Journal of the Korean Data & Information Science Society
2013, **24**(6), 1341–1348

http://dx.doi.org/10.7465/jkdi.2013.24.6.1341
한국데이터정보과학회지

데일리 렌즈 데이터를 사용한 데이터마이닝 기법 비교[†]

석경하[1] · 이태우[2]

[12]인제대학교 데이터정보학과

접수 2013년 9월 30일, 수정 2013년 10월 18일, 게재확정 2013년 10월 24일

요 약

데이터베이스 마케팅과 시장예측 등의 분야에서 분류문제를 해결하기 위해 다양한 데이터마이닝 기법들이 적용되고 있다. 본 연구에서는 데일리 렌즈 고객들의 거래 데이터를 기반으로 의사결정나무, 로지스틱 회귀모형과 같은 기존의 통계적 분류기법과 최근에 개발된 배깅, 부스팅, 라소, 랜덤 포리스트 그리고 지지벡터기계의 분류 성능을 비교하고자 한다. 비교 실험을 위해 데이터 정제, 탐색, 파생변수 생성, 그리고 변수 선택과정을 거쳤다. 실험결과 정분류율 측면에서는 지지벡터기계가 다른 모형보다 근소하게 높았지만 표준편차가 크게 나왔다. 정분류율과 표준편차의 관점에서는 랜덤 포리스트가 가장 좋은 결과를 보였다. 그러나 모형의 해석, 간명성 그리고 학습에 걸리는 시간을 고려하였을 때 라소모형이 적합하다는 결론을 내렸다.

주요용어: 데이터마이닝, 라소, 랜덤 포리스트, 로지스틱회귀모형, 배깅, 부스팅, 의사결정나무, 지지벡터기계.

1. 서론

많은 기업들에 의해 생산되는 정보는 넘쳐나며, 이를 사용할 수 있는 능력보다 빠르게 증가하고 있다. 성공한 기업들이란 자신들이 접근할 수 있는 엄청난 양의 데이터를 효과적으로 이용하는 기업들이다. 하지만 동일한 방법으로 정보를 사용할 경우 경쟁력을 가질 수 없을지도 모른다. 따라서 자신들에게 유리한 방향으로 가능한 많은 이익을 얻기 위해서는 최근에 개발된 데이터마이닝 기법을 활용하는 것이 필수적이라고 할 수 있다.

국내 다양한 업종에서 데이터마이닝을 활용하여 고객의 특성을 파악하고 있는데, 대표적인 업무로는 이용실적에 따른 고객의 세분화와 이탈 가능 고객의 예측 등을 통한 다양한 캠페인 등을 들 수 있다. 모든 고객에게 마케팅 활동을 펼치기에는 시간과 예산의 제약이 생긴다. 이러한 이유로 많은 기업들은 재구매 가능성이 높은 고객들을 대상으로 타겟 마케팅을 전개하고 있으며, 고객들의 구매 데이터를 기반으로 구축된 재구매 예측 모형은 타겟 마케팅에 활용성이 높다고 할 수 있다. 이러한 시대의 요구에 따라, 다양한 데이터마이닝 기법들을 바탕으로 기업의 의사결정에 필요한 예측모형에 관한 연구가 활발히 이루어지고 있다.

본 논문은 2006년 8월 1일부터 2011년 7월 31일까지 총 6년간 현장에서 얻어진 데일리 렌즈 구매 데이터를 기반으로 데이터마이닝 기법을 이용하여 고객들의 재구매 유도를 위한 재구매 예측모형을 설계

[†] 이 논문은 2012년도 인제대학교 학술연구조성비 보조에 의한 것임.
[1] 교신저자: (621-749) 경남 김해시 인제로 197, 인제대학교 데이터정보학과, 통계정보연구소 교수.
E-mail: statskh@inje.ac.kr
[2] (621-749) 경남 김해시 인제로 197, 인제대학교 데이터정보학과, 석사졸업.

하고 정분류율을 비교하고자 한다. 이를 위해 얻어진 원 자료를 정제하였고 탐색하며 파생변수를 생성하는 사전 작업을 진행하였다. 그리고 유의한 변수를 선택하는 과정을 거쳤다. 본 논문에서는 배깅 (bagging; bootstrap aggregating)과 적응부스트 (Adaboost; adaptive boosting), 라소 (LASSO; least absolute shrinkage and selection operator), 랜덤 포리스트 (random forest) 그리고 지지벡터기계 (support vector machine) 모형을 이용하여 재구매 예측모형을 구축하였으며, 각 모형들의 분류성능과 효용성을 비교 할 것이다. 또한 기존의 통계적 분류 기법인 의사결정나무와 로지스틱회귀모형과도 비교하고자 한다.

본 논문의 구성은 다음과 같다. 1절에서는 연구배경 및 목적에 대해 서술하였고, 2절에서는 본 논문에 제시된 데이터마이닝 기법들에 대해 간단히 소개하였다. 그리고 3절에서는 분석 자료에 대한 소개와 실험방법 및 설계에 대해서 서술하였고, 4절에서는 2절에서 소개한 기법들을 이용한 실험결과 및 모형 비교에 대해 정리하였다. 5절에서는 결론 및 향후 과제에 대해서 논의하였다.

2. 데이터마이닝 기법

2.1. 배깅

배깅은 Breiman (1996)에 의해 처음 소개된 알고리즘으로 붓스트랩 (bootstrap) 방법을 이용한 앙상블모형 (ensemble model)이다. 학습 데이터로부터 추출된 여러 개의 붓스트랩 샘플을 통하여 다수의 분류기들을 생성하고 그 결과의 단순 평균이나 다수 투표 (majority voting)를 통해 최종적인 하나의 분류결과를 얻어내는 방법이다. 특히, 배깅은 데이터의 변화가 분류기에 미치는 영향이 큰 경우, 즉 불안정한 분류기의 분산을 줄여 예측력을 높여준다는 장점과 설명변수 사이의 연관성으로 인한 분산을 줄이는 장점을 가지고 있다 (Hastie 등, 2009).

2.2. 부스팅

부스팅 기법은 1984년 Valiant에 의해 처음으로 개념이 소개되고 난 이후 많은 분야에서 적용을 하여 좋은 결과를 얻었다 (Kim 등, 2012). Freund와 Shapire (1996)에 의해 제안된 적응부스트는 잘못 분류된 개체에 더 관심을 가지고 이들을 더 잘 분류하는 것에 그 목적을 둔다고 할 수 있다. 본 연구에서 사용된 적응부스트는 각 분류기들을 생성하고 그들을 결합하여 하나의 분류기를 생성한다는 점이 배깅 알고리즘과 비슷해 보이지만 분류기들을 독립적으로 생성하는 배깅 알고리즘과 달리 적응부스트는 분류기들을 순차적으로 생성한다.

2.3. 라소

데이터마이닝 모형에서 주어진 데이터에 적절한 모형을 구축하는 것도 중요하지만 설명변수를 선택하는 것도 중요하다. 회귀분석에서의 일반 최소제곱법은 설명변수의 수가 증가하면 설명변수들 사이의 강한 상관관계로 인한 다중공선성이 존재할 수 있다. 다중공선성이 존재할 때 회귀계수의 추정량은 높은 분산으로 인해 신뢰도가 매우 낮다. 이는 곧 훈련용 데이터에서의 예측력은 높은 반면, 검증용과 시험용 데이터에서의 예측력은 크게 저하되는 일반화오류로 이어질 수 있다.

최소제곱법의 문제점이 보완된 방법으로 부분집합선택방법과 능형회귀 등이 있는데 능형회귀처럼 회귀 계수들의 크기에 대한 제약 조건을 주어 영향력이 없는 변수를 제거하여 변수선택을 가능하게 하는 라소가 개발되었다 (Tibshirani, 1996). 라소는 데이터의 변화에 불안정한 회귀분석의 단점을 극복하고자 나온 축소기법인데 회귀계수 추정치들의 크기를 축소시킴과 동시에 변수선택도 할 수 있는 장점을 지니고 있어 많은 분야에서 응용되고 있다.

2.4. 의사결정나무

Breiman 등 (1984)에 의해 처음 제안된 의사결정나무는 쉽게 분석결과를 이용할 수 있기 때문에 많이 사용되는 데이터마이닝 기법이다. 분리기준과 정지규칙에 따른 형성과정이 도표화되기 때문에 나무가 비교적 큰 경우에도 어려움 없이 규칙을 이해할 수 있다는 장점이 있다. 의사결정나무에 사용되는 알고리즘에는 CHAID, CART, C4.5가 있으며 최근에는 이들의 장점을 결합하여 보다 개선된 알고리즘들이 제안되고 있다.

분리기준과 정지규칙을 통해 의사결정나무가 완성되면 오류율을 증가시키는 규칙을 가지고 있는 가지를 제거한 후, 최종적으로 이익도표나 교차타당성 평가를 이용하여 나무 모형을 평가함으로써 해석 및 예측이 가능해진다. 의사결정나무는 이상치에 덜 민감하며 학습속도가 빠르고 과적합되지 않는다는 장점이 있지만, 연속형 변수의 값을 예측하거나 범주가 많은 경우에는 성능이 떨어지는 것으로 알려져 있다.

2.5. 랜덤 포리스트

랜덤 포리스트는 의사결정나무를 기반으로 설계된 앙상블 기법으로 Breiman (2001)에 의해 개발된 알고리즘이다. 랜덤 포리스트는 앙상블 접근법들 중에서도 가장 널리 사용되는 방법이며, 단일 의사결정나무를 사용하는 것 보다 정확성과 안정성 측면에서 더 나은 결과를 보이는 것으로 알려져 있다. 또한, 중요 변수를 추정하는데 용이하다는 점으로 유전자 연구분야나 의학계통 분야에서 많은 관심의 대상이 되고 있다.

랜덤 포리스트는 배깅과 유사하며, 원데이터로부터 생성된 동일한 크기의 부스트랩 샘플들에 의해 만들어진 의사결정나무로 구성된다. 각각의 의사결정나무는 랜덤하게 선택된 설명변수와 해당 부스트랩 샘플에 의해 생성되며, 가치치기를 하지 않고 나무를 최대한 확장한다. 이 과정을 통해 생성된 다수의 의사결정나무에 의한 예측결과를 가중 투표를 통하여 결합함으로써 최종 분류가 이루어진다.

2.6. 지지벡터기계

최근 패턴분류에 있어서 각광을 받고 있는 지지벡터기계는 Vapnik (1996)이 개발하였는데, 숫자 인식, 문서 범주화 그리고 얼굴 인식 등과 같은 패턴 인식 응용 분야에서 우수함이 입증되었다. 지지벡터기계의 기본 아이디어는 두 부류 사이에 존재하는 여백 (margin)이라는 개념을 분류기 설계에 도입하고 여백을 최대화하는 초평면을 찾는 것으로써 기존의 분류 방법들과는 기본 원리가 크게 다르다. 즉, 신경망을 포함한 기존의 통계적 분류 방법들은 오류율을 최소화하려는 목적으로 설계되었지만 지지벡터기계는 두 부류 사이에 존재하는 여백을 최대화 하여 일반화 성능 향상에 목적을 두고 있다. 지지벡터기계는 일반화 측면에서 다른 분류기와 비교하면 대등하거나 우수한 것으로 알려져 있다 (Park, 2011; Pi, 2013; Hwang 등, 2006). 입력변수가 상대적으로 많은 데이터에 대해서도 비교적 높은 일반화 성능을 유지할 수 있는 것으로 알려져 있다.

3. 자료분석

3.1. 데이터 소개

원 자료는 데일리렌즈 구매 고객들에 대한 데이터로 2004년 9월부터 2011년 7월까지의 총 고객수 100,000명에 대한 총 거래건수 796,845건의 자료이다 (Kim 등, 2012). Table 3.1은 원 자료의 변수 설명을 나타낸다.

Table 3.1 Variables description

Variable	Description
Seq	member ID
Kind	membership type (M : full member, S : associate member)
Gender	gender (M : male, F : female)
Age	age
Product	product type
Sales	unit price of product
Point	used/saved membership mileage (- : used, + : saved)
Pack_Kind	type of pack purchased
Pack_Count	number of packs purchased
Lens_Count	number of lenses purchased
Buy_Date	purchased date

3.2. 데이터 정제

거래기간과 재구매 기간 등을 고려하여 모형 구축을 위해 최종 선택된 데이터는 2006년 8월 1일부터 2011년 7월 31일까지 총 6년간의 구매데이터이다. 2004년도와 2005년도의 거래건수는 49건과 303건으로 다른 년도의 거래건수보다 현저히 적어 제외하였다. 원자료에서 age 변수는 자료를 만든 시점의 나이로 되어있기 때문에 분석을 위해 구매시점의 나이로 다시 계산하였고 상품에 대한 가격 변동이 있는 경우에는 최근 가격으로 동일하게 수정하였다. 그리고 현금구매의 포인트 값이 음수인 관측치들은 제거하였고 하루에 여러 번 구매가 일어났을 경우 하나의 구매로 처리하였다. 위와 같이 정제를 거치 후 거래건수는 215,893개가 줄어 580,952개로 되었다.

3.3. 변수 생성

분석에 사용할 데이터는 기준시점을 정하고, 기준시점 이전 1년간의 구매행태를 비롯한 여러 변수와 기준시점 이후 6개월 동안의 재구매 여부를 나타내는 목표변수 (re)로 구성하였다. 기준시점은 2007년 8월 1일부터 2010년 2월 1일까지 6개월 단위로 정하였으며 모두 8개 시점이다. 8개의 기준시점에 따라 만들어진 분석데이터의 거래건수는 294,475건이며 분석에 사용된 변수는 Table 3.2 와 같다. 마지막 거래일자와의 차이 (distance)는 기준시점과 직전 거래시점과의 차이에 해당하는 값이다. 평균구매간격 (interval)은 기준 시점 이전 1년 중에 마지막 거래와 처음 거래의 차이를 (거래횟수-1)로 나눈 것이며, 구매 기간 안에 한 번의 거래만 있는 경우는 일괄적으로 365일로 부여하였다. 모형 추정에 사용된 294,475건 중 re=1인 거래는 152,271건으로 전체의 51.71%로 나타났고, re=0인 거래는 142,204건으로 전체의 48.29%로 나타났다.

Table 3.2 Variables used in analysis

Variable	Description
Age	customer's age
Interval	average of purchase intervals
Count	number of purchases
Totprice	total amount of purchases
Distance	distance of base time and final purchase time
Totlens	cumulative number of purchased lenses
Re	repurchase

3.4. 변수 탐색

Table 3.3 은 1년간의 구매형태를 기반으로 고객을 분류하고 재구매를 예측하기 위하여 기준시점으로부터 지난 1년간의 구매데이터를 이용하여 다음 6개월 동안 재구매가 일어난 비율을 살펴본 결과이다. 기준시점이 최근으로 이동할수록 재구매 비율이 40.97%∼59.33%로 조금씩 증가하다가 마지막 기준시점에서 약간 감소하고 있음을 알 수 있다. 이러한 경향은 구매기간과 재구매 기간의 간격을 달리하였을 때도 비슷한 결과를 보였다.

Table 3.4는 재구매 유무에 따른 각 변수들의 기술통계량이다. 재구매를 한 고객들의 평균 구매 간격 (interval)과 마지막 구매일자와의 차이 (distance)가 재구매를 하지 않은 고객들보다 더 짧음을 알 수 있다. 구매횟수 (count), 총 구매금액 (totprice) 그리고 총 구매 렌즈 수 (totlens)의 평균은 재구매를 한 고객들이 더 많음을 알 수 있다.

Table 3.3 Repurchase rate at each base time

re	base time								total
	2007.08.01	2008.02.01	2008.08.01	2009.02.01	2009.08.01	2010.02.01	2010.08.01	2011.02.01	
0	19068 (59.0)	20788 (58.2)	21545 (54.8)	18259 (50.1)	14322 (42.4)	14200 (40.8)	15886 (40.6)	18136 (42.2)	142204 (48.29)
1	13234 (41.0)	14979 (41.8)	17765 (45.2)	18206 (49.9)	19502 (57.6)	20589 (59.2)	23176 (59.4)	24820 (57.8)	152271 (51.71)
total	32302 (10.9)	35767 (12.2)	39310 (13.4)	36465 (12.4)	33824 (11.5)	34789(11.8)	39062(13.3)	42956(14.5)	294475(100.0)

Table 3.4 Basic statistics of variables

Variable	repurchase	N	mean	std	min	max
Interval	0	86180	176.47	195.14	1.00	1597.00
	1	152271	106.70	91.29	1.00	545.00
Count	0	142204	3.80	4.94	1.00	192.00
	1	152271	9.93	10.43	2.00	386.00
Totprice	0	142204	292804	559298	15000	53622000
	1	152271	751329	985870	30000	53622000
Distance	0	142204	167.37	106.59	0.05	365.57
	1	152271	71.91	75.31	0.02	365.52
Totlens	0	142204	109.25	119.31	30.00	4864.00
	1	152271	225.41	229.97	30.00	8880.00

3.5. 변수 변환 및 이상치 제거

변수변환을 위해 재구매율 (p)에 대한 로짓값 ($\log(p/(1-p))$)과 변수들의 관계를 살펴보고 그 관계가 선형이 될 수 있도록 변수를 변환하였다. 그 결과 age, count, totprice과 totlens를 로그 변환한 변수 l_age, l_count, l_totprice, l_totlens가 생성되었다. 그리고 distance는 제곱근 변환된 변수 s_distance를 생성하였다. Interval은 로짓값과의 33이상인 구간과 미만인 2개 구간 에서 기울기가 다른 형태를 보이므로, interval>33 이면 X_1=interval, 아니면 X_1=0, interval<33 이면 X_2=interval, 아니면 X_2=0 인 두 개의 변수와 interval>33이면 ind =1, 아니면 ind=0 인 더미 (dummy)변수를 생성하였다. 상위와 하위 각각 0.5%에 포함되는 관측치는 이상치로 간주하여 제거하였다. 필터링 결과, 총 3108개 (2.6%)의 관측치가 제거되었다.

3.6. 모형 설계

모형 비교를 위해 총 294,475건 중 100,000건을 랜덤 샘플링하여 분석을 실시하였다. 샘플링에 의해 생기는 오차를 고려하기 위해 100번 반복하였으며, 배깅, 적응부스팅, 라소, 랜덤 포리스트 그리고 지지벡터기계는 자료를 훈련용 (training)과 평가용 (test)으로 7:3으로 분할하였다.

배깅과 적응부스팅은 반복적인 랜덤 샘플링 기법을 통해 다중 분류기를 생성한다. Opitz와 Maclin (1999)는 반복횟수가 25가 될 때까지는 오분류율이 개선되지만 그 이후에는 별 영향을 받지 않는다는

결과를 보였다. 따라서 본 논문에서도 배깅과 적응부스팅 모형을 구축할 때에는 반복 횟수를 25로 결정하였다.

라소의 조절모수는 교차타당성 (cross validation)방법으로 결정하였다. 그리고 랜덤 포리스트 모형의 각 노드에서 사용되는 변수의 수는 $\sqrt{\text{모든 변수의 수}}$로 하는 것이 일반적이지만, 본 연구에서는 변수의 수를 1부터 8까지 다양하게 조절하여 모형선택에 사용되지 않은 자료 (out of bag sample)의 오류율이 가장 최소로 되는 수를 결정하였으며, 분리기준으로는 지니계수를 사용하였다.

지지벡터기계에 사용된 커널함수는 원형기준함수 (radial basis function)이고, 커널모수와 정칙모수 (regularization parameter)는 교차타당성 방법으로 결정하였다

4. 실험결과

본 연구에서는 정분류율을 기준으로 각 분류방법을 비교하였다. 크기가 100,000인 랜덤표본을 훈련용과 평가용으로 7:3의 비율로 나누어 훈련용으로 모형을 만들고 평가용으로 비교하였다. 이러한 작업을 100회 반복하여 정분류율의 평균과 표준편차를 계산하여 Table 4.1에 나타내었다. 전체적으로 정분류율의 차이는 크에 나타나지 않았다.

Table 4.1 Average and standard deviation of correct classification rate

Model	average (standard deviation) with training data	average (standard deviation) with test data
Bagging	0.7335 (0.0011)	0.7120 (0.0010)
AdaBoost	0.7307 (0.0015)	0.7212 (0.0022)
LASSO	0.7281 (0.0015)	0.7310 (0.0023)
Support Vector Machine	0.7449 (0.0028)	0.7388 (0.0029)
Random Forest	0.7434 (0.0008)	0.7321 (0.0016)
Logistic regression	0.7267 (0.0013)	0.7312 (0.0020)
Decision Tree	0.7303 (0.0019)	0.7283 (0.0015)

훈련용 데이터에서는 지지벡터기계의 정분류율이 74.49%로 다른 방법들 보다 약간 높은 것으로 나타났다. 그 다음으로 랜덤 포리스트 (74.34%)와 배깅 (73.35)이 근소한 차이를 보이고 있다. 평가용 데이터에서도 지지벡터기계의 정분류율이 73.88%로 다른 모형들에 비해 분류 성능이 근소하게 높았음을 알 수 있었다. 그 다음으로는 랜덤 포리스트 (73.21%), 로지스틱 회귀모형 (73.12%) 그리고 라소 (73.10%) 등이 있다.

랜덤 포리스트와 배깅은 훈련용자료와 평가용자료에서 표준편차가 각각 0.0008, 0.001로 가장 적은 값을 가진다. 정분류율에서 좋은 결과를 보인 지지벡터기계는 훈련용과 평가용자료에서 0.0028과 0.0029로 가장 큰 표준편차를 가지는 것을 알 수 있다.

데이터마이닝 기법 중 지지벡터기계의 정분류율이 근소하게 높은 것으로 나타났지만, 다른 모형에 비해 학습시간이 상당히 오래 걸렸으며, 커널모수에 영향을 많이 받는 것으로 나타났다. 반면, 라소는 다른 모형보다는 비교적 적은 학습시간을 소모하였으며, 축소기법을 통해 총 8개의 설명변수 중에서 X_1, X_2, Ind를 제외한 나머지 5개의 변수를 선택하였다.

5. 결론 및 향후 과제

본 연구는 데일리 렌즈 구매데이터를 이용하여 과거 1년간의 구매행태로부터 향후 6개월 이내의 재구매 여부에 대한 모형을 비교하였다. 비교에 사용된 모형은 비교적 최근에 개발된 배깅, 적응부스팅, 라

소, 랜덤 포리스트, 지지벡터기계와 기존의 통계적 분류 기법인 로지스틱 회귀모형과 의사결정나무 등이다.

사용된 데이터는 총 6년간의 데일리 렌즈 고객들의 구매데이터이다. 각 변수들에 대한 특징을 살펴보고, 재구매 여부에 따른 평균비교를 통하여 재구매에 유의한 영향을 주는 변수들을 파악하였다. 각 변수에 대한 적절한 변환과 정제작업을 통하여 모형 구축에 필요한 변수들을 생성하였다.

정분류율을 살펴보면 지지벡터의 정분류율이 훈련용자료와 평가용자료에서 각각 74.30%, 73.88%로 다른 모형에 비해 근소하게 높았음을 알 수 있었다. 그러나 훈련시간이 많이 걸리고 표준편차가 크다는 것을 알 수 있었다. 정분률율과 표준편차의 관점에서는 랜덤 포리스트가 가장 좋은 결과를 보였다. 그러나 모형의 해석, 간명성 그리고 학습에 걸리는 시간을 고려하였을 때 라소모형이 적합하다는 결론을 내렸다.

각 모형에 사용되는 모수에 의해 정분류율이 결정되므로 좀 더 정확하고 객관적인 방법으로 모수를 선택하는 방법에 대한 연구가 있어야 할 것이다.

References

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

Breiman, L. (2001). Random forests. *Machine Learning*, **45**, 5-32.

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

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

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The element of statistical learning: Data mining, inference, and prediction*, New York, Spring Verlag.

Hwang, J,. Lee, J. and Kim, J. (2006). A comparison study of multiclass SVM methods in microarray data. *Journal of the Korean Data & Information Science Society*, **17**, 311-324.

Kim, A., Kim, J. and Kim, H. (2012). The guideline for choosing the right-size of tree for boosting algorithm. *Journal of the Korean Data & Information Science Society*, **23**, 949-959.

Kim, B., Cho, D., Lee, J., Lee, T., Hyun, J. and Kim, S. (2012). Comparison of two repurchase models using logistic regression and memory based reasoning. *Journal of the Korean Data Analysis Society*, **14**, 1301 - 1314.

Opitz, D. and Maclin, R. A. (1999). Popular ensemble methods : An empirical study. *Journal of Artificial Intelligence Research*, **11**, 169-198.

Park, H. (2011). Online abnormal events detection with online support vector machine. *Journal of the Korean Data & Information Science Society*, **22**, 197-206.

Pi, S. (2013). Self-diagnostic system for smartphone addictionusing multiclass SVM. *Journal of the Korean Data & Information Science Society*, **24**, 13-22.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society B*, **58**, 267-288.

Vapnik, V. N. (1996). *The nature of statistical learning theory*, Springer, New York.

Journal of the Korean Data & Information Science Society
2013, **24**(6), 1341–1348

http://dx.doi.org/10.7465/jkdi.2013.24.6.1341
한국데이터정보과학회지

Comparison of data mining methods with daily lens data[†]

Kyungha Seok[1] · Taewoo Lee[2]

[12]Department of Data Science, Inje University

Received 30 September 2013, revised 18 October 2013, accepted 24 October 2013

Abstract

To solve the classification problems, various data mining techniques have been applied to database marketing, credit scoring and market forecasting. In this paper, we compare various techniques such as bagging, boosting, LASSO, random forest and support vector machine with the daily lens transaction data. The classical techniques-decision tree, logistic regression-are used too. The experiment shows that the random forest has a little smaller misclassification rate and standard error than those of other methods. The performance of the SVM is good in the sense of misclassfication rate and bad in the sense of standard error. Taking the model interpretation and computing time into consideration, we conclude that the LASSO gives the best result.

Keywords: Bagging, boosting, data mining, decision tree, LASSO, logistic regression, support vector machine.

[†] This work was supported by grant from Inje University 2012.
[1] Corresponding author: Professor, Department of Data Science, Institute of Statistical Information, Inje University, Kimhae 621-749, Korea. E-mail: statskh@inje.ac.kr
[2] Master of Data Science, Department of Data Science, Inje University, Kimhae 621-749, Korea.