의사결정나무를 이용한 개인휴대통신 해지자 분석

최종후 · 서두성
고려대학교 정보통계학과

Abstract

본 논문에서는 최근 데이터마이닝의 도구로 활발하게 소개되고 있는 의사결정나무 분석을 이용하여 개인휴대통신의 해지자 분석을 실시한다. 또한 로지스틱 회귀모형을 이용하여 가입고객의 해지 가능성에 대한 점수화를 시도한다.

1. 서론

본 연구에서는 개인휴대통신 고객의 해지특성이 어떤 가입자 속성변인에 의존하는지를 분석하고 해지 가능성에 대한 점수화(scoring)를 시도한다.

고객의 해지특성을 알아보기 위하여 최근 데이터마이닝의 도구로 널리 이용되고 있는 의사결정나무(decision tree)분석을 이용하였으며, 해지 가능성에 대한 점수화는 로지스틱 회귀모형을 이용한다.

본 연구에서 분석의 전 과정은 SAS/E-Miner를 이용하여 분석하였다.

2. 의사결정나무(Decision Tree)

의사결정나무는 의사결정규칙(decision rule)을 나무구조로 도표화하여 분류(classification)와 예측(prediction)을 수행하는 분석방법으로 분류 또는 예측의 과정이 나무구조에 의한 추론규칙(induction rule)에 의해서 표현되기 때문에 분석자가 그 과정을 쉽게 이해하고 설명할 수 있다는 장점을 가지고 있다. 의사결정나무는 판별분석(discrimination analysis) 또는 회귀분석(regression analysis) 등에서 분석에 필요한 변수를 찾아내고 모형에 포함되어야 할 교호효과를 찾아내는 데에 사용될 수도 있으며, 그 자체가 분류 또는 예측 모형으로 사용될 수도 있다.

일반적으로 의사결정나무 분석은 다음과 같은 단계를 거친다(Berry and Linoff:1997; 강현철, 서두성, 최종후:1998).

- 의사결정나무의 형성: 분석의 목적과 자료구조에 따라서 적절한 분리기준(split criterion)과 정지규칙(stopping rule)을 지정하여 의사결정나무를 얻는다.
- 가지치기: 분류오류(classification error)를 크게 할 위험(risk)이 높거나 부적절한 추론규칙을 가지고 있는 가지(branch)를 제거한다.
- 타당성 평가: 이익도표(gains chart)나 위험도표(risk chart) 또는 검정용 자료(test data)에 의한 교차타당성(cross validation) 등을 이용하여 의사결정나무를 평가한다.
- 해석 및 예측: 의사결정나무를 해석하고 분류 및 예측모형을 설정한다.

2.1 이산형 목표변수에 대한 분리 기준

이산형 목표변수(target variable)에 대한 분리기준으로는 카이제곱 통계량, 지니 지수, 엔트로피 지수 등이 이용된다. 이러한 분리 기준으로 형성된 의사결정나무를 분류나무(classification tree)라고 한다.

- 카이제곱 통계량(Chi-Square statistic)

목표변수와 설명변수의 관측도수로 이루어진 $r \times c$ 분할표로부터 계산되며, 이때 카이제곱 통계량값은 다음과 같다.

$$\chi^2 = \sum_{i,j} \frac{(f_{ij} - E_{ij})^2}{E_{ij}}$$

$$\text{단, } E_{ij} = \frac{f_{i.} \times f_{.j}}{f_{..}}$$

분리기준을 카이제곱 통계량으로 한다는 것은 p값이 가장 작은 설명변수와 그때의 최적분리에 의해서 자식마디가 형성되게 한다는 것을 의미한다.

- 지니 지수(Gini index)

지니 지수는 n개의 원소 중에서 임의로 2개를 추출하였을 때, 2개가 서로 다른 그룹에 속해있을 확률로, 다음과 같이 표현된다(Breiman and etc.:1984).

$$G = \sum_{j=1}^{c} P(j)(1 - P(j))$$

$$= 1 - \sum_{j=1}^{c} P(j)^2 = 1 - \sum_{j=1}^{c} (n_j/n)^2$$

즉, 지니 지수는 각 마디에서의 불순도(impurity)를 재는 측도인데, 이 지니 지수를 가장 감소시키는 설명변수와 그 변수의 최적분리를 자식

마디로 선택한다.

· 엔트로피 지수(Entropy index)

지니 지수와 유사한 분리기준으로, 다항분포(multinomial distribution)에서의 우도비 검정통계량(likelihood ratio test statistic)을 사용하는 것과 같다(Quinlan:1993).

$$E = -\sum_{i=1}^{c} P(i)\ \log_2 P(i)$$

2.2 연속형 목표변수에 대한 분리 기준

연속형 목표변수에 대한 분리 기준은 F 통계량, 분산의 감소량(variance reduction) 등이 이용된다. 이러한 분리 기준으로 형성된 의사결정나무를 회귀나무(regression tree)라 한다.

· F 통계량

y_{ij}를 i 번째 설명변수의 범주에 속하는 j 번째 관측개체의 목표변수의 값이라고 하고, $\bar{y}_i$를 i 번째 범주의 평균, $\bar{y}$를 전체평균이라고 할 때, F 통계량은 다음과 같다.

$$F = \frac{\sum_{i=1}^{r} n_i(\bar{y}_i - \bar{y})^2/(r-1)}{\sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2/(n-r)}$$

이 통계량은 자유도 $(r-1,\ n-r)$인 F-분포를 따르며, F 통계량이 매우 작다는 것은 설명변수에 따른 목표변수의 평균차이가 유의하지 않다는 것을 의미한다.

· 분산의 감소량(variance reduction)

각 마디의 다양도(diversity)를 재는 측도로 다음과 같은 분산을 고려할 수 있다.

$$V = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

3. 로지스틱 회귀모형(Logistic Regression)

로지스틱 회귀분석은 목표변수가 명목척도로 측정되어있는 경우에 목표변수와 설명변수 간의 관계를 추정하기 위하여 적용되는 통계기법의 하나이다. 로지스틱 회귀분석의 사용은 판별분석을 사용하는 것과 마찬가지로 두 집단으로 구분된 개체에 대해 각 개체가 속하는 집단을 예측하거나, 집단의 구분에서는 어느 설명변수가 중요한지를 알아내는데 사용된다. 본 분석에서는 로지스틱 회귀모형을 통하여 고객의 해지 가능성에 대한 점수화를 시도한다.

일반적으로 설명변수의 수가 p, 목적변수 Y가 1 혹은 2인 로지스틱 회귀모형은 다음과 같다(허명회:1995).

$$\log\frac{P(Y=1 \mid x_1,\cdots,x_p)}{P(Y=2 \mid x_1,\cdots,x_p)} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

$$P(Y=1 \mid x_1,\cdots,x_p) = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}$$

4. 개인휴대통신 해지자 분석

고객DB를 이용한 고객 세분화(segment)로 이동통신 가입고객의 해지특성이 어떠한 패턴을 이루고 있는지를 알아보기 위하여 의사결정나무 분석을 실시하였다. 이러한 분석은 고객 이탈율(defection rate)을 감소시키는 고객유지 마케팅(retention marketing)의 일환으로 이용될 수 있다.

분석에 사용된 자료는 A통신회사의 서울지역 고객DB를 이용하여 랜덤추출로 2,500개의 표본을 획득한 것이다. 분석표본의 해지율은 13.2%이다. 목표변수로는 해지여부이며 기타 고객속성 변수가 설명변수로 사용되었다. 변수의 내용은 <표 1>과 같다.

<표 1> 분석에 사용된 변수

변 수 명	범 주
해지여부	정상사용/일반해지
고객계정상태	개통사용중/최종청구/정상해지
최근 4개월간 사용료	없음/1만5천원미만/1만5천원~2만7천원미만/2만7천원~4만원미만/4만원~5만5천원미만/5만5천원~7만1천원미만/7만1천원~9만2천원미만/9만2천원~11만8천원미만/11만8천원~15만7천원미만/15만7천원~22만8천원미만/22만8천원이상
최근 1년간 미납여부	없음/있음
납입방법	자동이체/카드이체/지로납부/중앙불
가입경력	6개월미만/6~10개월/11~12개월/13개월/14~18개월/19~22개월/23~26개월/27~33개월/34~46개월/47개월이상
디지털 유무	아날로그/디지탈
총 불만건수	없음/1번/2번/3번이상
요금계획	일반요금/비지니스/일반요금(VMS)/예치요금/예치요금(VMS)/프리미엄/프리미엄(VMS)/이코노미/이코노미(VMS)
성별	남자/여자
연령대	10대/20대/30대/40대/50대/60대/70대이상

의사결정나무 분석의 타당성을 위하여 자료를 분석용 자료(training data)와 타당성 평가용 자료(validation data)로 나누어 분석하였다.

<그림 1>은 의사결정나무 모형에서 CHAID 방법(Kass:1980)을 이용한 다중 나무구조(multi-tree structure)의 분류결과이다. 총 8개의 최종마디(leaf node)로 이루어진 나무구조가 형성되었다. 맨 위에

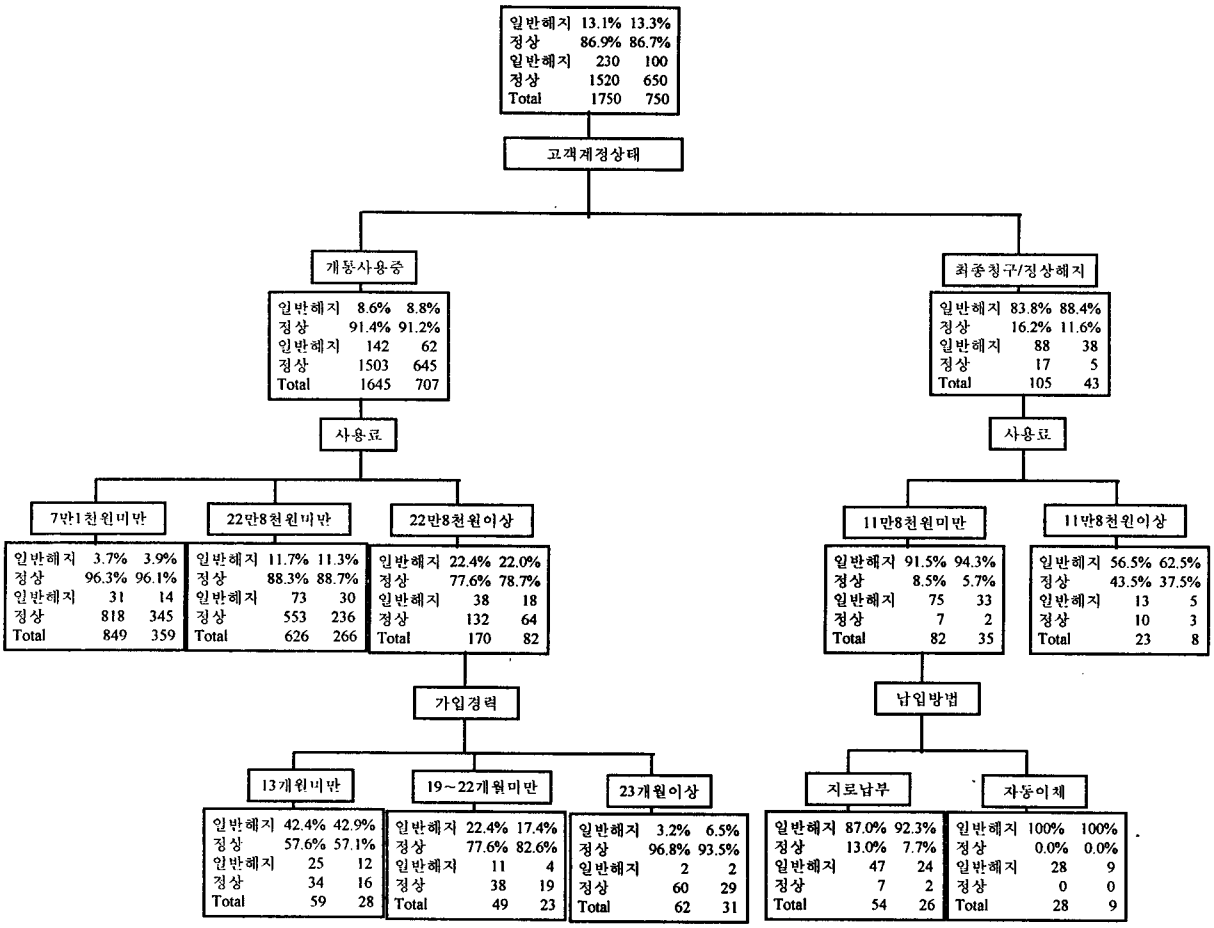

그림 1 의사결정나무

있는 뿌리마디(root node)는 2,500개의 관측치로 분석용 자료와 타당성 평가용 자료의 해지율이 각각 13.1%, 13.3%로 나타나고 있다.

가입고객의 해지를 결정하는 제일 중요한 변수로는 고객계정상태이며, 두 번째로는 최근 4개월간 사용료, 세 번째로는 가입경력과 납입방법 등이다. 이중 가입고객의 고객계정상태가 '최종청구/정상해지'인 경우에 해지율이 83.8%, 88.4%로 높아짐을 볼 수 있으며, 다음으로 가입고객이 고객계정상태가 '개통사용중'이면서 최근 4개월간 사용료가 '22만8천원이상'인 경우 해지율이 22.4%, 22.0%로 높아짐을 볼 수 있다. 특히, 가입고객의 고객계정상태가 '개통사용중'이면서 최근 4개월간 사용료가 '22만8천원이상'이면서 가입경력이 '13개월미만'의 경우 해지율이 42.4%, 42.9%로 높아짐을 볼 수 있다.

<표 2>는 의사결정나무 분석의 오분류 테이블이다. 오분류율(error rate)과 정확도(accuracy)가 각각 0.0904, 0.9096으로 잘 분류되어진 것 같으나 민감도(sensitivity)가 0.3818로 일반해지를 일반해지로 예측하는 예측력이 상당히 떨어짐을 볼 수가 있다.

<표 2> 의사결정나무 분석의 오분류 테이블

		예측		계
		일반해지	정상	
실제	일반해지	126 5.04%	204 8.16%	330
	정상	22 0.88%	2148 85.92%	2170
계		148	2352	2500
Error rate=0.0904, Accuracy=0.9096 Sensitivity=0.3818, Specificity=0.8592				

5. 해지가능성에 대한 고객 점수화

본 분석은 개인휴대통신 고객의 해지를 사전에 예측할 수 있는 모형을 구축하기 위하여 전체 자료를 로지스틱 회귀모형을 이용하였다. 여기서 추정되어지는 해지율을 통하여 해지 가능성에 대한 점수화를 시도하였다.

로지스틱 회귀모형에서 변수선택법을 적용한 결과 선택된 설명변수는 고객계정상태, 연령대, 최근 4개월간 사용료, 납입방법, 총 불만건수, 디지털 유무, 가입경력이 선택되었다. <표 3>은 전체 2,500개의 자료를 통하여 얻어진 로지스틱 회귀모형에 의한 오분류 테이블이다.

<표 3> 전체 자료 로짓모형의 오분류 테이블

		예측		계
		일반해지	정상	
실제	일반해지	127 5.08%	203 8.12%	330
	정상	24 0.96%	2146 85.84%	2170
계		151	2349	2500
Error rate=0.0908, Accuracy=0.9092 Sensitivity=0.3848, Specificity=0.8584				

<표 3>에서 민감도가 0.3848로 일반해지를 일반해지로 예측하는 예측력이 상당히 떨어짐을 볼 수가 있다. 이것은 이 모형이 해지 가능성에 대한 점수화에 적합한 모형이 아니라는 점을 시사하는 것이다. 따라서 의사결정나무분석에서 해지율이 13.2%보다 높은 가지로, 고객계정상태가 '최종청구/정상해지'이거나 고객계정상태가 '정상사용중'이면서 사용료가 '22만8천원이상'인 고객DB만을 따로 추출한 자료를 이용하여 로지스틱 회귀모형을 구축하였다. 로지스틱 회귀모형에서 선택되어진 변수로는 연령대, 디지털유무, 가입경력, 총 불만건수, 최근 4개월간 사용료, 성별이 선택되었다. <표 4>는 로지스틱 회귀모형의 오분류 테이블이다.

<표 4> 로지스틱 모형의 오분류 테이블

		예측		계
		일반해지	정상	
실제	일반해지	139 34.75%	43 10.75%	182
	정상	22 5.50%	196 49.00%	218
계		161	239	400
Error rate=0.1625, Accuracy=0.8375 Sensitivity=0.7637, Specificity=0.8991				

<표 4>에서 민감도가 0.7637로 일반해지를 일반해지로 예측하는 예측력이 <표 3>보다는 높으므로 로지스틱 회귀모형에서 추정된 확률값을 이용하여 해지 가능성에 대한 점수화를 실시하였다. <표 5>는 개인휴대통신 가입고객의 해지 가능성에 대한 점수표 중 일부이다.

6. 토의

지금까지 개인휴대통신 고객의 해지특성이 어떤 가입자 속성변인에 의존하는지에 대한 고객 해지패턴을 분석하였고 해지점수를 구하여 고객의 해지 유무를 알아보았다. 이러한 해지점수를 이용하여 이탈확신 고객, 이탈가능 고객, 이탈잠재 고객, 유지가능 고객, 유지확신 고객과 같이 고객을 그룹화하여 목표 마케팅(target marketing) 전략을 세울 수 있다.

이처럼 CHAID 기법과 같은 의사결정나무는

<표 5> 해지 가능성 점수

아날로그/디지탈	성별	총 불만 건수	연령대	가입경력	사용료	해지 유무	해지 유무 예측	해지 점수
아날로그	여자	0	40대	23~26개월	1만5천원 미만	일반 해지	일반 해지	76.01
아날로그	남자	1	20대	14~18개월	5만5천~ 7만1천원	일반 해지	일반 해지	51.89
아날로그	남자	0	10대	6개월미만	5만5천~ 7만1천원	일반 해지	일반 해지	82.55
아날로그	남자	1	30대	19~22개월	11만8천~ 15만7천원	일반 해지	일반 해지	74.01
아날로그	남자	0	30대	6개월미만	22만원8천 원이상	일반 해지	정상	45.91
아날로그	남자	0	30대	19~22개월	22만원8천 원이상	정상	정상	32.74
아날로그	남자	1	20대	6~10개월	22만원8천 원이상	정상	일반 해지	51.99
디지털	여자	0	20대	6~10개월	5만5천~ 7만1천원	일반 해지	일반 해지	85.65
디지털	남자	0	30대	14~18개월	22만원8천 원이상	일반 해지	정상	5.65
디지털	남자	0	20대	11~12개월	22만원8천 원이상	정상	정상	13.33
디지털	여자	1	30대	47개월이상	22만원8천 원이상	정상	정상	1.77

특정한 분포의 가정없이 분석하는 비모수적인 방법으로 분석결과의 나무구조그림만으로 모형을 쉽게 이해 할 수 있다는 장점이 있으며 추가적인 분석방법에 유용한 방법이다. 최근에는 SAS/E-MINER[1]뿐만 아니라 SPSS AnswerTree[2], CART[3] 등 상용화된 데이터마이닝 솔루션에서 의사결정나무를 손쉽게 사용할 수 있다.

참고문헌

(1) 강현철, 서두성, 최종후 (1998), Enterprise Minier의 의사결정나무분석 알고리즘, *SUGI-K '98*.

(2) 박찬욱 (1996), *데이터베이스 마케팅*, 연암사.

(3) 허명회 (1995), *SAS 범주형 데이터 분석*, 자유아카데미.

(4) Berry, M. J. A. and Linoff, G. S. (1997), *Data Mining Techniques*, New York: John Wiley & Sons, Inc.

(5) Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone. (1984). *Classification and regression trees*. Belmont: Wadsworth.

(6) Kass, G. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*. 29:2, 119-129.

(7) Quinlan, J. R. (1993). *C4.5 Programs for machine learning*. San Mateo: Morgan Kaufmann.

1) http://www.sas.com/software/data_mining/
2) http://www.spss.com/datamine/
3) http://www.salford-systems.com/