의사결정나무를 이용한 15대 대선 선거예측조사 사례분석

최 종 후* · 서 두 성** · 김 유 진***

Ⅰ. 분석의 목적과 자료

본 사례분석의 목적은 선거예측조사에서 흔히 발생하는 무응답층(의사결정유보층)에 대한 지지율을 예측함으로써 전체집단에서의 지지율을 추정(예측)하고자 하는 것이다.

분석에 사용된 자료는 (주)리서치앤리서치의 1997년 제15대 대통령 선거에서 각 후보의 지지율의 파악을 위해 실시된 전화조사에 의해 얻어진 것이다. 조사항목 중에서 '투표유무' 항목에 대해 '반드시 투표할 것이다', '아마 투표할 것이다'라고 응답한 응답자에 대해서만 분석을 시도하였는데, 자료의 수는 979개이다. <표 1>이 조사항목이다

<표 1> 조 사 항 목

변수이름	형 태	변 수 값
거주지역	명목형	서울, 부산, 인천, 대구, 광주, 대전, 울산, 경기, 강원, 충북, 충남, 전북, 전남, 경북, 경남, 제주
나 이	명목형	20대이하, 30대, 40대, 50대, 60대이상
성 별	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
학 력	순서형	국졸이하, 중졸, 고졸, 대재이상

*(339-700) 충남 연기군 조치원읍 고려대학교 정보통계학과 부교수, jchoi@tiger.korea.ac.kr
**(137-751) 서울시 서초구 서초2동 1337-20 아태빌딩 한국신용정보(주), dsseo@nice.co.kr
***(339-700) 충남 연기군 조치원읍 고려대학교 대학원 정보통계학과, sasw@naver.com

변수이름	형 태	변 수 값
직 업	명목형	농/임/어업, 자영업, 판매/서비스직, 기능/숙련공, 일반작업직, 사무/기술직, 경영/관리직, 전문/자유직, 주부, 학생, 무직, 기타
소 득	순서형	70만원이하, 71~100만원, 101~150만원, 151~200만원, 201~250만원, 251~300만원, 301만원이상
원 적 지	명목형	서울, 부산, 인천, 대구, 광주, 대전, 울산, 경기, 강원, 충북, 충남, 전북, 전남, 경북, 경남, 제주

Ⅱ. 분석과정

각 후보의 지지율을 예측하기 위해서는 우선 무응답층에 대한 예측이 선행되어야 하는데 이를 위한 분석과정은 다음과 같다.

단계 1 : 먼저 전체 자료를 '지지후보' 변수에 대해 '무응답' 범주인 관측치들(이하 무응답층)과 그것이 아닌 관측치, 즉 '지지후보' 변수에 대해 응답한 관측치들(이하 응답층)로 나눈다.

단계 2 : 응답층으로부터 의사결정나무구조모형을 구축한다.

단계 3 : 구축된 모형을 무응답층에 적용하고, 이를 통한 '지지후보' 변수의 각 범주에 대한 비율을 계산한다.

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

본 자료의 응답층의 관찰치 개수는 690개, 무응답층의 관찰치 개수는 289이다.

Ⅲ. 의사결정나무

의사결정나무(decision tree)는 관심대상이 되는 집단을 몇 개의 소집단으로 분류하거나 예측을 수행하기 위해서 사용되는 분석기법 중의 하나이다. 특히, 의사결정규칙(decision rule)이 나무구조로 표현되기 때문에 분류 또는 예측을 수행하는 다른 방법들에 비해서 연구자가 분석과정을 쉽게 이해하고 설명할 수 있다는 장점을 가지고 있다(최종후 외, 1998). 의사결정나무분석은 예측모형 자체로 사용될 뿐만 아니라 이상치를

검색하거나 분석에 필요한 변수 또는 교호효과를 찾아내는데 많이 이용되고 있다.

의사결정나무 분석이 유용하게 활용되는 응용분야는 다음과 같다.

- 세분화(Segmentation) : 데이터를 비슷한 특성을 갖는 몇 개의 그룹으로 분할하여 각 그룹별 특성을 발견하는 경우.
- 분류(Classification) : 관측치(observation)를 여러 예측변수들에 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 하는 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 이산화 하고자 하는 경우.

일반적으로 의사결정나무 분석은 다음과 같은 단계를 거친다(Berry and Linoff, 1997; 최종후 외, 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.

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

Ⅳ. CHAID 알고리즘

의사결정나무분석을 수행하기 위해 CHAID(Kass, 1980), CART(Breiman et. al., (1984) ; Quinlan(1993)), QUEST(Loh & Shih, 1997) 등의 알고리즘들이 제안되어 있으며, 지금도 많은 연구자들에 의해서 다양하게 개선된 알고리즘들이 개발 및 제안되고 있다.

CHAID(Chi-squared Automatic Interaction Detection; Kass, 1980)는 카이-제곱검정(이산형 목표변수), 또는 F-검정(연속형 목표변수)을 이용하여 다지분리(multiway split)를 수행하는 알고리즘이다.

CHAID 알고리즘은 목표변수가 이산형일 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.

Pearson의 카이제곱 통계량은

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

과 같이 정의되고, 우도비 카이제곱 통계량은

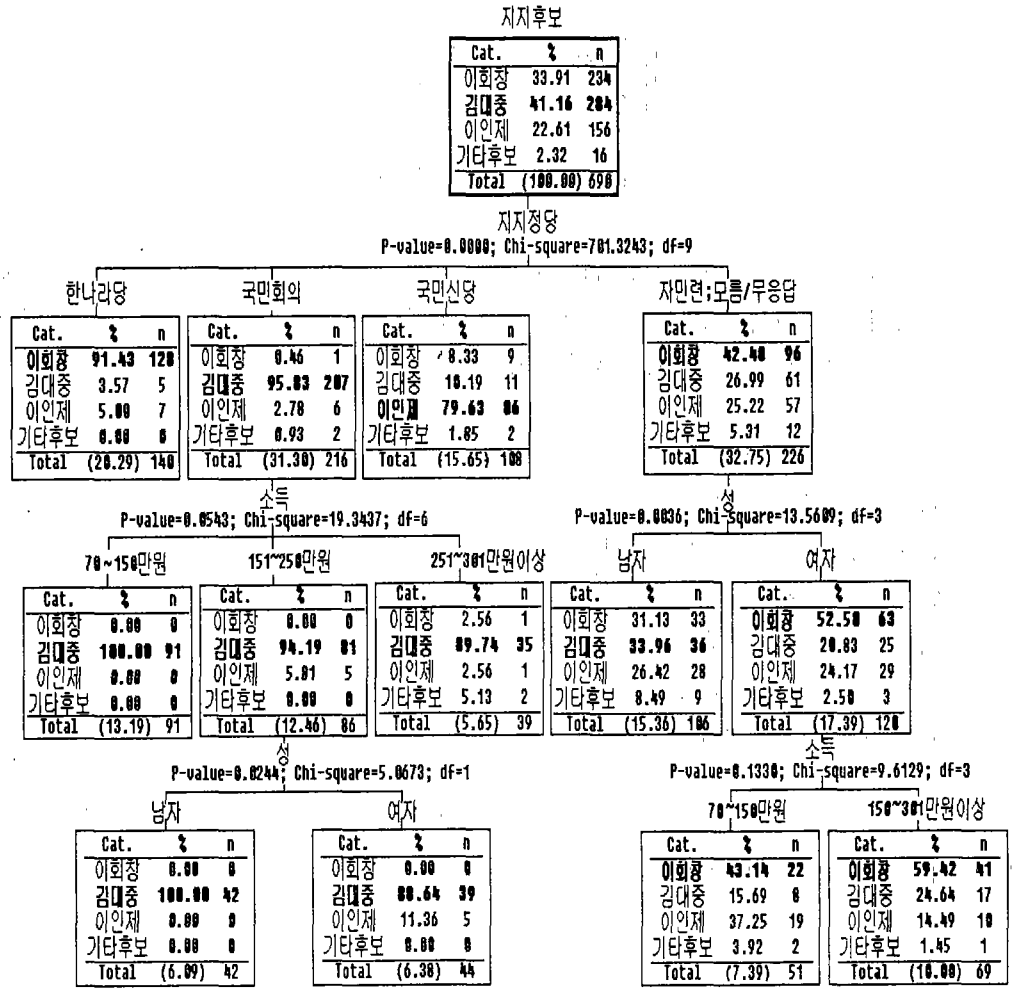
$$\chi^2 = 2 \sum_{i,j} f_{ij} \times \log_e \left(\frac{f_{ij}}{e_{ij}} \right)$$

로 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)이다. f_{ij} 는 관측도수이다.

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.

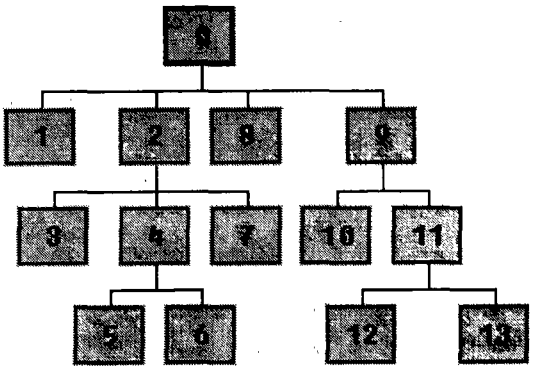
V. 의사결정나무의 평가

<그림1>은 위의 자료에 대한 CHAID 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 1> 지지후보에 대한 의사결정나무 모형 평가

분석결과 지지후보를 결정하는데 제일 중요한 변수로는 지지정당, 다음으로는 소득 및 성별로 나타났다.



<그림 2> 의사결정나무 마디

의사결정나무에서 이익도표(gains chart)는 이산형 목표변수(target variable)의 특정 범주가 각 마디에서 획득한 백분율을 나타낸다. <표3>~<표5>는 각 후보들의 지지성향을 보기 위한 이익도표이다. 표에 나타나는 통계량은 다음과 같다.

- Node : 마디의 번호
- Node(n) : 개체의 수
- Node(%) : (개체의 수)/(전체 개체의 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)

<표2>~<표4>는 각 후보들의 이익지수와 관련된 값들을 정리한 표이다. 해당 마디에서의 지지율은 이익값(Gains)이 된다.

<표 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 Gains가 가장 높은 마디가 마디 1임을 알 수 있다. <그림1>에서 볼 수 있듯이 마디 1은 지지정당이 '한나라당'임을 알 수 있다. 다음으로 높은 Gains를 획득한 마디는 13으로 지지정당이 '자민련'이거나 '모름/무응답'인 범주 중에서 성별이 '여자'이면서 소득이 '150-300만원'임을 알 수 있다. 마디 1의 Index는 269.59이므로 이는 전국에서 획득한 지지율인 33.91%보다 마디 1에 해당하는 집단에 대해서 2.69배나 높은 지지율을 얻었다는 것을 보여준다.

<표 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 Gains가 가장 높은 마디가 마디 5와 3임을 알 수 있다. 마디 5는 지지정당이 '국민회의'이면서 소득이 '150-250만원'이면서 성별이 '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 대해서 2.42배나 높은 지지율을 얻었다는 것을 보여준다.

<표 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	37.25	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	11.34
3	91	13.19	0	0.00	0.00	0.00
5	42	6.09	0	0.00	0.00	0.00

이인재후보의 경우 Gains가 가장 높은 마디가 마디 8임을 알 수 있다. 마디 8은 지지정당이 ' 국민신당'임을 알 수 있다. 다음으로 높은 Gains를 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 5>는 의사결정나무 모형의 오분류 테이블이다.

<표 5> 오분류 테이블

		실제결과				
		이회창	김대중	이인제	기타후보	total
예측결과	이회창	191	30	36	3	260
	김대중	34	243	34	11	322
	이인제	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 0.016이었다.

Ⅵ. 추정 결과

이제까지 응답층에 대한 지지후보의 의사결정나무 모형을 구축하였다. 이렇게 구축된 의사결정나무모형 결과를 무응답층(관찰치 289개)에 적용하여 예측한 빈도가 <표 6>이다.

<표 6> 무응답층의 예측빈도

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

<표7>은 응답층의 실제빈도와 무응답층의 예측빈도를 통합하여 지지율의 추정치를 얻은 표이다.

<표 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)

Ⅶ. 토 의

현재 의사결정나무 모형은 SAS/EMINER[1], SPSS AnswerTree[2], CART[3] 등 상용화된 데이터마이닝 솔루션에서 손쉽게 사용할 수 있다.

본 연구에서 적용한 CHAID 알고리즘 이외에 CART(Breiman et. al., (1984); Quinlan(1993)), QUEST(Loh(1997)) 를 적용하여 분석한 결과는 <표 8>과 같다. CHAID와 CART는 대동소이한 결과를 보이며, QUEST는 다소 상이한 결과를 보여주고 있다. 이는 의사결정나무모형의 적용에 있어 여러 가지 알고리즘을 동시에 비교 검토하고 이에 대한 종합적인 결론을 얻는 것이 필요하다는 점을 시사하고 있다.

국내에서 선거예측조사의 무응답층 분류에 의사결정나무모형을 적용한 사례는 최종후 외(1998), 이태림 · 박무익(1999) 등이 있다. 이태림 · 박무익(1999)의 경우 한국갤럽에서 행한 14대 대선 선거예측조사 자료를 토대로 하여 판별분석(Discriminant Analysis), 의사결정나무모형의 CART, RBF(Radial Basis Function) 신경망모형(Neural Network Model)에 의한 예측결과를 비교하였다. 이들에 의한 분석결과에 의하면 의사결정나무모형의 예측의 정확도가 가장 높게 나타났다.

<표 8> 지지후보에 대한 전체 추정치

	이회창	김대중	이인제	기타후보	전 체
CHAID	381 (38.92)	407 (41.57)	175 (17.87)	16 (1.63)	979 (100)
CART	378 (38.61)	410 (41.88)	175 (17.87)	16 (1.63)	979 (100)
QUEST	408 (41.67)	384 (39.2)	171 (17.47)	16 (1.63)	979 (100)
실제 결과	(38.7)	(40.3)	(19.2)	(1.8)	(100)

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

<참고문헌>

이태림, 박무익 (1999), "선거예측을 위한 분류모형 연구: 14대 대선을 중심으로", 한국분류학회지, 3: pp.66~80.

최종후, 한상태, 강현철, 김은석, (1998), 《AnswerTree를 이용한 데이터마이닝 의사결정나무분석》, SPSS아카데미.

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

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

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

Loh, W., Shih, Y.(1997). "Split selection methods for classification trees", *Statistica Sinica*, 7, pp.815~840.

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

http://www.sas.com/software/data_mining/

http://www.spss.com/datamine/

http://www.salford-systems.com/