Journal of the Korean
Data & Information Science Society
2001, Vol. 12, No. 2, pp. 113 ~ 124

데이터마이닝기법상에서 적합된 예측모형의 평가
-4개분류예측모형의 오분류율 및 훈련시간 비교평가 중심으로

이상복 [1]

요 약

의사결정나무모형 가운데 하나인 CHAID, 로지스틱 회귀모형, 이들을 이용한 각각의 베깅모형 등 4가지 예측분류모형에 대한 오분류율과 훈련시간을 표본크기별로 계산하고, 이들 모형에 대한 모의실험 비교를 통하여 주어진 알고리즘들의 효율성을 평가하였다. 베깅 의사결정나무모형은 오분류율은 낮았으나 상대적으로 훈련시간이 가장 길었다.

주제어: bagging, data mining, CHAID, logistic regression, error rate, model efficiency, training time.

1. 서론

대용량 정보자료집합을 가진 마케팅과 경영분석의 필요성에서 시작된 데이터마이닝 기법들은 현재 경제, 정보기술, 생명공학분야 등 폭 넓은 분야에서 그 유용성과 다양한 활용을 위해 많은 연구.개발이 이루어지고 있다. 에이시앰(ACM)의 데이터베이스지식발견 국제연구회(SIGKDD) 2000년 보고서에 따르면, 2000년 데이터마이닝 경연대회에서 대표적으로 활용된 마이닝기법으로는 의사나무결정, 선형 혹은 로지스틱 회귀모형, 의사결정구칙, 베깅/부스팅(bagging/boosting), 연관성규칙, 베이지안분석, 순차분석, 스포트 벡터 기계학습, 인공신경망, 군집분석의 순이었다. 그리고, 데이터마이닝 기법과 알고리즘의 연구에 있어 최근 관심분야를 다음 세가지 가지 측면으로 설명할 수 있다.

데이터마이닝 기법 종류는 분류(classification), 추정(estimation), 예측(prediction), 유사통합(affinity grouping), 군집화(clustering), 기술(description) 문제 등으로 대별할 수 있다. 대용량 자료에 있어 가장 흔히 나타나는 범주형 변수의 분류와 예측의 문제는 성별, 연령대별 구분이나, 신문기사 데이터베이스의 핵심어를 사용한 텍스트 마이닝, 금융기관의 고객 신용평가, 웹 사이트, 웹 로그 분류 등에 많이 이용되고 있다. 분류를 위한 분석용 자료분할은 학습용 자료와 확인용 자료로 구별하여 모형을 구축하며, 일반적으로 로지스틱 회귀분석, 의사나무결정, 인공신경망, 판별분석기법이 일반적으로 사용된다. 또한, 목표 변수가 포함된 자료를 분류 및 예측하는 분석기법을 지도예측(supervised prediction)이라고 하며, 목표값이 범주형인 경우에 지도분류(supervised classfication)라고도 한다. 판별분석, 선형 혹은 로지스틱 회귀분석, 인공신경망, 시계열분석 등은 지도예측모형으로 분류된다. 이에 반해, 군집분석, 연관성규칙발견은 자율예측기법이다.

그러나, 이러한 여러 가지 마이닝 기법들의 단점 가운데 하나는 분류 및 예측의 불안정성에 있었으며, 초기의 데이터마이닝 연구에서는 안정된 정확한 분류 및 예측문제가 주된 관심

[1](712-702) 경북 경산시 하양읍 금락 1리 330 대구가톨릭대학교 응용과학부 교수

분야였다. 분류예측의 안정성 해결을 위한 다중분류문제의 예로서, 베깅알고리즘(Breiman, 1996)은 분석용 자료의 붓스트랩 표본들을 생성하고, 각 표본에 대해 분류예측모형를 만들어 목표변수가 구간척도일 경우에 이들의 평균값을 이용하거나, 혹은 범주형 예측변수를 가질 경우 복식투표(plural voting)를 하여 예측분류하는 방법이다. 또 다른 다중분류방법으로 분류 예측치를 각 분류나무예측치의 가중투표(weighted voting)에 의해 결정하는 부스팅기법 등이 있다.

또한, 데이터마이닝에 있어 마이닝 분석결과의 해석은 상당한 어려움이 뒤따른다. 데이터마이닝 대상자체가 메가바이트급 이상의 대용량의 데이터를 다룰 뿐만 아니라, 변수간의 관계가 잘 알려져 있지 않은 불확실한 성격의 변수들을 포함한 많은 변수들을 사용하기 때문에, 분석결과의 해석이 쉽지 않은 경우가 많다. 마이닝 분석결과의 적용사례에 맞는 해석적 어려움과 아울러, 주어진 사례에 안정적이며 효과적인 최적예측모형의 선정이나, 나아가 모형의 일반화문제는 데이터마이닝의 현안 과제로 여전히 남아 있다. 특별한 경우, 금융시장에 있어 고객 신용도에 따른 고객분류 및 예측문제는 법적요구 문제로 발전되기도 한다(강현철외 2인, 2000).

그리고, 데이터마이닝에 필요한 컴퓨팅시간은 적지 않은 계산비용 문제를 유발시키고 있다. 현재까지 주어진 데이터베이스에 대한 균일적으로 가장 효율적인 분류. 예측알고리즘이 없음은 이미 잘 알려져 있으며, 그러한 알고리즘에 있어서 과도한 훈련시간은 당연히 바람직하지 않다(Hand, 1997). 최근에는 이러한 처리시간과 관련된 연구들도 많이 진행되고 있다(Lim 외 1인,2000).

본 논문에서는 의사결정나무모형 가운데 하나인 CHAID, 로지스틱 회귀모형, 이들을 이용한 각각의 베깅모형 등 4가지 예측분류모형에 대한 오분류율과 훈련처리시간을 표본크기별로 계산하고, 이들의 비교를 통한 주어진 알고리즘들의 효율성을 평가하였다. 2절에서는 알고리즘 개요를 소개하였고, 3절에서는 모의실험설계와 자료에 대한 설명을 하였으며, 금융시장에 있어 고객 신용분석을 위한 SAS 시스템상의 가상자료에 대하여 SAS 이 마이너(e-miner) 통계패키지 4.0를 사용하여 의사나무결정, 로지스틱 회귀모형, 베깅모형에 대한 오분류율과 훈련시간을 4절에서는 계산하였다. 5절 결론부분에서는 이들 분류예측모형에 대한 평가 및 추후연구를 위한 가지 제언을 하였다.

2. 평가를 위해 사용된 데이터마이닝 분류예측모형 및 알고리즘

이 논문에서는 분류예측모형평가를 위한 알고리즘은 Lim(2000)에서 비교하지 않은 다음 4가지를 사용하였다

CHAID :

명목척도 혹은 서열척도가 사용된 범주형 예측변수의 분류를 위하여 Kass(1980) 에 의해 제안된 의사결정나무모형으로, 분리기준으로는 범주형(혹은 구간형) 목표변수에 대한 카이제곱(F분포) 검정통계량의 최적분리가 수정된 최소 p 값에 따라 결정된다. 정지규칙으로는 주어진 입력변수에 대하여 분리를 위한 더 이상의 범주들의 결합이 없거나, 임계값(threshold) p 에 의해 결정된다. 가지치기 규칙으로 주어진 입력변수들의 범주들에 대한 p 값들을 비교한 결과, 최소 p 값을 가진 입력변수가 분할 변수로 선택되며, 이 p 값이 임계값보다 작을 경우 자식마디가 형성된다.

LogisticRegression :
로지스틱 판별분석을 의미하며 Agresti(1990)가 제안한 로지스틱 다항회귀모형을 사용하였으며, 변수선택으로 단계별 방법을 사용했다. 추정된 회귀계수 a, b_1, $b_2,\cdots$, b_p를 이용한 사후확률계산은 다음 식과 같다.

$\hat{p}(y=1|x_1,x_2,\cdots,x_p) = \exp(a+b_1x_1+\cdots+b_px_p)/[1+\exp(a+b_1x_1+\cdots+b_px_p)]$ 이렇게 계산한 사후확률은 0과 1사이의 값을 가지므로 훈련용 자료집합에서 얻어진 적절한 절단값에 의해 각 개체는 분류된다.

Bagging :
분류예측변수의 예측안정성과 정확성 개선을 위한 붓스트랩 합산평균법 혹은 복식투표에 의한 일반화된 다중분류예측기법으로 Breiman(1996)이 제안하였다. 이 논문에서는 5회 반복 재표본에 대하여 각각의 CHAID 혹은 로지스틱 회귀모형을 구한 다음 복식투표에 의한 분류를 하였다.

이들 모형의 분류예측 정확성 평가를 위한 판정기준으로 본 연구에서는 확인용 자료집합에 대하여 다음과 같은 오분류율 식을 사용하였다.
오분류율=[(실제 0, 예측1)의 도수+(실제 1, 예측0)의 도수]/표본크기

3. 모의실험 설계 및 자료집합

주어진 데이터베이스에 대하여, 앞 2절에 제시된 알고리즘의 오분류율과 분류시간을 추정하기 위하여 표본 크기(1000, 2000, 3000)별로 50개의 데이터 세트를 반복 추출하였다. 표본별로 훈련용 자료집합과 확인용 자료집합을 6 : 4로 배정하여 모의실험하였으며, CHAID와 로지스틱을 각각 5회 배깅을 실시하였으며, 배깅은 10회 반복하였다. 분리기준으로 CHAID 알고리즘의 경우에 카이제곱 검정의 유의수준 0.2, 로지스틱 모형의 경우 단계별 변수선택과 최소 오분류율을 사용하여 유의수준 0.05를 적용하였다. 정지규칙으로는 CHAID 경우에 Kass 기준을 적용하였으며, 최소 오분류율 로지스틱의 경우 정지규칙이 따로 없다. 모형평가의 측도로 오분류율이 정확한 분류측도를 선택하였다. 온라인 SAS 기술보고서에 따르면 의사결정모형 계산을 위한 최소 RAM 크기는 8 x 표본크기이며, 계산속도는 전체 자료집합이 임시기억하기 위해서는 8 x 표본크기 x (입력변수의 수 + 1)에 비례한다. 로지스틱 회귀모형에서 최소 RAM 크기는 8 x 입력항의 수 x 2 이고 , 처리시간은 변수선택 단계의 수 x 입력항의 수 x 표본크기 x 입력항의 수에 비례한다.
본 연구의 모의실험계산을 위한 컴퓨터 사양은 다음 표 1과 같다.

항목	사양	비고
OS	Win 98 2nd.	
CPU	pantium3 330mhz	
Software	SAS 8.e · E-miner 4.0	각프로시져별
RAM size	128MB	512KB 사용

표 1. 모의 실험을 위한 컴퓨팅환경

데이터로는 SAS 시스템의 가상자료로 HMEQ를 선택하였다. 5960개의 레코드를 가졌으며, 변수는 모두 13개이고 이 가운데 범주형 변수가 3개이다. 예측변수로 대출금의 상환여부를 묻는 이가범주형 변수BAD로 정했다. 나머지 12개의 변수를 입력변수로 산입하는 모형으로 설정하였다. 변수에 대한 설명은 다음 표 2와 같다(강현철 외 5인 1998). 이 자료의 미상환율은 19.95 %이다.

변수명	내 용
BAD	대출상환여부 : 0 (상환), 1(미상환)
LOAN	총대출액
MORTDU	저당액
VALUE	재산액
REASON	대출사유 : 빚정리(Debtcom), 주택개량(HomeImp)
JOB	직업 : 노동자(Mgr) 사무원(Office) 숙련기술자(ProfExe) 기타(other) 판매원(Sales) 자영업(Self)
Yoj	직장 근무년수(year)
DEBTINC	대출금 대 수입의 비율(%)
DEROG	신용거래 중 불량사유 보고회수(회)
DELINQ	체납회수(회)
CLAGE	최장 대출기간(일)
NINQ	최근 신용거래 요청 회수(회)
CLNO	신용거래 회수(회)

표 2 . HMEQ 자료집합의 변수 및 변수내용

4. 모의실험

프로그램작성은 SAS 8. e 와 E-miner 4.0을 사용하여 모의실험을 수행하였다.

4.1 오분류율에 대한 분석

앞 절에서 제안한 예측모형의 정확성 판정기준으로 오분류율을 계산한 결과 다음 표 3과 같았다.

표 3에서는 각 알고리즘에 대하여 표본크기별 오분류율을 계산 하였다. Lim(2000)에서와 마찬가지로 표본의 크기가 클수록, 또한 의사결정나무모형이 회귀모형보다 오분류율이 작았다. 배깅기법이 Breiman (1996)에서와 같이 대체로 우수하였다.

Descriptive Statistics of error rates as sample sizes

sizes		N	Minimum	Maximum	Mean	Std. Deviation
1000	trees	50	.10361	.12836	.117067	5.0E-03
	logistic	50	.16149	.19883	.182119	1.0E-02
	bagging trees	10	.11065	.12441	.117004	3.9E-03
	bagging regression	10	.17122	.20512	.183925	1.0E-02
2000	trees	50	.10403	.12696	.115591	5.2E-03
	logistic	50	.16779	.19547	.183013	6.7E-03
	bagging trees	10	.10688	.12290	.115183	5.0E-03
	bagging regression	10	.17500	.19631	.184967	6.5E-03
3000	trees	50	.10780	.12248	.114239	3.6E-03
	logistic	50	.16569	.19463	.181688	7.8E-03
	bagging trees	10	.10990	.11745	.113994	3.1E-03
	bagging regression	10	.17022	.18993	.181686	6.6E-03

표 4 알고리즘별 오분류율 분산분석의 결과, 알고리즘에 따라 오분류율은 모두 유의하였다. 또한 표 3에서 표본크기 3000일 때 베깅 의사결정나무모형 오분류율은 0.113994으로 가장 작았으며, 베깅 회귀모형의 경우 전체적으로 오분류율이 컸다.

ANOVA of error rates among 4 models as sample sizes

		Sum of		Mean		
size 3000	Between Groups	.137	3	4.555E-02	1298.9	.000
	Within Groups	4.068E-03	116	3.507E-05		
	Total	.141	119			
size 2000	Between Groups	.138	3	4.600E-02	1300.3	.000
	Within Groups	4.104E-03	116	3.538E-05		
	Total	.142	119			
size 1000	Between Groups	.128	3	4.273E-02	659.03	.000
	Within Groups	7.522E-03	116	6.484E-05		
	Total	.136	119			

표 4 . 표본크기에 따른 예측 알고리즘별 오분류율 분산분석

표본크기에 따른 각 알고리즘의 오분류율 분산분석을 한 결과인 표 5에서는 CHAID 모형의 p 값이 0.011이었으며, 다른 나머지 모형들은 표본크기에 유의하지 않았다.

ANOVA of error rates among 3 sample sizes

		Sum of Squares	df	Mean Square	F	Sig.
CHAID	Between Groups	2.001E-04	2	1.001E-04	4.623	.011
	Within Groups	3.182E-03	147	2.165E-05		
	Total	3.382E-03	149			
Logistic	Between Groups	4.564E-05	2	2.282E-05	.324	.724
	Within Groups	1.037E-02	147	7.051E-05		
	Total	1.041E-02	149			
Bagging trees	Between Groups	4.597E-05	2	2.298E-05	1.395	.265
	Within Groups	4.449E-04	27	1.648E-05		
	Total	4.909E-04	29			
Bagging regression	Between Groups	5.621E-05	2	2.811E-05	.446	.645
	Within Groups	1.701E-03	27	6.300E-05		
	Total	1.757E-03	29			

표 5 . 알고리즘의 표본크기별 오분류율 분산분석

4.2 예측 알고리즘의 훈련시간

표 6에서는 각 알고리즘별 표본크기에 따른 훈련시간을 요약정리 하였다. SAS 기술보고서에서 보고한 바와 같이 표본의 크기가 커질수록, 훈련시간은 증가하였다. 입력항이 더 늘어나는 베깅의 경우도 훈련시간이 대 체로 증가하였으나, 기술보고서에서 지적한 대로의 완전한 비례관계는 보이지 않았다.

표 7에서는 각 알고리즘의 훈련시간은 표본크기에 따라 유의함을 보였다.

그림 1은 CHAID 모형일 경우, 훈련시간이 표본의 크기에 따라 다름을 보여 주고 있다. 표본크기 3000일 때 가장 훈련시간이 길다.

훈련시간의 중앙값을 사용하여 알고리즘별 분산분석을 한 결과 표 8과 같았다. 표본크기에 따라, 4 개 예측알고리즘의 훈련시간은 유의하였다. 그림 2와 Duncan 사후 검정결과 표본크기 3000일 때, 훈련시간평균은 로지스틱(79.25초), CHAID(90.00초), 베깅회귀모형(183.22초), 베깅 의사결정나무모형(190.12초)순이었다.

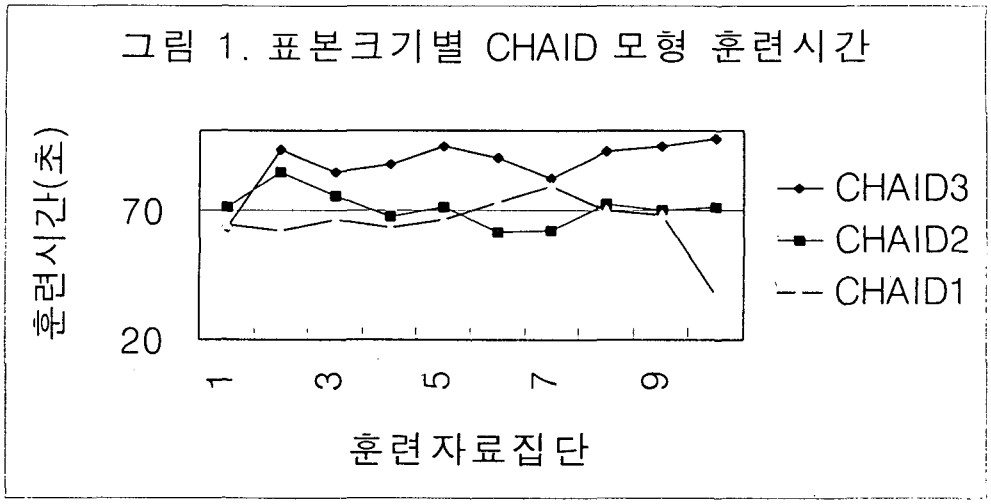
표본크기\알고리즘		반복회수	최소값	최대값	평균	표준편차
1000	CHAID	50	37.00	79.00	64.80	11.02
	Logistic	50	9.00	72.00	35.80	25.63
	Bagging Trees	10	77.35	164.56	136.94	23.35
	Bagging regression	10	20.16	153.92	77.39	53.97
2000	CHAID	50	61.00	84.00	64.80	6.50
	Logistic	50	61.00	81.00	69.80	6.16
	Bagging Trees	10	125.14	175.12	146.84	13.86
	Bagging regression	10	132.10	168.48	147.83	12.86
3000	CHAID	50	63.00	97.00	87.60	9.86
	Logistic	50	37.00	119.00	73.00	22.16
	Bagging Trees	10	134.25	202.05	183.23	19.76
	Bagging regression	10	85.46	175.12	157.84	21.09

표 6. 표본크기와 알고리즘별 훈련시간

ANOVA of run times

		Sum of Squares	df	Mean Square	F	Sig.
CHAID	Between Groups	18004.840	2	9002.420	5.167	.007
	Within Groups	256114.9	147	1742.278		
	Total	274119.7	149			
logistic	Between Groups	57430.493	2	28715.247	16.03	.000
	Within Groups	263358.1	147	1791.552		
	Total	320788.6	149			
bagging trees	Between Groups	19290.969	2	9645.485	4.822	.009
	Within Groups	292046.7	146	2000.320		
	Total	311337.6	148			
bagging regression	Between Groups	71673.100	2	35836.550	17.72	.000
	Within Groups	285217.2	141	2022.817		
	Total	356890.3	143			

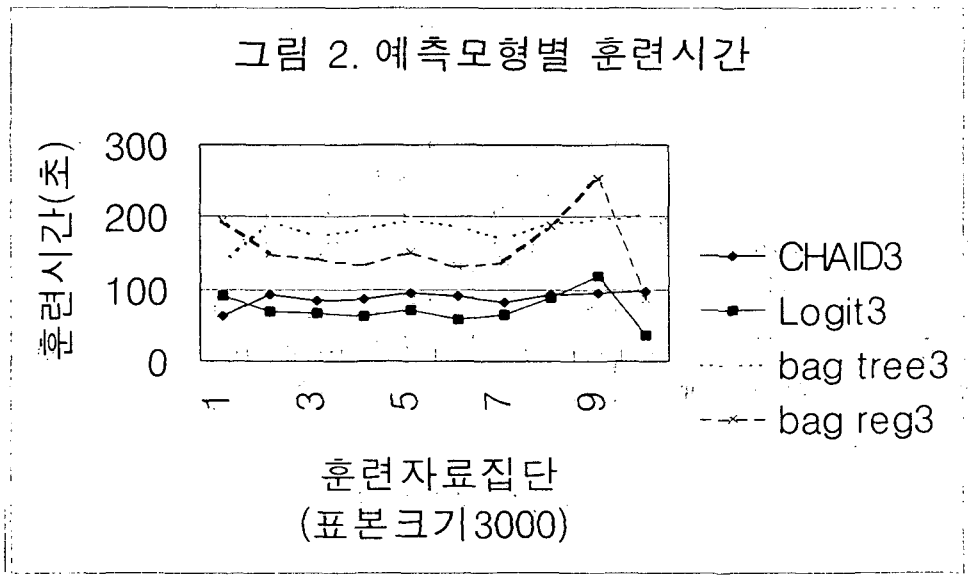
표 7. 표본크기에 대한 훈련시간 분산분석



ANOVA of training time among 4 models as sample sizes

		Sum of Squares	df	Mean Square	F	Sig.
size 3000	Between Groups	107138.2	3	35712.74	125.7	.000
	Within Groups	10229.739	36	284.159		
	Total	117367.9	39			
size 2000	Between Groups	61451.709	3	20483.90	154.2	.000
	Within Groups	4648.127	35	132.804		
	Total	66099.836	38			
size 1000	Between Groups	52965.906	3	17655.30	16.154	.000
	Within Groups	39344.753	36	1092.910		
	Total	92310.659	39			

표 8. 표본크기에 따른 4 알고리즘의 훈련시간 분산분석



4.3 오분류율과 훈련시간

그림 3, 4, 5는 10개 훈련시간의 평균 순위에 따른 오분류율을 대응시킨 그림이다. 표본크기와 훈련시간이 증가하여도 오분류율은 별 차이가 없음을 확인 할 수 있다. 전체적으로는 4. 2에서 이미 밝힌 바와 같이 CHAID와 베깅 의사결정나무모형이 로지스틱모형이나 베깅 회귀모형보다 오분류율이 작다는 것을 알 수 있다.

그림 3. 오분류율과 훈련시간(표본크기1000)

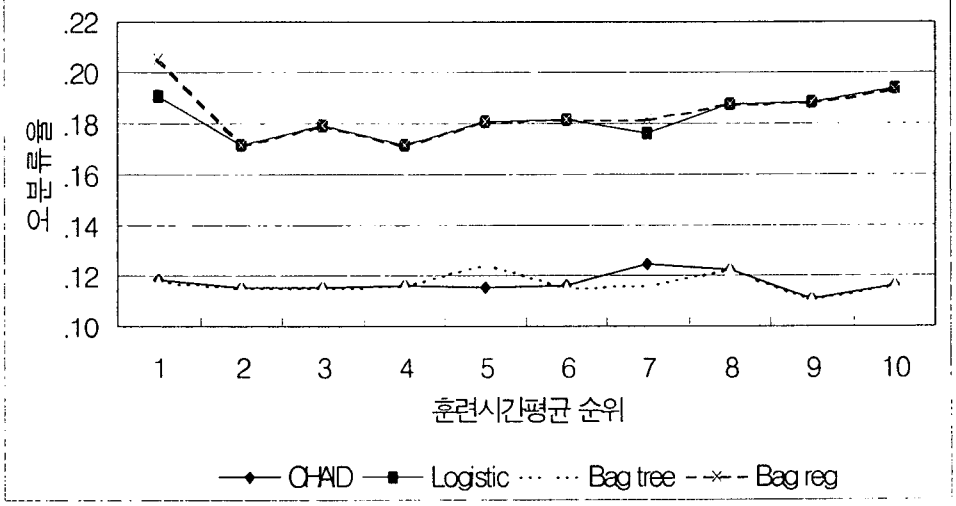
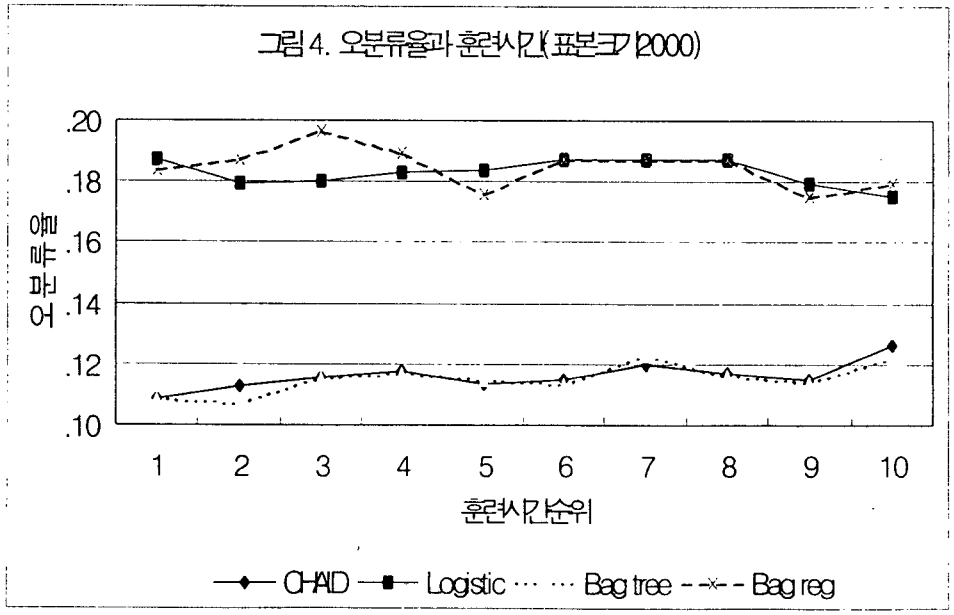
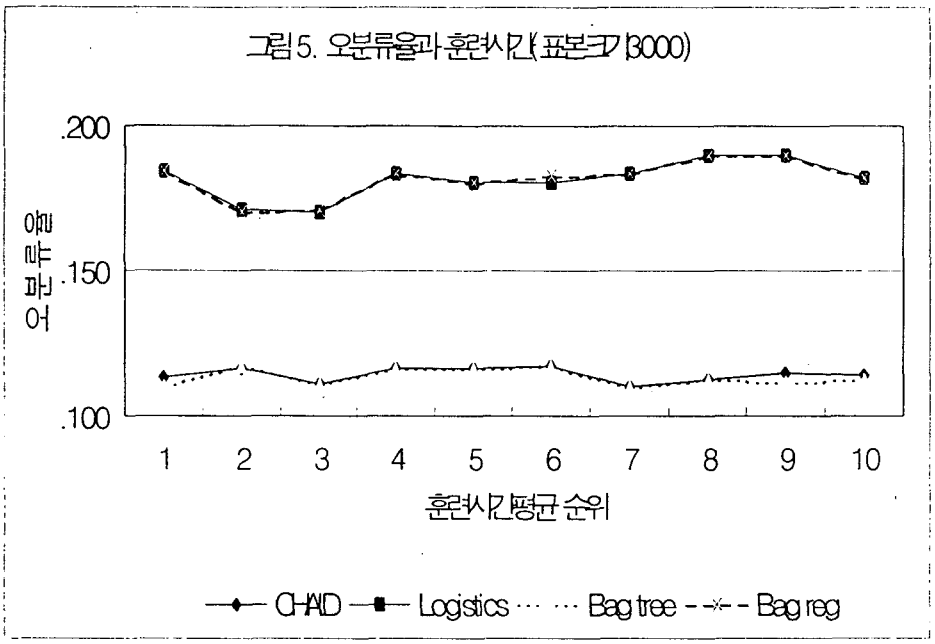


그림 4. 오분류율과 훈련시간(표본크기2000)





5. 결론

데이터마이닝의 분류예측모형의 정확성 평가기준 가운데 하나인 오분류율이 작은 알고리즘을 선택하는 것이 물론 중요하나, 본 논문의 모의실험에서 살펴 보았듯이, 알고리즘의 훈련시간에 대한 고려도 중요함을 알 수 있었다. 따라서, 균일적으로 효율성이 좋은 예측모형이 없음으로, 오분류율이 작은 의사결정나무모형을 적용할 것인가 혹은 훈련시간이 보다 작은 회귀모형을 사용할 것인지, 아니면 훈련시간은 다소 증가하지만 비교적 안정성을 가진 베깅모형으로 결정할 것인가하는 문제는 쉬운 것이 아님을 보여 주었다. 따라서 베이터베이스의 구조와 데이터마이닝의 목적에 따른 적절한 예측모형 선택을 해야 한다.

그리고, 데이터마이닝의 또 다른 중요한 면은 현실 산업의 현장적용이라는 측면이다. 인터넷 등의 정보통신망상으로 연결된 데이터베이스를 처리하기 위해서는 지금보다 훨씬 처리속도가 빠른 마이닝 알고리즘과 마이닝기법의 개발이 필요한 것이 현실이다. 기존 모형의 다양한 비교 평가는 새로운 모형을 제시해 주는 하나의 시금석으로써, 앞으로 연구개발용 혹은 상업용 마이닝 툴 사용자 입장의 입출력과정을 포함한 전과정 처리시간에 대한 비교연구도 할 필요성이 있다.

참 고 문 헌

1. 1. 강현철, 한상태, 최종후(2000). 의사결정나무를 활용한 데이터 마이닝 예측모형 해석, Proceedings of the Spring Conference, 39-44, Korean Statistical Society.

2. 2. 강현철,한상태,최종후,김창용,김은석,김미경(1998). SAS Enterprise Miner를 이용한 데이터마이닝, 방법론 및 활용, 자유아카데미.

3. 3. Elder, J. F. & Abbott, D.W.(1998). A Comparison of Leading Data Mining Tools, 4th International Conference on Knowledge Discovery and Data Mining, 1998, KDD-98, N.Y.

4. 4. Breiman, L. (1996). Bagging Predictors, Machine Learning, 24, 123-140, Kluwer Academics Plublishers, Boston.

5. 5. Breiman, L. (2000). Randomizing Outputs to Increase Prediction Accuracy, Machine Learning, 40(3), 229-242, Kluwer Academics Plublishers, Boston.

6. 6. Dietterich, T.G. (2000). An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization, Machine Learning, 40(2), 139-157, Kluwer Academics Plublishers, Boston.

7. 7. Hand, D. J.(1997). Construction and Assessment of Classification Rules, John Wiley & Sons.

8. 8. Kass, G. V.(1980). An Exploratory Technique for investigating large quantities of categorical data, G.V. Applied Statistics 29, 2, 119-127.

9. 9. Kohavi, Brodley, C. E., Frasca, B. Mason, L. and Zheng, Z.(2000). Reports from KDD-2000, KDD-Cup 2000 Organizer's Report: Peeling the Onion, R. SIGKDD Explorations, Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining, Volume 2, Issue 2.

10. 10. Lim, T. S., Loh, W. Y., and Shih, Y. S.(2000). A comparison of prediction accuracy. complexity, and training time of thirty three old and new classification algorithms, Machine Learning, 40, 203-228, 2000, Kluwer Academic Publishers.

11. 11. http://www.sas.com/service/techsup

Evaluations of predicted models fitted for data mining

- comparisons of classification accuracy and training time for 4 algorithms

Sang-Bock Lee [2]

Abstract

CHAID, logistic regression, bagging trees , and bagging trees are compared on SAS artificial data set as HMEQ in terms of classification accuracy and training time. In error rates, bagging trees is at the top, although its run time is slower than those of others. The run time of logistic regression is best among given models, but there is no uniformly efficient model satisfied in both criteria.

[2]Professor, Department of Informational Statistics, Catholic University of Daegu,712-702, Korea: sang-bock@cuth.cataegu.ac.kr