앙상블기법을 이용한 다양한 데이터마이닝 성능향상 연구

정연해^a, 어수행^a, 문호석^b, 조형준^{1,a}

^a고려대학교 통계학과, ^b육군사관학교 전자공학과

요약

본 논문은 8가지 방법의 데이터 마이닝 알고리즘(CART, QUEST, CRUISE, 로지스틱 회귀분석, 선형판별분석, 이차판별분석, 신경망분석, 서포트 벡터 머신) 기법과 단일 알고리즘에 2가지 앙상블기법(배깅, 부스팅)을 적용한 16가지 방법을 바탕으로 총 24가지의 방법을 비교하였다. 알고리즘의 성능 비교를 위하여 13개의 이항반응변수로 구성된 데이터를 사용하였다. 비교 기준은 민감도, 특이도 및 오분류율을 사용하여 데이터 마이닝 기법의 성능향상에 대해 평가하였다.

주요용어: 앙상블기법, 배깅, 부스팅, 데이터 마이닝.

1. 서론

대용량 데이터에서 유용한 정보와 관계를 탐색하고 모형화하여 지식을 발견하고자 하는 일련의 과정을 데이터 마이닝(data mining)이라고 한다. 고전적 통계 모형인 선형판별분석(linear discriminant analysis), 로지스틱 회귀분석(logistic regression) 등에서부터 최근 신경망분석(neural network), 서포트 벡터 머신(support vector machine)에 이르기까지 다양한 알고리즘(algorithm)이 데이터 마이닝에 사용되고, 지속적으로 연구되고 있다. 데이터 마이닝 기법을 이용하여 모형을 구축할 때 주어진 데이터를 이용하여 목표변수(target variable)를 가장 잘 예측할 수 있는 분류자(classifier)를 형성하는 것이 모든 분류기법의 목적이다.

분류자를 형성하는데 있어서 단일 분류자(single classifier)보다 다중 분류자(multiple classifiers)를 앙상블 방법에 의해 결합한 분류자가 경험적으로 더 정확하고 좋은 결과가 나온다는 것이 여러 연구자들에 의해 제안하였다 (Breiman, 1996; Clemen, 1989; Perrone, 1993; Wolpert, 1992).

앙상블기법 중에는 대표적으로 배깅(bagging)과 부스팅(boosting) 알고리즘이 있다. 이 알고리즘들은 분석용(training) 데이터에서 재표본(resampling)기법으로 얻어진 데이터에 의해 각 분류자가 형성된다. 배깅의 경우 관측값이 동일한 가중치를 갖는 분포에서 추출된 데이터에서 분류자가 형성되지만 부스팅은 관측값의 가중치가 변하는 분포에서 추출된 데이터에서 분류자가 형성된다. 이러한 앙상블 기법은 종종 단일 데이터 마이닝 기법보다 더 좋은 성능을 갖는 것이 실험적으로 보여져왔다 (Bauer와 Kohavi, 1999; Drucker와 Cortes, 1996; Breiman, 1996; Freund와 Schapire, 1996; Quinlan, 1996). 특히 부스팅의 경우 분류하기 힘든 관측값에 대해서 높은 가중치를 주게되어 분류를 잘하도록 유도를 하기 때문에 일반적으로 부스팅은 배깅과 비교했을 때 매우 효과적으로 오차를 감소시킨다고 알려져 있다.

본 논문에서는 다양한 데이터 마이닝 기법(8가지)에 앙상블기법인 배깅과 부스팅 알고리즘을 적용시켜 이항반응변수를 갖는 13개의 데이터의 분석결과를 통해 포괄적인 평가를 하고자 한다. 논문의 구

고려대학교 정경대학 교수 연구비에 의하여 수행되었음.

¹ 교신저자: (136-701) 서울 성북구 안암동 5가 1번지, 고려대학교 정경대학 통계학과, 부교수.

E-mail: hj4cho@korea.ac.kr

표 1: 논문에 사용된 프로그램과 알고리즘

Methods		Program
Decision Tree	RPART	RPART
	QUEST	R with FORTRAN/95
	CRUISE	R with FORTRAN/95
Classification	LDA	LDA
	QDA	QDA
Logistic Regression		GLM
Neural Networks		NNET
SVM		SVM

성은 다음과 같다. 2장에서는 본 논문에 사용된 8가지의 기본 데이터 마이닝 알고리즘에 대해 소개하고 각 방법들에 대해 간략하게 살펴본다. 3장에서는 2가지의 앙상블기법을 소개하기로 한다. 4장에서는 이항반응변수를 갖는 총 13개의 데이터들을 소개하기로 한다. 5장에서는 2, 3장에 소개한 알고리즘 및 앙상블 기법을 중심으로 4장에서 소개한 예제데이터들을 오분류율(misclassification rate), 민감도(sensitivity), 특이도(specificity)를 기준으로 비교하기로 한다. 마지막으로 6장에서는 위 결과를 종합하여 앙상블 알고리즘의 성능을 평가해보고 향후 연구방향에 대해서 논의하도록 하겠다.

2. 데이터마이닝 알고리즘

본 논문에서는 R (version 2.6.2) 또는 FORTRAN/95으로 만들어진 프로그램을 사용하였다. 표 1에 본 논문에서 사용된 프로그램과 분류 옵션을 포함한 총 8가지 알고리즘이 제시되어 있다. FORTRAN/95으로 짜여진 QUEST와 CRUISE는 R 패키지 상에서 FORTRAN/95과 연동을 시키는 프로그램을 만들어 반복시행을 해주었다. 각 알고리즘에 대하여 간략하게 살펴보면 다음과 같고, 세부 내용은 알고리즘에서 인용된 논문을 참고하기 바란다.

2.1. 의사결정나무

의사결정나무분석은 가정 사항이나 모형에 대한 제약을 덜 받기 때문에 어떠한 자료의 형태나 구조에서도 쉽게 사용이 가능하며, 모형이 나무구조로 표현되어 해석의 용이하다. 또한 설명변수들 간의 교호작용을 쉽게 발견할 수 있으며, 이상치(outlier)의 영향을 덜 받고, 결측치(missing value)를 간단하게 제어할 수 있는 장점이 있다.

의사결정나무의 형태로는 반응변수의 형식에 따라 분류나무(classification tree)와 회귀나무(regression tree)로 나눌 수 있다. 반응변수가 이산형인 경우 분류나무를 구성하고 연속형인 경우는 회귀나무를 구성한다. 본 논문에서는 반응변수가 이항변수이므로 분류나무를 중점적으로 다루기로 한다. 현재까지 많은 의사결정나무를 기반으로 한 기법들이 제안되었으며, 대표적으로 CHAID (Kass, 1980) CART (Breiman 등, 1984), C4.5 (Quinlan, 1992), QUEST (Loh와 Shih, 1997), CRUISE (Kim과 Loh, 2001) 등이 있으며, 최근에는 이들의 장점을 결합하여 보다 개선된 알고리즘들이 제안되고 상용화되고 있다. 자세한 내용은 참고문헌을 참조하기 바란다.

2.2. 로지스틱 회귀분석

로지스틱 회귀분석은 반응변수가 범주형 또는 순서형일 때 확률에 대한 로짓변환(logit transformation)을 하여 분석하는 방법으로, 설명변수 x_i 와 반응변수 Y 사이에 아래와 같은 선형관계가 있음을 가정한다 (김규곤, 2003; 허명희와 서혜선, 2001). 식 (2.1)을 로짓 모형(logit model)이라고 정의하며, 일

반적으로 로지스틱 회귀분석의 결과는 선형판별분석과 비슷한 결과를 갖게 된다.

$$f(x) = \log \frac{P(Y = 1 | x_1, x_2, \dots, x_p)}{P(Y = 0 | x_1, x_2, \dots, x_p)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p. \quad (2.1)$$

2.3. 선형/이차 판별 분석

판별분석은 2개 이상의 범주를 갖는 목표변수 Y 의 집단 간의 차이를 식별하는데 사용되는 방법이며 그 기준으로 판별함수를 사용한다. 판별함수는 각 범주를 가장 잘 구분 짓는 변수 x_i ($i = 1, \dots, p$)들의 선형 또는 비선형 함수를 도출하고, 이를 기반으로 한 분류규칙을 제시한다. 이때 다변량 정규분포를 사용하게 되는데, 정규분포의 분산-공분산 행렬 동일성 여부에 따라 선형판별분석 및 이차판별분석으로 나뉘게 된다 (김기영과 전명식, 1994).

2.4. 신경망 분석

신경망 분석은 인간의 두뇌를 모델로 등장하였으며, 복잡한 구조를 가진 자료에서의 예측 문제를 해결하기 위해서 사용되는 비선형모델(nonlinear model)이다. 신경망 분석은 자료로부터 반복적인 학습과정을 거쳐 숨어있는 패턴을 찾아내는 모델링 기법으로 은닉마디(hidden units)라는 독특한 구성요소를 가진다. 신경망 분석은 입력층(input layer), 은닉층(hidden layer), 출력층(output layer)으로 구성되어 있다. 입력층과 출력층은 통계학적 용어로 설명변수와 종속변수로 해석할 수 있으며, 종속변수가 설명변수와 관계가 있고 그 두 변수 사이에 은닉층(hidden layer)이 있는 모형을 가정한다. 또한, 설명변수가 은닉층 안의 모든 노드에 영향을 주게 되며 종속변수는 은닉층 안의 모든 노드에 의해 영향을 받는 것을 가정한다. 은닉층의 수는 클 수 있지만, 대체적으로 하나 또는 두 개의 은닉층이 사용된다.

2.5. 서포트 벡터 머신

서포트 벡터 머신 (Vapnik, 1979)은 비선형 최적화 기법으로 Vapnik이 제안한 학습 이론에 기반을 둔 분류방법의 하나이다. 일반적인 통계적 학습 방법에서의 경험적 리스크 최소화(ERM; Empirical Risk Minimize)와는 다른 구조적 리스크 최소화(SRM; Structural Risk Minimize)를 통해 오류를 최소화시키는 방법을 이용한다. 일반적 통계적 방법을 통한 학습에 실제 정확률은 학습에 이용된 데이터에 의해 제약을 받게 된다. 따라서 수집된 데이터의 양과 분포에 따라 함수의 성능이 좌우된다고 볼 수 있다. 이와 같은 ERM방법을 이용한 전통적인 통계적 학습 방법들의 단점을 보완하기 위해 서포트 벡터 머신 알고리즘이 이용될 수 있다.

서포트 벡터 머신의 알고리즘은 다음과 같이 요약할 수 있다. 학습 자료로 주어진 n 차원의 벡터공간에서 분류 공간 간에 모든 점 사이의 거리를 최대화하도록 만들어 하나의 평면을 구해내는 것을 목표로 한다. 이 선형 평면 분류 경계를 최대마진분리경계면(OSH; Optimal Separating Hyperplane)라고 하며, 최대마진분리경계면에 가장 가까운 점들을 서포트 벡터(support vector)라고 부른다. 즉, 서포트 벡터 머신의 근본적인 원리는 최대 마진(maximum margin)을 갖는 결정경계(decision boundary)를 찾는 것이다. 마진(margin)은 가장 가까운 데이터 포인트로 이루어진 결정경계 사이의 거리를 의미한다.

지금까지 2장에서 살펴본 8가지 알고리즘의 장점과 단점은 표 2와 같다.

3. 양상불 기법 : 배경과 부스팅

데이터가 불안정하고 분류자의 변동성이 큰 경우에는 데이터가 조금이라도 바뀐 상태에서 예측자의 변동성을 감소시키고자 붓스트랩 방법을 통해 분류자를 얻을 수 있다. 이러한 방법을 배경 알고리

표 2: 단일 알고리즘 요약 정리

알고리즘	장점	단점
의사결정나무	해석의 용이성 상호작용 효과의 해석 비모수적 모형	비연속성 비안정성
로지스틱 회귀 분석	실제성과 현실성 해석상의 편리	선형관계만 분류 상호작용의 결여
선형판별 분석	구현의 용이성 설명력이 뛰어남	편의가 많이 발생
이차판별 분석	자료가 비선형성일 때 유용	편의가 많이 발생
신경망 분석	예측에 유용 비선형 형태를 가질 때 유용	해석의 어려움 중요한 설명변수 알 수 없음 비전문가 사용의 어려움
서포트 벡터 머신	구조적 위험을 최소화 우수한 일반화 성능	계산의 복잡성 시스템 구현의 어려움

즘(bagging algorithm)이라 하며, 배깅은 붓스트랩(bootstrap)방법 (Efron과 Tibshirani, 1993)을 이용한 앙상블기법으로 Breiman (1996)에 의해 제안되었다.

그림 1은 배깅 알고리즘에서 분석용 데이터로부터 붓스트랩 방법을 이용하여 서로 다른 붓스트랩 데이터를 만드는 과정이다 (이영섭 등, 2005). 서로 다른 붓스트랩 데이터들은 원래의 분석용 데이터에서 무작위로 관측값을 선택하여 만들어지기 때문에 중복되는 관측값도 있을 수 있고 한 번도 선택되지 않는 관측값도 있을 수 있다. 이렇게 무작위로 다르게 샘플링 된 붓스트랩 데이터를 가지고 각각 데이터 마이닝기법을 통해 단일 분류자를 생성시켜 그 집합으로 앙상블을 형성한다. 이렇게 만들어진 배깅 분류자는 그것을 구성하는 단일 분류자들의 다수결(Majority Voting)로 최종 분류 결정을 내린다. Breiman (1996)에 의하면 분석용 데이터가 불안정(unstable) 하다면, 배깅 분류자는 결함을 통해서 향상되어진다. 그러나 분석용 데이터가 안정적(stable)이라면, 배깅 과정을 통해서 얻어진 배깅 분류자는 분석용 데이터에서 얻어진 단일 분류자와 비슷함을 알 수 있다.

부스팅은 Valiant (1984)의 PAC(probably approximately correct) 학습모델에 기반을 둔 알고리즘으로, Kearns와 Valiant (1994)는 처음으로 약 분류(weak classification) 알고리즘에 부스팅 적용을 제안하였다. 이 알고리즘은 PAC 모델에 있어서 단순추측(random guessing)하는 것보다 약간 더 수행을 잘하는 약 분류자(weak classifier)는 강 분류(strong classifier) 알고리즘으로 부스트(boosted) 될 수 있음을 보여 준다.

Schapire (1990)는 약 분류자들은 부스팅을 통해 강 분류자로 변형할 수 있음을 보였으며, 그 후에 Freund (1995)는 Schapire 알고리즘의 결점에도 불구하고 더 효율적인 부스팅 알고리즘을 제안하였다. 부스팅 알고리즘을 처음으로 실제 데이터에 적용한 것은 OCR(optical character recognition)으로 Drucker 등 (1993)에 의해 수행되었다.

위에서 설명한 기본 부스팅을 변화시킨 여러 가지 버전이 존재한다. 그 중에서 가장 유명하고 본 논문에서 정확도 비교 대상으로 선택한 AdaBoost(adaptive boosting)에 대해 설명하도록 하겠다. Freund와 Schapire (1996)에 의해 처음 소개된 AdaBoost는 이전의 부스팅 알고리즘의 어려움을 많이 해결하였으며, Schapire와 Singer (1999)에 의해 일반화되었다.

그림 2를 통해 살펴보면 부스팅 방법의 초점은 분류자를 순차적으로 생성하고, 분석용 데이터는 이전의 분류자의 수행을 기반으로 하여 추출된 각 관측값을 사용한다. 처음 분석용 데이터 관측값의 가중치는 동일한 상태에서 시작하며, 분류자에 의해 오분류된 관측값은 높은 가중치를 주고, 정분류된 관측값은 반대로 낮은 가중치를 부여함으로써 관측값들의 가중치가 재조정된다. 이러한 과정을 통해 가

그림 1: 배깅 알고리즘

그림 2: 부스팅 알고리즘

중치가 재조정된 관측값들에 의해 새로운 분석용 데이터가 형성되는데, 가중치가 증가한 관측값이 많이 선택됨으로써 분류하기 힘든 관측값을 더 잘 분류하도록 한다 (Freund와 Schapire, 1996).

배깅과 부스팅을 실험적으로 비교한 논문에 의하면 배깅은 대부분의 데이터셋에서 단일 분류기보다 정확도가 더 높았으나 부스팅보다 정확도가 훨씬 낮은 경우도 있었다. 부스팅은 정확도를 크게 상승시키기도 하지만 일부 데이터 셋에서 단일 분류기보다도 낮은 경우도 있었다. 이는 부스팅이 데이터 셋에 좀 더 의존적임을 의미하는 것이다 (Bauer와 Kohavi, 1999; Opitz와 Maclin, 1999).

4. 데이터마이닝 알고리즘 및 양상불기법 성능 비교 결과

본 논문에서는 반응변수로 이항변수(binary variable)을 가지는 13개의 데이터를 분석에 사용하였다. 각 데이터에 대한 설명은 표 3에 요약되어 있다. 표에서 Data는 분석에 사용된 데이터의 이름,

표 3: 분석에 사용된 데이터셋

Data	Sample Size	Class Dist.	Description about data	# of predictors	
				Num	Cat
Chile	1,703	49.10%	Intension of a vote	6	1
Greene	384	29.70%	Judgement in Canada congress	1	5
Long	915	69.90%	The research results of biochemists	2	3
Mroz	753	56.80%	Participant in labor in American women	5	4
Powers	1,643	70.10%	Graduation from high school	3	4
WorkingMoms	2,293	55.40%	The attitude of working mothers	3	3
German	1,000	70.00%	German credit data	24	0
Austra	690	44.40%	Australian credit approval data	6	8
Heart	270	44.40%	A heart disease data	10	3
Ionosphere	351	64.10%	Radar returns from the ionosphere	34	0
Iris	100	50.00%	Iris data	4	0
Promoter	106	50.00%	Promoter gene sequences (DNA) with partial domain data	0	57
BreastCancer	699	65.50%	Original Wisconsin Breast Cancer Data	9	0

Sample size는 각 데이터 셋의 관측치의 수, Class Dist.은 전체 관측치에서 반응변수의 두 개의 값 중 1이 가지는 비율을 말한다. 그리고 # of predictors는 설명변수의 수를 말하는데, Num은 연속형 설명변수의 개수, Cat는 이산형 설명변수의 개수를 나타낸 것이다. 분석에 사용된 모든 데이터 셋은 UCI repository (Frank와 Asuncion, 2010)에서 다운로드 받을 수 있다.

알고리즘의 비교 기준은 2장에서 기술한 8가지의 단일 데이터 마이닝 알고리즘을 3장의 배경과 부스팅 알고리즘에 결합하여 총 24가지 방법(단일 알고리즘 8가지, 앙상블 방법 16가지)을 비교하였다. 각 알고리즘의 정확도에 대한 비교 기준으로 오분류율과 민감도, 특이도의 평균 및 표준오차를 살펴보고 알고리즘의 속도에 대한 비교를 위해 각각의 알고리즘의 시간을 측정하여 살펴보고 분류기준값으로는 0.5를 사용하였다. 또한, 각 데이터를 분석용 데이터(training data)와 테스트 데이터(test data)의 비율이 7:3이 되도록 랜덤하게 분할하여 알고리즘의 정확도 및 속도를 비교하였고, 정확한 값을 얻기 위하여 100번의 독립된 반복 실험을 행하였다.

그림 3, 4는 분석에 사용된 데이터의 결과이다. 그래프에서 x축은 알고리즘을 y축은 오분류율을 의미한다. 여기서 실선은 Single 모델의 오분류율, 긴 점선은 Bagging의 오분류율, 짧은 점선은 Boosting의 오분류율이다. 각 데이터에 대한 자세한 분석 결과는 홈페이지에 공개하였다(<http://www.korea.ac.kr/~stat2242/Ensemble>). 먼저 Chile 데이터의 결과를 보면, 대체적으로 평균 0.079의 낮은 오분류율 및 0.918의 높은 민감도와 0.922의 높은 특이도 값으로 모형이 적절하게 적합 되었음을 볼 수 있다. 선형판별분석(LDA)의 알고리즘에 부스팅을 적용한 경우 오분류율이 0.074(0.011)로 가장 낮은 값을 갖는 것으로 나타났다. 이와는 반대로 이차판별분석(QDA)의 알고리즘에 부스팅을 적용하였을 경우 0.085(0.011)로 가장 높은 오분류율 값을 갖고 있었다. 하지만 각 알고리즘의 전반적인 오분류율 값 차이는 대체적으로 미미한 수준을 나타냈으며, 민감도와 특이도 역시 전반적으로 비슷한 수치를 보였다. 전반적인 결과를 살펴보았을 때 단일 알고리즘을 사용한 경우와 앙상블 기법인 배경과 부스팅을 적용하였을 경우의 오분류율이 비슷한 값을 보였다.

Greene 데이터의 분석결과를 보면, 전반적으로 평균 0.278의 오분류율을 보이며, 민감도와 특이도의 평균값은 각각 0.761, 0.572로 대부분의 알고리즘에서 민감도의 값이 더 높았다. 의사결정나무 기법 중 CRUISE 알고리즘에 부스팅기법을 적용하여 50번 붓스트랩 반복을 행한 경우, 0.259(0.032)의 오분류율로 비교실험을 한 알고리즘들 중에서 가장 낮은 값을 가졌다. 이와는 반대로 같은 의사결정나무 기법 중 RPART의 단일 알고리즘의 경우 0.31(0.031)로 가장 높은 오분류율 값을 갖고 있었다. 전반적인 결과를 살펴보았을 때 단일 알고리즘을 사용한 경우보다 앙상블 기법인 배경과 부스팅을 적용하였

그림 3: 개별 데이터 셋의 분석 결과1: 각 데이터 셋별로 Single(실선), Bagging(긴 점선), Boosting(짧은 점선)의 오분류율을 나타내었다. 데이터셋의 경우, 좌측열은 위에서부터 순서대로 Chile, Heart, Long, German 데이터의 분석 결과이고 우측열은 위에서부터 Greene, Ionosphere, Mroz, Australian Credit 데이터의 분석결과이다.

그림 4: 개별 데이터 셋의 분석 결과2: 각 데이터 셋별로 Single(실선), Bagging(긴 점선), Boosting(짧은 점선)의 오분류율을 나타내었다. 데이터셋의 경우, 좌측열은 위에서부터 순서대로 Iris, Powers 데이터의 분석 결과이고 우측열은 위에서부터 Promoter, WorkingMoms 데이터, 하단 가운데는 Breast Cancer 데이터의 분석결과이다.

을 때 오분류율이 감소하는 것을 볼 수 있다.

Heart 데이터의 분석결과를 보면, 대체적으로 평균 0.202의 오분류율을 보이며, 민감도와 특이도의 평균값은 각각 0.801, 0.79로 전반적으로 비슷한 값을 가지는 것을 살펴볼 수 있다. 선형관별분석(LDA)의 단일 알고리즘의 경우 오분류율이 0.155(0.035)으로 가장 낮은 값을 갖는 것으로 나타났다. 이와는 반대로 의사결정나무의 RPART의 단일 알고리즘의 경우 0.271(0.039)로 가장 높은 오분류율을 갖고 있었다. 전반적인 결과를 살펴보았을 때 의사결정나무 알고리즘의 경우 단일 알고리즘을 사용한 경우보다 앙상블 기법인 배깅과 부스팅을 적용하였을 경우 대부분의 오분류율이 감소하는 것을 볼 수 있으며, 그 외의 알고리즘의 경우에는 오히려 대부분의 경우 오분류율이 증가하는 것을 볼 수 있

다.

Ionosphere 데이터의 분석결과를 보면, 대체적으로 평균 0.116의 오분류율을 보이며, 민감도와 특이도의 평균값은 각각 0.903, 0.844로 전반적으로 비슷한 값을 가지는 것을 살펴볼 수 있다. 의사결정나무의 CRUISE 알고리즘에 배깅을 적용한 경우 오분류율이 0.076(0.022)으로 가장 낮은 값을 갖는 것으로 나타났다. 이와는 반대로 의사결정나무 QUEST의 단일 알고리즘의 경우 0.144(0.031)로 가장 높은 오분류율 값을 갖고 있었다. 전반적인 결과를 살펴보았을 때 의사결정나무와 신경망분석(NNET) 알고리즘의 경우 단일 알고리즘을 사용한 경우보다 앙상블 기법인 배깅과 부스팅을 적용하였을 경우 대부분의 오분류율이 감소하는 것을 볼 수 있으며, 그 외의 알고리즘의 경우에는 배깅과 부스팅의 효과가 없는 것을 볼 수 있다.

Long 데이터의 분석결과를 보면, 대체적으로 평균 0.312의 오분류율을 보이며, 민감도와 특이도의 평균값은 각각 0.513, 0.716로 전반적으로 특이도의 값이 더 높을 것을 살펴볼 수 있다. 신경망분석(NNET)의 알고리즘에 배깅을 적용한 경우 오분류율이 0.287(0.024)로 가장 낮은 값을 갖는 것으로 나타났다. 이와는 반대로 선형판별분석(LDA)의 알고리즘에 배깅을 적용하여 50번 붓스트랩 반복을 한 경우 0.367(0.149)로 가장 높은 오분류율 값을 갖고 있었다. 하지만 전반적인 결과를 살펴보았을 때 신경망분석을 제외한 단일 알고리즘을 적용한 경우보다 앙상블 기법인 배깅과 부스팅을 적용하였을 때 오분류율이 증가하는 경향을 볼 수 있다.

Mroz 데이터의 분석결과를 보면, 대체적으로 평균 0.308의 오분류율을 보이며, 민감도와 특이도의 평균값은 각각 0.654, 0.729로 전반적으로 특이도의 값이 더 높을 것을 살펴볼 수 있다. 의사결정나무의 QUEST 알고리즘에 배깅을 적용하여 50번 붓스트랩 반복 한 경우 오분류율이 0.261(0.027)으로 가장 낮은 값을 갖는 것으로 나타났다. 이와는 반대로 신경망분석(NNET)의 단일 알고리즘의 경우 0.335(0.025)로 가장 높은 오분류율 값을 갖고 있었다. 전반적인 결과를 살펴보았을 때 단일 알고리즘을 적용한 경우보다 앙상블 기법인 배깅과 부스팅을 적용하였을 때 오분류율이 감소하는 것을 볼 수 있다.

German Credit 데이터의 분석결과를 보면, 대체적으로 평균 0.28의 오분류율을 보이며, 민감도와 특이도의 평균값은 각각 0.757, 0.568로 전반적으로 민감도의 값이 더 높을 것을 살펴볼 수 있다. 의사결정나무의 CRUISE 알고리즘에 부스팅을 적용하여 50번 붓스트랩 반복을 한 경우 오분류율이 0.259(0.032)으로 가장 낮은 값을 갖는 것으로 나타났다. 이와는 반대로 의사결정나무의 RPART의 단일 알고리즘의 경우 0.305(0.031)로 가장 높은 오분류율 값을 갖고 있었다. 전반적인 결과를 살펴보았을 때 단일 알고리즘을 적용한 경우보다 앙상블 기법인 배깅과 부스팅을 적용하였을 때 오분류율이 감소하는 것을 볼 수 있다.

Australian Credit 데이터의 분석결과를 보면, 대체적으로 평균 0.158의 오분류율을 보이며, 민감도와 특이도의 평균값은 각각 0.861, 0.829로 전반적으로 비슷한 값을 가지는 것을 살펴볼 수 있다. 의사결정나무의 CRUISE 알고리즘에 부스팅을 적용하여 50번 붓스트랩 반복을 한 경우 오분류율이 0.132(0.021)으로 가장 낮은 값을 갖는 것으로 나타났다. 이와는 반대로 신경망분석(NNET)의 단일 알고리즘의 경우 0.262(0.104)로 가장 높은 오분류율 값을 갖고 있었다. 전반적인 결과를 살펴보았을 때 단일 알고리즘을 적용한 경우보다 앙상블 기법인 배깅과 부스팅을 적용하였을 때 오분류율이 감소하는 것을 볼 수 있다.

Iris 데이터의 분석결과를 보면, 대체적으로 평균 0.06의 낮은 오분류율 및 0.935의 높은 민감도와 0.948의 높은 특이도 값으로 모형이 적절하게 적합 되었음을 살펴볼 수 있다. 선형판별분석(LDA)의 단일 알고리즘의 경우 오분류율이 0.036(0.029)으로 가장 낮은 값을 갖는 것으로 나타났다. 이와는 반대로 의사결정나무의 RPART의 단일 알고리즘의 경우 0.098(0.044)로 가장 높은 오분류율 값을 갖고 있었다. 전반적인 결과를 살펴보았을 때 의사결정나무의 RPART, QUEST, CRUISE의 경우 단일 알고

리즘을 사용한 경우보다 앙상블 기법인 배깅과 부스팅을 적용하였을 경우 대부분의 오분류율이 감소하는 것을 볼 수 있으며, 그 외의 알고리즘의 경우에는 오히려 대부분의 경우 오분류율이 증가하는 것을 볼 수 있다. 또한, 선형판별분석(LDA)와 이차판별분석(QDA)의 경우 다른 알고리즘에 비해 탁월한 효과가 있음을 알 수 있다.

Promoter 데이터의 분석결과를 보면, 대체적으로 평균 0.21의 오분류율을 보이며, 민감도와 특이도의 평균값은 각각 0.792, 0.802로 전반적으로 비슷한 값을 가지는 것을 살펴볼 수 있다. 의사결정나무의 RPART 알고리즘에 배깅을 적용하여 50번 붓스트랩 반복을 한 경우 오분류율이 0.127(0.075)으로 가장 낮은 값을 갖는 것으로 나타났다. 이와는 반대로 이차판별분석(QDA)의 단일 알고리즘의 경우 0.299(0.068)로 가장 높은 오분류율 값을 갖고 있었다. 전반적인 결과를 살펴보았을 때 단일 알고리즘을 사용한 경우보다 앙상블 기법인 배깅과 부스팅을 적용하였을 때 오분류율이 감소하는 것을 볼 수 있다.

Powers data 분석결과를 보면, 대체적으로 평균 0.279의 오분류율을 보이며, 민감도와 특이도의 평균값은 각각 0.564, 0.755로 전반적으로 특이도의 값이 더 높을 것을 살펴볼 수 있다. 로지스틱회귀분석(GLM)의 알고리즘에 부스팅을 적용하여 50번 붓스트랩 반복 한 경우 오분류율이 0.262(0.018)로 가장 낮은 값을 갖는 것으로 나타났다. 이와는 반대로 의사결정나무의 RPART 단일 알고리즘과 이차판별분석(QDA)의 단일 알고리즘의 경우 0.3으로 가장 높은 오분류율 값을 갖고 있었다. 전반적인 결과를 살펴보았을 때 단일 알고리즘을 적용한 경우보다 앙상블 기법인 배깅과 부스팅을 적용하였을 때 오분류율이 감소하는 것을 볼 수 있다.

WorkingMoms 데이터의 분석결과를 보면, 대체적으로 평균 0.371의 오분류율을 보이며, 민감도와 특이도의 평균값은 각각 0.573, 0.655로 전반적으로 특이도의 값이 더 높을 것을 살펴볼 수 있다. 선형판별분석(LDA)의 알고리즘에 부스팅을 적용하여 50번 붓스트랩 반복을 한 경우 오분류율이 0.35(0.015)로 가장 낮은 값을 갖는 것으로 나타났다. 이와는 반대로 신경망분석(NNET)의 Single 알고리즘 경우 0.43(0.096)으로 가장 높은 오분류율 값을 갖고 있었다. 전반적인 결과를 살펴보았을 때 앙상블 기법인 배깅과 부스팅을 적용하였을 때의 오분류율이 감소하는 경우와 증가하는 경우가 있는 것을 볼 수 있다.

BreastCancer 데이터의 분석결과를 보면, 대체적으로 평균 0.041의 낮은 오분류율 및 0.972의 높은 민감도와 0.935의 높은 특이도 값으로 모형이 적절하게 적합 되었음을 살펴볼 수 있다. 서포트 벡터 머신(SVM)의 알고리즘에 배깅을 적용하여 50번 붓스트랩 반복을 한 경우 오분류율이 0.03(0.011)으로 가장 낮은 값을 갖는 것으로 나타났다. 이와는 반대로 의사결정나무의 RPART의 단일 알고리즘의 경우 0.079(0.015)로 가장 높은 오분류율 값을 갖고 있었다. 하지만 각 알고리즘의 전반적인 오분류율 값 차이는 대체적으로 미미한 수준을 나타냈다. 민감도와 특이도 값 역시 전반적으로 비슷한 수치를 보였다.

데이터들에 대한 전반적인 분석결과를 보면, 이차판별분석(QDA)와 로지스틱회귀분석(GLM)의 경우를 제외하고는 앙상블 기법인 배깅과 부스팅을 적용하였을 경우 Single 경우보다 오분류율이 약간 감소하는 것을 볼 수 있다.

표 4는 각 데이터의 알고리즘별 최저 오분류율 및 최적 앙상블모형, 붓스트랩 반복수를 보여준다. 위에서 언급한 24가지 방법을 가지고 각 데이터에 맞는 최적의 알고리즘을 찾아본 결과, 단일 알고리즘을 최적의 방법으로 택한 데이터가 2개(Heart, Iris data), 배깅인 경우 5개(Long, Mroz, Ionosphere, Promoter, Breast Cancer data), 부스팅인 경우가 6개(Chile, Greene, Powers, Working Moms, German Credit, Australian data)를 보였다. 즉, 알고리즘의 비교 기준을 오분류율로 정한다면, 단일 알고리즘을 이용하여 분석을 하는 것보다는 배깅과 부스팅같은 앙상블 기법을 이용하여 분석하였을 때 보다 나은 결과를 가져온다고 할 수 있다. 계산 비용을 알고리즘의 비교기준으로 보면, 붓스트랩 반복을 25번 실행

표 4: 분석 결과 요약

Data	RPART	QUEST	CRUISE	LDA	QDA	GLM	NNET	SVM
Chile	0.078	0.076	0.077	0.074	0.084	0.076	0.079	0.079
	Bagging	Bagging	Boosting	Boosting	Bagging	Boosting	Boosting	Bagging
Greene	0.264	0.263	0.259	0.268	0.290	0.262	0.274	0.276
	Bagging	Boosting	Boosting	Boosting	Boosting	Boosting	Boosting	Bagging
Long	0.300	0.298	0.298	0.298	0.342	0.292	0.287	0.297
	Single	Bagging	Single	Single	Boosting	Boosting	Bagging	Single
Mroz	0.272	0.261	0.266	0.317	0.323	0.321	0.315	0.278
	Bagging	Bagging	Bagging	Boosting	Boosting	Boosting	Boosting	Boosting
Powers	0.270	0.271	0.277	0.264	0.285	0.262	0.267	0.275
	Boosting	Boosting	Boosting	Boosting	Boosting	Boosting	Boosting	Boosting
Working Moms	0.372	0.372	0.369	0.350	0.355	0.351	0.358	0.357
	Bagging	Boosting	Bagging	Boosting	Boosting	Boosting	Boosting	Bagging
German Credit	0.264	0.265	0.259	0.268	0.290	0.262	0.273	0.276
	Bagging	Boosting	Boosting	Boosting	Boosting	Boosting	Boosting	Bagging
Australian	0.137	0.135	0.132	0.136	0.175	0.133	0.187	0.141
	Bagging	Boosting	Boosting	Boosting	Boosting	Single	Boosting	Bagging
Heart	0.163	0.189	0.183	0.155	0.190	0.165	0.231	0.172
	Boosting	Boosting	Boosting	Single	Bagging	Bagging	Boosting	Bagging
Ionosphere	0.086	0.123	0.076	0.131	0.124	0.129	0.117	0.079
	Boosting	Boosting	Bagging	Bagging	Bagging	Bagging	Boosting	Single
Iris	0.078	0.062	0.063	0.036	0.038	0.057	0.042	0.058
	Boosting	Boosting	Bagging	Single	Bagging	Single	Bagging	Single
Promoter	0.127	0.148	0.140	0.281	0.248	0.219	0.132	0.149
	Bagging	Bagging	Boosting	Boosting	Boosting	Boosting	Boosting	Bagging
Breast Cancer	0.036	0.033	0.032	0.036	0.043	0.034	0.033	0.030
	Boosting	Boosting	Boosting	Boosting	Single	Single	Bagging	Bagging

표 5: 분석에 사용된 알고리즘 비교 정리

단일 알고리즘	단일 알고리즘보다 우수한 경우	
	배깅	부스팅
RPART	12(92.3%)	12(92.3%)
QUEST	9(69.2%)	11(84.6%)
CRUISE	12(92.3%)	10(76.9%)
LDA	6(46.2%)	9(69.2%)
QDA	7(53.8%)	8(61.5%)
GLM	4(30.8%)	7(53.8%)
NNET	12(92.3%)	13(100%)
SVM	8(61.5%)	4(30.8%)
Total	67.3%	71.1%

행한 경우와 50번 실행한 경우의 결과차이가 거의 없음을 볼 수 있다. 이는 앙상블 기법을 사용한 일반적인 분석에서는 25번의 붓스트랩 반복으로도 적절한 결과를 가져올 수 있음을 뜻한다.

표 5은 각각의 단일 알고리즘별 앙상블 기법의 효과를 비교한 표이다. 총 13개의 데이터를 가지고 붓스트랩 반복 25번인 경우를 기준으로 하였다. 첫 번째로 의사결정나무의 QUEST, 선형관별분석(LDA), 이차관별분석(QDA), 로지스틱회귀분석(GLM), 신경망분석(NNET)의 결과를 살펴보면, 부스팅이 배깅에 비해 상대적으로 향상된 결과를 보이며, 반대로 의사결정나무의 CRUISE, 서포트 벡터 머신(SVM)의 경우에는 배깅의 경우에서 부스팅에 비해 상대적으로 향상된 결과가 많았다.

두 번째로 의사결정나무의 RPART, QUEST, CRUISE와 신경망분석(NNET)의 경우를 보면 앙상블 기법의 적용이 탁월한 효과를 보이는 것을 알 수 있다.

세 번째로 배깅과 부스팅의 성능을 총괄적으로 비교하여 보았을 때 배깅의 경우 67.3% 향상된 경우가 있었으며, 부스팅의 경우 71.1%로 부스팅이 배깅보다 단일 알고리즘에 적용되었을 때 향상된 결과를 보이는 것을 알 수 있다. 따라서 본 논문의 13개 데이터를 분석결과를 종합해 보면 부스팅의 성능이 좀 더 우수하다고 말할 수 있다.

5. 결론 및 향후과제

본 논문에서는 13가지 이항반응변수 데이터에 대해 오분류율 및 민감도 그리고 특이도를 각 알고리즘 별로 계산하여 그 성능을 분석하였다. 대부분의 모든 데이터에서 배깅의 오분류율은 8가지의 단일 데이터 마이닝의 오분류율보다 약간 감소하였고, 일부 데이터에서는 부스팅의 오분류율은 단일 데이터 마이닝기법보다 크게 감소하였지만 반대로 오분류율이 증가한 데이터도 있다. 예를 들어 Heart data 분석결과와 이차판별분석(QDA) 알고리즘을 살펴보면 단일기법일 경우 오분류율이 0.192인데 반해 부스팅을 적용한 경우 0.209로 오히려 오분류율이 높아진 경우를 볼 수 있다.

결론적으로 본 논문의 13개 데이터를 분석결과를 종합해 보면 단일 데이터 마이닝 기법보다 앙상블 기법인 배깅과 부스팅을 적용하였을 경우 성능이 향상된다는 것을 알 수 있다. 또한, 2가지 앙상블 기법 중 부스팅의 경우가 배깅보다 좀 더 성능이 좋은 것을 알 수 있다. 데이터 셋의 관측값 수, 목표변수, 입력변수의 크기가 클수록 부스팅 알고리즘이 적합함을 알 수 있고 이와 반대로 관측값 수, 목표변수, 입력변수의 크기가 작을수록 배깅의 알고리즘이 적합함을 알 수 있다. 이것은 대용량의 데이터와 데이터의 속성이 복잡해질수록 부스팅의 알고리즘이 적합하고, 데이터의 양이 적고 데이터 속성이 단순할수록 배깅 알고리즘이 적합하다고 할 수 있다.

본 논문은 기존에 앙상블기법의 분류자로 사용되면 의사결정나무의 CART 방법 외에 7가지 다양한 분류자를 사용하여 앙상블기법을 적용하여 본 결과 성능향상이 있다는 점에서 의미가 있다고 할 수 있다. 그러나 배깅과 부스팅에 성능향상을 13개의 데이터를 가지고 일반화를 시킨다는 것은 어려움이 따른다. 따라서 더 많은 데이터를 가지고 앙상블 알고리즘을 적용하여 배깅과 부스팅의 일반적인 특성을 파악해야 할 것이다. 또한, 배깅과 부스팅 외에도 랜덤 포레스트(Random Forest) 알고리즘을 적용하여 성능을 살펴보는 것을 향후 연구과제로 생각할 수 있다.

참고 문헌

- 김규곤 (2003). 데이터 마이닝에서 분류방법에 관한 연구, *Journal of the Korean Data Analysis Society*, **5**, 101-112.
- 김기영, 전명식 (1994). <다변량 통계자료분석>, 자유아카데미, 서울.
- 이영섭, 오현정, 김미경 (2005). 데이터 마이닝에서 배깅, 부스팅, SVM 분류 알고리즘 비교 분석, <응용통계연구>, **18**, 343-354.
- 허명희, 서혜선 (2001). <SAS 회귀분석>, 자유아카데미, 서울.
- Bauer, E. and Kohavi, R. (1999). An empirical comparison of voting classification algorithms: Bagging, Boosting and variants, *Machine Learning*, **36**, 105-139.
- Breiman, L. (1996). Bagging predictors, *Machine Learning*, **26**, 123-140.
- Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification and Regression Trees*, Chapman & Hall, New York.
- Clemen, R. (1989). Combining forecasts: A review and annotated bibliography, *Journal of Forecasting*, **5**, 559-583.

- Drucker, H. and Cortes, C. (1996). Boosting decision trees, *Neural Information Processing Systems*, **8**, 470–485.
- Druker, H., Schapire, R. and Simard, P. (1993). Boosting performance in neural networks, *International Journal of Pattern Recognition and Artificial Intelligence*, **7**, 705–719.
- Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*, Chapman & Hall, New York.
- Frank, A. and Asuncion, A. (2010). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
- Freund, Y. (1995). Boosting a weak learning algorithm by majority, *Information and Computation*, **121**, 256–285.
- Freund, Y. and Schapire, R. (1996). Experiments with a new boosting algorithm, *Proceedings of the Thirteenth International Conference on Machine Learning*, 148–156.
- Kass, G. V. (1980). An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **29**, 119–127.
- Kearns, M. and Valiant, L. G. (1994). Cryptographic limitations on learning Boolean formulae and finite automata, *Journal of the Association for Computing Machinery*, **41**, 67–95.
- Kim, H. J. and Loh, W. Y. (2001). Classification trees with unbiased multiway splits, *Journal of the American Statistical Association*, **96**, 598–604.
- Loh, W. Y. and Shih, Y. S. (1997). Split selection method for classification trees, *Statistica Sinica*, **7**, 815–840.
- Opitz, D. and Maclin, R. (1999). Popular ensemble methods: An empirical study, *Journal of the Artificial Intelligence Research*, **11**, 169–198.
- Perrone, M. (1993). Improving regression estimation: Averaging methods for variance reduction with extensions to general convex measure optimization, *Doctoral dissertation*, Department of Physics, Brown University.
- Quinlan, J. R. (1992). *C4.5 : Programming with Machine Learning*, Morgan Kaufmann Publishers.
- Quinlan, J. R. (1996). Bagging, boosting, and C4.5, *Proceedings of the Fourteenth National Conference on Machine Learning*, 725–730.
- Schapire, R. E. (1990). The strength of weak learnability, *Machine Learning*, **5**, 197–227.
- Schapire, R. E. and Singer, Y. (1999). Improved boosting algorithms using confidence-rated predictions, *Machine Learning*, **37**, 297–336.
- Valiant, L. G. (1984). A theory of the learnable, *Communication of the ACM*, **27**, 1134–1142.
- Vapnik, V. (1979). *Estimation of Dependences Based on Empirical Data*, Nauka, Moscow.
- Wolpert, D. (1992). Stacked generalization, *Neural Network*, **5**, 241–259.

A Study for Improving the Performance of Data Mining Using Ensemble Techniques

Yon Hae Jung^a, Soo-Heang Eo^a, Ho Seok Moon^b, HyungJun Cho^{1,a}

^aDepartment of Statistics, Korea University

^bDepartment of Computer and Information, Korea Military Academy

Abstract

We studied the performance of 8 data mining algorithms including decision trees, logistic regression, LDA, QDA, Neral network, and SVM and their combinations of 2 ensemble techniques, bagging and boosting. In this study, we utilized 13 data sets with binary responses. Sensitivity, Specificity and missclassificate error were used as criteria for comparison.

Keywords: Ensemble, bagging, boosting, data mining.

¹ Corresponding author: Associate Professor, Department of Statistics, Korea University, Anam-Dong, Seongbuk-Gu, Seoul 136-701, Korea. E-mail : hj4cho@korea.ac.kr