Journal of the Korean Data & Information Science Society
2010, **21**(5), 917–926

한국데이터정보과학회지

연속형 반응변수를 위한 데이터마이닝 방법 성능 향상 연구[†]

최진수[1] · 이석형[2] · 조형준[3]

[123]고려대학교 통계학과

접수 2010년 7월 25일, 수정 2010년 9월 17일, 게재확정 2010년 9월 23일

요 약

배깅과 부스팅의 기법은 예측력을 향상 시킨다고 알려져 있다. 이는 비교 실험을 통하여 성능이 검증 되었는데, 목표변수가 범주형인 경우에 특정 의사결정나무 알고리즘인 회귀분류나무만 주로 고려되었다. 본 논문에서는 의사결정나무 외에도 다른 데이터마이닝 방법도 고려하여 목표변수가 연속형인 경우에 배깅과 부스팅 기법의 성능 검증을 위한 비교 실험을 실시하였다. 구체적으로, 데이터마이닝 알고리즘 기법인 선형회귀, 의사결정나무, 신경망에 배깅 및 부스팅 앙상블 기법을 결합하여 8개의 데이터를 비교 분석하였다. 실험 결과로 연속형 자료에 대한 여러 데이터마이닝 알고리즘에도 배깅과 부스팅의 기법이 성능 향상에 도움이 되는 것으로 확인되었다.

주요용어: 배깅, 부스팅, 앙상블, 의사결정나무.

1. 서론

사회가 더욱 복잡 다양해짐에 따라 사회 현상을 설명하기 위해 축적한 데이터의 양이 기하급수적으로 증가하고 있다. 컴퓨터 과학 기술의 발달로 대용량의 데이터를 저장하는 기술도 발달하고 이로부터 유용한 정보를 발견하고자 하는 데이터마이닝이 더욱 중요해지고 있다. 데이터마이닝에서는 유용한 정보를 얻기 위해 적절한 모형 설정하는 것이 무엇보다도 중요하다. 따라서 고전적 통계 모형인 선형판별분석 (linear discriminant analysis), 로지스틱 회귀분석 (logistic regression) 등에서 부터 최근 신경망분석 (neural network), 서포트 벡터 머신 (support vector machine)에 이르기까지 우수한 모형 구축을 위한 다양한 알고리즘 (algorithm)에 대한 지속적으로 연구되고 있다 (조영준과 이용구, 2004; 석경하와 류태욱, 2002). 그 중에도 가장 두드러진 연구가 단일 알고리즘을 다양한 형태로 반복 이용하는 앙상블 (ensemble)기법이라 할 수 있다. 앙상블 기법으로 데이터마이닝 성능향상을 위해 가장 많이 연구되고 있는 것이 목표변수가 범주형 (특히, 범주가 두 개)인 자료를 의사결정나무이다. 의사결정나무 중에서도 CART라고 불리는 특정 알고리즘만이 주로 고려되었다. 하지만 목표변수가 연속형인 자료는 상대적으로 연구가 미흡한 것이 사실이다. 이는 성공, 실패와 같은 두 개의 범주를 가진 이진 반응 자료가 흔하고 앙상블 기법의 성능 향상을 설명하기 쉬운 때문일 것이다. 또한, 연속형 반응 변수를 포함한 다른 형태의 자료에도 유사한 결과를 보일 것이라는 예상하기 때문일 것이다.

[†] 이 논문은 2009년도 정부 (교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (2009-0087564).
[1] (136-701) 서울특별시 성북구 안암동 5가 1번지, 고려대학교 정경대학 통계학과, 석사.
[2] (136-701) 서울특별시 성북구 안암동 5가 1번지, 고려대학교 정경대학 통계학과, 석박사 통합과정.
[3] 교신저자: (136-701) 서울특별시 성북구 안암동 5가 1번지, 고려대학교 정경대학 통계학과, 교수.
E-mail: hj4cho@korea.ac.kr

본 논문에서는 CART외에 다른 의사결정나무와 고전적인 회귀모형, 복잡한 결과를 주는 신경망모형도 연구 대상에 포함하고 상대적으로 연구가 미흡한 연속형 반응 자료에 대하여 앙상블 기법의 성능을 검증한다. 우리는 R을 사용하여 각 알고리즘의 예측력을 향상시킬 수 있는 배깅 (bagging) (Breiman, 1996)과 부스팅 (boosting) (Shestha와 Solomatine, 2004)의 성능 향상 실험을 실시한다.

2. 알고리즘 소개

본 논문에서는 목표변수가 연속형인 각 알고리즘에 대하여 3가지 앙상블기법을 R 프로그램을 사용하여 성능 비교하였다. 각 알고리즘을 간략하게 살펴보면 다음과 같다.

2.1. 선형회귀

선형회귀 (linear regression)는 목표변수가 연속형인 경우에 가장 일반적으로 사용하는 회귀분석이다. 이 분석을 선호하는 이유는 여러 함수형태 중에서 직선이 가장 단순하여 다루기가 쉽고, 모든 함수는 독립변수의 구간이 작을 때에 직선으로 근사하게 나타낼 수 있으며, 이론적으로 X와 Y의 결합분포가 이변량정규분포를 따른다면 Y의 조건부 기대치 $E(Y|X)$는 X의 선형함수 즉 직선이 되기 때문이다. 모형은 다음과 같이 나타낸다.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon \tag{2.1}$$

여기서 β_0는 절편 (상수항)이고, β_i는 각 독립변수의 계수이며, p는 선형 회귀로 추정되는 모수의 개수이다. 또한 오차항의 평균은 0이고, 오차항들은 독립이며, 등분산이고, 정규분포를 따라야한다는 가정을 충족시켜야 이 모형을 쓸 수 있다.

2.2. 의사결정나무

의사결정나무란 의사결정규칙을 도표화하여 관심대상이 되는 집단을 몇 개의 소집단으로 분류하거나 예측을 수행하는 데이터마이닝 기법 중 하나이다 (박희창과 조광현, 2004). 분석과정이 나무구조에 의해 표현되기 때문에 분석과정을 쉽게 이해하고 설명할 수 있다는 장점을 가지고 있다.

의사결정나무에서 분류 또는 예측을 하고자 하는 대상을 목표변수라고 하며, 목표변수와 관련된 여러 변수들을 예측변수라고 한다. 즉 p개의 예측변수를 $X_1, X_2, \cdots, X_p$라고 하고 목표변수를 Y라고 하면, 예측변수들은 목표변수의 분포를 설명하는 변수들이 된다. 의사결정나무는 목표변수의 성격 (범주형 또는 연속형)에 따라 분류 나무 (classification tree)와 회귀 나무 (regression tree)로 구분된다. 분류 나무는 목표변수가 범주형인 경우 생성되는 의사결정나무로서, 목표변수의 각 범주에 속하는 사례의 빈도에 기초하여 분류를 수행한다. 이에 반해 회귀 나무는 목표변수가 연속형인 경우에 생성되는 의사결정나무로서 목표변수의 평균에 의해 분리를 수행한다. 본 논문에서는 목표변수가 연속형인 경우인 회귀 나무를 실시한다. 이런 회귀나무에는 CART, GUIDE 등이 쓰이고 있다.

CART: CART (Classification And Regression Trees)는 Breiman *et al.* (1984)에 의해 발표되었으며, 가장 널리 알려지고 사용되고 있는 알고리즘이다. 목표변수가 연속형인 경우에는 분산의 감소량 또는 F-통계량을 이용하여 부모 노드 (parent node)로부터 자식 노드 (child node)를 2개만 형성하는 이진분리 (binary split)를 수행하는 알고리즘이다. 목표변수의 F-통계량에 의해 분리마디를 결정하는 경우에는 목표변수와 설명변수들 간에 가장 설명력이 큰 변수를 먼저 선택하여 분리한다. F-통계량을 계산하는 절차는 많이 알려진 통계분석법인 분산분석 (ANOVA)와 동일한 방법이다. 특정한 마디를 분

리함으로써 분산이 줄어든다는 것은 동질적 성격을 가지고 있는 개체들은 같은 집단으로 이질적 성격을 가지고 있는 개체들은 다른 집단으로 나뉘었다는 것을 의미하므로, 분산을 가장 크게 감소시키는 분리조합이 최선의 분리를 이끌어 내는 것이다. 이를 알아보기 위해 CART의 R버전인 rpart 함수를 사용하였다.

GUIDE: GUIDE (Generalized, Unbiased Interaction Detection and Estimation)는 Loh (2002)에 의해 발표된 알고리즘이다. GUIDE의 특징은 잔차와 부스트랩으로 교정된 유의확률의 카이제곱 검정을 이용한 결과 변수선택의 편향이 무시해도 될 정도이며, 변수들 사이의 2가지 (curvature, interactions)검정으로 실시되고, 범주형 예측변수와 순서가 있는 범주형 변수를 포함한다. 또한 각 예측변수를 분리점을 찾는 것에만 사용하는 변수, 모델링에만 사용하는 변수, 분리점을 찾고 모델링에도 사용하는 3가지 변수로 역할을 결정하는 것이 특징이다.

GUIDE는 분리변수를 선택하기 위해서 상수 모형 또는 선형회귀모형 적합 후의 잔차를 0을 기준으로 2개로 분류하고 독립변수를 연속형인 경우에는 사분위수를 이용하고, 범주형인 경우는 모든 범주별로 나누어 분할표를 만든다. 이 분할표를 기준으로 χ^2 통계량의 p-value를 구한다. 이것은 목표변수를 하나의 독립변수로 설명하는 것 (curvature test)이고, 두 개의 독립변수로 설명하는 경우도 생각을 하여 결합시킬 수 있는 모든 경우 (interaction test)를 생각해 본다. 모두 연속형인 경우에는 잔차를 대신하여 각 변수의 표본평균값으로 나누어 분류하고, 하나의 경우에만 연속형인 경우는 연속형 변수는 표본평균값으로 나눈 후 범주형 변수인 경우 각 범주별로 나눈다. 마지막으로 모두 범주형인 경우에는 각 범주별로 분할하여 χ^2 통계량의 p-value를 구한다. 이때 구해진 모든 p-value를 비교하여 가장 작은 값을 가지는 것을 분리변수로 선택한다. 여기서 만약 하나의 독립변수로 목표변수를 설명하는 p-value값이 가장 작으면 구해진 변수를 분리변수로 선택한다. 하지만 두 개의 독립변수로 목표변수를 설명하는 p-value값이 작으면 이때 구해진 두 개의 변수 중에서 하나의 변수를 선택하여야 한다. 두 변수가 연속형 변수인 경우에는 각각의 SSE값을 비교하여 더 작은 값을 가지는 변수를 선택하고, 적어도 하나의 변수가 범주형 변수인 경우에는 앞에서 구한 하나의 독립변수로 목표변수를 설명하는 p-value값이 작은 변수로 선택한다.

2.3. 신경망

신경망 (neural network)은 사람 뇌의 구조를 모방한 비선형 기법이다. 신경망은 입력층 (input layer), 은닉층 (hidden layer) 그리고 출력층 (output layer)으로 구성되어 있다. 입력층으로부터 전달되는 변수값들의 선형결합을 은닉층에서 비선형함수로 처리하여 출력층으로 전달하는 것이다. 신경망에서 쓰이는 계산요소 또는 노드는 비선형적이며 대개 아날로그인데 현재의 디지털 회로에 비하여 속도면에서 매우 느리다. 가장 간단한 노드는 N개의 입력을 받아 N개의 연결강도의 벡터들과 곱해져서 특정한 출력함수 (activation fuction, transfer fuction)를 거쳐 출력을 내게 된다. 인공 신경망 모델에서 뉴런의 주요 기능은 입력과 연결 강도의 가중합을 구한 다음 활성화 함수에 의해 출력을 내보내는 것이다. 따라서, 어떤 활성화 함수를 선택하느냐에 따라 뉴런의 출력이 달라질 수 있다.

이런 신경망 분석은 비선형 기법이지만 비선형 회귀분석처럼 특정한 설정이 필요없다. 또한 상효 작용 효과가 뚜렷이 파악된다. 그러나 얻어진 모형이 복잡하여 해석하기가 힘들다는 단점이 있다.

3. 앙상블 기법 소개

3.1. 배깅

데이터가 조금이라도 변하는 상황에서 분류자의 변동이 큰 경우에는 예측 결과의 변동성을 감소시키고자 부스트랩 (boostrap) 방법 (Efron과 Tibshirani, 1993)을 통해 분류기를 얻을 수 있다. 이러한 방법을 배깅 (bagging) 알고리즘이라 하며 Breiman (1996)에 의해 제안 되었다 (이상복, 2001).

배깅 앙상블 과정은 모집단으로부터 추출된 분석용 데이터에서 랜덤하게 복원추출로 관측값이 n개인 부스트랩 표본을 M개 생성한다. 생성된 M개의 부스트랩 표본에서 각각의 단일 분류기를 형성하여 단일 분류기 집합을 얻는다. 이러한 여러 단일 분류기를 통합하는 방법은 목표변수가 연속형일 때와 범주형일 때 다르게 적용된다. 목표변수가 연속형 변수이면 평균을 이용하고, 범주형 변수이면 다중 투표 (majority vote) 방법을 사용한다. 이렇게 결합되어 형성된 분류기를 배깅 분류기라고 한다. 분석용 데이터가 불안정 (unstable)하다면, 배깅 분류기는 결합을 통해서 그 예측 정확도가 향상된다. 본 논문에서는 목표변수가 연속형일 경우를 실시한다.

3.2. 부스팅

부스팅 (boosting)은 이전 분류기의 수행을 기반으로 재조정된 분석용 자료를 이용해 분류기를 순차적으로 생성한다. 분류기에 의해 오분류된 관측치는 가중치를 높게 주고, 정분류된 관측치는 가중치를 낮게 줌으로써 가중치가 재조정된다. 가중치가 재조정된 관측치들에 의해 새로운 분석용 자료가 형성되는데 이때 높은 가중치를 가지는 관측치가 많이 선택됨으로써 오분류되기 쉬운 관측치를 올바르게 분류되도록 한다 (Freund와 Schapire, 1996; Dietterich, 2000).

최초로 사용되는 가중치는 동일하게 시작하며 이전 모형의 성능에 근거하여 분석용 자료의 분포를 계속 변화시키는 기법이라고 할 수 있다. 본 논문에서는 회귀 (regression)에 적합한 부스팅을 하기 위하여 Adaboost.RT (regression threshold)를 이용하였다 (Schrestha와 Solomatine, 2004).

4. 데이터 소개

이 장에서는 본 논문에 사용된 데이터를 소개하도록 한다. 목표변수가 연속형인 변수를 갖는 총 8개의 데이터가 분석에 사용되었으며, 전체자료를 70%의 훈련용 (training) 자료와 30%의 검정용 (testing) 자료로 분할하여 분석하였다. 각 데이터에 대해 간략하게 살펴보면 다음과 같다.

Machine data: 1987년 CPU 성능과 관련된 데이터 (Ein-Dor와 Feldmesser, 1987)로서, 종속변수는 공인된 CPU의 성능을 나타낸 연속형 값이다. 설명변수는 9개 (범주형 설명변수 2, 연속형 설명변수 7)이고, 208개의 관측치로 이루어져 있다. 이 데이터는 CPU 성능을 올리기 위하여 관련이 되는 변수를 찾기 위한 데이터이다.

Dmabase data: 1986년 야구선수들의 연봉과 관련된 기록 데이터로서, 종속변수는 1987년의 연봉의 Log변환한 연속형 값이다. 설명변수는 20개 (범주형 설명변수 4, 연속형 설명변수 16)이고, 263개의 관측치로 이루어져 있다. 이 데이터는 1987년 야구선수들의 연봉을 이용하여 연봉을 높게 받기 위하여 어떤 것이 중요한 자료가 되는지 알아보고자 하는 데이터이다.

Auto-mpg data: 1993년 자동차와 관련된 데이터 (Quinlan, 1993)로서, 종속변수는 1갤런당 가는 마일 (miles per gallon)을 나타낸 연속형 값이다. 설명변수는 8개 (범주형 설명변수 3, 연속형 설명변

수 5)이고 392개의 관측치로 이루어져 있다.

Housing data: 주택가격을 결정하는 요인의 관한 데이터 (Harrison과 Rubinfeld, 1978)로서, 종속변수는 평균 주택가격 ($1,000)을 나타내는 연속형 값이다. 설명변수는 13개 (범주형 설명변수 0, 연속형 설명변수 13)이고 506개의 관측치로 이루어져 있다.

Forestfires data: 산불이 일어난 지역에 관한 데이터 (Cortez와 Morais, 2007)로서, 종속변수는 불이 난 지역을 0.0에서 기울어진 정도를 log 변환하여 나타낸 연속형 값이다. 설명변수는 12개 (범주형 설명변수 2, 연속형 설명변수 10)이고 517개의 관측치로 이루어져 있다.

Concrete data: 콘크리트 압축 강도에 관한 데이터 (I-Cheng, 1999)로서, 종속변수는 콘크리트 압축 강도로서 연속형 값이다. 설명변수는 8개 (범주형 설명변수 0, 연속형 설명변수 8)이고 1,030개의 관측치로 이루어져 있다.

Wage data: 임금을 결정하는 요인들의 관한 데이터 (Berndt, 1991)로 1978년, 1985년 두 시기의 조사된 데이터이다. 종속변수는 시간당 평균임금을 Log변환한 값인 연속형 값이다. 19개의 설명변수 (범주형 설명변수 15, 연속형 설명변수 4)와 1,084개의 관측치로 이루어져 있다.

Buytest data: 물건을 구입한 사람들의 신상 데이터로서, 종속변수는 DM에 의한 구입총액을 나타내는 연속형 값이다. 16개의 설명변수 (범주형 설명변수 12, 연속형 설명변수 4)와 10,000개의 관측치로 이루어져 있다.

표 4.1 데이터요약

데이터	종속변수 (Y)	관측치 (N)	설명변수의 수	
			범주형	연속형
Machine	연속형	208	2	7
Dmabase	연속형	263	4	16
Auto-mpg	연속형	392	3	5
Housing	연속형	506	0	13
Forestfires	연속형	517	2	10
Concrete	연속형	1,030	0	8
Wage	연속형	1,084	15	4
Buytest	연속형	10,000	12	4

5. 성능 비교 결과

본 절에서는 4장에서 소개된 데이터들에 대해 각 알고리즘 분석 결과를 살펴보기로 한다. 알고리즘의 비교 기준으로는 MSE (Mean Square Error)값이 사용되었다. 전체 반복실험횟수는 100번으로 실시하였다.

5.1. 데이터별 성능 비교

표 5.1 Machine data의 분석 결과를 보면, rpart에서는 부스팅을 25번 반복한 MSE값이 7306.5로 가

장 좋은 값으로 나타났다. 선형회귀와 신경망 역시 부스팅에서 가장 좋은 값을 가지는 것으로 나타났다. 약간의 차이점으로는 반복이 50번했을 때 3567.1, 4759.4로 가장 좋은 값을 가지는 것을 알 수 있다. 이와는 다르게 GUIDE 에는 각각 1489.9인 값을 가지는 배깅에서 25번 반복했을 때가 가장 좋은 값을 가지는 것으로 나타났다.

표 5.1 Machine data의 분석 결과: MSE

		Linear Regression	CART (rpart)	GUIDE	Neural network
Base		4087.9	10944.1	1947.4	10256.7
Bagging	25	4273.0	7384.2	**1489.9**	5565.1
	50	3617.7	7466.4	1526.4	5434.7
Boosting	25	4117.8	**7306.5**	1865.7	5335.5
(Adaboost.RT)	50	**3567.1**	7699.2	1764.8	**4759.4**

표 5.2 Dmabase data의 분석 결과를 보면 rpart에서는 배깅을 50번 반복한 MSE값과 배깅을 50번 반복한 MSE값과 부스팅을 50번 반복한 MSE값이 0.262로 같은 값으로 가장 좋은 값으로 나타났다. GUIDE와 신경망 경우에는 부스팅을 50번 반복한 값이 0.149, 0.325로 가장 좋은 값으로 나타났다. 이와는 다르게 선형회귀에서는 배깅을 50번 반복한 값이 0.385로 가장 좋은 값으로 나타났다.

표 5.2 Dmabase data의 분석 결과

		Linear Regression	CART (rpart)	GUIDE	Neural network
Base		0.389	0.332	0.176	0.352
Bagging	25	0.397	0.275	0.158	0.337
	50	**0.385**	**0.262**	0.155	0.338
Boosting	25	0.402	0.275	0.159	0.327
(Adaboost.RT)	50	0.392	**0.262**	**0.149**	**0.325**

표5.3 Auto-mpg data의 분석 결과를 보면 rpart에서는 부스팅을 50번 반복하였을 때의 값이 17.9로 가장 좋은 값으로 나타났다. 이와는 다르게 GUIDE와 선형회귀에서는 배깅에서 가장 좋은 값으로 나타났는데 반복수가 각각 50번, 25번일 때 7.9, 11.3로 가장 좋은 값으로 나타났다. 신경망에서는 배깅을 25번 했을 때의 값과 부스팅을 50번 했을 때의 값이 동일하게 12.0로 가장 좋은 값으로 나왔다.

표 5.3 Auto-mpg data의 분석 결과

		Linear Regression	CART (rpart)	GUIDE	Neural network
Base		11.6	27.6	10.7	12.1
Bagging	25	**11.3**	27.7	8.1	**12.0**
	50	11.4	28.1	**7.9**	12.8
Boosting	25	11.5	18.2	8.2	12.2
(Adaboost.RT)	50	11.7	**17.9**	8.1	**12.0**

표 5.4 Housing data의 분석결과를 보면 rpart, GUIDE, 선형회귀, 신경망 모든 경우에서 부스팅에서 가장 좋은 결과를 나타냈다. 먼저 rpart를 보면 50번 반복한 결과 25.8로 가장 좋은 값을 나타냈다. 나머지 GUIDE, 선형회귀, 신경망에서는 25번 반복한 결과 15.1, 23.7, 19.2로 가장 좋은 값을 나타냈다.

표 5.4 Housing data의 분석 결과

		Linear Regression	CART (rpart)	GUIDE	Neural network
Base		23.8	36.2	23.1	25.2
Bagging	25	24.0	36.9	15.3	22.4
	50	24.1	36.7	16.1	21.0
Boosting	25	**23.7**	25.9	**15.1**	**19.2**
(Adaboost.RT)	50	24.3	**25.8**	15.9	20.3

표 5.5 Forestfires data의 분석 결과를 보면 rpart와 신경망에서는 부스팅을 50번 반복한 결과 3728.1, 3639.4로 가장 좋은 결과를 나타냈다. 이와는 다르게 GUIDE와 선형회귀는 배깅을 50번 반복한 결과 3664.0, 3749.1로 가장 좋은 결과를 나타냈다.

표 5.5 Forestfires data의 분석 결과

		Linear Regression	CART (rpart)	GUIDE	Neural network
Base		4328.2	3824.6	4045.7	4647.8
Bagging	25	4086.4	4091.7	3669.6	4040.8
	50	**3749.1**	4424.5	**3664.0**	4625.9
Boosting	25	4314.5	4342.4	4027.0	4472.9
(Adaboost.RT)	50	4220.9	**3728.1**	4092.1	**3639.4**

표 5.6 Concrete data의 분석 결과를 보면 rpart, GUIDE, 신경망에서 동일하게 부스팅을 50번 반복한 결과 각 202.9, 27.3, 152.1로 가장 좋은 결과를 나타냈다. 이와는 다르게 선형회귀인 경우 배깅을 25번 반복한 결과 155.9가 가장 좋은 결과를 나타냈다.

표 5.6 Concrete data의 분석 결과

		Linear Regression	CART (rpart)	GUIDE	Neural network
Base		156.1	220.9	39.6	157.9
Bagging	25	**155.9**	203.8	28.0	153.0
	50	156.8	203.1	28.1	153.2
Boosting	25	156.4	203.7	28.4	152.3
(Adaboost.RT)	50	158.8	**202.9**	**27.3**	**152.1**

표 5.7 Wage의 분석 결과를 보면 rpart, 선형회귀, 신경망에서 동일하게 부스팅을 50번 반복했을 때가 가장 좋은 결과로 나타났다. 그 값들은 0.241, 0.162, 0.274이다. 이와는 다르게 GUIDE 경우에는 배깅을 25번 반복한 결과 0.162로 가장 좋은 결과로 나타났다.

표 5.7 Wage data의 분석 결과

		Linear Regression	CART (rpart)	GUIDE	Neural network
Base		0.163	0.267	0.167	0.311
Bagging	25	0.167	0.243	**0.162**	0.285
	50	0.163	0.243	0.163	0.279
Boosting	25	0.164	0.245	0.164	0.276
(Adaboost.RT)	50	**0.162**	**0.241**	0.165	**0.274**

표 5.8 Buytest data의 분석 결과를 보면 rpart, 선형회귀는 동일하게 배깅에서 가장 좋은 결과를 나타냈다. rpart와 선형회귀는 50번 반복에서 각각 713.1, 692.2로 가장 좋은 결과를 나타냈다. 이와는 다르게 GUIDE와 신경망의 경우 부스팅에서 가장 좋은 결과를 나타냈다. 둘 모두 50번 반복한 결과 684.2, 708.3이 가장 좋은 결과를 나타냈다.

표 5.8 Buytest data의 분석 결과

		Linear Regression	CART (rpart)	GUIDE	Neural network
Base		708.1	724.1	700.1	726.6
Bagging	25	712.8	730.5	711.1	709.0
	50	**692.2**	**713.1**	698.3	719.8
Boosting	25	707.6	714.9	699.5	711.2
(Adaboost.RT)	50	701.9	721.1	**684.2**	**708.3**

6. 결과 및 토의

본 논문에서는 앙상블 기법인 배깅, 부스팅이 연속형 반응변수를 위한 데이터마이닝 방법의 성능을 향상시키는 것에 대해 연구하였다. 실제 8가지 연속형 목표변수 데이터에 대해 평균제곱오차 및 시간을 각 알고리즘 및 기법에 대해 그 성능을 분석하였다. 그 결과 모든 데이터에 대해서 의사결정나무인 rpart (CART의 R버전)에서는 마지막 데이터를 제외하고 부스팅이 가장 좋은 방법으로 선택된 것을 알 수 있었다. 또한 의사결정나무인 GUIDE의 경우는 배깅과 부스팅 중에서 어느 한쪽이 좋은 방법으로 선택되기 보다는 경우에 따라서 더 좋은 방법으로 선택된 것을 알 수 있었다. 선형회귀의 경우는 부스팅보다 배깅이 더 좋은 방법으로 많이 선택된 것을 알 수 있었다. 하지만 그 격차가 거의 나지 않기 때문에 이 역시 어느 한쪽이 좋은 방법으로 선택되었다고 말할 수 없다. 마지막으로 신경망인 경우는 모든 데이터에 대해서 부스팅이 가장 좋은 방법으로 선택된 것을 알 수 있었다.

부스팅은 분류가 잘되지 않는 데이터를 중점으로 분류를 잘할 수 있도록 하는 알고리즘이고, 이에 반해 배깅은 분류가 잘되지 않는 데이터의 중점을 두는 것이 아니라 전반적으로 분류를 잘 할 수 있도록 유도하는 것으로 여러번 실시 후 평균 값으로 하는 것이기 때문에 배깅보다는 부스팅이 평균제곱오차의 값이 더 작게 나와 예측력이 더 좋은 것이다.

본 논문의 결과, 어느 경우에서라도 배깅과 부스팅은 성능 향상의 도움이 된다는 것을 확인했다. 하지만 본 논문의 실험으로만 배깅과 부스팅의 성능을 단정 짓기에는 무리가 있을 수 있다. 더욱 다양한 형태의 자료로 비교 실험이 요구된다고 할 수 있다.

참고문헌

박희창, 조광현 (2004). 의사결정나무기법에 의한 환경조사 모형화. <한국데이터정보과학회지>, **15**, 759-771

석경하, 류태욱 (2002). The efficiency of boosting on SVM. <한국데이터정보과학회지>, **13**, 55-64

이상복 (2001). 데이터마이닝기법상에서 적합된 예측모형의 평가 - 4개 분류예측모형의 오분류율 및 훈련시간 비교평가 중심으로. <한국데이터정보과학회지>, **12**, 113-124

조영준, 이용구 (2004). 단층퍼셉트론 모형에서 초기치 최적화 방법에 관한 연구. <한국데이터정보과학회지>, **15**, 331-337

Berndt, E. (1991). *The practice of economics: Classic and contemporary, reading*, Mass, Addison-Wesley.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification and regression trees*, New York, Chapman and Hall.

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

Cortez, P. and Morais, A. (2007). A data mining approach to predict forest fires using meteorological data, In Neves, J.M. and Santos, M.F. and Machado J.M.. *New Trends in Artificial Intelligence: Proceedings of the 13th EPIA 2007* - Portuguese Conference on Artificial Intelligence, December, Guimaraes, Portugal, 512-523.

Dietterich, T. G.(2000). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting and randomization. *Machine Learning*, **40**, 139-158.

Efron, B and Tibshirani, R. J (1994). *Nonparametric regression and generalized linear models*, New York, Chapman and Hall.

Ein-Dor, P. and Feldmesser, J. (1987). Attributes of the performance of central processing units: A relative performance prediction model. *Communications of the ACM*, **30**, 308-317.

Freund, Y. and Schapire, R. E. (1996). Experiments with a new boosting algorithm. *Machine Learning*, Proceedings of the Thirteenth International Conference 148-156. Morgan Kauffman, San Francisco.

Harrison, D. and Rubinfeld, D. L. (1978). Hedonic prices and the demand for clean air. *Journal of Environmental Economics & Management*, **5**, 81-102.

I-Cheng, Y. (1999). Design of high-performance concrete mixture using neural networks and nonlinear programming. *Journal of Computing in Civil Engineering*, **13**, 36-42.

Loh, W. Y. (2002). Regression trees with unbiased variable selection and interaction detection. *Statistica Sinica*, **12**, 361-386.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning, San Mateo*, CA Morgan Kaufmann.

Shestha, D. L. and Solomatine, D. P. (2004). *AdaBoost.RT: A boosting algorithm for regression problems,* International Joint Conference on Neural Networks, Budapest, Hungary.

Journal of the Korean Data & Information Science Society
2010, **21**(5), 917–926

한국데이터정보과학회지

A study for improving data mining methods for continuous response variables[†]

Jin Soo Choi[1] · Seok Hyung Lee[2] · HyungJun Cho[3]

[123]Department Statistics, Korea University

Received 25 July 2010, revised 17 September 2010, accepted 23 September 2010

Abstract

It is known that bagging and boosting techniques improve the performance in classification problem. A number of researchers have proved the high performance of bagging and boosting through experiments for categorical response but not for continuous response. We study whether bagging and boosting improve data mining methods for continuous responses such as linear regression, decision tree, neural network through bagging and boosting. The analysis of eight real data sets prove the high performance of bagging and boosting empirically.

Keywords: Bagging, boosting, decision tree, ensemble.

[†] This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2009-0087564).
[1] Master, Department of Statistics, Korea University, Seoul 136-701, Korea.
[2] Combined master's and doctorate program, Department of Statistics, Korea University, Seoul 136-701, Korea.
[3] Corresponding author: Professor, Department. of Statistics, Korea University, Seoul 136-701, Korea. E-mail: hj4cho@korea.ac.kr