데이터 마이닝에서 배깅과 부스팅 알고리즘 비교 분석

이영섭[1] 오현정[2]

요약

데이터 마이닝의 여러 기법중 모형의 변동성을 줄이고 정확도가 높은 분류자를 형성하기 위하여 다양한 앙상블 기법이 연구되고 있다. 그 중에서 배깅과 부스팅 방법이 가장 널리 알려져 있다. 여러 가지 데이터에 이 두 방법을 적용하여 오분류율을 구하여 비교한 후 각 데이터 특성을 입력변수로 하고 배깅과 부스팅 중 더 낮은 오분류율을 갖는 알고리즘을 목표변수로 하여 의사결정나무를 형성하였다. 이를 통해서 배깅과 부스팅 알고리즘이 어떠한 데이터 특성의 패턴이 존재하는지 분석한 결과 부스팅 알고리즘은 관측치, 입력변수, 목표변수 수가 큰 것이 적합하고 반면에 배깅 알고리즘은 관측치, 입력변수, 목표변수 수의크기가 작은 것이 적합함을 알 수 있었다.

주요용어: 의사결정나무, 배깅, 부스팅, 데이터 마이닝

1.서론

지난 수십 년 간 데이터의 양은 기하급수적으로 증가하고, 우리가 원하는 정보를 찾아내는 일을 보다 어렵게 만들고 있다. 우리는 이러한 대용량의 데이터로부터 의미 있는 지식을 찾아내는 것을 KDD(knowledge discovery in database)과정이라 하며, KDD 과정 중 데이터 탐사 및 분석단계를 데이터 마이닝이라 한다. 데이터 마이닝 기법은 다양하고, 이런 다양한 기법을 통해서 데이터에 적합한 모델을 설정하여 우리가 원하는 정보를 얻어내는 것이 중요하다.

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. 다중 분류자를 결합하는 앙상블 방법 중 대표적으로 배깅(bagging)과 부스팅(boosting) 알고리즘이 있다. 이 알고리즘들은 분석용(training) 데이터에서 재표본(resampling)기법으로 얻어진 데이터에 의해 각 분류자가 형성된다. 이때 배깅은 관측값이 동일한 가중치를 갖는 분포에서 추출된 데이터에서 분류자가 형성되지만 부스팅은 관측값의 가중치가 변하는 분포에서 추출된 데이터에서 분류자가 형성된다.

데이터 마이닝 기법 중 하나인 의사결정나무(decision trees)는 데이터가 조금이라도 변하게

1) (100-715)서울특별시 중구 필동 3가 , 동국대학교 통계학과 , 조교수
E-mail : yung@dongguk.edu

2) (130-810)서울특별시 동대문구 용두2동 , DNI consulting , 컨설턴트
E-mail : ulgi@dni.co.kr

되면 모델이 쉽게 변하는 단점을 가지고 있다. 이때 배깅과 부스팅 알고리즘을 의사결정나무에 적용시키면 의사결정나무의 단점을 보안하여 정확한 예측을 할 수 있는 분류자를 형성하는데 매우 효과적임을 알 수 있다. (Drucker 와 Cortes, 1996)

본 논문에서는 다양한 데이터를 의사결정나무에 배깅과 부스팅 알고리즘을 적용시켜 포괄적인 평가를 한다. 특히 배깅과 부스팅 알고리즘에 적합한 데이터 특성이 무엇인지를 알아보기 위해 배깅과 부스팅을 목표변수로 하고 데이터 특성들을 입력변수로 하여 의사결정나무를 형성해보기로 한다. 그리고 이 모델을 통해서 배깅과 부스팅 알고리즘에는 어떠한 데이터 특성 패턴을 가지는지 알아보기로 한다.

2. 의사결정나무

의사결정나무는 의사결정규칙(decision rule)을 나무구조로 도표화하여 분류(classification)와 예측(prediction)을 수행하는 분석방법이다. 이 방법은 분류 또는 예측의 과정이 나무구조에 의한 추론규칙에 의해서 표현되기 때문에 다른 방법들에 비해서 연구자가 그 과정을 쉽게 이해하고 설명할 수 있다.

범주형 목표변수에 적용하는 분류 의사결정 나무의 알고리즘에는 분리 기준에 따라 Kass(1980)가 제안한 CHAID(chi-squared automatic interaction detection), Breiman 등 (1984)이 제안한 CART(classification and regression trees), Quinlan(1993) 이 제안한 C4.5가 있다. 본 논문에서는 배깅과 부스팅 알고리즘과 비교하기 위한 기존의 의사결정나무 알고리즘으로 가장 널리 쓰이는 방법인 CART 알고리즘을 사용하였다.

3. 배깅 분류자

데이터가 불안정한 경우에 , 즉 데이터가 조금이라도 바뀐 상태에서 분류자의 변동성이 큰 경우에는 예측자의 변동성을 감소시키고자 붓스트랩 방법을 통해 분류자를 얻을 수 있다. 이러한 방법을 배깅 알고리즘이라 하며, 배깅은 붓스트랩(bootstrap) 방법(Efron 과 Tibshirani , 1993)을 이용한 앙상블 기법으로 Breiman(1996) 에 의해 소개되었다.

배깅 알고리즘 과정은 모집단으로부터 추출된 분석용 데이터에서 복원 단순 임의추출에 의해 붓스트랩 분석용 데이터를 생성한다. 생성된 분석용 데이터 집합에서 각각의 단일 분류자를 형성하여 단일 분류자 집합을 얻는다. 이러한 여러 단일 분류자 집합을 결합하는데 목표변수가 연속형일 때 평균, 범주형일 때는 다중 투표(majority vote)를 사용하여 얻어진 분류자를 배깅 분류자라 한다.(Breiman,1996)

4. 부스팅 분류자

기계학습(machine learning) 연구에서 이론적인 시초는 Valiant (1984)의 PAC(probably approximately correct) 학습모델(learning model) 이며, Kearns 과 Valiant (1994) 는 처음으로 약 분류(weak classification) 알고리즘을 제안하였다. Freund 과 Schapire(1996)에 의해 처음 소개된 AdaBoost(adaptive boosting)는 이전의 부스팅 알고리즘의 어려움을 많이 해결하였으며 Schapire 와 Singer(1999)에 의해 일반화되었다.

부스팅 방법의 초점은 분류자를 순차적(sequentially)으로 생성하고, 각각의 분석용 데이터는

이전의 분류자의 수행을 기반으로 하여 추출된 각 관측값을 사용한다. 처음 분석용 데이터 관측값의 가중치는 동일한 상태에서 시작하며 이때 형성된 분류자에 의해 오분류된 관측값은 다음번 관측값에 높은 가중치를 주고, 정분류된 관측값은 반대로 낮은 가중치를 부여한다. 관측값들의 가중치가 재조정된 새로운 분석용 데이터를 생성하여 오분류율이 일정수준에 도달할때까지 위의 과정을 반복함으로써 최종적으로 형성된 분류자를 부스팅 분류자라고 한다. .(Freund와 Schapire, 1996)

5. 실제 데이터를 이용한 배깅과 부스팅 알고리즘 비교

본 논문에서 사용되는 데이터는 목표변수가 범주형인 경우만 다루고 사용된 27개 데이터는 기계학습에서 사용되어지는 공유 데이터들의 저장소(Wang,2002)에서 추출한 실제 데이터이다. (http://www.cs.sfu.ca/~wangk/ucidata/dataset)

표5.1은 27개 데이터에 대한 CART, 배깅과 부스팅의 오분류율을 측정한 값이다. 각 알고리즘의 오분류율을 비교하면 배깅 알고리즘의 오분류율은 dna 데이터를 제외하고 CART보다 전체적으로 약간이라도 낮음을 알 수 있다. 배깅과 달리 부스팅의 오분류율은 auto, sonar, vehicle, promoter, car, segmentation, krvskp, letter 과 connect4 의 9개 데이터에 대해서는 CART 보다 작지만, labor, wine, inosphere과 dna 는 CART보다 부스팅의 오분류율이 더 크다. 이는 전체적으로 오분류율을 조금씩이라도 감소시키는 배깅과 달리 부스팅의 경우 일부 데이터에서는 CART보다 아주 작은 오분류율을 가지고 일부 데이터에서는 CART 보다 큰 오분류율을 가짐을 알 수 있다.

표 5.1: CART, 배깅과 부스팅 오분류율 비교

데이터	CART	배깅	부스팅	데이터	CART	배깅	부스팅
connect4	0.307	0.296	0.225	crx	0.166	0.138	0.147
krkopt	0.609	0.597	0.541	balance	0.226	0.162	0.200
letter	0.147	0.074	0.039	ionosphere	0.102	0.081	0.120
led7	0.273	0.267	0.270	cleve	0.226	0.207	0.215
krvskp	0.063	0.057	0.024	heart	0.245	0.200	0.206
splice	0.070	0.06	0.050	glass	0.257	0.173	0.200
segmentation	0.030	0.03	0.010	sonar	0.295	0.254	0.223
car	0.080	0.065	0.028	auto	0.244	0.209	0.187
dna	0.280	0.300	0.330	wine	0.062	0.047	0.071
german	0.310	0.251	0.254	hepatities	0.233	0.189	0.217
promoter	0.141	0.101	0.065	iris	0.069	0.064	0.062
vehicle	0.305	0.258	0.223	zoo	0.127	0.110	0.094
breasts	0.072	0.043	0.051	labor	0.165	0.105	0.183
austra	0.173	0.141	0.144				

표 5.1에서 제시한 27개 데이터의 배깅과 부스팅의 오분류를 비교하여 배깅 오분류가 낮은 14개의 데이터에 대해서는 배깅으로 지정하고 부스팅 오분류가 낮은 13개의 데이터에 대해서도 부스팅으로 지정하였다. 이렇게 정해진 목표변수(배깅 또는 부스팅)와 관측값 개수, 목표변수의 계급 수, 입력변수 개수, 입력변수의 연속형 비율, 입력변수의 범주형 비율과 결측치 존재 유뮤를 입력변수한 값들은 표 5.2에 제시되어 있다.

표 5.2 : 의사결정나무 형성의 위한 입력변수 및 목표변수 값

데이터	입력변수						목표변수*	데이터	입력변수						목표변수
	관측값	목표변수수	입력변수수	연속형비율	범주형비율	결측치			관측값	목표변수수	입력변수수	연속형비율	범주형비율	결측치	
austra	690	2	14	0.43	0.57	x	0	hepatities	155	2	19	0.32	0.68	o	0
auto	205	5	25	0.6	0.4	o	1	ionosphere	351	2	34	1	0	x	0
balance	625	1	4	0	1	x	0	iris	150	3	4	1	0	x	1
breasts	699	6	9	0	1	x	0	krkopt	28056	17	6	0	1	x	1
car	1354	4	6	0	1	x	1	krvskp	3196	2	36	0	1	x	1
cleve	303	5	13	0.38	0.62	o	0	labor	57	2	16	0.5	0.5	o	0
connect-4	67557	3	42	0	1	x	1	led7	3200	9	7	0	1	x	0
crx	690	2	15	0.4	0.6	o	0	letter	20000	26	16	1	0	x	1
dna	1000	3	180	0	1	o	0	promoter	936	2	57	0	1	x	1
german	1000	2	20	0.35	0.65	x	0	segmentation	2310	7	19	1	0	x	1
glass	214	7	9	1	0	x	0	sonar	208	2	60	1	0	x	1
heart	270	2	13	0.46	0.54	x	0	splice	3190	3	60	0	1	x	1
vehicle	846	4	18	1	0	x	1	zoo	101	7	16	0	1	x	1
wine	178	3	13	1	0	x	0								

*0:배깅 1:부스팅

입력변수와 목표변수를 이용하여 그림 5.1과 같이 의사결정나무를 형성할 수 있다. 이렇게 형성된 의사결정나무를 통해 배깅과 부스팅 알고리즘으로 분류하는데 표5.2에서 제시한 입력변수 중 관측값 개수 , 입력변수 개수와 목표변수의 계급 수가 중요한 변수로 작용했다.

그림 5.1의 의사결정나무을 보면 1번 노드에 27개 데이터를 잘 반영하는 관측값 개수가 1177개에 의해 2번, 3번 노드로 분리되었다. 여기서 3번 노드는 관측값 개수가 1177개 보다 큰 데이터다. 특히 4번 노드는 관측값 개수가 1177개보다 작은 데이터에 중 입력변수 개수가 15.5개보다 작은 데이터이다. 또한 관측값 개수가 1177개보다 작은 데이터 중에서 입력변수 개수가 15.5개보다 큰 데이터 들 중에서 목표변수의 계급 수가 3.5개보다 작은 데이터들은 6번 노드이며, 목표변수의 계급 수가 3.5개보다 큰 노드는 7번 노드이다. 배깅의 알고리즘의 오분류가 낮은 데이터들은 대체적으로 4번, 6번 노드로, 부스팅의 경우에는 3번, 7번 노드로 분류됨을 알 수 있다. 여기에서 4번, 6번 노드는 관측치 개수, 입력변수 개수, 목표변수의 계급 수가 작을수록, 3번 7번 노드는 관측치수, 입력변수 수, 목표변수의 계급 수가 큰 데이터로 이루어져 있음을 알 수 있다.

6. 결론 및 제언

본 논문에서는 의사결정나무에 배깅과 부스팅 알고리점을 적용하여 오분류율를 측정하여 이를 비교하였다. 대부분의 모든 데이터에서 배깅의 오분류율은 CART의 오분류율보다 약간 작았으며, 일부 데이터에서는 부스팅의 오분류율은 CART보다 아주 작았지만 반대로 CART보다 오분류율이 큰 데이터도 있었다.

배깅과 부스팅 알고리즘을 특징을 알아보기 위하여 데이터 특성에 의해 형성된 의사결정나무는 관측값 개수, 목표변수의 계급 수, 입력변수 개수에 따라서 배깅과 부스팅 알고리즘으로 분류됨을 알 수 있다. 의사결정나무에서 관측값 수, 목표변수, 입력변수의 크기가 클수록 부스팅 알고리즘이 적합하고 이와 반대로 관측값 수, 목표변수, 입력변수의 크기가 작을수록 배깅의 알고리즘이 적합함을 알 수 있다. 이것은 대용량의 데이터와 데이터의 속성이 복잡해질수록 부스

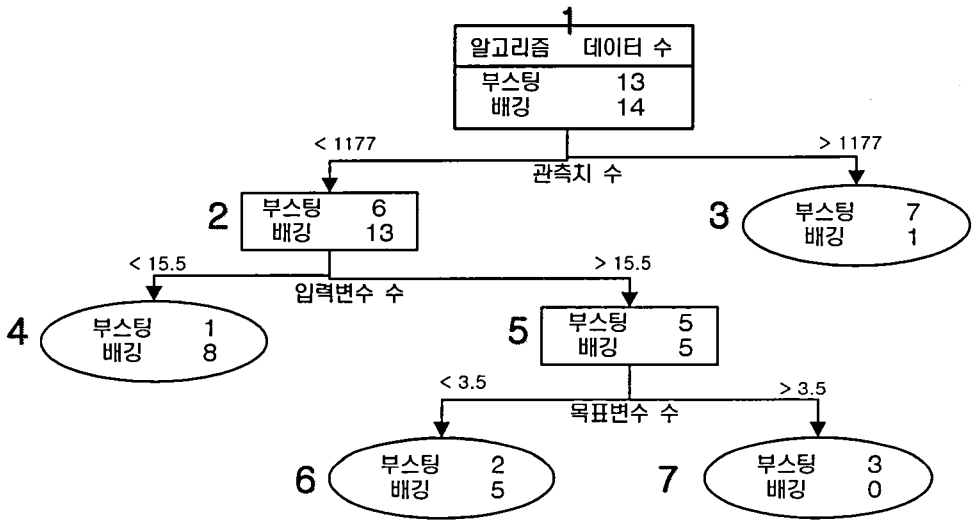

그림 5.1 배깅과 부스팅 비교연구를 위한 의사결정나무

팅의 알고리즘이 적합하고, 데이터의 양이 적고 데이터 속성이 단순할수록 배깅 알고리즘이 적합하다고 할 수 있다.

그러나 배깅과 부스팅에 적합한 데이터 특성을 의사결정나무를 형성하여 27개의 데이터를 가지고 일반화를 시킨다는 것은 어려움이 따른다. 따라서 더 많은 데이터을 사용하여 일반화를 시켜볼 수 도 있다. 또한 알고리즘에 적합한 데이터의 특성을 찾기 위해서 의사결정나무 이외에도 로지스틱 회귀분석이나, 배깅과 부스팅의 오분류율을 목표변수로 한 회귀분석 등 다양한 기법을 이용한 비교가 시행되어야 할 것이다.

참고문헌

Breiman, L.(1996) , "Bagging Predictor" , *Machine Learning.* 26, 123-140

Breiman, L. , Friedman, J. H. , Olshen, R. A. and Stone, C. J. (1984) *Classification and Regression Trees*, Chapman and Hall.

Drucker H. and Cortes C. (1996) , "Boosting Decision Trees" , *Neural Information Processing* 8, 470-485.

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

Freund, Y. and Schapire, R. (1996), "Experiments with a new Boosting algorithm" , *In Proceedings of the Thirteenth International Conference on Machine Learning*, 148-156

Kass , G.V, (1980) "An exploratory technique for investing large quantities of categorical data." , *Applied Statistics*, 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(1) ,67-95.

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

Schapire, R. and Singer. Y. (1999), "Improved Boosting algorithms using confidence-rated predictions", *Machine Learning*, 37(3) , 297-336

Valiant. L. G. (1984), "A theory of the learnable" , *Communication of the ACM*. 27(11) , 1134-1142,

Wang K.(2002) UCI repository of machine learning data base (http://www.cs.sfu.ca/~wangk/ucidata/dataset)