The Korean Communications in Statistics
Vol. 13 No. 1, 2006, pp. 151-165

Estimating Prediction Errors in Binary Classification Problem: Cross-Validation versus Bootstrap

Ji-Hyun Kim[1] and Eun-Song Cha[2]

Abstract

It is important to estimate the true misclassification rate of a given classifier when an independent set of test data is not available. Cross-validation and bootstrap are two possible approaches in this case. In related literature bootstrap estimators of the true misclassification rate were asserted to have better performance for small samples than cross-validation estimators. We compare the two estimators empirically when the classification rule is so adaptive to training data that its apparent misclassification rate is close to zero. We confirm that bootstrap estimators have better performance for small samples because of small variance, and we have found a new fact that their bias tends to be significant even for moderate to large samples, in which case cross-validation estimators have better performance with less computation.

Keywords : Generalization Error; Prediction Accuracy; Classification Tree; Boosting

1. 서론

신용카드의 부정사용 여부를 가리거나, 환자의 질병을 진단하는 문제와 같은 분류(classification) 문제에서, 좋은 분류규칙 또는 분류기(classifier)를 얻기 위해서는 분류기의 성능을 알아야만 한다. 분류기의 성능을 판단하는 기준을 미래의 자료에 대한 예측오차(prediction error) 또는 참오분류율(true misclassification error)로 정했을 때, 훈련자료(training data)와 독립적인 검증자료(test data)가 있다면 이 자료를 이용하여 참오분류율을 추정할 수 있다. 하지만 자료가 충분하지 않을 때 별도의 검증자료를 갖기가 힘들며, 이런 경우 참오분류율을 추정하기 위한 전통적인 방법은 교차타당성(cross-validation, 이하 CV로 표시)에 의한 방법이다. 한편 Efron (1983)과 Efron and Tibshirani (1997)는 붓스트랩(bootstrap)을 이용한 추정방법을 제안하였다.

이 연구에서는 검증자료가 없는 상황에서 분류기의 성능을 평가하고자 할 때, 교차타당성과 붓스트랩 추정량 중 어떤 것을 쓰는 것이 좋은가를 실험적(empirically)으로 알아보고자 한다. 교차타당성 추정량은 표본의 크기가 작을 때 분산이 크다고 알려져 있다

1) Professor, Dept. of Statistics, Soongsil University, Dongjak-Ku Sangdo-Dong, Seoul 156-743, Korea. Correspondence : jhkim@stat.soongsil.ac.kr
2) Graduate student, Dept. of Statistics, Soongsil University, Dongjak-Ku Sangdo-Dong, Seoul 156-743, Korea.

(Efron (1983)). 붓스트랩 추정량은 표본크기가 작을 때 분산이 작아 교차타당성 추정량에 비해 경쟁력을 가진다는 실험적인 증거가 제시되었다(Efron (1983), Efron and Tibshirani (1997), Crawford (1989), Merler and Furlanello (1997)). 그러나 이론적 근거가 약하고 다양한 표본크기와 분포에서 충분한 실험이 이루어지지 않았다. 그리고 기존의 연구들은 선형모형이나 나무모형을 고려하였으며, 부스팅(boosting)과 같이 예측력이 높으며 과적합이 쉽게 일어나지 않는 분류기의 성능을 평가하는 방법에 대한 연구가 지금까지 없었다. 부스팅 기법에 의한 분류기를 쓰면 훈련자료에 대한 오분류율이 대부분 0에 가깝게 되는데, 이럴 때 분류기의 참오분류율을 추정하기 위해 교차타당성과 붓스트랩 추정방법 중에서 어떤 것을 써야 하는가에 대해 실험적 판단근거를 제시하고자 한다.

논문의 구성은 다음과 같다. 2절에서 연구주제를 설명하고 참오분류율을 추정하는 방법에 대해 정리하였고, 3절에서 모의자료를, 4절에서 실제자료를 이용하여 실험하였다. 5절에서 결과를 요약하고 향후 연구 과제를 제시하였다.

2. 예측오차와 추정방법

반응변수와 설명변수에 대한 통계적 모형 $y=f(x)+\epsilon$ 을 가정하고, 관측자료 (y_i, x_i), $i=1,\cdots,n$ 로부터 적합모형(fitted model) $\hat{y}=\hat{f}(x)$ 을 얻었다고 하자. 이 때 적합모형의 성능을 평가하는 측도로 예측오차(prediction error) $E(Y_0-\hat{Y}_0)^2$ 를 쓰기로 한다. 위 식에서 (X_0, Y_0) 는 $\hat{f}$ 을 얻을 때 쓰인 자료가 아닌 독립적인 자료이며 $\widehat{Y_0}=\hat{f}(X_0)$ 이다. 그리고 기대값 E 는 $\hat{f}$ 이 고정되었을 때 새로운 관측값 (X_0, Y_0) 에 대한 기대값이다. 관측값 Y_0 와 적합값 $\widehat{Y_0}$ 가 0 또는 1이 되는 두 계급 분류 문제에서는 $E(Y_0-\widehat{Y_0})^2=P(Y_0\neq\widehat{Y_0})$ 로서, 예측오차가 곧 참오분류율(true misclassification rate)이 된다. 지금부터 두 계급 분류 문제에 대해서만 논의하기로 한다.

모형적합과 모형의 성능평가에 모두 같은 자료를 써서 얻어지는 겉보기오분류율(apparent misclassification rate)은 참오분류율을 과소추정한다고 알려져 있다. 이를 해결하기 위해 가장 널리 쓰이는 것이 교차타당성에 의한 추정방법인데 보통 주어진 자료를 10등분하는 10겹(10-fold) CV를 쓴다 (Kohavi (1995)). 그 외 붓스트랩 표본을 이용하는 추정방법이 있는데 이를 간략히 설명한다.

참오분류율과 겉보기오분류율을 각각 Err, err^0 로 표시할 때 $Err=err^0+(Err-err^0)$ 라고 쓸 수 있다. Efron (1983)은 겉보기오분류율의 편향을 다음과 같이 수정한 추정량을 제안하고 .632 붓스트랩 추정량이라고 이름 지었다.

$$\begin{aligned} err^{.632} &= err^0+.632(\epsilon_0-err^0) \\ &= .368err^0+.632\epsilon_0 . \end{aligned}$$

위 식에서 ϵ_0 는 붓스트랩 표본을 이용하여 참오분류율을 추정한 값이다(정확한 식과 계산방법은 Efron (1983)과 Efron and Tibshirani (1997)를 참조. 이 연구에서는 재추출하는 붓스트랩 표본의 수를 Efron and Tibshirani (1997)의 제안에 따라 50으로 하였다).

각 붓스트랩 표본의 크기는 비록 n 이지만 서로 다른 자료의 개수는 평균적으로 $.632n$ 개이다. 따라서 붓스트랩 표본에 의해 추정된 모형은 서로 다른 n 개의 자료로 추정된 모형보다 정확도가 떨어지게 되어 ϵ_0 는 Err 를 과대추정하게 되며, $(\epsilon_0 - err^0)$는 $(Err - err^0)$보다 큰 값을 갖게 된다. Efron (1983)은 그 크기를 .632배만큼 줄여준 추정량인 $err^{.632}$ 를 제안하였다. 그러나 최대깊이까지 허용하는 나무모형이나 부스팅 기법에 의한 분류기와 같이, 겉보기오분류율 err^0 이 0에 가까운 분류기를 쓸 때 $err^{.632}$ 가 참오분류율을 과소추정하게 되는 문제가 발생한다. 이에 Efron and Tibshirani (1997)는 err^0 와 ϵ_0 에 대한 가중값을 가변적으로 조정하는 다음과 같은 추정량을 제안하고 .632+ 붓스트랩 추정량이라고 이름 지었다.

$$err^{.632+} = (1-\hat{w})err^0 + \hat{w}\,\epsilon_0 .$$

위 식에서 $\hat{w} = .632/(1-.368\hat{R})$, $\hat{R} = (\epsilon_0 - err^0)/(\hat{\gamma} - err^0)$ 이다. $\hat{\gamma}$ 는 아무런 정보가 없을 때, 즉, X 와 Y 가 전혀 관련성이 없을 때의 오분류율에 대한 추정값으로서, 두 계급 분류 문제에서는 $\hat{p}$ 를 n 개의 관측값 y 중에서 1의 비율이라고 하고 $\hat{q}$ 를 추정값 $\hat{y}$ 중에서 1의 비율이라고 할 때 $\hat{\gamma} = \hat{p}(1-\hat{q}) + (1-\hat{p})\hat{q}$ 이 된다. 가중값 $\hat{w}$ 는 0.632부터 1 사이의 값을 가지는데, 정보가 없어 ϵ_0 가 $\hat{\gamma}$ 에 가까워질수록 1에 가까운 값을 갖게 되어 결과적으로 $err^{.632+}$ 가 ϵ_0 에 가깝게 된다. 반대로 $\hat{R}$ 이 0일 때는 $err^{.632+}$ 는 $err^{.632}$ 와 같게 된다.

교차타당성 추정량의 편향은 문제가 되지 않으나 작은 표본에서 분산이 커진다고 알려져 있다(2겹이나 5겹 교차타당성 추정량의 편향은 문제가 된다(차은송 (2005)). 한편 .632 붓스트랩 추정량 $err^{.632}$ 는 겉보기오분류율 err^0 이 0에 가까운 값을 가지면 참오분류율 Err 을 과소추정하게 되며, 이를 수정한 .632+ 추정량 $err^{.632+}$ 는 $err^{.632}$ 에 비해 편향은 작아지지만 분산이 커진다는 것이 관찰되었다(Efron and Tibshirani (1997)). 두 붓스트랩 추정량의 점근적 성질에 대한 증명은 제시되지 못했다.

본 연구에서는 부스팅(boosting) 기법을 이용한 분류기의 성능을 추정하는 문제에 초점을 맞추었다. 기계학습(machine learning) 전문가들이 개발한 부스팅 기법은 약한 분류기(weak classifiers)들을 결합하여 강한 분류기를 만들어내는 기법이다(Freund and Schapire (1997)). AdaBoost 또는 이산형 AdaBoost (Discrete AdaBoost)라고도 불리는 이 알고리즘은, 주어진 자료의 각 관측값에 다른 가중값을 부여하여 약한 분류기를 적합시키는데, 이전 분류기가 잘못 분류한 관측값에 보다 큰 가중값을 부여한다. 이렇게 축차적으로 만들어진 약한 분류기의 가중 선형결합으로 최종분류기를 생성하는데, 약한 분류기의 분류성능에 비례하여 가중값을 둔다. 약한 분류기로서 나무모형을 이용하면, 부스팅 기법에 의해 생성되는 최종분류기는 많은 나무모형들의 가중 선형결합으로 표현된다. 이 최종분류기는 하나의 나무모형에 비해 분류성능을 효과적으로 높여주는 것으로 알려져 있다(Bauer and Kohavi (1999)).

부스팅 기법에 의한 분류기의 특성 중 하나는 훈련표본에 대한 오분류율, 즉 겉보기오분류율이 아주 낮다는 것인데, 본 연구의 목적이 겉보기오분류율이 심하게 편향되어 있을 때 참오분류율의 추정량들을 비교하고자 하는 데에 있기 때문에 부스팅 기법에 의한

분류기를 실험대상으로 선택하였다. 겉보기오분류율이 낮은 또 다른 분류기인 최대깊이를 허용하는 나무모형도 실험대상에 포함시켜 비슷한 결과가 나오는지를 살펴보았다.

3. 모의실험

모의실험을 통해 다음 사항들을 알아보고자 하였다.

- 기존 연구(Efron (1983), Efron and Tibshirani (1997), Crawford (1989))의 제한된 모의실험 결과를 보면, 소표본에서 선형모형이나 나무모형에 의한 분류기의 오분류율을 추정할 때 붓스트랩 추정량이 교차타당성 추정량보다 대체로 더 우수한 성능을 갖는다. 부스팅 기법에 의한 분류기에 대해서도 과연 그러할까?
- 표본의 크기가 커지면 어떻게 되는가? 성능의 차이가 줄어들게 되는가?
- 궁극적으로, 부스팅 기법을 이용하여 구축한 분류기의 참오분류율을 추정할 때, n겹 교차타당성(leave-one-out CV)과 10겹 교차타당성 추정량, 그리고 .632와 632+ 붓스트랩 추정량 중에서 어떤 추정량을 쓰는 것이 좋은가?

모의실험을 위해 R(R Development Core Team (2004))을 이용하였다. 나무모형을 적용하기 위해 rpart package (Therneau and Atkinson (1997))를, 부스팅 기법에 의한 분류기는 Discrete Adaboost 알고리즘(Freund and Schapire (1996))을 썼으며, 붓스트랩 추정량은 bootstrap package (Efron and Tibshirani (1993))의 입출력 부분을 수정하여 적용하였다.

3.1 모의실험 I

먼저 두 개의 설명변수가 있는 비교적 간단한 경우를 고려하였다. 서로 독립인 두 설명변수 X_1과 X_2는 구간 (0,1)에서 균일분포를 따르며, 두 설명변수의 값이 모두 0.5보다 작거나 모두 0.5보다 클 때 반응변수 Y의 값은 1을 갖고, 그렇지 않은 경우에는 값 0을 갖는다. 이 자료는 어느 한 변수만으로는 분류가 되지 않는다는 특성을 갖고 있다.

불확실성을 주기 위해 정해진 영역 (X_1, X_2)에서 정해진 Y의 값이 될 확률 p를 0.5부터 1.0 사이의 값을 갖도록 하여($p = .5, .6, .7, .8, .9, 1.0$) 자료를 생성하였으며, 자료의 크기 n은 20부터 400까지 변화시켰다($n = 20, 50, 100, 200, 300, 400$).

구현된 R 프로그램은

(1) 훈련자료를 생성하는 부분

(2) 만든 훈련자료에 나무모형과 부스팅 기법을 적용하여 분류기를 구축하는 부분

(3) 분류기의 참오분류율을 추정하는 부분, 즉 (2)에서 생성한 분류기의 참오분류율을 독립적인 검증자료없이 교차타당성이나 붓스트랩에 의해 추정하는 부분

(4) 추정량의 성능을 비교하기 위해서는 참오분류율을 알아야 한다. 이를 위해 크기 5000의 독립적인 검증자료를 생성하여 (2)에서 생성한 분류기의 참오분류율을 구하는 부분으로 이루어져 있으며, 각각의 p와 n에서 100번씩 반복한다.

여러 추정량의 성능을 훈련자료에 따른 $(\widehat{Err} - Err)$의 분포를 이용하여 비교하고자

한다. 이 때 Err 은 n 개의 특정한 훈련자료가 주어졌을 때의 조건부 참오분류율로서, 훈련자료가 바뀔 때마다 달라진다. $\widehat{Err}$ 은 특정한 훈련자료가 주어졌을 때 Err 의 추정량으로서, 겉보기오분류율이나 교차타당성 또는 붓스트랩 추정량을 나타낸다. Braga-Neto and Dougherty (2004)는 $\widehat{Err}$ 의 성능을 비교하려면 $\widehat{Err}$ 이 아닌 $(\widehat{Err}-Err)$ 의 분포를 고려해야 함을 지적하고, 이 분포를 편차분포(deviation distribution)라 불렀다.

$(\widehat{Err}-Err)$ 을 하나의 확률변수로 간주했을 때, 다음 식

$$Var(\widehat{Err}-Err)=E[(\widehat{Err}-Err)^2]-[E(\widehat{Err}-Err)]^2 \tag{3.1}$$

이 성립하는데, 여기서 Var 와 E 는 훈련자료의 분포에 따른 분산과 기대값을 나타낸다. 식 (3.1)의 각 항을 모의실험에서 어떻게 추정할 수 있는가를 알아보자. t $(=1,2,\cdots,100)$ 번째 주어지는 n 개의 훈련자료에 대한 Err 과 $\widehat{Err}$ 을 각각 Err_t과 $\widehat{Err}_t$ 로 표현할 때, $Err_t \approx \sum_{i=1}^{5000} I(\hat{y}_{0i} \neq y_{0i})/5000$ 와 같이 추정한다. 여기서 $\{(y_{0i}, x_{0i}),\ i=1,\cdots,5000\}$은 t 번째 훈련자료와 독립적으로 추출한 5000개의 검증자료를 나타내며, $\hat{y}_{0i}$ 는 t 번째 훈련자료를 이용하여 생성한 분류기에 x_{0i} 를 대입했을 때 분류기에 의해 예측되는 값을 나타낸다. $\widehat{Err}_t$ 의 계산식은 추정량의 종류에 따라 달라진다. 그리고,

$$Var(\widehat{Err}-Err) \approx \sum_{t=1}^{100}[(\widehat{Err}_t-Err_t)-E(\widehat{Err}-Err)]^2/(100-1)$$

$$E(\widehat{Err}-Err) \approx \sum_{t=1}^{100}(\widehat{Err}_t-Err_t)/100$$

와 같이 추정한다. 훈련자료를 반복해서 추출하는 횟수를 크게 할수록 보다 정확하게 추정할 수 있겠지만 실행시간을 고려하여 100으로 정하였다. 식 (3.1)에서 $E[(\widehat{Err}-Err)^2]$ 의 제곱근인 $\sqrt{E[(\widehat{Err}-Err)^2]}$ 을 평균제곱오차의 제곱근(Root Mean Squared error)으로 정의하고 RMS로 부르기로 한다. 식 (3.1)을

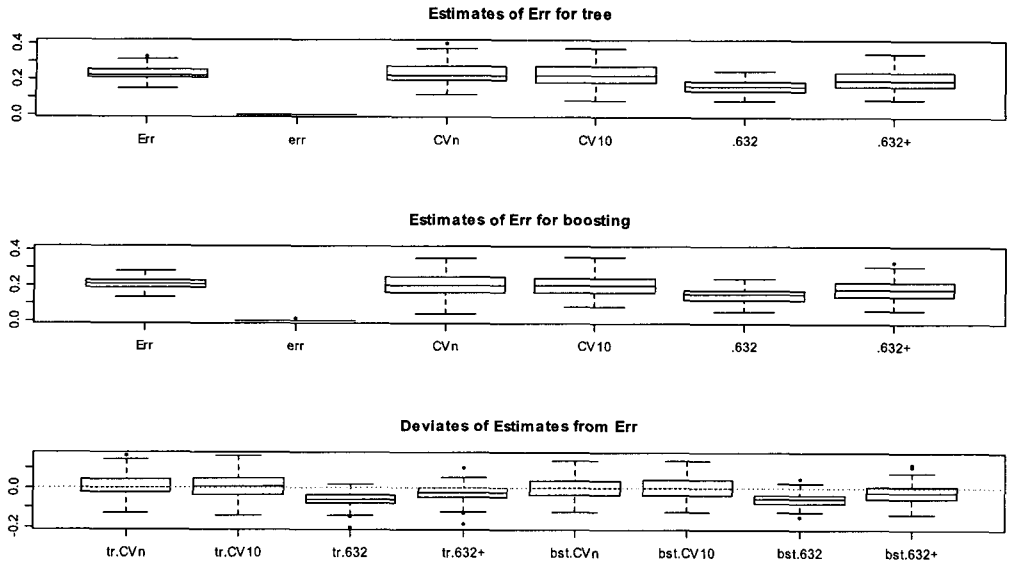
$$E[(\widehat{Err}-Err)^2]=Var(\widehat{Err}-Err)+[E(\widehat{Err}-Err)]^2$$

으로 다시 쓸 수 있으며, $E(\widehat{Err}-Err)$ 을 추정량 $\widehat{Err}$ 의 편향(bias)을 나타내는 항으로 볼 수 있으므로, 식 (3.1)은 결국

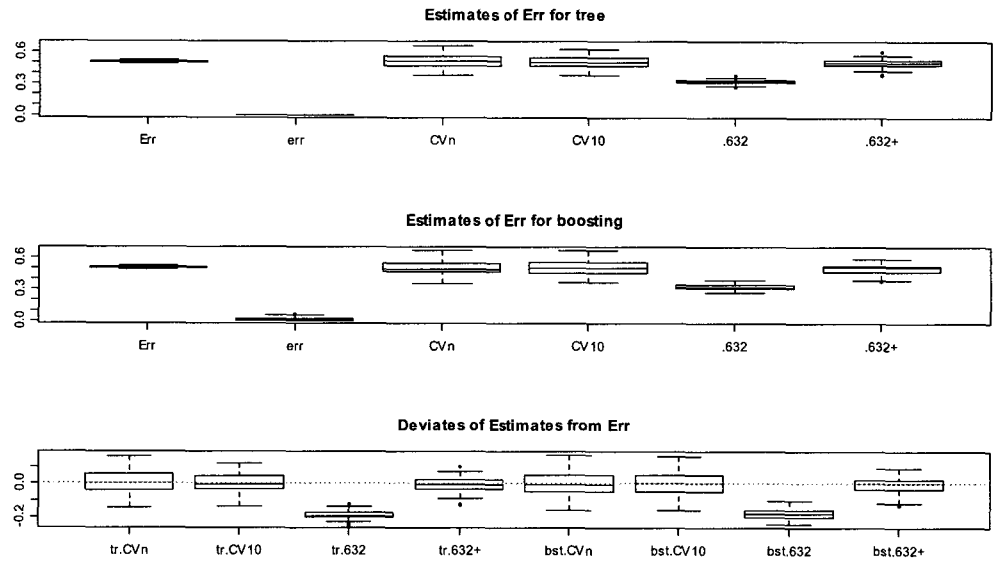
$$RMS^2=Var(\widehat{Err}-Err)+bias^2$$

이 된다. 모의실험에서 RMS와 함께 편향과 표준편차 $\sqrt{Var(\widehat{Err}-Err)}$ 를 각각 계산하고, 이 값들로 추정량들을 비교하고자 한다.

먼저, 표본크기가 100이고 불확실성을 나타내는 값 p 가 0.9일 때 실험한 결과를 보자. 가지치기(pruning) 없이 최대깊이 30까지 허용한 하나의 나무모형과, 깊이 2인 나무모형 100개의 선형결합을 분류기로 쓰는 부스팅 기법 모두 이 실험에서는 비슷한 결과를 보인다. <그림 3.1>의 위 두 그림에서 알 수 있듯이 겉보기오분류율은 100번의 모의실험에서 거의 모두 0 이었다. 교차타당성 추정량은 편향이 없어 보이나 붓스트랩 추정량에 비해 (특히 .632 붓스트랩 추정량에 비해) 분산이 커 보임을 알 수 있다. 제일 아래 그림은 각 100번의 반복실험에서 얻어진 $(\widehat{Err}-Err)$의 분포를 나타내는데, 붓스트랩 추정량



<그림 3.1> $n=100$, $p=0.9$ 일 때 참오분류율과 그 추정량의 분포: (100번의 모의실험 결과로 상자그림을 그렸으며 E_{rr} 은 참오분류율, err 은 겉보기오분류율. 제일 위 그림은 나무모형, 가운데 그림은 부스팅 기법에 의한 추정량, 아래 그림은 추정량과 참오분류율의 차에 관한 분포; tr은 나무모형, bst는 부스팅을 나타냄)



<그림 3.2> $n=100$, $p=0.5$ 일 때 참오분류율과 그 추정량의 분포

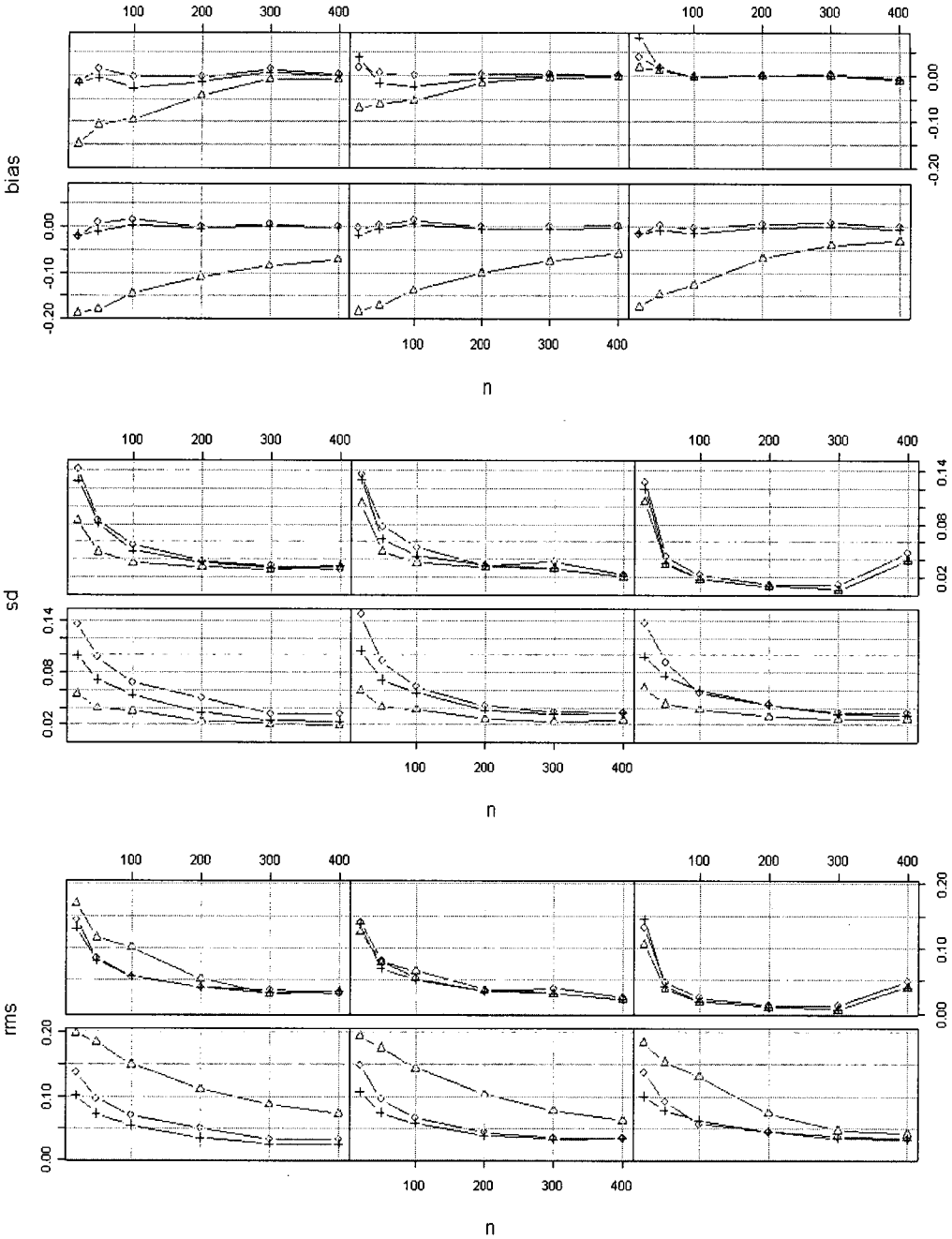
의 편향은 유의하였다(참오분류율과 그 추정량의 차에 대한 양측검정 결과 p 값은 .632와 632+ 추정량 모두 소수 셋째 자리까지 0). 분산의 경우 교차타당성 추정량과 .632 붓스트랩 추정량 사이에는 유의한 차이가 있었으나(각 100개로 이루어진 두 추정량의 표본이 독립이라고 가정하고 F검정 실시), 교차타당성 추정량과 .632+ 붓스트랩 추정량 사이에는 분산의 차이가 유의하지 않았다. 불확실성의 정도를 나타내는 p 가 0.9일 때의 결과를 요약하면, 두 붓스트랩 추정량은 모두 아래로 편향되었으며, .632 추정량의 분산은 교차타당성 추정량에 비해 작다는 것을 알 수 있다.

$p=0.5$ 에 대해서도 실험하였다(<그림 3.2>). p 가 0.5이면 자료에 아무런 규칙이 없으므로 분류기의 예측력 또한 있을 수가 없다. 따라서 참오분류율은 0.5이다. 하지만 최대깊이를 허용한 나무모형이나 부스팅 기법을 이용한 분류기의 겉보기오분류율은 0 또는 0에 가까운 값을 가졌다. 이 실험에서 분류기의 참오분류율에 대한 교차타당성 추정량에는 p 가 0.9일 때와 마찬가지로 편향이 없었으나, .632 붓스트랩 추정량에는 더욱 심각한 편향이 있음을 알 수 있다. 하지만 .632+추정량의 편향은 유의하지 않았다. 분산의 경우 두 붓스트랩 추정량의 분산 모두 교차타당성 추정량의 분산에 비해 유의하게 작았다. 불확실성의 정도를 나타내는 p 가 0.5일 때의 결과를 요약하면, 두 붓스트랩 추정량 중 .632 붓스트랩 추정량만 아래로 편향되었으며, 두 붓스트랩 추정량의 분산은 교차타당성 추정량에 비해 작다는 것을 알 수 있다.

n 겹 교차타당성 추정량 err^{CVn} 은 표본의 크기가 커질수록 실행시간이 길어진다. 10겹 교차타당성 추정량 err^{CV10} 과 성능에서 차이가 없다면 실행시간이 짧은 err^{CV10} 으로 대치하는 것이 좋다. err^{CVn} 과 err^{CV10} 의 성능에 유의한 차이가 있나 알아본 결과 차이가 없었다. (100번의 반복실험에서 나온 결과를 이용하여 $E(err^{CVn}-err^{CV10})=0$ 이라는 가설에 대해 대응 t검정을 실시한 결과, p 가 0.9일 때 나무모형의 경우 p 값은 0.82, 부스팅의 경우 0.10 이었다. p 가 0.5일 때 p 값은 각각 0.34, 0.52 이었다.)

앞의 $n=100$, $p=0.9$ 일 때의 실험에서 교차타당성 추정량에 비해 붓스트랩 추정량의 편향이 크고 분산(또는 표준편차)이 작은 경향이 있음을 알 수 있었다. 이러한 현상이 표본의 크기가 달라짐에 따라 ($n=20, 50, 100, 200, 300, 400$) 그리고 불확실성의 정도가 달라짐에 따라 ($p=0.5, 0.6, 0.7, 0.8, 0.9, 1.0$) 일관성있게 관찰되는지를 살펴보기로 하자. 대규모의 모의실험이므로 실행시간이 문제가 되는 n 겹 교차타당성 추정량은 고려 대상에서 제외하였다. n 과 p 의 각 조건에서 20번 반복 실험하였으며 전체 실험을 5번 반복하여 평균한 결과를 구하였다. 나무모형과 부스팅의 결과가 비슷하여 부스팅의 결과만 보고하기로 한다.

<그림 3.3>에서 추정량들의 성능을 나타내는 편향, 표준편차, 그리고 RMS를 차례대로 비교해보았다. 자료에 잡음이 전혀 없는 경우, 즉 $p=1$ 일 때에는 표본의 크기가 아주 작지 않으면 ($n>50$) 추정량들의 성능에 차이가 없었다. 보다 현실적으로, 잡음이 있는 경우 추정량들의 성능에 차이가 나는데, <그림 3.3>에서 먼저 편향의 그림을 보자. y 축의 값 0에 가까운 추정량이 좋은 추정량이다. 10겹 교차타당성 추정량 err^{CV10} 은 모든 n 과 p 의 값에 대해 0에 가까운 값을 보여주고 있다. 반면에 붓스트랩 추정량, 특히 .632 붓스트랩 추정량 $err^{.632}$ 의 편향은 p 가 0.5에 가까울수록 심각하다. 편향의 크기



<그림 3.3> 모의실험 I 자료에 대한 세 추정량의 성능(편향, 표준편차, RMS) 비교 (-o-, -△-, -+- 은 각각 10겹 교차타당성, .632, .632+ 붓스트랩 추정량을 나타냄; 세 그림 각각의 하단 왼쪽부터 오른쪽으로 $p = 0.5, 0.6, 0.7$, 상단 왼쪽부터 오른쪽으로 $p = 0.8, 0.9, 1.0$ 일 때의 그림)

는 전체적으로

$$|bias(err^{.632})| > |bias(err^{.632+})| > |bias(err^{CV10})| \approx 0$$

의 관계를 보인다.

추정량들의 변동을 비교하기 위해 표준편차(sd로 표시)를 보면, 전체적으로

$$sd(err^{.632}) < sd(err^{.632+}) < sd(err^{CV10})$$

의 관계가 있음을 알 수 있다. 또한 교차타당성 추정량과 붓스트랩 추정량의 표준편차의 차이는 표본의 크기가 커질수록 줄어든다는 사실도 알 수 있다.

편향과 분산을 같이 고려하는 RMS를 비교하여 보자. $p=1$ 인 경우를 제외하면 .632 추정량은 RMS가 커 좋은 추정량이 아님을 알 수 있다. 이는 편향이 크기 때문이며 .632 추정량의 표준편차가 비록 작지만 편향을 상쇄하지 못하기 때문이다. err^{CV10} 과 $err^{.632+}$ 를 비교하면, 표본의 크기가 작을 때는 $err^{.632+}$ 의 RMS가 작지만, 표본의 크기가 아주 작지 않고($n \geq 100$) p 가 0.5가 아닐 때에는 err^{CV10} 의 RMS가 $err^{.632+}$ 와 차이가 없거나 더 작아짐을 알 수 있다. 이 역시 $err^{.632+}$ 의 편향 때문이다.

붓스트랩 추정량에 관한 기존의 연구에서 표본의 크기가 작을 때 붓스트랩 추정량의 경쟁력을 강조하였으나, 표본의 크기가 클 때에 대해서는 다루고 있지 않다(암묵적으로 차이가 없어짐을 가정하고 있는 듯하다). 이 모의실험에서 붓스트랩 추정량의 편향은 표본의 크기가 커지더라도 문제가 되며, 편향과 분산을 같이 고려하는 RMS 기준에서도 붓스트랩 추정량은 교차타당성 추정량에 비해 경쟁력이 떨어질 수 있음을 알 수 있다.

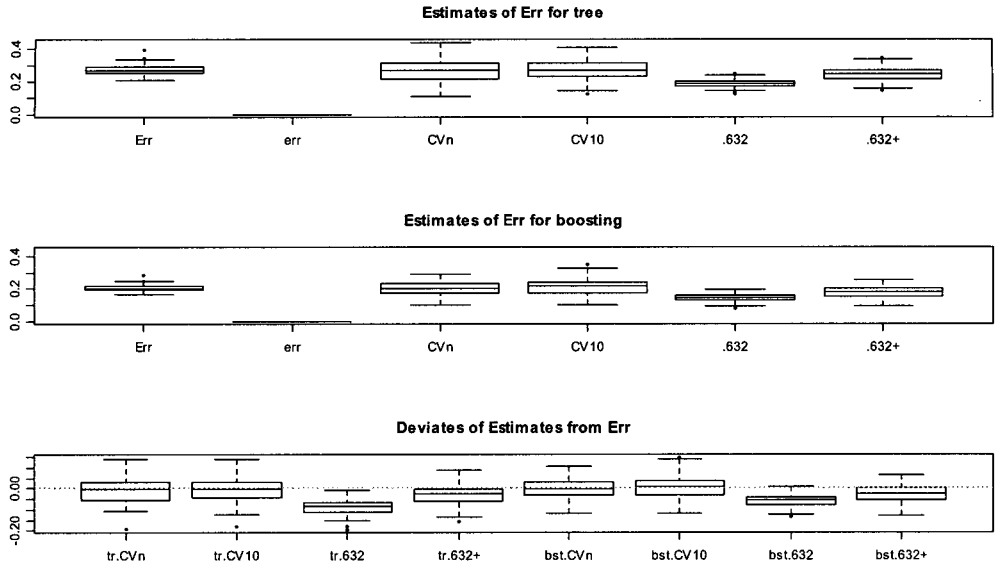
3.2 모의실험 II

모의실험 I에서는 설명변수가 2개인 간단한 모형을 고려하였다. 이번에는 설명변수가 보다 많은 다음과 같은 모형을 고려한다.

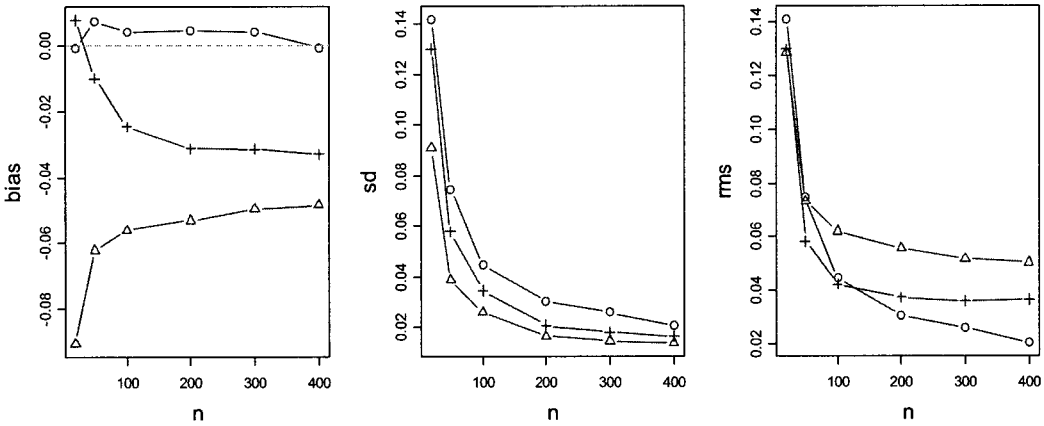
$$F(\boldsymbol{x}) = -10 + 10\sin(\pi x_1 x_2) + 5(x_3 - 1/2)^2 + 5x_4 + 2x_5 .$$

처음 4개의 설명변수들은 구간 (0,1)에서 균일분포를 따르며, x_5 의 분포는 이산형 균일분포로서 가능한 값은 $\{1, 2, 3\}$이다. 분석을 위한 자료에는 위 5개의 설명변수에 반응변수와 아무 관련이 없는 5개의 설명변수를 추가하였다. 추가적인 변수들의 분포는 참모형 $F(\boldsymbol{x})$ 에 있는 5개의 설명변수와 같은 분포를 갖도록 하였다. 반응변수 y 는 이항분포 $b(1, \mu)$, $\mu = 1/[1+\exp(-F(\boldsymbol{x}))]$ 에서 생성하였다.

먼저, 표본크기가 100인 자료에 대해 실험한 결과를 보자. 최대깊이 30까지 허용한 하나의 나무모형과, 깊이 2인 나무모형 50개의 선형결합을 분류기로 쓰는 부스팅 기법 모두 이 실험에서는 비슷한 결과를 보인다. <그림 3.4>의 위 두 그림에서 참오분류율(Err)의 값을 비교해보면 알 수 있듯이 부스팅에 의한 분류기가 하나의 나무모형에 의한 분류기보다 더 나은 성능을 보인다. 그리고 100번의 모의실험에서 두 분류기 모두 겉보기오분류율은 0이었다. 100번의 반복실험에서 나타난 참오분류율과 그 추정량의 차($\widehat{Err} - Err$)에 관한 분포를 나타내는 제일 아래 그림에서 알 수 있듯이, 두 붓스트랩 추정량의 편향은 모두 유의하였으며, 분산의 경우 10겹 교차타당성 추정량과 두 붓스트랩 추정량 사이에 각각 유의한 차이가 있었다(p 값은 모두 소수 셋째 자리까지 0).



<그림 3.4> 모의실험 II 자료에 대한 참오분류율과 그 추정량의 분포: $n=100$ (100번의 모의실험 결과로 상자그림을 그렸으며 E_{rr} 은 참오분류율, err 은 겉보기오분류율. 제일 위 그림은 나무모형, 가운데 그림은 부스팅 기법에 의한 추정량, 아래 그림은 추정량과 참오분류율의 차에 관한 분포; tr은 나무모형, bst는 부스팅을 나타냄)



<그림 3.5> 모의실험 II 자료에 대한 세 추정량의 성능(편향, 표준편차, RMS) 비교 (-o-, -△-, -+- 은 각각 10겹 교차타당성, .632, .632+ 붓스트랩 추정량을 나타냄)

표본의 크기에 따른 효과를 보기 위해 훈련표본의 크기(n)를 20, 50, 100, 200, 300, 400으로 변화시켜가며 100번씩 반복하여 실험하였다. <그림 3.5>에서 추정량들의 성능을 나타내는 편향, 표준편차, 그리고 RMS를 차례대로 비교해보았다. 세 추정량의 편향과 표준편차에는 모의실험 I에서와 같이 모든 표본크기에서

$$|bias(err^{.632})| > |bias(err^{.632+})| > |bias(err^{CV10})| \approx 0$$
$$sd(err^{.632}) < sd(err^{.632+}) < sd(err^{CV10})$$

의 관계가 있으며, 표본크기가 커질수록 err^{CV10} 의 표준편차는 작아지는 반면 붓스트랩 추정량의 편향은 계속 문제가 되기 때문에 err^{CV10} 의 RMS가 붓스트랩 추정량의 RMS보다 작아짐을 알 수 있다. Efron and Tibshirani (1997)는 $err^{.632}$ 의 편향을 교정하는 추정량으로 $err^{.632+}$ 을 제안하였으나, $err^{.632+}$ 에도 편향의 문제가 여전히 남아있음을 이 모의실험을 통해 알 수 있다.

4. 실제 자료를 이용한 실험

실제 자료를 이용하여 참오분류율의 추정량들을 비교하였다. 자료는 캘리포니아 주립대학 자료 저장소(UCI Repository, Blake and Merz (1998))에 있는 것 중에서 비교적 크기가 큰 것(자료의 크기가 설명변수의 수의 20배 이상이 되는 것)을 이용하였다. 일부 자료는 (solar 자료와 letter 자료) 반응변수가 다항범주이지만 범주들을 묶어 이항자료로 만들었다. 자료의 특성을 <표 4.1>에 정리하였다.

<표 4.1> 실제자료 설명

자료 이름	자료의 크기	문자형 설명변수	연속형 설명변수	전체 설명변수의 수
heart	920	9	5	14
pima	768	0	8	8
solar	1389	8	2	10
credit	690	9	6	15
glass	214	0	9	9
breast	569	0	30	30
adult	32561	8	6	14
voting	435	16	0	16
letter	10000	0	16	16

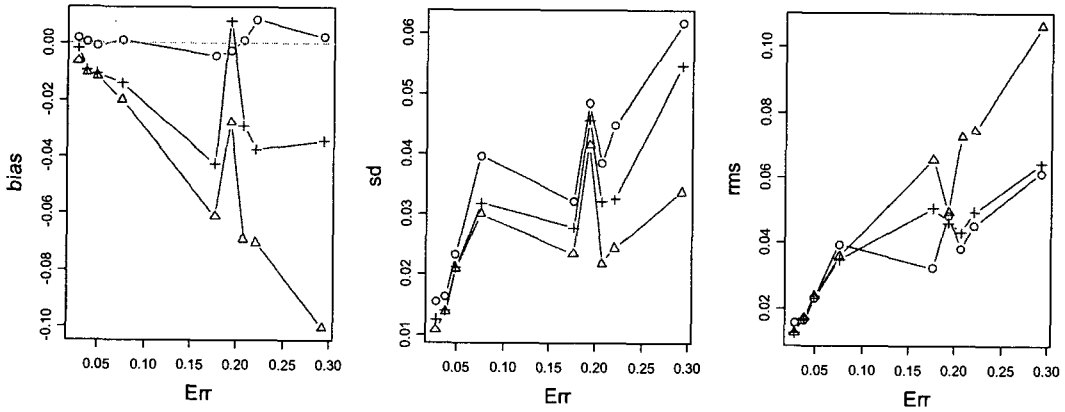
전체 자료 중에서 설명변수의 수의 10배 만큼의 자료를 임의로 추출하여 훈련자료로 사용하고 나머지는 검증자료로 사용하였으며, 이 과정을 100번 반복하여 실험하였다.

앞 절의 모의실험 결과 .632 추정량은 심각한 편향 때문에 겉보기오분류율이 0에 가까

<표 4.2> 10겹 교차타당성 추정량과 .632+ 붓스트랩 추정량의 성능 비교

자료 이름	E_{rr}	err	BIAS		SD		RMS	
			CV10	632+	CV10	632+	CV10	632+
heart	0.2207	0.0012	0.0083	-0.0374	0.0448	0.0325	0.0454	0.0494
pima	0.2930	0.0000	0.0021	-0.0344	0.0616	0.0545	0.0614	0.0642
solar	0.1929	0.0896	-0.0025	0.0076	0.0484	0.0458	0.0483	0.0462
credit	0.1762	0.0000	-0.0045	-0.0426	0.0321	0.0278	0.0323	0.0508
glass	0.0756	0.0000	0.0011	-0.0142	0.0397	0.0317	0.0395	0.0346
breast	0.0375	0.0000	0.0004	-0.0092	0.0163	0.0139	0.0162	0.0166
adult	0.2069	0.0000	0.0012	-0.0290	0.0384	0.0321	0.0383	0.0431
voting	0.0490	0.0126	-0.0007	-0.0107	0.0232	0.0211	0.0231	0.0236
letter	0.0276	0.0000	0.0017	-0.0016	0.0155	0.0125	0.0155	0.0125

운 분류기의 참오분류율을 추정하는 데에 쓸 수 없는 추정량임을 확인하였다. <표 4.2>에서 10겹 교차타당성 추정량과 .632+ 붓스트랩 추정량의 성능을 비교하였다. 하나의 나무모형과 부스팅의 결과가 비슷하므로 부스팅의 결과만 보고하기로 한다. 참오분류율 Err 의 값이 아주 작고 편향의 크기도 무시할 수 있을 만큼 작은 letter 자료를 제외하면 앞 절의 모의실험에서와 같이 err^{CV10} 의 편향이 $err^{.632+}$ 의 편향보다 작다. 또한 err^{CV10} 의 분산이 $err^{.632+}$ 의 분산보다 크다. 편향과 분산을 같이 고려하는 RMS 기준에서의 우열은 자료마다 다르다. err^{CV10} 와 $err^{.632+}$, $err^{.632}$ 의 성능을 비교하기 쉽게 <그림 4.1>에 나타내었다.

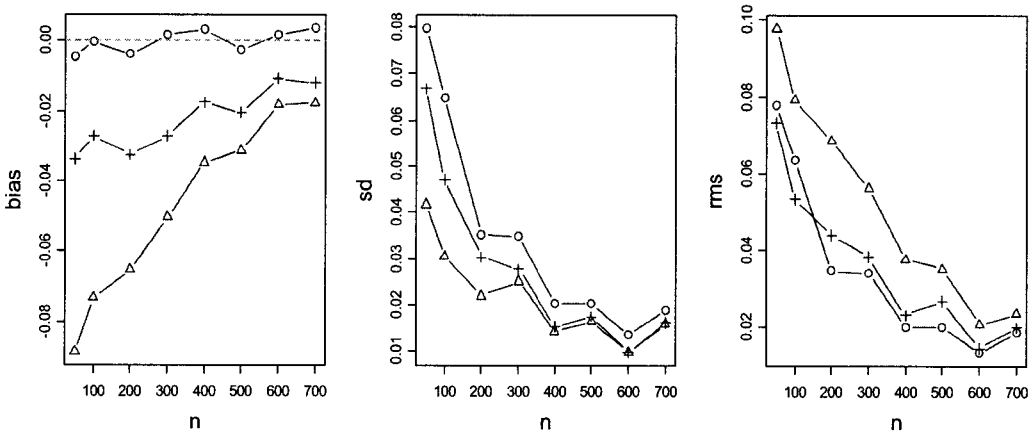


<그림 4.1> 9개의 실제자료로 비교한 세 추정량의 성능(편향, 표준편차, RMS) (-o-, -△-, -+- 은 각각 10겹 교차타당성, .632, .632+ 붓스트랩 추정량을 나타냄. 점을 선으로 연결한 것은 추정량별로 구분하기 위한 것이지 x 축의 값의 크기에 따른 경향을 보고자 한 것은 아님)

크기가 제일 큰 adult 자료를 선택하여 표본의 크기에 따른 효과를 살펴보았다. 훈련자료의 크기를 변화시켜가며 ($n = 50, 100, 200, 300, 400, 500, 600, 700$), 각 표본크기에서 20번 반복실험하여 세 추정량의 편향과 표준편차, RMS를 추정하였다. 앞에서의 결과와 마찬가지로

$$|bias(err^{.632})| > |bias(err^{.632+})| > |bias(err^{CV10})| \approx 0$$
$$sd(err^{.632}) < sd(err^{.632+}) < sd(err^{CV10})$$

의 관계가 일관되게 관측되었으며, 표본의 크기가 커지면 교차추정량의 RMS가 붓스트랩 추정량의 RMS보다 작은 값을 가졌다. 특히 붓스트랩 추정량의 편향은 표본이 커지더라도 줄어들지 않았는데, $n = 700$ 일 때, 20번의 반복실험에서 얻은 20개의 $err^{.632+} - Err$ 값으로 편향의 유의성을 검정한 결과 $E(err^{.632+} - Err)$ 이 유의하게 0보다 작았다(p값 $= 0.0035$).



<그림 4.2> adult 자료에서 훈련표본 크기를 변화시켜가며 본 세 추정량의 성능 (편향, 표준편차, RMS)

(-o-, -△-, -+- 은 각각 10겹 교차타당성, .632, .632+ 붓스트랩 추정량을 나타냄)

5. 결론

훈련자료만 있고 독립적인 검증자료가 없을 때 분류기의 참오분류율을 추정하는 문제는 중요하다. 이 문제에 대한 실험적인 연구 결과를 제시하고자 하였다. 교차타당성 방법과 붓스트랩 방법을 모의자료와 실제자료를 이용하여 비교한 결과 교차타당성에 의한 추정량이 작은 편향을, 붓스트랩에 의한 추정량이 작은 분산을 가지며, 작은 표본에서 붓스트랩에 의한 추정량이 편향과 분산을 같이 고려하는 평균제곱오차의 기준에서 다나은 성능을 갖는다는 기존의 연구결과가 이 연구에서 고려한 실험조건에서도 성립함을 확인할 수 있었다.

추가적으로 이 연구에서는, 부스팅 기법에 의한 분류기와 같이 겉보기오분류율이 작은 값을 가질 때, 붓스트랩에 의한 추정량이 표본의 크기가 커지더라도 편향의 크기가 줄지 않는 반면 교차타당성 추정량의 분산은 표본의 크기가 커질 때 분산이 작아져, 결국 평균제곱오차의 기준에서는 표본의 크기가 커질수록 교차타당성 추정량의 성능이 붓스트랩 추정량보다 나은 경향이 있다는 사실을 알 수 있었다. 재추출하는 붓스트랩 표본의 수가 50인 붓스트랩 추정량은 10겹 교차타당성 추정량에 비해 실행시간이 약 5배 더 걸린다는 점을 고려하면, 큰 표본에서 교차타당성 추정량의 장점은 더 커진다. 붓스트랩 추정량, 특히 편향 문제를 개선한 .632+ 추정량의 편향의 크기가 표본이 커지더라도 문제가 된다는 사실은 다른 연구에서 지적되지 않은 새로운 사실이다.

겉보기오분류율이 0에 가까운 값을 가지는 분류기의 참오분류율을 추정할 때 교차타당성에 의한 추정량은 분산의 문제를, 붓스트랩에 의한 추정량은 편향의 문제를 안고 있으므로 새로운 대안이 필요하다. Efron and Tibshirani (1997)는 .632+ 추정량에 대한 연구에서 붓스트랩 추정량의 편향을 교정하려고 하면 분산이 커진다는 사실을 지적하였다. 우리는 붓스트랩 추정량을 개선하는 대신, 편향이 문제가 되지 않는 교차타당성 추정량의 분산을 줄이는 방법을 대안으로 생각하고 있으며, 몇 번의 교차타당성 추정량의 평균을 내는 방법과 또 다른 대안들의 효과에 대해 연구를 진행하고 있는 중이다.

참고문헌

[1] Cha, E.S. (2005). 예측오차 추정방법에 대한 비교연구, 석사학위논문, 숭실대학교.

[2] Bauer, E. and Kohavi, R. (1999). An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine Learning*, Vol. 36, 105-139.

[3] Blake, C.L. and Merz, C.J. (1998). UCI Repository of machine learning databases. University of California in Irvine, Department of Information and Computer Science.

[4] Braga-Neto, U.M. and Dougherty, E.R. (2004). Is cross-validation valid for small-sample microarray classification? *Bioinformatics*, Vol. 20, 374-380.

[5] Crawford, S.L. (1989). Extensions to the CART algorithm, *International Journal of Man-Machine Studies*, Vol. 31, 197-217.

[6] Efron, B. (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association*, Vol. 78, 316-331.

[7] Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*, Chapman and Hall.

[8] Efron, B. and Tibshirani, R. (1997). Improvements on cross-validation: The .632+ bootstrap method. *Journal of the American Statistical Association*, Vol. 92. 548-560.

[9] Freund, Y. and Schapire, R. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, Vol. 55, 119-139.

[10] Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Technical Report*, Stanford University, Department of Computer Sciences.

[11] Merler, S. and Furlanello, C. (1997). Selection of tree-based classifiers with the bootstrap 632+ rule. *RIST Technical Report*: TR-9605-01, revised Jan 97.

[12] R Development Core Team (2004). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. Available from http://www.R-project.org.

[13] Therneau, T.M. and Atkinson, E.J. (1997). An introduction to recursive partitioning using the RPART routines. *Technical Report*, Mayo Foundation.

[Received November 2005, Accepted February 2006]