응용통계연구 제 17권 3호, 2004년, pp. 445-457

계급불균형자료의 분류: 훈련표본 구성방법에 따른 효과 *

김지현[1] 정종빈[2]

요 약

두 계급의 분류문제에서 두 계급의 관측 개체수가 심하게 불균형을 이룬 자료를 분석할 때, 흔히 인위적으로 두 계급의 크기를 비슷하게 해준 다음 분석한다. 본 연구에서는 이런 훈련표본 구성방법의 타당성에 대해 알아보았다. 또한 훈련표본의 구성방법이 부스팅에 미치는 효과에 대해서도 알아보았다. 12개의 실제 자료에 대한 실험 결과 나무모형으로 부스팅 기법을 적용할 때는 훈련표본을 그대로 둔 채 분석하는 것이 좋다는 결론을 얻었다.

주요용어: 나무모형, 부스팅(boosting), 정규화기대비용

1. 서론

두 계급의 분류문제에서 한 계급의 관측 개체수가 다른 계급에 비해 현저하게 적을 때가 흔히 있다. 금융관련 자료에서 대출자를 연체자와 비연체자로 분류할 때나 환경관련 자료에서 관측지역을 오염지역과 비오염지역으로 분류할 때, 그리고 의료관련 자료에서 검진자를 특정질병환자와 정상인으로 구분할 때 한 계급의 관측 개체수는 나머지 계급의 관측 개체수에 비해 적은 경우가 보통이다. 두 계급의 관측 개체수가 심하게 불균형을 이룬 자료(이하 '계급불균형자료'로 칭함)를 분석할 때, 흔히 현장에서는 인위적으로 두 계급의 크기를 비슷하게 해준 다음 분석하는데, 본 연구에서는 이렇게 하면 원래 크기를 유지한 채로 분석했을 때와 어떤 차이가 있는가에 대해 알아보고 대안을 제시하고자 한다.

나무모형(tree model, Breiman *et al.* (1984))은 계급불균형자료가 주어졌을 때 소수계급(minority class)을 다수계급(majority class)으로 오분류하는 비율이 높아지는 경향이 있다. 왜냐하면, 훈련표본(training sample)에 다수계급의 자료가 많으면 전체 오분류율을 낮추기 위해 다수계급을 노드(node)의 대표값으로 분류를 많이 하게 되고, 따라서 이들 노드에 속하는 소수계급은 다수계급으로 오분류가 되어 소수계급의 오분류율이 높아지게 되기 때문이다(Weiss & Provost (2001) 6.1절 참조). 이 때 인위적으로 소수계급의 비율과 다수계급의 비율을 맞추어 훈련표본을 구성하면 전체적인 오분류율은 높아지나 소수계급의 오분류율은 보통 낮아진다. 만약 소수계급의 오분류 비용이 다수계급의 오분류 비용보다 월등히 크다면, 전체적인 오분류율이 낮지만 소수계급의 오분류율이 높은 방법이, 전체적인

* 본 연구는 숭실대학교 교내연구비 지원으로 이루어졌음

1) (156-743) 서울시 동작구 상도동 1-1 숭실대학교 자연과학대학 정보통계학과 교수
E-mail: jhkim@stat.ssu.ac.kr
2) (156-743) 서울시 동작구 상도동 1-1 숭실대학교 자연과학대학 정보통계학과 대학원 석사과정
E-mail: jbjeong@stat.ssu.ac.kr

오분류율이 다소 높지만 소수계급의 오분류율이 낮은 방법에 비해 기대비용이 클 수 있다. 실제로 연체자를 비연체자로 잘못 분류했을 때 발생하는 비용이나, 특정질병을 가진 환자를 정상인으로 잘못 분류할 때의 비용이 상대적으로 훨씬 큰 경우가 많다. 이런 경우 분류기(classifier)의 성능평가기준을 오분류율로 하는 것은 합리적이지 못하다.

계급불균형자료의 문제점에 대한 해결방법에 대해서는 이미 많은 연구가 있었고 현재도 진행되고 있다 (Boz, 2001). 이 연구들에서 '비용에 민감한 분류기(cost-sensitive classifier)'란 키워드를 쓰는데, 오분류의 종류에 따라 비용이 다르다면 비용에 민감한 분류기가 필요하다. 엄격히 얘기하면 계급불균형(class imbalance) 문제와 비용민감성(cost-sensitiveness)의 문제가 같지는 않다. 왜냐하면 계급불균형은 없어도 두 계급 사이의 오분류 비용에 큰 차이가 있어 비용을 최소화하는, 즉 비용에 민감한 모형을 찾을 필요가 있기 때문이다. 하지만 계급불균형이 있을 때 소수계급의 오분류 비용이 더 높은 경우가 대부분이며 오분류율이 아닌 총비용을 평가기준으로 선택한다면, 결국 비용을 최소화하는 분류방법을 찾는 것이 해결방법이므로, 해결방법의 관점에서는 두 문제를 구분하지 않아도 될 것이다. 해결방법으로서 비용을 최소화하는 부스팅(cost-sensitive boosting)을 이용하는 방법을 포함하여 다양한 방법들이 있고 이들 방법들에 대한 비교연구도 이루어졌다(Ting, 2000). 하지만 과소표집이나 과대표집 등을 통한 훈련표본의 구성방법에 따른 효과를 비용의 관점에서 살펴본 연구는 아직 없다. 또한 훈련표본의 구성방법에 따라 부스팅의 효과가 어떻게 달라지는지에 대한 연구도 아직 행해지지 않았다.

본 연구에서는 나무모형을 적용할 때 두 가지 훈련표본 구성방법, 즉, 계급 비율을 그대로 두는 것과 인위적으로 균형을 이루도록 하는 것 중 어느 것이 더 좋은가에 초점을 맞추었다. 제2절에서 여러 가지 가능한 훈련표본 구성방법들을 설명하였고, 구성방법들을 비교하기 위한 평가기준도 제시하였다. 제3절에서 하나의 나무모형(single classifier)을 적용할 때의 결과들을 비교하였고, 제4절에서는 여러 개의 나무모형을 결합(composite classifier)하는 부스팅 기법을 적용했을 때의 결과를 비교하였다.

2. 성능평가기준과 훈련표본 구성방법

계급불균형자료에서 오분류율은 좋은 성능평가기준이 되지 못하는데, 이 때 고려할 수 있는 다른 기준들을 정의하고, 계급불균형 문제에 대처하기 위한 여러 가지 훈련표본 구성방법들에 대해 설명한다.

흔히 분류규칙 또는 분류모형의 정확도를 평가할 때 표 2.1 같은 오분류표를 작성한다. 소수계급을 양의 계급이라고 하고 다수계급을 음의 계급이라고 할 때, 오분류율(misclassification rate)은 예측이 잘못될 확률로서 '거짓−'와 '거짓+'의 수를 총 자료의 수로 나눈 값이다. 오분류율을 모형의 성능평가기준으로 이용할 때에는 두 계급의 오분류비용이 같다는 것을 전제로 한다. 하지만, 계급불균형자료의 경우 소수계급에 대한 오분류비용이 더 큰 경우가 많으며, 이 때 오분류율은 적절한 기준이 되지 못한다.

고비용오분류율(high-cost error rate)은 오분류비용이 큰 소수계급의 오분류에만 관심을 둔 기준으로서 '거짓−'의 수를 총 자료의 수로 나눈 값이다. 소수계급의 오분류에만 관

표 2.1: 오분류표

	실제−	실제+
−로 예측	참− (True Positive)	거짓− (False Negative)
+로 예측	거짓+ (False Positive)	참+ (True Positive)

심을 두는 것에는 문제가 있지만 대처방법들의 특성을 비교할 때 도움이 된다.

기대오분류비용(expected cost of misclassification)은 두 계급의 오분류율과 오분류비용을 함께 고려한 기준으로서, $p(+)$과 $p(-)$를 두 계급의 확률이라고 하고, FN과 FP를 각각 주어진 계급에서 음과 양으로 오분류될 조건부확률이라고 하며, $C(+\mid -)$와 $C(-\mid +)$를 각각 다수계급과 소수계급에 대한 오분류비용이라고 할 때, $FNp(+)C(-\mid +)+FPp(-)C(+\mid -)$으로 정의된다. 모형의 성능을 평가하는 기준으로서 기대오분류비용이 오분류율보다 좀더 합리적이지만 실제로 적용할 때 오분류비용인 $C(+\mid -)$과 $C(-\mid +)$의 값이나 또는 두 값의 비를 정확하게 알 수 없다는 문제점이 있다. 본 연구에서는 기대오분류비용을 0과 1 사이의 값을 갖도록 표준화한 정규화기대비용(normalized expected cost)을 이용하는데(Drummond & Holte, 2000), 두 오분류비용의 비$(C(-\mid +)/C(+\mid -))$를 r이라고 할 때, 정규화기대비용은

$$\frac{FNp(+)r + FPp(-)}{p(+)r + p(-)}$$

으로 정의된다.

모형의 평가기준에서 모형의 간결성 또한 중요한 요소가 될 수 있다. 나무모형에서 다른 평가기준들이 비슷한 값을 갖는다면 좀더 간결한 모형이 좋은 모형이라고 할 수 있을 것이다. 나무모형에서 모형의 크기(tree size)를 나타낼 때 모든 노드(node)의 수나 최종노드(leaf)의 수를 쓸 수 있는데, 최종노드의 수는 (모든 노드의 수 + 1)/2 와 같으므로 어떤 것을 쓰든 마찬가지다. 본 논문에서는 최종노드의 크기를 이용하여 모형의 크기를 나타내기로 한다.

계급불균형으로 인해 소수계급의 오분류율이 높아지는 문제를 해결하기 위해 본 논문에서 고려한 몇 가지 훈련표본 구성방법에 대해 먼저 설명하고, 앞에서 설명한 4가지 성능평가기준을 이용하여 방법별로 비교한 결과를 다음 절에서 보고하기로 한다.

방법1: 인위적으로 두 계급의 크기가 균형을 이루도록 훈련표본의 계급 비율을 조정하는 방법이다. 소수계급에서 복원으로 과대추출(over-sampling)하여 두 계급의 비율이 같아지도록 한다. 훈련표본의 크기가 너무 커지는 문제가 발생하면 다수계급에서 과소추출도 병행한다.

방법2: 훈련표본의 계급 비율을 조정하지 않고 그대로 둔다. 이 방법은 계급불균형 문제에 대처하는 방법은 아니지만 다른 방법과의 비교를 위해 고려하였다.

방법3: 훈련표본의 소수계급에서 두 오분류비용의 비 r배만큼 복원으로 과대추출하는 방법이다. 방법1은 방법3의 특수한 경우이다.

방법4: 과대표집이나 과소표집을 통하여 비율을 조정하는 대신 소수계급의 관측값에

r배만큼 가중값을 부여하는 방법이다. 주어진 훈련표본에는 변함이 없고 나무모형을 적용할 때 훈련표본에 부여하는 가중값을 계급에 따라 다르게 하는 방법이다.

본 연구에서는 위 4가지 방법을 적용한 훈련표본으로부터 만들어지는 나무모형, 즉 분류기(classifier)들을 앞에서 설명한 여러 가지 성능평가기준으로 비교하고자 한다. 이들 방법들에 의해 만들어지는 분류기는 각각 하나의 나무모형으로 표현되는데, 본 연구에서는 많은 나무들의 선형결합으로 표현되는 모형을 생성하는 부스팅(boosting) 기법도 적용시켜 보았다. 먼저 하나의 나무모형에 대한 실험결과에 대해 알아보자.

3. 나무모형을 적용한 실험

분류 문제 연구에 자주 쓰이는 자료들 중에서 불균형 정도가 심한 12개의 자료들을 이용하여 앞 절에서 설명한 4가지 훈련표본 구성방법에 대한 비교분석을 실시하였다(표 3.1). 국내 A-보험사 자료와 Mammography 자료(Woods *et al.* 1993)를 제외한 자료는 캘리포니아 주립대학 자료저장소(UCI Repository, Blake and Merz (1998))에 있는 것을 이용하였다. 자료의 이름에 *표시가 있는 것들은 본래 다항형 자료이나 본 실험을 위해 인위적으로 이항형으로 만든 것들이다. .

표 3.1: 실험에 쓰인 실제자료

	문자형 변수	연속형 변수	총변수의 크기	소수계급의 크기 (비율)	전체 자료의 수	CV횟수
Discordant	15	7	22	45 (1.6%)	2800	5
Mammography	0	6	6	260 (2.32%)	11183	10
국내A-보험사	1	10	11	778 (3.89%)	20000	10
Letter-A *	0	16	16	789 (3.94%)	20000	10
Hyper-Thyroid	18	6	24	151 (4.80%)	3163	10
Forest cover *	17	10	27	2747 (7.13%)	38501	5
Balance-scale *	0	4	4	49 (7.84%)	625	5
Satimage *	0	36	36	626 (9.73%)	6435	10
Page-Block *	0	10	10	560 (10.23%)	5473	10
Image-Cement *	0	19	19	330 (14.28%)	2310	10
Solar-flare	8	2	10	218 (15.7%)	1389	10
German	17	3	20	300 (30.0%)	1000	10

공개된 통계분석용 프로그래밍언어인 R의 rpart package(Therneau & Atkinson, 1997)를 이용하여 실험하였다. 주어진 실험자료에 대해 rpart로 나무모형을 생성하였으며 분리기준(splitting criteria)과 정지규칙(stopping rule) 등은 자동설정값을 이용하였고 가지치기(pruning)는 하지 않았다.

교차타당성(cross-validation)을 통하여 구한 오분류표를 이용하여 방법별로 오분류율과 고비용오분류율, 정규화기대비용 등을 계산하였고, 모형의 크기도 출력하였다. 평가방법은 10중 교차타당성(10-fold cross validation)을 기본으로 하되, 소수계급의 관측 개체수가 100보다 적은 경우는 5중 교차타당성을 적용하였고, 같은 실험을 10번 반복한 결과의 평균을 구하였다. 이 때 훈련표본은 방법별로 계급 비율이 변할 수 있지만, 검증표본(test sample)은 원래 비율을 그대로 유지한다. 다시 말해, 방법별로 훈련표본의 계급 비율을 변화시켜 모형을 얻되, 얻어진 모형의 성능을 추정하기 위해 검증표본을 적용할 때에는 원래 비율을 그대로 유지한다.

본 실험에서는 오분류비용의 비(r)가 알려져 있다고 가정하였다. 소수계급에 대한 오분류비용이 다수계급에 대한 오분류비용보다 크다는 가정 하에 r을 다음과 같이 10개의 값으로 변화시켜가며 실험하였다.

$$r = \frac{p(+) + 0.1t(p(-) - p(+))}{p(+)}, \quad t = 1, 2, ..., 10$$

그림 3.1은 방법1과 방법2를 비교한 것으로서, 앞 절에서 설명한 4가지 평가기준 각각에 대해 두 방법을 적용한 결과의 로그비(log(방법1의 값/방법2의 값))를 나타낸 그래프이다. 가로축의 각 t의 값에서 12개의 자료에 대응되는 점이 나타난다. 세로축은 로그비의 값인데 0을 기준으로 적용방법간의 우위를 쉽게 파악할 수 있다. 예를 들어 오분류율의 로그비는 모두 양수로서, 예상대로 방법1의 오분류율이 방법2보다 높음을 알 수 있다. 반면 고비용오분류율에 있어서는 모두 음수로서 방법1이 방법2보다 소수계급을 잘 분류함을 의미한다. (방법1과 방법2의 오분류율과 고비용오분류율, 그리고 모형의 크기는 t에 영향을 받지 않는다. 그래프에서 자료의 이름을 표시하지는 않았지만 12개 자료에 대응되는 점들의 순위는 t값이 변하더라도 바뀌지 않았고, t값이 달라짐에 따라 생기는 변동은 랜덤변동(random fluctuation)으로 볼 수 있다.) 앞의 두 그래프에서 보았듯이 방법1은 방법2에 비해 오분류율은 높지만 고비용오분류율은 낮다. 따라서 두 오분류비용의 비 r이 크게 되면 소수계급의 오분류율이 낮은 방법1이 기대비용의 관점에서 우월해질 수 있는데, 정규화기대비용의 그래프는 그러한 현상이 일어나고 있음을 보여준다. 한편, 소수계급의 관측값을 과대표집하는 방법1을 적용하면 모형의 크기가 더 커지는 경향이 있음을 알 수 있다. 이것은, 방법2와 같이 소수계급 관측값이 많지 않으면 소수계급을 다수계급으로 잘못 분류하고 나무 생성 과정을 일찍 정지하게 되어 크기가 작아지나, 방법1과 같이 인위적으로 소수계급 관측값을 많게 하면 두 계급을 모두 잘 분류하게 될 때까지 나무를 생성하므로 크기가 커지는 것이라고 설명할 수 있다.

그림 3.2는 방법1과 방법3을 비교한 것이다. 실험에서 비교의 공정을 기하기 위하여 주어진 자료의 크기를 일정하게 유지하였으므로, 소수계급에서 과대표집을 하면 그 크기만큼 다수계급에서 과소표집을 하였다. 따라서 t가 5가 되면 방법1과 방법3은 같은 방법이 된다. 기대비용의 기준으로 보면 5보다 작은 t에 대해서 방법3이 더 우월한 방법임을 알 수 있다. t가 5보다 크게 되면 과대표집으로 인해 소수계급이 오히려 다수계급이 됨을 의미하는데, 실제 현장에서 이렇게까지 과대표집을 하지는 않을 것이나 과대표집의 효과를 보기 위하여 실험에서 적용해 보았다. 이 때 방법3은 지나친 과대표집으로 인해 소수계급에 대한

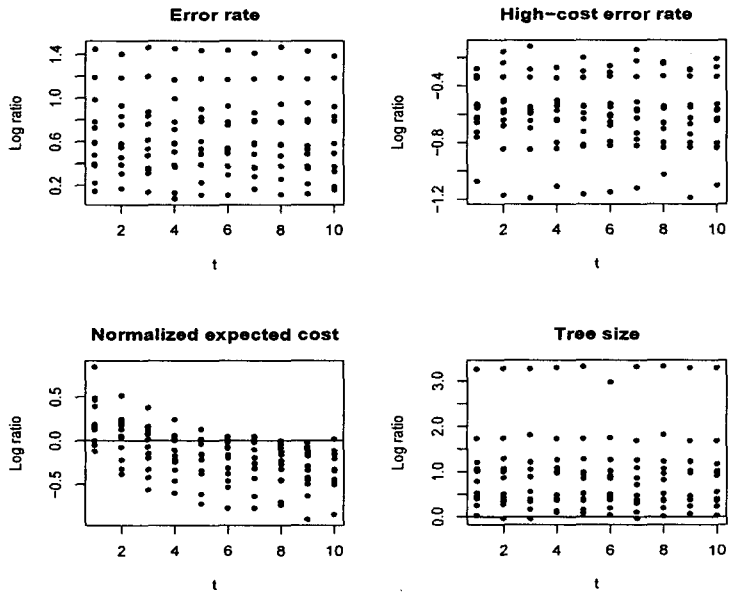

그림 3.1: 방법1과 방법2의 모형평가기준별 로그비 (log(방법1/방법2))

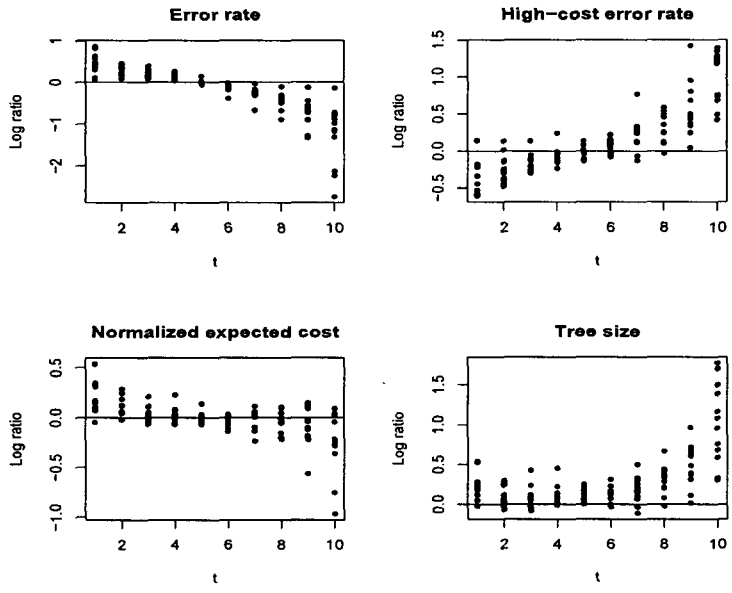

그림 3.2: 방법1과 방법3의 모형평가기준별 로그비 (log(방법1/방법3))

과적합(over-fitting)이 일어나게 되어 기대비용의 관점에서 방법1에 비해 열등하게 됨을 알 수 있다.

부록에 수록한 방법1과 방법4를 비교한 그래프를 살펴보면 t가 5보다 작을 때, 방법4도 방법3과 비슷한 결과를 보이지만, t가 5보다 커지더라도 방법 3에서 발생했던 과적합의 정도가 크지 않음을 알 수 있다. (방법4를 적용할 때, 방법3과 공정한 비교가 되게 하기 위하여 소수계급의 가중값을 1에서 r로 높여주고 다수계급의 가중값은 전체 가중값의 합이 훈련표본의 크기와 같아지도록 낮춰주었다.)

이상의 결과를 요약하면, 고비용오분류율이나 기대비용을 낮추고자 한다면, 인위적으로 두 계급의 크기를 같게 해주는 방법1이 아무런 조치를 취하지 않는 방법2보다 나은 방법이라고 할 수 있다. 그리고 두 오분류비용의 비 r을 아는 경우에는 이를 활용하는 방법3과 방법4를 권장할 수 있으며, 인위적으로 과대표집을 하는 대신 가중값만 조정하는 방법4가 소수계급에 대한 과적합을 줄일 수 있다는 장점이 있어 좀더 나은 방법이다.

4. 부스팅 기법을 적용한 실험

기계학습(machine learning) 이론가들이 개발한 부스팅 기법은 약한 분류기(weak classifiers)들을 결합하여 강한 분류기를 만들어내는 기법이다 (Shapire (1990), Freund & Schapire (1997)). AdaBoost 또는 이산형 AdaBoost (Discrete AdaBoost)라고도 불리는 이 알고리즘은, 주어진 자료의 각 관측값에 다른 가중값을 부여하여 약한 분류기를 축차적으로 적합시키는데, 이전 분류기가 잘못 분류한 관측값에 보다 큰 가중값을 부여하여 다음 분류기를 적합시킨다. 이렇게 축차적으로 만들어진 약한 분류기들의 선형결합으로 최종분류기를 생성한다. 약한 분류기로서 나무모형을 이용하면, 부스팅 기법에 의해 생성되는 최종분류기는 많은 나무모형들의 선형결합으로 표현된다. 이 최종분류기는 하나의 나무모형에 비해 오분류율을 효과적으로 낮춰 주는 것으로 알려져 있다(Bauer & Kohabi, 1999).

부스팅 기법을 계급불균형 문제 해결에 이용하고자 하는 연구도 있었다. 대표적 연구로 Ting & Zheng (1998), Ting (2000) 등을 들 수 있다. 하지만 이들 연구는 부스팅 기법에 의해 만들어진 분류기와 나무모형에 의한 분류기의 성능을 비교하거나 주어진 자료를 바탕으로 여러 가능한 부스팅 기법들을 서로 비교하고 있는데, 소수계급의 관측값을 과대추출하는 방법을 비롯하여 앞 절에서 살펴본 여러 방법들에 부스팅을 적용시킨 연구는 아직 없다. 본 절에서는 계급불균형자료의 경우 훈련표본 구성방법에 따라 부스팅 기법의 효과에 어떤 차이가 있는지 알아본다.

실험을 위한 이산형 AdaBoost 알고리즘(Friedman *et al.* (2000))을 R로 작성하였다. 약한 분류기로 rpart에 의한 나무모형을 썼는데 나무의 깊이(depth)는 5로, 최종모형을 구성하는 나무의 수는 100개로 정하였다. 모형의 성능평가를 위해 앞 절에서와 같이 교차타당성을 이용하였으며, 안정된 결과를 얻기 위해 각 자료에 대해 실험을 5번 반복하였다. 부스팅 기법에서는 나무의 깊이와 나무의 수를 미리 정해주므로 모형의 크기가 별 의미가 없어 성능평가기준에서 제외하였다.

방법별 비교를 하기 전에 부스팅의 효과를 먼저 살펴보았다. 그림 4.1은 방법2를 적용

했을 때의 그림이다. 오분류율, 고비용오분류율, 정규화기대비용 등을 살펴보면 부스팅 후에 전체적으로 크게 낮아져, 부스팅의 효과가 분명함을 알 수 있다. 반면에 방법1(부록 그림 A.3)을 비롯하여 방법3과 방법4 등은 부스팅 후에 소수계급의 오분류율이 오히려 높아지는 경향이 있음을 알 수 있다.

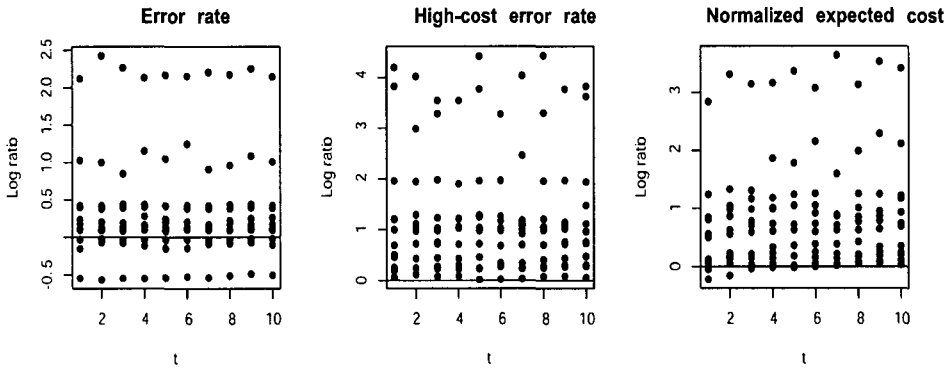

그림 4.1: 방법2에 대한 부스팅의 효과 (log(부스팅 전/부스팅 후))

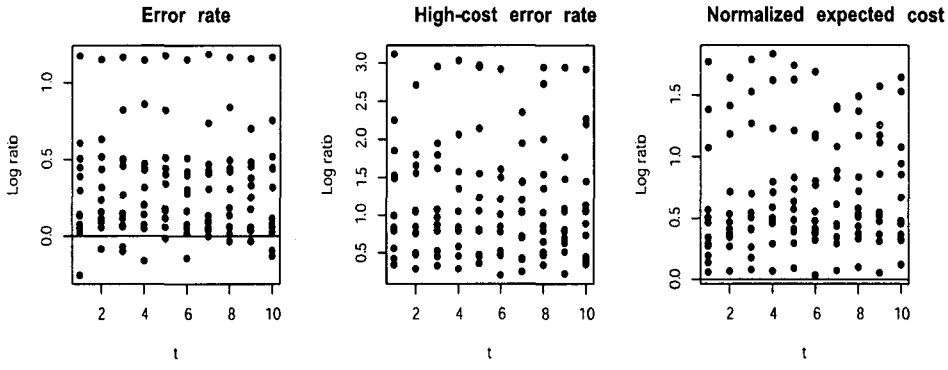

그림 4.2: 부스팅 적용시 방법1과 방법2의 비교 (log(방법1/방법2))

방법1과 방법2를 비교해본 결과(그림 4.2), 부스팅을 하기 전에는 방법1이 방법2에 비해 전체적인 오분류율은 높지만 고비용오분류율은 낮았는데, 부스팅을 적용하면 원래 자료를 있는 그대로 이용하는 방법2가 모든 성능평가기준에서 우월한 결과를 보였다. 이는 매우 흥미있는 결과로서, 부스팅을 하면 소수계급의 오분류된 관측값에 계속 높은 가중값을 자체적으로 부여하면서 소수계급의 오분류율을 효과적으로 낮춰준다는 것을 알 수 있다. 오히려 방법1과 같이 인위적으로 과대표집을 하면 역효과가 난다는 것을 실험결과가 보여준다. 방법2를 방법3과 방법4와 각각 비교한 결과에서도 대체적으로 비슷한 현상이 관측되었

으나, 인위적 표집 대신 가중값만 조정하는 방법4의 경우 방법2보다 우월한 결과를 보이는 자료가 방법3에 비해 몇 개 더 관측되었다. (부록 그림 A.4, 그림 A.5 참조).

5. 토의 및 결론

두 계급의 분류 문제에서 두 계급의 크기가 현저히 불균형일 때, 관행적으로 다수계급에서 과소표집을 하거나 소수계급에서 과대표집을 하여 계급 크기가 균형을 이루도록 한 다음 분류 규칙을 찾는다. 본 연구에서는 이런 관행의 타당성을 알아보고자 하였다.

방법1과 같이 표집을 통하여 인위적으로 두 계급이 균형을 이루도록 한 다음 나무모형을 적합시키면 고비용오분류율이 낮아지는 경향이 있는데, 소수계급의 오분류비용이 상대적으로 높으면 이로 인해 기대비용이 작아진다는 사실을 확인할 수 있었다. 즉 계급 크기가 균형을 이루도록 한 다음 분석하는 관행은 나름대로의 타당성이 있음을 확인할 수 있었다. 하지만 두 오분류비용의 비(r)에 대한 정보가 있다면, 표집을 통해 무조건 균형을 이루도록 하는 방법1보다는 이 정보를 이용하는 방법3이나 방법4가 기대비용의 기준에서 더 나으며, 이 중 소수계급의 가중값을 조정해주는 방법4가 과적합을 줄일 수 있어 더 선호되어야 할 것으로 본다.

오분류율을 낮춰주는 효과적 기법인 부스팅 기법을 계급불균형 문제에 적용해 보았다. 부스팅을 하면 나무모형이 갖는 좋은 해석력을 잃어버리게 되는 단점이 있으나, 해석보다는 예측이 목적이라면 일반적으로 부스팅 기법을 적용하는 것이 바람직하다. 본 연구에서 계급불균형자료에 부스팅 기법을 적용할 때 주어진 자료를 그대로 이용하는 것이 좋다는 유용한 결과를 얻었다.

참고문헌

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

Blake, C. L. and Merz, C. J. (1998). UCI Repository of machine learning databases. http://www.ics.uci.edu/~mlearn/MLRepository.html, University of California in Irvine Department of Information and Computer Science.

Boz O. (2001). Cost Sensitive Learning Bibliography. http://home.ptd.net/~olcay/cost-sensitive.html.

Breiman, L., Friedman, J. H., Olshen, J. A., and Stone, C. J. (1984). *Classification and Regression Trees.* Belmont, CA, Wadsworth.

Drummond, C. and Holte, R. (2000). Explicitly representing expected cost: An alternative to ROC representation. Technical Report, School of Information Technology and Engineering, University of Ottawa.

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*, **55**, 119-139.

Friedman, J., Hastie, T., and Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting. *The Annals of Statistics*, **28**, 337-374.

Shapire, R. E. (1990). The strength of weak learnability. *Machine Learning*, **5**, 197-227.

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

Ting, K. M. (2000). A comparative study of cost-sensitive boosting algorithms. *Proceedings of the 17th International Conference on Machine Learning*, 983-990.

Ting, K. M. and Zheng, Z. (1998). Boosting cost-sensitive trees. *Proceedings of the First International Conference on Discovery Science*, 244-255.

Weiss, G. M. and Provost, F. (2001). The effect of class distribution on classifier learning. Technical Report, Department of Computer Science, Rutgers University.

Woods, K., Doss, C., Bowyer, K., Solka, J., Priebe, C., and Kegelmeyer, P. (1993). Comparative evaluation of pattern recognition technique for detection of microcalcifications in mammography. *International Journal of Pattern Recognition and Artificial Intelligence*, 1417-1436.

[2003년 7월 접수, 2004년 3월 채택]

부록

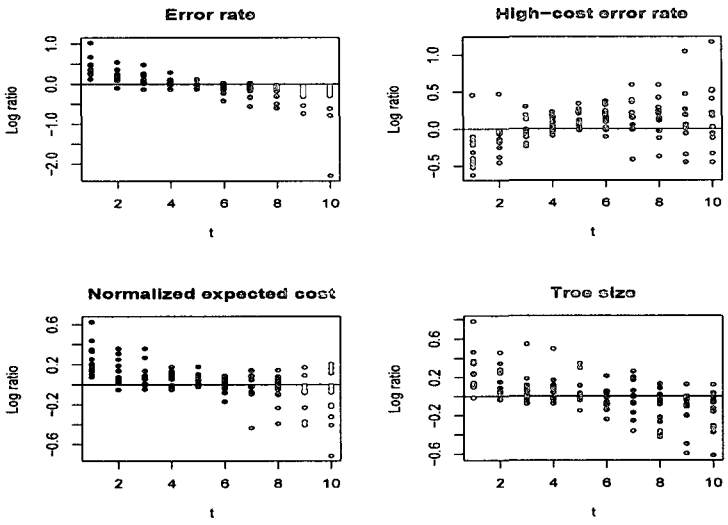

그림 A.1: 방법1과 방법4의 모형평가기준별 로그비 (log(방법1/방법4))

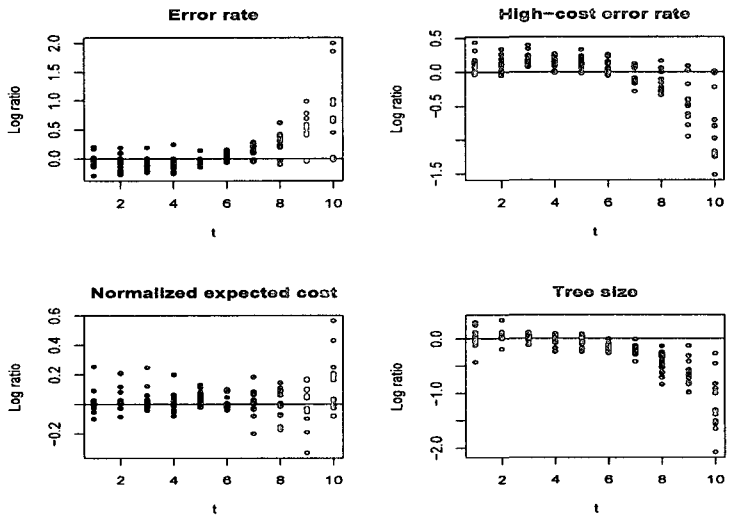

그림 A.2: 방법3과 방법4의 모형평가기준별 로그비 (log(방법3/방법4))

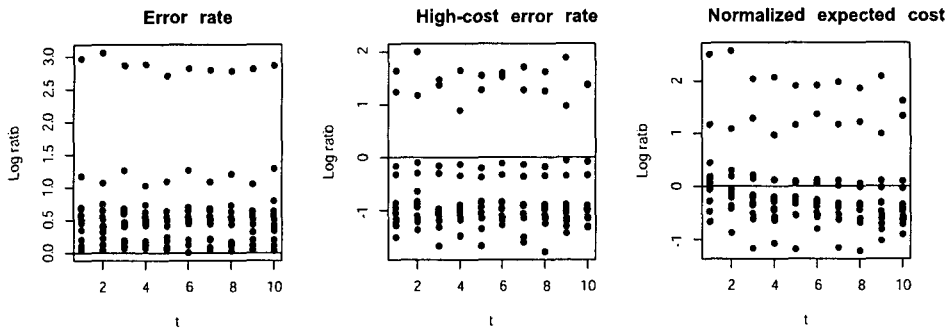

그림 A.3: 방법1에 대한 부스팅의 효과 (log(부스팅 전/부스팅 후))

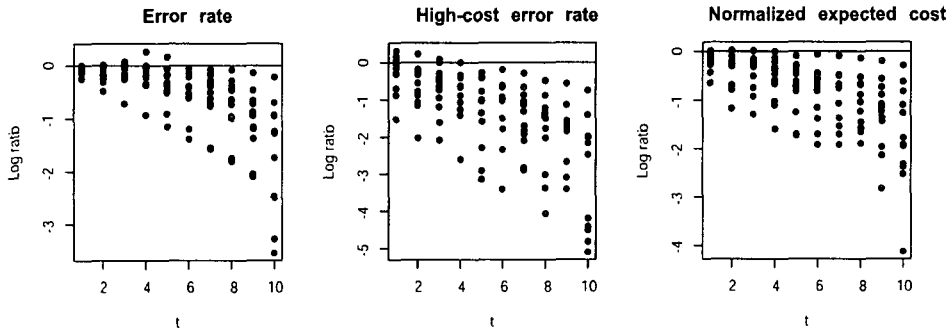

그림 A.4: 부스팅 적용시 방법2와 방법3의 비교 (log(방법2/방법3))

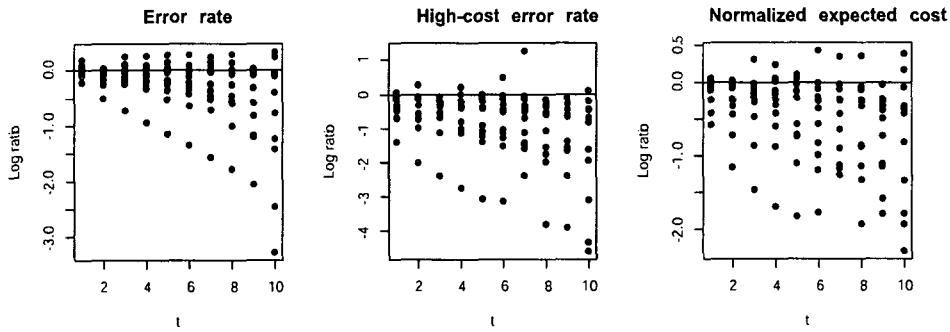

그림 A.5: 부스팅 적용시 방법2와 방법4의 비교 (log(방법2/방법4))

Classification of Class-Imbalanced Data: Effect of Over-sampling and Under-sampling of Training Data *

Ji-Hyun Kim [1] Jongbin Jeong [2]

ABSTRACT

Given class-imbalanced data in two-class classification problem, we often do over-sampling and/or under-sampling of training data to make it balanced. We investigate the validity of such practice. Also we study the effect of such sampling practice on boosting of classification trees. Through experiments on twelve real datasets it is observed that keeping the natural distribution of training data is the best way if you plan to apply boosting methods to class-imbalanced data.

Keywords: Tree model, Boosting, Cost-sensitive learning

* This research was supported by the Soongsil University Research Fund.

1) Professor, Dept. of Statistics, Soongsil University, Sangdo-dong 1-1 Dongjak-ku, Seoul 156-743, KOREA.
E-mail: jhkim@stat.ssu.ac.kr

2) Graduate Student, Dept. of Statistics, Soongsil University, Sangdo-dong 1-1 Dongjak-ku, Seoul 156-743, KOREA.
E-mail: jbjeong@stat.ssu.ac.kr