

Bagging-MLP의 결과에서는 패턴 선택 후가 선택 전의 결과보다 실험평균상 우위하다는 결론을 얻진 못했다(T-test P-value: 0.8465, 0.3466). 그러나 실험편차상에서는 보다 안정되거나 적어도 비슷한 안정성을 보임을 알 수 있었다(F-test P-value: 0.0001, 0.6260).

표 3. 실제 예측문제에 대한 실험결과

Mean Squared Error		Boston Housing		Ozone	
# of Patterns		354	287	230	199
Single MLP	Avg ($x10^{-4}$)	131.277	81.56	257.47	204.20
	Std ($x10^{-4}$)	75.75	25.70	113.99	59.63
	P-value	(T-test) 0.0017 (F-test) 0.0001		(T-test) 0.0283 (F-test) 0.0008	
Bagging MLP	Avg ($x10^{-4}$)	65.83	65.19	154.09	148.66
	Std ($x10^{-4}$)	15.73	8.97	21.14	23.16
	P-value	(T-test) 0.8465 (F-test) 0.0001		(T-test) 0.3466 (F-test) 0.6260	

6. 결론 및 추후 연구과제

본 연구에서는 앙상블 출력값의 편기와 분산을 이용하여 분류(classification) 및 예측(regression) 문제에 유용한 학습 패턴을 찾아내는 방법을 제안하였다. 유용한 학습 패턴이란 정확한 패턴들 중 특히 모델의 학습에 많은 양의 정보를 전달하는 패턴들이다. 분류문제에서는 클래스들간의 분류경계면에 근접한 패턴들이 클래스의 내부에 분포한 패턴들보다는 유용하며, 예측문제에 있어서는 추정면에 근접한 패턴들이 멀리 떨어져 있는 패턴들보다는 유용하다.

본 연구의 2절에서는 패턴의 유용성을 효용지수로 정의하는 방법에 대하여 소개하였다. 효용지수는 근접도와 정확도로 구성된다. 근접도는 패턴이 분류경계면 또는 추정면에 가까운 정도를 측정하며, 정확도는 패턴이 정상인지에 대한 척도이다. 분류문제에서의 근접도는 패턴에 대한 앙상블 출력값들의 분산이 클수록 증가한다. 예측문제에서의 근접도는 앙상블 출력값들의 분산이 작을수록 증가한다. 정확도는 두 문제의 경우 공통적으로, 패턴에 대한 앙상블 출력값들의 편기가 작을수록 크다.

3절에서는 효용지수를 근거로 하여 전체 학습 셋으로부터 일부 패턴을 선정하는 방법에 대하여 기술하였다. 이는 효용지수의 분포를 양분하는 drilldown T-test에 의해 이루어진다. 제안한 방법은 (1) 선정된 학습 데이터의 수가 지나치게 많거나 적게 되는 것을 방지하면서 (2) 분리된 두 그룹을 최대한 상이하게 만드는 원칙 하에 고안되었다.

4절에서는 제안된 패턴 선정방법이 각 문제에 대하여 의도한 패턴들을 발췌하는가를 확인하기 위하여 합성문제에 대한 실험을 실시하였다.

5절에서는 패턴 선택 전, 후의 일반화 성능을 비교하기 위한 실제문제에 대하여 실험하였다. 그 결과 첫째, 제안된 방법은 분류문제에서는 분류경계면 부근의 정상 패턴들을, 예측문제에서는 이상 패턴들이 제거된 추정면 부근의 패턴들을 선정하는 방법임이 검증되었다. 둘째, 선정된 학습 패턴 셋으로 학습하는 경우에는 과도하게 복잡한 학습 모델이라도 과적합을 하지 않는다는 것을 보였다. 이는 패턴 수 감소로 인하여 검증 셋의 설정이 어려울 수도 있다는 우려를 해소시킨다. 동시에 문제의 복잡도에 적합한 모델의 구조나 학습 파라미터 결정을 위하여 시행착오를 하지 않아도 된다는 의미이기도 하다. 셋째, 제안된 방법은 불안정한 학습 모델의 성능을 상당히 개선

시켰다. 특히 반복실험의 실험편차가 크게 안정되었으며, 이는 모델의 신뢰도가 향상될 수 있음을 의미한다. 마지막으로 제안된 방법은 보다 적은 수의 학습 패턴으로, 패턴 선정 전의 모델 성능을 개선시켰거나 적어도 동등한 결과를 산출함을 보였다. 이는 제안한 전처리 과정을 통하여 이후 학습에 소요될 메모리 및 시간, 계산의 복잡도를 감소시킬 수 있다는 것을 말한다. 특히 SVM과 같이 패턴의 수와 모델의 성능이 밀접하게 관련되어 있다거나, 여러 모델들간의 성능을 비교한다거나 반복 실험하는 경우에는 제안한 방법의 활용도가 증대된다.

제안된 패턴 선정방법은 효용지수 산정 과정에서 사용한 신경망 알고리듬 대신 학습 시간이 짧은 decision tree 등의 알고리듬을 사용하면 보다 나은 성과를 얻을 수 있으리라 기대된다. 또 하나의 추후 연구과제로는 효용지수에 대한 근접도와 정확도의 반영방법과 관련된다. 본 연구에서 제안한 선형조합 방법 외의 다른 대안으로서, 각각을 순차적으로 적용하는 방법을 생각할 수 있다.

참고문헌

Bishop, C. M. (1995), Neural Networks For Pattern Recognition, Oxford University Press, New York, 386-439.

Breiman, L. (1996a), Bagging Predictors, *Machine Learning*, **24**, 123-140.

Breiman, L. (1996b), Bias, Variance, and Arcing Classifiers, Technical Report 460, Department of Statistics, University of California, Berkeley, CA.

Burges, C.J.C (1998), A Tutorial on Support Vector Machines for Pattern Recognition , *Data Mining and Knowledge Discovery*, **2**, 121-167.

Cachin, C. (1994), Pedagogical Pattern Selection Strategies, *Neural Networks*, **7**(1), 175-181.

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

Drucker, E. (1997), Improving Regressors Using Boosting Techniques, *The Fourteenth International Conference on Machine Learning*, 107-115.

Drucker, E. (1999), Boosting Using Neural Networks, In Amanda J. C. Sharkey (Eds), Combining Artificial Neural Nets :Ensemble and Modular Learning, Springer-Verlag, 51-77.

Foody, G. M. (1999), The Significance of Border Training Patterns in Classification by a Feedforward Neural Network Using Back Propagation Learning, *International Journal of Remote Sensing*, **20**(18), 3549-3562.

Freund, Y., Schapire, R. E. (1997), A Decision-Theoretic Generalization of On-line Learning and an Application to Boosting, *Journal of Computer and System Sciences*, **55**(1), 119-139.

Gunn, S. (1998), Support Vector Machines for Classification and Regression, ISIS Technical Report.

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

Haykin, S. (1999), Neural Networks: A Comprehensive Foundation, Macmilan, New York, 351-390.

Hearst, M. A. (1998), Support Vector Machines, *IEEE INTELLIGENT SYSTEM*, 167-179.

Krogh, A. and Vedelsby, J. (1995), Neural Network Ensembles, Cross Validation, and Active Learning, In: Tesauro, G., Touretzky, D. S. and Leen. T. K. (Eds), *Advances in Neural Information Processing Systems 7*, Cambridge, MA: MIT Press, 231-238.

Kwok, J. T. (1999), Moderating the Outputs of Support Vector Machine Classifiers, *IEEE Transactions on Neural Networks*, **10**(5), 1018-1031.

Lee, C. and Landgrebe, D. A. (1997), Decision Boundary Feature Extraction for Neural Networks, *IEEE Transactions on Neural Networks*, **8**(1), 75-83.

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

Mackay, D. J. C. (1992), Bayesian Interpolation, *Neural Computation*, **4**, 415-447.

Mitchell, T. M. (1997), Machine Learning, McGRAW-HILL International Editions (Computer Science Series), 81-127.

Parmanto, B., Munro, P. W. and Doyle, H. R. (1996), Reducing Variance of Committee Prediction with Resampling Techniques, *Connetion Science*, **8**, 405-425.

Perrone, M. P. (1993a), Improving Regression Estimation: Averaging Methods for Variance Reduction with Extension to General Convex Measure Optimization, PhD Thesis, Department of Physics, Brown University, Providence, RI.

Perrone, M. P. and Cooper, L. N. (1993b), When Networks Disagree: Ensemble Methods for Hybrid Neural Networks, Artificial Neural Networks for Speech and Vision, Chapman and Hall, London.

Plutowski, M. and White, H. (1993), Selecting Concise Training Sets from Clean Data, *IEEE Transactions on Neural Networks*, **4**(2), 305-318.

Plutowski, M. (1994), Selecting Training Examplars for Neural Network Learning, Ph.D. Dissertation, Univ. California, San Diego.

Qu, D., Wong, P. M., Cho, S. and Gedeon, T. D. (2001), A Hybrid Intelligent System for Improved Petrophysical Predictions, *to appear in ICONIP proceedings*.

Röbel, A. (1994), The Dynamic Pattern Selection Algorithm: Effective Training and Controlled Generalization of BackPropagation Neural Networks, Technische Univ. Berlin, Germany, Technical Report.

Sharkey, A. J. C. (1996), On Combining Artificial Neural Nets, *Connection Science*, **8**, 299-313.

Sharkey, A. J. C. (1997), Combining Diverse Neural Nets, *The Knowledge Engineering Review*, **12**(3), 231-247.

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

UCI Repository Of Machine Learning Databases, http://www.ics.uci. edu/~mlearn

Vincent, P. and Bengio, Y. (2000), A Neural Support Vector Network Architecture with Adaptive Kernels, *IEEE Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks*, 187-192.

Zhang, B. T. (1993), Learning by incremental Selection of Critical Examples, Arbeitspaper der GMD, No. 735, German National Research Center for Computer Science (GMD), St Augustin /Bonn.

Zhang, B. T. (1994), Accelerated Learning by Active Example Selection, *Incremental Journal of Neural Systems*, **5**(1), 67-75.