통계청「통계분석연구」제4권 제1호(99.봄) 61-83

데이터마이닝 의사결정나무의 응용

최종후* · 서두성**

본 논문의 목적은 최근 국내에서 활발하게 논의되고 있는 데이터마이닝의 주요한 도구인 의사결정나무를 정리, 소개하는 데에 있다. 본 논문에서는 1997에 실시된 체 15대 대통령선거예측조사 자료를 이용한 무 응답의 분류 및 예측문제와 개인휴대통신의 해지자 분석에 이를 적용한 결과를 보인다.

끝으로 효율적 통계조사를 위한 전략수립에 의사결정나무 활용 가능성을 검토한다.

< 차 례 >

* 고려대학교 정보통계학과 부교수, jchoi@tiger.korea.ac.kr

** 고려대학교 정보통계학과 석사과정

1. 서론

1.1 의사결정나무의 소개

의사결정나무는 의사결정규칙(decision rule)을 도표화하여 관심대상이 되는 집단을 몇 개의 소집단으로 분류(classification)하거나 예측(prediction)을 수행하는 분석방법이다. 분석과정이 나무구조에 의해서 표현되기 때문에 판별 분석(Discriminant Analysis), 회귀분석(Regression Analysis), 신경망(Neural Networks) 등과 같은 방법들에 비해 연구자가 분석과정을 쉽게 이해하고 설명할 수 있다는 장점을 가지고 있다.

의사결정나무는 분류 또는 예측을 목적으로 하는 어떤 경우에도 사용될 수 있으나 분석의 정확도보다는 분석과정의 설명이 필요한 경우에 더 유용하게 사용된다. 의사결정나무 분석이 활용될 수 있는 응용분야는 다음과 같다.(최종후 외:1998)

- 세분화(Segmentation) : 관측개체를 비슷한 특성을 갖는 몇 개의 그룹으로 분할하여 각 그룹별 특성을 발견하고자 하는 경우
- 분류(Classification) : 여러 예측변수(predicated variable)에 근거하여 목표변수(target variable)의 범주를 몇 개의 등급으로 분류하고자 하는 경우
- 예측(Prediction) : 자료로부터 규칙을 찾아내고 이를 이용하여 미래의 사건을 예측하고자 하는 경우
- 차원축소 및 변수선택(Data reduction and variable screening) : 매우 많은 수의 예측변수 중에서 목표변수에 큰 영향을 미치는 변수들을 골라내고자 하는 경우
- 교호작용효과의 파악(Interaction effect identification) : 여러 개의 예측변수들이 결합하여 목표변수에 작용하는 교호작용을 파악하고자 하

는 경우

- 범주의 병합 또는 연속형 변수의 이산화(Category merging and discretizing continuous variable) : 범주형 목표변수의 범주를 소수의 몇 개로 병합하거나, 연속형 목표변수를 몇 개의 등급으로 범주화 하고자 하는 경우

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 규칙을 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.

이상과 같은 과정에서 정지기준, 분리기준, 평가기준 등을 어떻게 지정하느냐에 따라서 서로 다른 의사결정나무가 형성된다.

1.2 의사결정나무의 알고리즘

1.2.1 CHAID 알고리즘

CHAID(Chi-squared Automatic Interaction Detection : Kass(1980))는 카이제곱 검정(범주형 목표변수) 또는 F-검정(연속형 목표변수)을 이용하여 다지분리(multiway split)를 수행하는 알고리즘이다.

CHAID 알고리즘은 목표변수가 범주형일 때, Pearson의 카이제곱 통계량 또는 우도비 카이제곱 통계량(likelihood ratio Chi-square statistic)을 분리기준으로 사용한다. 여기서 목표변수가 순서형 또는 사전그룹화된 연속형인 경우에는 우도비 카이제곱 통계량이 사용된다.

카이제곱 통계량은 관측도수(f_{ij})로 이루어진 $r \times c$ 분할표로부터 계산된다. 분할표의 구조는 <표 1.1>과 같다.

<표 1.1> 분할표의 구조

목표변수 / 설명변수	범주 1	범주 2	…	범주 c	합 계
범주 1	f_{11}	f_{12}	…	f_{1c}	$f_{1.}$
범주 2	f_{21}	f_{22}	…	f_{2c}	$f_{2.}$
…	…	…	…	…	…
범주 r	f_{r1}	f_{r2}	…	f_{rc}	$f_{r.}$
합 계	$f_{.1}$	$f_{.2}$	…	$f_{.c}$	$f_{..}$

<표 1.1>의 분할표로부터, Person의 카이제곱 통계량은

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

과 같이 정의되고, 우도비 카이제곱 통계량은

$$x^2 = 2\sum_{i,j} f_{i,j} \times \log_e(\frac{f_{ij}}{e_{ij}})$$

으로 정의된다. 이 때 두 통계량의 자유도(degree of freedom)는 (r-1)(c-1)로서 동일하다. 여기서 e_{ij}는 분포의 동일성 또는 독립성의 가설 하에서 계산된 기대도수(expected frequency)를 말하며, 아래에 주어진 식 $e_{ij} = \frac{f_i \times f_i}{f..}$ 과 같이 계산된다.

카이제곱 통계량이 자유도에 비해서 매우 작다는 것은 예측변수의 각 범주에 따른 목표변수의 분포가 서로 동일하다는 것을 의미한다. 따라서 예측변수가 목표변수의 분류에 영향을 주지 않는다고 결론지을 수 있다. 자유도에 대한 카이제곱 통계량 값의 크고 작음은 P-값으로 표현될 수 있는데, 카이제곱 통계량 값이 자유도에 비해서 작으면 P-값은 커지게 된다. 결국 분리기준을 카이제곱 통계량 값으로 한다는 것은 P-값이 가장 작은 예측변수와 그 때의 최적분리에 의해서 자식마디를 형성시킨다는 것을 의미한다.

1.2.2. CART 알고리즘

CART(Classification and Regression Trees, Breiman et al.(1984))는 지니 지수(범주형 목표변수인 경우 적용) 또는 분산의 감소량(연속형 목표변수인 경우 적용)을 이용하여 이지분리(,binary split)를 수행하는 알고리즘이다(Quinlan, 1993).

지니 지수(Gini Index)는 불순도(impurity)를 측정하는 하나의 지수이다. 임의의 한 개체가 목표변수의 i번째 범주로부터 추출되었고, 그 개체를

목표변수의 j번째 범주에 속한다고 오분류(misclassification)할 확률은 $P(i)P(j)$가 된다. 여기에서 $P(i)$는 각 마디에서 한 개체가 목표변수의 I번째 범주에 속할 확률이다. 이러한 오분류 확률은 모두 더하여

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

를 얻을 수 있고, 이는 위와 같은 분류규칙 하에서 오분류 확률의 추정치가 된다. 여기서 c는 목표변수의 범주의 수를 말한다.

일반적으로 CART는 범주형 목표변수에 대해서는 지니 지수를 분리기준으로 사용한다. 지니 지수는 각 마디에서의 불순도 또는 다양도(diversity)를 재는 측도 중의 하나로써

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

와 같이 표현될 수 있다. 여기에서 n은 그 마디에 포함되어 있는 관찰치 수를 말하고, n_i는 목표변수의 i번째 범주에 속하는 관찰치 수를 말한다. 지니 지수는 n개의 원소 중에서 임의로 2개를 추출하였을 때, 추출된 2개가 서로 다른 그룹에 속해있을 확률을 의미하며 Simpson의 다양도 지수(diversity index)로도 알려져 있다. 목표변수의 범주가 2개인 경우에는 지니 지수는 다음과 같이 표현될 수 있으며,

$$G \;=\; 2P(1)P(2) \;=\; 2\left(\frac{n_1}{n}\right)\left(\frac{n_2}{n}\right)$$

이는 카이제곱 통계량을 사용하는 것과 같은 결과를 갖는다.

CART 알고리즘은 지니 지수를 가장 감소시켜주는 예측변수와 그 변수

의 최적분리를 자식마디로 선택하는데, 지니 계수의 감소량은 다음과 같이 계산된다.

$$\Delta\ G = G - \frac{n_L}{n} G_L - \frac{n_R}{n} G_R.$$

여기서 n은 부모마디의 관측치 수를 말하고, n_R과 n_L는 각각 자식마디의 관측치 수를 의미한다. 즉, 자식마디로 분리되었을 때의 불순도가 가장 작도록 자식마디를 형성하는 것이다. 이는 다음과 같은 자식마디에서의 불순도 가중합을 최소화하는 것과 동일하다.

$$P(L)\ G_L\ +\ P(R)\ G_R = \frac{n_L}{n} G_L\ +\ \frac{n_R}{n} G_R.$$

2. 선거예측조사 무응답의 분류 및 예측

2.1 개요

선거예측조사에서 흔히 발생하는 문제로서 지지후보에 대한 유권자의 무응답현상을 들 수 있다. 선거에 임박한 예측조사에서 무응답이 다수 발생하는 경우 이러한 무응답층에 대한 분석은 선거예측조사의 성패의 관건이 된다.

2장에서는 CHAID 알고리즘을 이용하여 선거자료에서 흔히 발생하는 무응답자의 패턴을 분류하고 이들의 지지후보를 예측한다,.

2.1.1 자료설명

다음 자료는 리서치앤리서치社가 1997년 제 15대 대통령 선거를 앞두고

각 후보의 지지율조사를 위해 실시한 전화조사에 의해 얻어졌다[1]. 이 중 '투표유무'라는 항목에 대해 '반드시 투표할 것이다,' '아마 투표할 것이다' 라고 답한 응답자에 대해서만 분석을 시도하였는데 이러한 유효응답의 수는 총 979개이다.

<표 2.1>은 분석에 사용되는 변수에 대한 설명이다.

<표 2.1> 분석에 사용된 변수

변수이름	형 태	변 수 값
거주지역	명목형	서울, 부산, 인천, 대구, 광주, 대전, 울산, 경기, 강원, 충북, 충남, 전북, 전남, 경북, 경남, 제주
나이	명목형	20대이하, 30대, 40대, 50대, 60대이상
성별	명목형	남자, 여자
투표유무	순서형	반드시 투표할 것이다, 아마 투표할 것이다 아마 투표하지 않을 것이다, 전혀 투표할 생각이 없다.
지지후보	명목형	이회창, 김대중, 이인제, 기타 후보, 무응답
지지정당	명목형	한나라당, 국민회의, 국민신당, 자민련
학력	순서형	국졸이하, 중졸, 고졸, 대재이상
직업	명목형	농/임/어업, 자영업,판매/서비스직, 기능/숙련공, 일반작업직, 사무/기술직, 경영/관리직, 전문/자유직, 주부, 학생, 무직, 기타
월소득	순서형	70만원이하, 71~100만원, 101~150만원, 151~200만원, 201~250만원, 251~300만원, 301만원이상
원적지	명목형	서울, 부산, 인천, 대구, 광주, 대전, 울산, 경기, 강원, 충북, 충남, 전북, 전남, 경북, 경남, 제주

2.1.2 분석과정

각 후보의 지지율을 계산하는 과정은 다음과 같다.

단계 1 : 먼저 전체 자료를 변수 '지지후보'에 대해 범주 '무응답'인 관측치들(이하 무응답층)과 그렇지 않은 관측치, 즉 '지지후보' 변수에 대해 응답한 관측치들(이하 응답층)로 나눈다.

단계 2 : 응답층으로부터 나무구조모형을 구축한다.

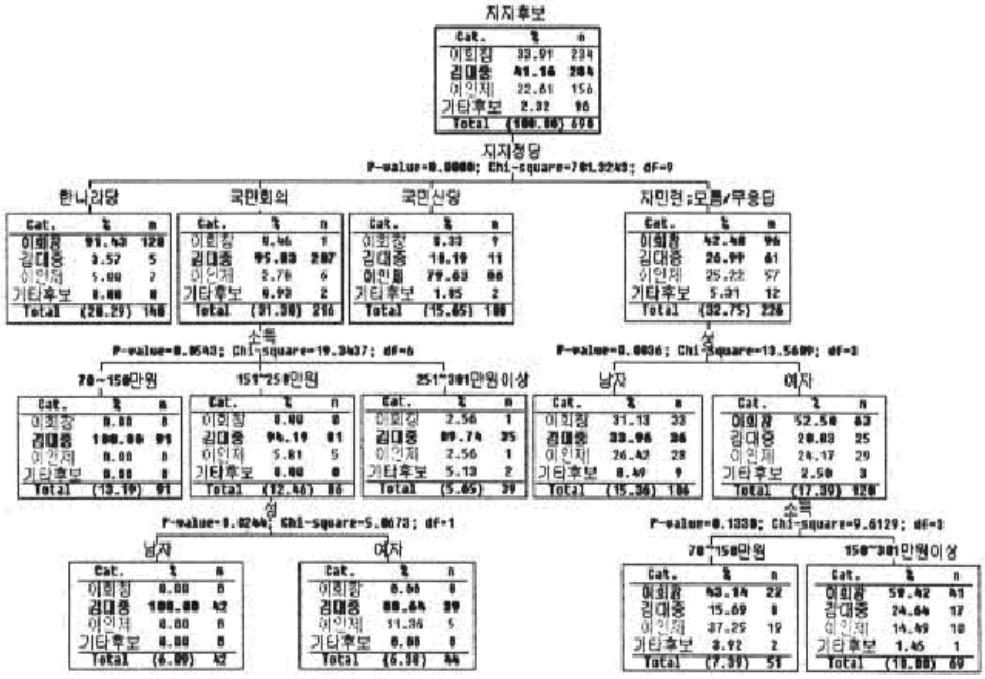
1) 리서치앤리서치社는 이 자료를 연구용으로 공개한 바 있다.

단계 3 : 구축된 모형을 무응답층에 적용하고, 이를 통해 '지지후보'의 범주별 지지율을 계산한다.

단계 4 : 응답층의 실제 지지율과 단계 3에서 얻은 무응답층의 비율을 더하여 전체 지지율을 예측한다.

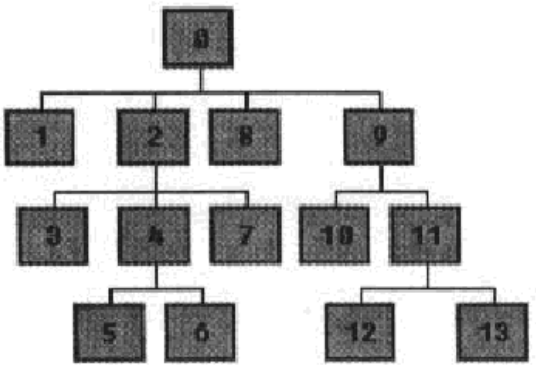
2.2 의사결정나무 결과

<그림2.1>은 의사결정나무 알고리즘을 이용한 다중나무구조(Multi-Tree Structure)의 분류결과이다. 총 9개의 최종마디로 이루어진 나무가 형성되었다. 맨 위에 있는 뿌리마디는 690개의 관측치로, 지지후보에 대한 비율은 각각 33.91%, 41.16%, 22.61%, 2.32%로 나타나고 있음을 볼 수 있다.



<그림 2.1> 지지후보에 대한 의사결정나무 모형

지지후보를 결정하는데 제일 중요한 변수로는 지지정당이며, 다음으로는 소득 및 성으로 구분된다.



<그림 2.2> 의사결정나무 마디의 번호

의사결정나무에서 이익도표(gains chart)는 범주형 목표변수(target variable)의 특정 범주가 각 마디에서 획득한 백분율을 나타낸다. <표 2.2>~<표 2.4>는 각 후보들의 지지성향을 보기 위한 이익도표이다.

<표>에 나타나는 통계량은 다음과 같다.

- Node : 마디의 번호
- Node(n) : 개체의 수
- Node(%) : (개체의 수)/(전체 개수의 수)
- Resp(n) : 목표범주의 개체의 수
- Resp(%) : (목표범주의 개체의 수)/(전체에서 목표범주의 개체의 수)
- Gain(%) : (목표범주의 개체의 수)/(개체의 수)
- Index(%) : (목표범주의 비율)/(전체 목표범주의 비율)

<표 2.2>~<표 2.4>는 각 후보들의 이익지수와 관련된 값들을 정리한 표이다.

<표 2.2> 이회창 후보의 이익도표

Node	Node: n	Node: %	Resp: n	Resp: %	Gain (%)	Index (%)
1	140	20.29	128	54.70	91.42	269.59
13	69	10.00	41	17.52	59.42	175.21
12	51	7.39	22	9.40	43.13	127.19
10	106	15.36	33	14.10	31.13	91.79
8	108	15.65	9	3.85	8.33	24.57
7	39	5.65	1	0.43	2.56	7.56
3	91	13.19	0	0.00	0.00	0.00
6	44	6.38	0	0.00	0.00	0.00
5	42	6.09	0	0.00	0.00	0.00

이회창후보의 경우 Gain이 가장 높은 마디가 마디 1임을 알 수 있다. <그림 2.2>에서 볼 수 있듯이 마디 1은 지지정당이 '한나라당'임을 알 수 있다. 다음으로 높은 Gain을 획득한 마디는 13으로 지지정당이 '자민련'이거나 '모름/무응답'인 범주 중에서 성별이 '여자'이면서 소득이 '150-300만원'임을 알 수 있다. 마디 1의 Index<표 2.3> 김대중 후보의 이익도표는 269.59이므로 이는 전국에서 획득한 지지율인 33.91%보다 마디 1에 해당하는 집단에 대해서 2.69배나 높은 지지율을 얻었다는 것을 보여준다.

<표 2.3> 김대중 후보의 이익도표

Node	Node: n	Node: %	Resp: n	Resp:(%)	Gain (%)	Index (%)
5	42	6.09	42	14.79	100.00	242.96
3	91	13.19	91	32.04	100.00	242.96
7	39	5.65	35	12.32	89.74	218.04
6	44	6.38	39	13.73	88.64	215.35
10	106	15.36	36	12.68	33.96	82.51
13	69	10.00	17	5.99	24.64	59.86
12	51	7.39	8	2.82	15.69	38.11
8	108	15.65	11	3.87	10.19	24.75
1	140	20.29	5	1.76	3.57	8.68

김대중후보의 경우 Gain이 가장 높은 마디가 마디 5와 3임을 알 수 있다. 마디 5는 지지정당이 '국민회의'이면서 소득이 '150-250만원'이면서 성

별이 '남자'임을 알 수 있으며, 마디 3은 지지정당이 '국민회의'이면서 소득이 '70-150만원'임을 알 수 있다. 마디 5와 3의 Index는 242.96으로 이는 전국에서 획득한 지지율인 41.16%보다 마디 5와 3에 해당하는 집단에 대행서 2.42배나 높은 지지율을 얻었다는 것을 보여준다.

<표 2.4> 이인제 후보의 이익도표

Node	Node: n	Node: %	Resp: n	Resp: %	Gain (%)	Index(%)
8	108	15.65	86	55.13	79.63	352.21
12	51	7.39	19	12.18	7.25	164.78
10	106	15.36	28	17.95	26.42	116.84
13	69	10.00	10	6.41	14.49	64.10
6	44	6.38	5	3.21	11.36	50.26
1	140	20.29	7	4.49	5.00	22.12
7	39	5.65	1	0.64	2.56	11.24
3	91	13.19	0	0.00	0.00	0.00
5	42	6.09	0	0.00	0.00	0.00

이인제후보의 경우 Gain이 가장 높은 마디가 마디 8임을 알 수 있다. 마디 8은 지지정당이 '국민신당'임을 알 수 있다. 다음으로 높은 Gain을 획득한 마디는 12로 지지정당이 '자민련'이거나 '모름/무응답'인 범주 중에서 성별이 '여자'이면서 소득이 '70-150만원'임을 알 수 있다. 마디 8의 Index는 352.21이므로 이는 전국에서 획득한 지지율인 22.61%보다 마디 8에 해당하는 집단에 대해서 3.52배나 높은 지지율을 얻었다는 것을 보여준다.

<표 2.5>는 의사결정나무 모형의 오분류 테이블이다.

<표 2.5> 오분류 테이블

		실제결과				
		이회창	김대중	이인제	기타후보	total
예측결과	이 회 창	191	30	36	3	260
	김 대 중	34	243	34	11	322
	이 인 제	9	11	86	2	108
	기타후보	0	0	0	0	0
	total	234	284	156	16	690
Risk Estimate		0.246377				
SE of Risk Estimate		0.016404				

전체적인 오분류율은 약 24.6%정도이며, 이에 대한 표준오차는 0.016이다.

2.3 선거예측 결과

이제까지 응답층에 대한 지지후보의 의사결정나무 모형을 구축하였다. 이렇게 구축된 나무모형 결과를 무응답층(관찰치 289개)에 적용하여 얻은 예측 빈도가 <표 2.6>이다

<표 2.6> 무응답층의 예측빈도

	이회창	김대중	이인제	기타후보	전 체
무응답층	147	123	19	0	289
예측빈도	(50.9)	(42.6)	(6.6)	(0)	(100)

<표 2.7>은 응답층의 실제빈도와 무응답층의 예측빈도를 더해서 지지율의 추정치를 얻은 표이다.

<표 2.7> 지지후보에 대한 전체 추정치

	이회창	김대중	이인제	기타후보	전 체
응답층의 실제빈도	234 (33.9)	284 (41.2)	156 (22.6)	16 (2.3)	690 (100)
무응답층 예측빈도	147 (50.9)	123 (42.6)	19 (6.6)	0 (0)	289 (100)
전체 추정치	381 (38.92)	407 (41.57)	175 (17.87)	16 (1.63)	979 (100)
실제 결과	(38.7)	(40.3)	(19.2)	(1.8)	(100)

지금까지 의사결정나무 알고리즘을 이용하여 응답층의 나무구조를 해석하고 응답층에 대한 나무구조를 이용하여 무응답층의 판별과 분류를 실시하였다.

선거 무응답층의 지지후보 예측에 관한 기존의 연구는 주로 판별분석에 의존해 왔다(박무익, 1998)[2].그런데 이 경우 판별변수가 되는 인구속성변수들은 주로 범주형 변수이기 때문에 판별분석에서 요구되는 가정(assumption)충족에서 문제가 발생한다.(예컨대 정규상의 가정)

전술한 이익도표는 각 후보의 지지패턴 분석이나 선거운동 전략에 유용하게 이용되리라 생각된다.

3. 개인휴대통신 해지자 분석

3.1 개요

3장에서는 개인휴대통신 고객의 해지특성이 어떤 가입자 속성변인에 의존하는지에 대한 해지패턴을 분석하고 해지 가능성에 대한 점수화

2) 한국갤럽은 1997년 실시된 15대 대통령선거의 선거예측조사에서 무응답층의 분석을 위하여 판별분석을 적용한 바 있는데 이때 고려했던 판별변수는 성, 연령, 교육수준, 원적이었다.

(scoring)를 시도한다. 고객의 해지패턴을 알아보기 위하여 의사결정나무(decision tree)분석을 이용하였으며, 해지 가능성에 대한 점수화는 로지스틱 회귀모형(Logistic Regression Model)을 이용한다.

고객 DB(Data Base)를 이용한 고객 세분화(segment)로 이동통신 가입고객의 해지특성이 어떠한 패턴을 이루고 있는지를 알아보기 위하여 의사결정나무 분석을 실시하였다. 이러한 분석은 고객 해지율(defection rate)을 감소시키는 고객유지 마케팅(retention marking)의 일환으로 이용될 수 있다.

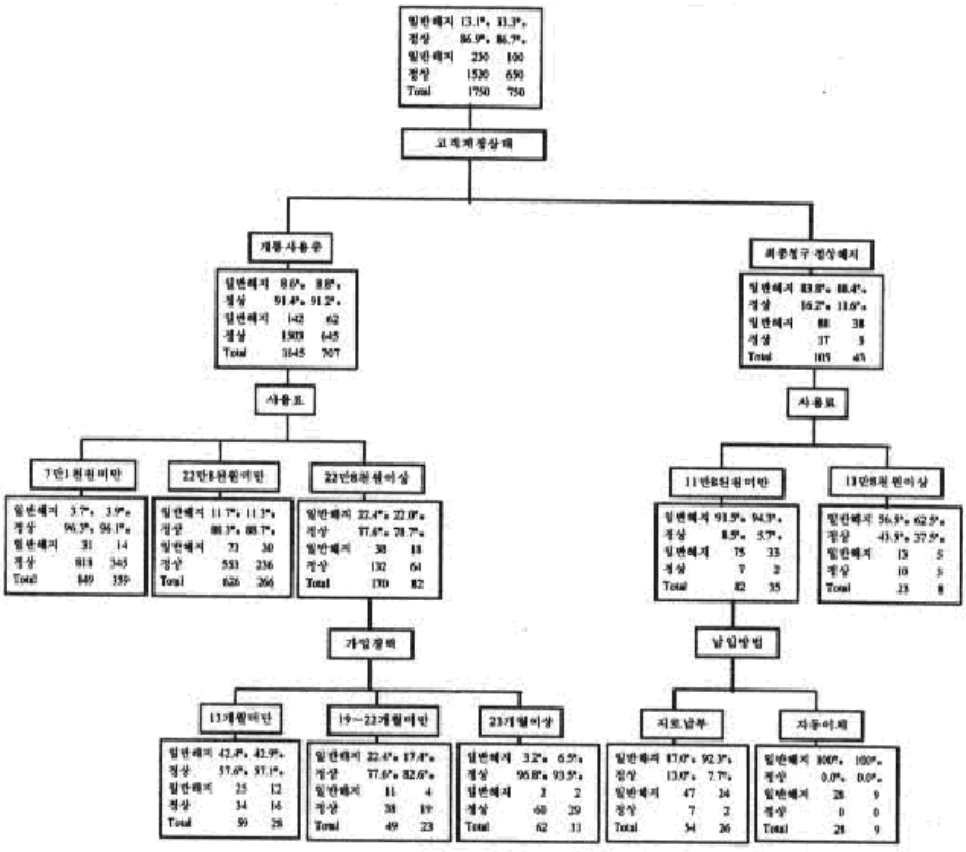
분석에 사용된 자료는 이동통신회사의 서울지역 고객DB를 이용하여 랜덤추출로 2,500개의표본을 획득한 것이다. 분석표본의 해지율은 13.2%이다. 목표변수로는 해지여부이며 기타 고객속성 변수가 설명변수이다. 변수의 내용은 <표 3.1>과 같다.

<표 3.1> 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~5만원천원미만/5만5천원~7만1천원미만/7만1천원~9만2천원미만/9만2천원~11만8천원미만/11만8천원~15만7천원미만/15만7천원~2 2만8천원미만/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
연령	10대/20대/30대/40대/50대/60대/70대이상

3.2 의사결정나무 결과

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



<그림 3.1> 해지유무에 대한 의사결정나무[3)]

<그림 3.1>은 의사결정나무 모형의 다중 나무구조의 분류결과이다. 총 8개의 최종마디로 이루어진 나무구조가 형성되었다. 맨 위에 있는 뿌리마

3) 각 마디의 분석 결과에서 왼쪽은 분석용 자료에 대한 값이고, 오른쪽은 타당성 평가용 자료에 대한 값이다.

디(root node)에서 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 납입방법으로 구분된다. 이중 가입고객의 고객계정상태가'최종청구/정상해지'인 경우에 해지율이 83.8%(분석용),88.4%(타당성 평가용)로 높아짐을 볼 수 있으며, 다음으로 가입고객이 고객계정상태가 '개통사용중'이면서 최근 4개월간 사용료가 '22만 8천원이상'인 경우 해지율이 22.4%, 22.0%로 높아짐을 볼 수 있다. 특히, 가입고객의 고객계정상태가 '개통사용중'이면서 최근 4개월간 사용료가 '22만 8천원이상'이면서 가입경력이 '13개월미만'의 경우 해지율이 42.4%(분석용), 42.9%(타당성 평가용)로 높아짐을 볼 수 있다.

<표 3.2> 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=0.9899				

<표 3.2>는 의사결정나무 분석의 오분류 테이블이다. 오분류율(error rate)과 정확도(accurany)가 각각 0.0904, 0.9096으로 잘 분류되어진 것 같으나 민감도(sensitivity)[4]가 0.3818로 떨어짐을 볼 수가 있다.

4) 민감도는 관심을 둔 사건을 제대로 예측할 확률이다. 이 경우에는 해지가 관심있는 사건이므로 일반해지를 일반해지로 예측한 확률이다. 여기서는 126/330.

3.3 고객 점수화

개인휴대통신 고개의 해지가능성 점수를 사전에 예측할 수 있는 모형을 구축하기 위하여 로지스틱 회귀모형을 이용한다.

로지스틱 회귀분석은 목표분석가 명목척도로 측정되어있는 경우에 목표변수와 설명변수 간의 관계를 분석하기 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.

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

$$\log \frac{P(Y=1 \mid x_1, \cdots, x_p)}{P(Y=2 \mid X_1, \cdots, x_p)} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

또는

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

<그림 3.1>의 의사결정나무에서 해지율이 13.2% 보다 높은 마디에 해당하는 가지로, 고객계정상태가 '최종청구/정상해지'이거나 고객계정상태가 '개통사용중'이면서 사용료가 '22만 8천원이상'인 자료(관찰치 400개)만을 이용하여 로지스틱 회귀모형을 구축한다. 단계적 로지스틱 회귀모형(Stepwise Logistic Regression Model)로 선택되어진 변수로는 연령, 디지털유무, 가입경력, 총 불만건수, 최근 4개월간 사용료, 성별이 선택되었다. <표 3.3>은 로지스틱 회귀모형에 의한 오분류 테이블이다.

<표 3.3> 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 Sensitvity=0.7637, Specificity=0.8991				

<표 3.3>에서 민감도가 0.7637로 일반해지를 일반해지로 예측하는 예측력이 높으므로 로지스틱 회귀모형에서 추정된 확률값을 이용하여 해지 가능성에 대한 점수화[5]를 실시한다. <표 3.4>는 개인휴대통신 가입고객의 해지 가능성에 대한 점수표 중 일부이다. 해지유무예측은 해지점수가 50점 이상인 경우를 일반해지로 예측한 경우이다[6].

지금까지 개인휴대통신 고객이 해지특성이 어떤 가입자 속성변인에 의존하는지에 대한 고객 해지패턴을 분석하였고 해지점수를 구하여 고객의 해지유무를 알아보았다. 이러한 해지점수를 이용하여 해지확신 고객, 해지가능 고객, 해지잠재 고객, 유지가능 고객, 유지확신 고객과 같이 고객을 그룹화하여 목표 마케팅(target marketing) 전략을 세울 수 있다.

<표 3.4> 해지가능성 점수(일부)

5) 해지가능성점수 = P(Y=해지)×100

6) 해지점수가 50점 이상인 경우를 일반해지로 예측한 이유는 일반해지고객과 정상고객의 해지가능성 점수의 분포를 그려보면 50점 근처에서 교차가 일어나기 때문이다.

아날로그/디지탈	성별	총 불만 건수	연령대	가입경력	사용료	해지 유무	해지 유무 예측	해지 점수
아날로그	여자	0	40대	23~26개월	1만5천원미만	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~7만1천	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	22만원8천원이상	일반 해지	정상	45.91
아날로그	남자	0	30대	19~22개월	22만원8천원이상	정상	정상	32.74
아날로그	남자	1	20대	6~10개월	22만원8천원이상	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	22만원8천원이상	일반 해지	정상	5.65
디지털	남자	0	20대	11~12개월	22만원8천원이상	정상	정상	13.33
디지털	여자	1	30대	47개월이상	22만원8천원이상	정상	정상	1.77

4. 토 의

지금까지 의사결정나무를 통하여 2장에서는 선거자료에서 발생하는 무응답자의 패턴을 분석하고 이러한 패턴을 이용하여 무응답자의 지지율을 예측하는 예측모형의 사례를 보였다. 또한 3장에서는 개인휴대통신 고객의 해지패턴을 분석하고 로지스틱을 회귀모형을 통하여 고객의 해지가능성 점수를 구하는 사례를 살펴 보았다.

의사결정나무는 판별분석, 분산분석, 회귀분석 등과 같은 전통적인 통계분석기법의 구현에 앞서 탐색적 절차에 다각도로 유용하게 활용될 수 있다. 또한 통계조사에서 조사의 성·패라는 목표변수를 피조사자의 인구

학적 속성을 통해 분석해 낸다면 그 결과는 효율적 통계조사 잔략수립에 도움을 주게 될 것이다.

현재 의사결정나무 모형은 데이터마이닝(data mining)의 주요기법으로 자리잡고 있으며 SAS/EMINER[7], SPSS AnswerTree[8], CART[9] 등 상용화된 데이터마이닝 솔루션 등에서 이를 사용할 수 있다.

<참 고 문 헌>

7) http://www.sas.com/software/data_mining/
8) http://www.spss.com/datamine/
9) http://www.salford-systems.com/

(1) 최종후, 한상태, 강현철, 김은석 (1998), AnswerTree를 이용한 데이터 마이닝 의삭결정나무분석, 서울 : SPSS아카데미.

(2) 강현철, 서두성, 최종후(1998), Enterprise Minier의 의사결정나무분석 알고리즘, SAS 사용자 컨퍼런스 발표자료집, 서울 : SAS-Korea, pp.169~186.

(3) 박무익 (1998), 한국의 제 15대 대통령선거와 선거예측조사, 한국통계학회 1998년 춘계학술발표회 논문집, pp.1-9.

(4) 허명회 (1995), SAS 범주형데이타분석, 서울 : 자유아카데미.

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

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

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

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

Decision Trees and Its Applications

Jonghoo Choi, Doosung Seo

Abstract

In this paper, we introduce and investigate the decision trees. Decision trees are charts that illustrate decision rules. If we have data divided into classes (e.g. subscribers or nonsubscribers, voters versus nonvoters), we can use decision trees as a classifier old or new cases with maximum accuracy. We explore the applications of decision trees based on two real examples.