응용통계연구 (2009)
22(5), 923–935

인체측정조사에서 측정곤란부위 예측을 위한 의사결정나무 추천 모형 탐지에 관한 연구

최종후[1] · 김선경[2]

[1]고려대학교 정보통계학과, [2](주)한국스코어링

(2009년 7월 접수, 2009년 8월 채택)

요 약

본 연구는 의사결정나무의 추천 모형 선택을 위한 비교실험에 초점을 두고 있다. 의사결정나무 모형은 구축된 모형에 기반을 두고 미래 관측치에 대한 예측 기능을 수행하게 될 것이므로 구축된 모형이 아무리 정치(精緻)하다고 하더라도 일반화의 성질을 충족시키지 못하면 실제성이 없게 된다. 따라서 본 연구는 교차타당성 검토를 통해 일반화의 성질을 충족시키면서 우수한 예측력을 갖는 추천 모형을 탐지하는 절차를 연구하는 데에 초점을 맞추고 있다. 사례 연구로 인체측정자료를 사용하여 측정곤란부위 예측을 위한 의사결정나무 추천 모형을 탐지한다. 그 결과 CART 모형이 추천 모형으로 탐지되었다.

주요용어: 의사결정나무, k-fold 교차타당법, CHAID, exhaustive CHAID, CART.

1. 서론

본 연구에서 추천 모형이라고 함은 경쟁 모형 중에서 일반화(Generalization)의 성질을 최대로 충족할 것으로 기대되는 모형을 의미한다. 주지하는 바대로 의사결정나무 모형은 구축된 모형에 기반을 두고 미래 관측치에 대한 예측 기능을 수행하게 된다. 따라서 구축된 모형이 아무리 정치하다 하더라도 일반화의 성질을 충족하지 못하면 실제성이 없게 된다. 따라서 본 연구는 교차타당성 검토를 통해 일반화의 성질을 충족시키면서 우수한 예측력을 갖는 추천 모형을 탐지하는 절차를 연구하는 데에 초점을 맞추고 있다.

본 연구에서는 의사결정나무를 형성하기 위하여 기존에 제안되어 있는 세 가지 알고리즘, CHAID(CHi-squared Automatic Interaction Detection; Kass, 1980), Exhaustive CHAID (Biggs 등, 1991), CART(Classification And Regression Tree; Breiman 등, 1984)에 기반하여 모형을 구축하고, 구축된 모형에 대하여 교차타당성 평가를 실시함으로써 일반화의 성질을 최대로 충족하는 모형을 찾고, 우수한 예측력을 갖는 추천 모형을 탐지한다. 본 연구에서는 사례 연구로 인체측정자료를 사용하여 측정곤란부위 예측을 위한 의사결정나무 추천 모형을 탐지하게 되는데, 그 결과 CART 모형이 추천 모형으로 탐지되었다. 2장에서는 의사결정나무 분리 알고리즘을 요약하였고, 3장에서는 비교실험을 위한 k-fold 교차타당법을 요약하였으며, 4장에서는 사례 연구 그리고 마지막으로 5장에서는 토의 및 결론을 정리하였다.

[1]교신저자: (339- 700) 충남 연기군 조치원읍 서창리 208, 고려대학교 정보통계학과, 교수.
E-mail: jhchoi@korea.ac.kr

2. 분리 알고리즘

의사결정나무는 다양한 알고리즘이 제안되고 있는데, 대표적으로 CHAID, CART, C4.5 (Quinlan, 1993), QUEST 등이 상용화되어 있으며 최근에 들어서는 이와 관련된 많은 연구와 소프트웨어의 개발로 말미암아 알고리즘이 개선되고 서로 결합되어 기능적인 측면에서 구별이 모호해지고 있는 실정이다 (강현철 등, 2006).

2.1. CHAID

CHIAD(CHi-squared Automatic Interaction Detection)는 카이제곱 검정통계량 또는 F 통계량의 p-값을 이용하여 예측변수를 선택하고, 분리 또는 병합(Merge)을 반복하면서 다지분리(Multiway Split)를 수행하는 알고리즘이다. 분리의 각 단계에서 목표변수와 가장 강하게 연관된 예측변수를 선택하여 분리한다.

목표변수의 척도에 따라 범주형인 경우 피어슨 카이제곱 통계량(Pearson Chi-squared Statistic) 또는 우도비 카이제곱 통계량(Likelihood-ratio Chi-squared statistic)의 p-값이 사용되는데 순서형이거나 사전그룹화된 연속형인 경우 우도비 카이제곱 통계량이 사용된다. 연속형 목표변수인 경우 F 통계량의 p-값을 분리기준으로 하여 나무를 형성한다.

2.2. Exhaustive CHAID

CHAID는 예측변수의 남아있는 범주가 통계적으로 유의하면 더 이상 병합을 하지 않기 때문에 최적의 분리를 하지 못하는 경우가 있을 수 있다. Exhaustive CHAID는 이러한 CHAID의 단점을 보완하여 개발된 알고리즘이다. 모든 가능한 조합을 탐색하여 최적분리를 찾기 때문에 각각의 예측변수에 대하여 최적의 분리를 찾을 뿐만 아니라 분리를 위한 최적의 예측변수를 선택한다. 그러나 이러한 이유로 자료의 수나 각 예측변수의 범주 수가 많으면 계산시간이 오래 걸리는 단점이 있다 (Biggs 등, 1991).

2.3. CART

CART(Classification And Regression Trees)는 Gini 지수나 Twoing 지수 또는 분산을 분리기준으로 사용하여 예측변수를 선택하고 이지분리(Binary Split)를 수행하는 알고리즘이다. 부모마디를 분리할 때 자식마디 내 목표변수 범주의 동질성, 즉 순수도를 최대로 하는 기준으로 분리한다 (Breiman 등, 1984).

3. k-fold 교차타당법

k-fold 교차타당법(k-fold cross-validation)은 데이터를 상호배반(Mutually Exclusive)인 k개의 집합(fold)으로 임의 분할한 후, i번째 집합을 제외한 나머지 집단을 '자료 i'라고 했을 때 생성되는 k개 자료 각각에 대한 위험추정치(Risk Estimate) M_i의 평균값을 k-fold 교차타당법에 의한 위험추정치로 사용한다. 흔히 범주형 목표변수의 경우 위험추정치는 k번의 반복으로부터 각각 도출된 오분류율의 평균값을, 연속형 목표변수의 경우 평균제곱오차(MSE)의 평균값을 위험추정치로 사용한다(최종후 등, 2002; Han과 Kamber, 2006). 이때 교차타당성 평가를 적용하지 않았을 때의 결과와 교차타당성 평가의 결과가 별다른 차이를 보이지 않을 때 구축모형이 안정적이며 일반화의 성질을 충족한다고 할 수 있다. 본 연구에서는 k-fold 중 $k = 10$이 우수하다는 선행연구에 따라 10-fold를 사용하였다 (Kohavi, 1995; 최종후 등, 2002).

표 **3.1.** k-fold 교차타당법

전체 자료
1
2
⋮
$k-1$
k

자료 1
1
2
⋮
$k-1$
k

자료 2
1
2
⋮
$k-1$
k

자료 $k-1$
1
2
⋮
$k-1$
k

자료 k
1
2
⋮
$k-1$
k

■ 위험추정치 $= \left[\sum_{i=1}^{k} M_i\right] / k$

M_1 M_2 M_{k-1} M_k

4. 사례 연구

정부(공업진흥청-현 중소기업청)는 1979년, 1986년에 이어 세 번째로 1992년에 '국민인체측정조사'를 수행한 바 있으며, 1997년에 네 번째 조사를 실시한 바 있다. 이후 4~5년 간격으로 정례적으로 조사가 실시되고 있다. 이 조사는 제반 산업설계의 기반이 되는 인체측정 자료를 산업체에 보급함으로써 의류, 가구류, 신발, 설비 등 산업제품이 보다 국민생활에 편리하게 만들어지게 하고자 하는데 그 목적이 있다 (박경수, 1993; 공업진흥청, 1992). 즉 인간공학(Ergonomics) 제품 설계를 위한 기초 데이터 획득의 의미를 담고 있다.

1992년 국민인체측정조사는 6세부터 50세까지 전 국민을 모집단으로 하여 표본설계에 따른 6,647개의 표본(이를 구현하는 과정에서 실제 측정된 피측정자의 수는 8,886명/ 남자 4,530명, 여자 4,356명이다)에 대하여 인체 84개 부위를 측정한 것이다. 이 조사결과는 이미 '국민표준체위 조사보고서'로 발간된 바 있다 (공업진흥청, 1992).

인체측정은 인체측정 기준인 KSA 7003과 KSA 7004에 따라 인체 84개 측정부위별 측정항목(부록 표 A.1)이 선정되어 있으며, 1992년 조사의 경우 측정요원은 마틴(Martin)자를 이용하여 피측정자의 측정항목을 직접 측정한 것이다. 측정부위별 측정단위는 밀리미터 단위이다. 이러한 직접측정방법은 비용이 적게 들며, 측정도구가 간단하여 이동 측정이 용이하나 측정과정에서 피측정자와 측정자 간에 신체접촉이 이루어지기 때문에 측정오차가 문제시된다. 특히 '젖꼭지간격', '밑가슴둘레', '회음높이' 등과 같은 측정항목의 경우 피측정자의 수치심을 유발시켜 정확한 측정에 어려움이 있다. 본 연구는 이러한 점에 착안하여 상대적으로 측정이 용이한 측정항목을 이용하여 측정이 곤란한 부위를 추정하기 위하여 의사결정나무 모형을 적용하고자 한다. 실제로 1979년, 1986년의 1, 2차 국민인체측정조사에서는 몇몇 측정항목에 대하여 회귀추정식을 이용하여 간접측정을 시도한 바 있다. 그러나 이 경우 지나치게 많은 부위를 단순한 선형회귀분석만을 적용하여 추정하였기 때문에 그 결과의 효용성에 대한 의문이 제기된 바 있다 (공업진흥청, 1992).

본 연구에서 측정이 곤란한 부위로 선택한 항목은 '젖꼭지간격'이다. '젖꼭지간격'의 측정방법은 피측정자가 서있는 자세에서 피측정자가 자연스럽게 숨을 들이마신 후 숨을 멈추듯이 할 때, 양쪽 젖꼭지점 사이의 직선거리를 앞쪽에서 측정한다 (공업진흥청, 1992). 본 연구에서는 '젖꼭지간격'을 목표변수로 두고 의사결정나무 모형을 적용하여 측정곤란 부위를 예측하기 위한 추천 모형을 탐지하고자 한다 (최종후와 소선하, 2005).

4.1. 데이터 및 변수

분석을 위한 데이터는 1992년 '국민인체측정조사' 자료 중에서 피측정자의 나이가 만 12세 이상(중학

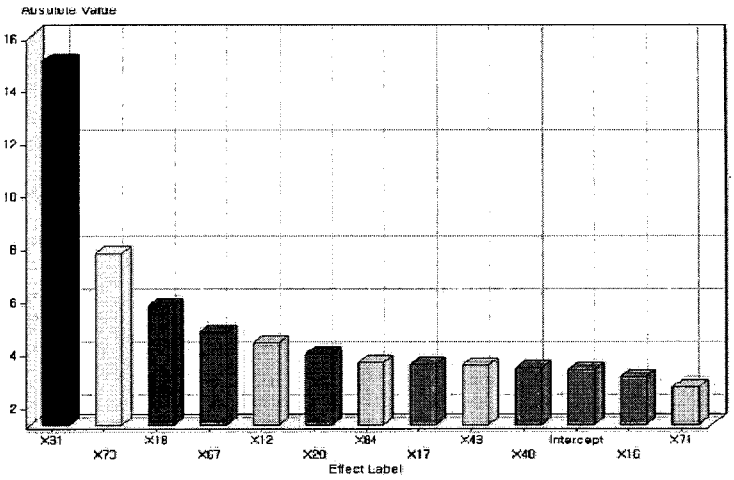

그림 **4.1.** 인체측정자료: 변수 선택

생 이상)인 여성 자료이다. 결측치 및 각 항목의 바깥 울타리 밖에 있는 특이치를 제거하여 실제 분석에 사용된 자료는 2,493개이다. 의사결정나무 모형 구축에 앞서서 목표변수가 되는 '젖꼭지간격'을 설명하는 유의한 측정항목을 선별하였다. 변수선택에는 젖꼭지간격을 제외한 83개 항목과 나이(Age), 총 84개 예측변수를 이용하여 단계적 회귀분석의 변수선택법(Stepwise Variable Selection; slstay = 0.01, slentry = 0.01)을 적용하였다. 그 결과 선택된 변수는 가슴둘레($X31$), 뒤품($X73$), 가슴너비($X18$), 손너비($X67$), 머리위로뻗은Grip높이($X12$), 엉덩이너비($X20$), 몸무게($X84$), 몸통너비($X17$), 앉은팔꿈치높이($X43$), 엉덩이오금길이($X48$), 어깨너비($X16$), 앞품($X71$)이다. 그림 4.1에서 통계적으로 유의한 변수 선택 결과를 보여주고 있는데, y축은 검정통계량 t 값이다.

4.2. 추천 모형 선택

의사결정나무는 사용되는 분리기준과 정지기준에 따라 다른 모형을 구축하며 구축된 여러 모형의 비교를 통해 적절한 의사결정나무를 선택하여 사용하게 된다. 본 논문에서 비교실험을 수행하는 취지는 의사결정나무 모형은 적용하는 기준에 따라 구축 모형이 다르게 나타난다는 점, 나무가 클수록 정확도는 높지만 일반화의 성질을 충족하기 어렵다는 점, 사용되는 정지기준이 분석용 자료에만 적용되어진다는 점에 착안해 교차타당성 평가를 함께 실시, 비교함으로써 바람직한 추천 모형을 선택하는 것이다.

표 4.1은 본 연구에서 적용한 의사결정나무의 세 가지 경쟁 모형, 즉 CHAID, Exhaustive CHAID, CART를 비교한 결과이다. 목표변수가 연속형이므로 사용된 위험 추정치는 MSE이다. CHAID의 경우에 하위 마디의 최소 개체수가 5이고 나무 깊이가 3일 때 구축된 모형은 모두 상위 마디의 최소 개체수가 20이상이므로 상위 마디의 개체수가 10, 15, 20은 동일한 나무가 구축된다. 따라서 이 중 가장 단순한 모형인 상위 마디 개체수 20만을 남기고 셀을 병합하였다. 이 경우 MSE는 1.259이다. 볼드체로 표시한 부분은 10-fold 교차타당법에 의한 MSE인데 서로 다른 10개의 분석용 자료와 겹치지 않는 검증용 자료를 통해 각 자료에서 구축된 임의의 모형에 의해서 계산된 MSE들의 평균값이다. 회귀나무는 끝마디에서의 목표변수 평균을 예측하는 것이 목적이므로 10-fold 교차타당법에 의한 MSE가 작은 모형을 후보 모형과 추천 모형으로 선택한다. 셀에서 밑줄 친 것에 해당하는 모형이 각 알고리즘에서 선택된 후보 모형이다.

표 4.1. 인체측정자료: 경쟁 모형의 MSE

CHAID							Exhaustive CHAID				
개체수		나무 깊이					나무 깊이				
하위	상위	3	4	5	6	7	3	4	5	6	7
5	10	1.259	1.203	1.170	1.150	1.142	1.282	1.255	1.249	1.246	
			1.510	**1.532**	**1.526**	**1.521**					
	15		1.205	1.174	1.154	1.146					
			1.503	**1.542**	**1.520**	**1.528**				**1.516**	
	20		1.209	1.179	1.166	1.159					
		1.477	**1.481**	**1.487**	**1.508**	**1.507**	**1.480**	**1.499**	**1.478**		
10	20	1.262	1.220	1.207	1.203	1.201	1.285	1.258			
			1.442	**1.470**	**1.511**	**1.492**		**1.508**			
	30		1.228	1.215	1.211	1.209		1.262			
			1.471	**1.507**	**1.483**	**1.486**		**1.507**			
	40		1.236	1.223	1.219	1.218		1.270			
		1.484	**1.462**	**1.451**	**1.487**	**1.500**	**1.457**	**1.483**			
20	50	1.295	1.270	1.267	1.263		1.298	1.286			
		1.448	**1.454**	**1.460**	**1.473**						
	80	1.299	1.276	1.273							
		1.433	**1.478**	**1.470**			**1.451**	**1.483**			
	100	1.310	1.287				1.309	1.298			
		1.457	**1.462**				**1.466**	**1.454**			

CART						
개체수		나무 깊이				
하위	상위	3	4	5	6	7
5	10	1.397	1.306	1.229	1.183	1.138
				1.452	**1.484**	**1.470**
	15			1.232	1.186	1.142
	20			**1.458**	**1.496**	**1.480**
10	20			1.236	1.204	1.177
	30					
	40		**1.436**	**1.425**	**1.428**	**1.449**
20	50		1.318	1.252	1.226	1.207
					1.441	**1.457**
	80				1.229	1.210
				1.417	**1.387**	**1.415**
	100			1.258	1.242	1.223
		1.487	**1.428**	**1.422**	**1.413**	**1.398**

그 결과 CHAID와 Exhaustive CHAID에서는 하위 마디의 최소 개체수 20, 상위 마디의 최소 개체수 80, 나무 깊이가 3인 모형이 후보 모형으로 선택되었고, CART에서는 하위 마디의 최소 개체수가 20이고 상위 마디의 최소 개체수가 80, 나무 깊이가 6인 모형이 선택되었다.

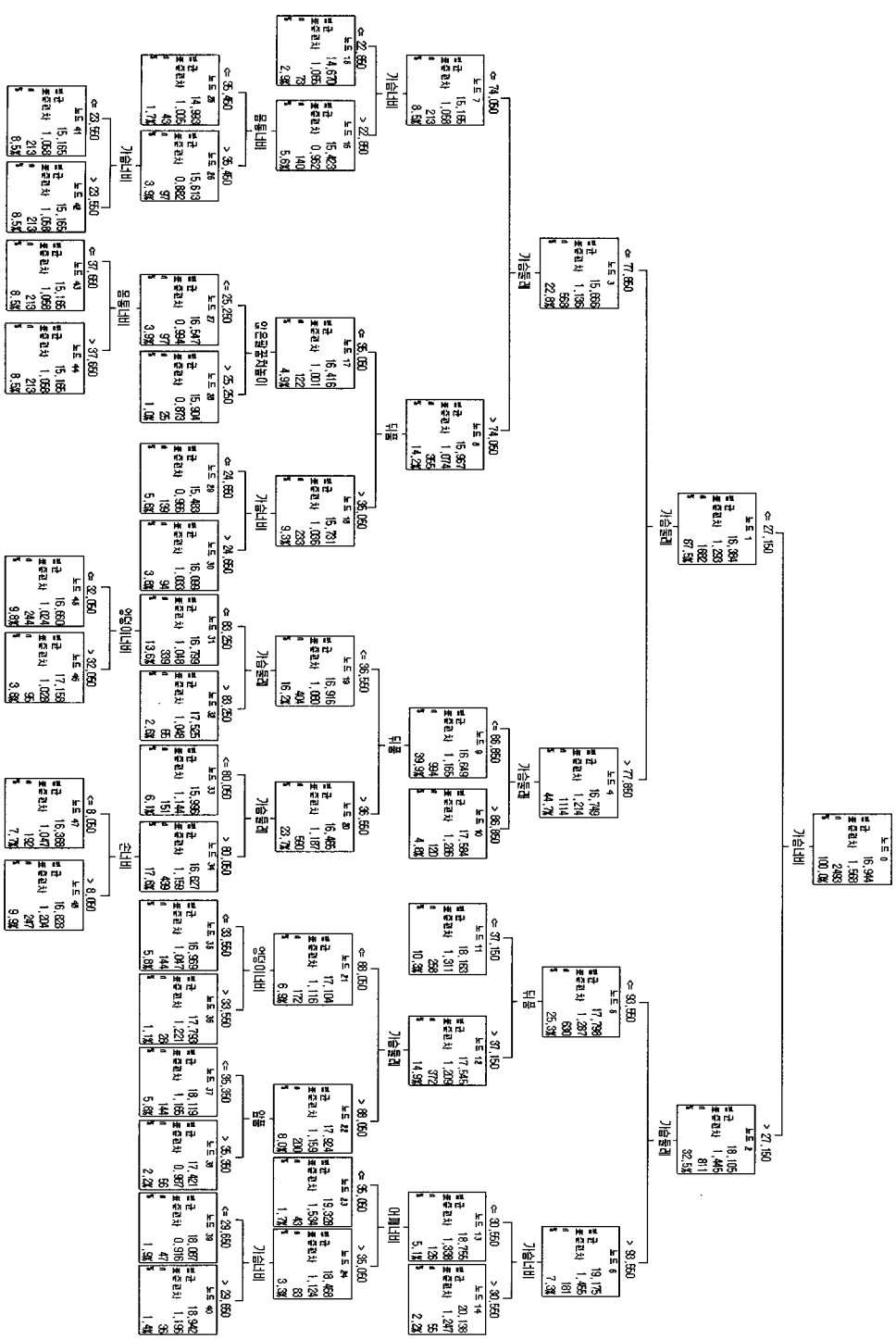

그림 **4.2.** 추천 모형의 의사결정나무 구조

표 4.2. 인체측정자료: 끝마디와 예측변수

마디 번호	예측 평균	예측변수
14	20.1382	30.550 < 가슴너비, 93.550 < 가슴둘레
23	19.3279	27.150 < 가슴너비 ≤ 30.550, 93.550 < 가슴둘레, 어깨너비 ≤ 35.050
40	18.9417	29.650 < 가슴너비 ≤ 30.550, 93.550 < 가슴둘레, 35.050 < 어깨너비
11	18.1628	27.150 < 가슴너비, 가슴둘레 ≤ 93.550, 뒤품 ≤ 37.150
37	18.1194	27.150 < 가슴너비, 88.050 < 가슴둘레 ≤ 93.550, 37.150 < 뒤품, 앞품 ≤ 35.350
39	18.0872	27.150 < 가슴너비 ≤ 29.650, 93.550 < 가슴둘레, 35.050 < 어깨너비
36	17.7929	27.150 < 가슴너비, 가슴둘레¡=88.050, 37.150¡뒤품, 33.550¡엉덩이너비
10	17.5842	가슴너비 ≤ 27.150, 86.850 < 가슴둘레
32	17.5246	가슴너비 ≤ 27.150, 83.250 < 가슴둘레 ≤ 86.850, 뒤품 ≤ 6.550
38	17.4214	27.150 < 가슴너비, 88.050 < 가슴둘레 ≤ 93.550, 37.150 < 뒤품, 35.350 < 앞품
46	17.1579	가슴너비 ≤ 27.150, 77.850 < 가슴둘레 ≤ 83.250, 뒤품 ≤ 36.550, 32.050 < 엉덩이너비
35	16.9694	27.150 < 가슴너비, 가슴둘레 ≤ 88.050, 37.150 < 뒤품, 엉덩이너비 ≤ 33.550
48	16.8283	가슴너비 ≤ 27.150, 80.050 < 가슴둘레 ≤ 86.850, 36.550 < 뒤품, 8.050 < 손너비
44	16.7965	가슴너비 ≤ 27.150, 74.050 < 가슴둘레 ≤ 77.850, 뒤품 ≤ 35.050, 앉은팔꿈치높이 ≤ 25.250, 37.650 < 몸통너비
45	16.6598	가슴너비 ≤ 27.150, 77.850 < 가슴둘레 ≤ 83.250, 뒤품 ≤ 36.550, 엉덩이너비 ≤ 32.050
47	16.3677	가슴너비 ≤ 27.150, 80.050 < 가슴둘레 ≤ 86.850, 36.550 < 뒤품, 손너비 ≤ 8.050
43	16.1925	가슴너비 ≤ 27.150, 74.050 < 가슴둘레 ≤ 77.850, 뒤품 ≤ 35.050, 앉은팔꿈치높이 ≤ 25.250, 몸통너비 ≤ 37.650
30	16.0989	24.650 < 가슴너비 ≤ 27.150, 74.050¡가슴둘레 ≤ 77.850, 35.050 < 뒤품
33	15.9960	가슴너비 ≤ 27.150, 77.850 < 가슴둘레 ≤ 80.050, 36.050 < 뒤품
28	15.9040	가슴너비 ≤ 27.050, 74.050 < 가슴둘레 ≤ 77.850, 뒤품 ≤ 35.050, 25.250 < 앉은팔꿈치높이
42	15.8029	23.550 < 가슴너비 ≤ 27.150, 가슴둘레 ≤ 74.050, 35.450 < 몸통너비
29	15.4827	가슴너비 ≤ 24.650, 74.050 < 가슴둘레 ≤ 77.850, 35.050 < 뒤품
41	15.1464	22.850 < 가슴너비 ≤ 23.550, 가슴둘레 ≤ 74.050, 35.450 < 몸통너비
25	14.9930	22.850 < 가슴너비 ≤ 27.150, 가슴둘레 ≤ 74.050, 몸통너비 ≤ 35.450
15	14.6699	가슴너비 ≤ 22.850, 가슴둘레 ≤ 74.050

최종적으로 후보 모형 군에서 MSE값이 가장 작은 CART를 추천 모형으로 선택한다. 선택된 추천 모형은 CART에 기반한 하위 마디의 최소 개체수 20, 상위 마디의 최소 개체수 80, 나무 깊이가 6인 모형이다. 가슴둘레, 가슴너비, 몸통너비, 엉덩이너비, 어깨너비, 손너비, 뒤품, 앞품, 앉은팔꿈치높이가 예측변수로 사용되며 그 결과는 그림 4.2, 부록의 표 A.2와 같다. 부록의 표 A.2는 모형의 모든 마디(Node)를 나타낸 표이고 음영으로 표시된 마디는 끝마디이다. 1차 예측변수는 상위 마디로부터 최종 분할된 변수와 분리점이다. 예를 들어 마디 10은 상위 마디 4로부터 가슴둘레가 86.850보다 큰 경우에 분리된 마디이다. 또한 마디 4는 마디 1로부터 가슴둘레가 77.850보다 큰 경우에 분리되었으며 마디 1은 뿌리마디인 마디 0으로부터 가슴너비 27.150이하인 경우에 분리된 마디이다.

그림 4.2를 살펴보면 가슴너비, 가슴둘레, 엉덩이너비, 몸통너비, 손너비가 클수록 젖꼭지간격이 크고, 어깨너비, 뒤품, 앞품, 앉은팔꿈치높이가 작을수록 젖꼭지간격이 크게 나타난다.

표 4.2는 끝마디들의 예측된 평균값을 내림차순으로 정리한 것이다. 예측변수 열은 분리되는 순서에 따른 것이 아니라 최종적으로 분리된 예측변수와 범주이다. 예를 들어, 그림 4.2에서 14번 마디는 가슴너

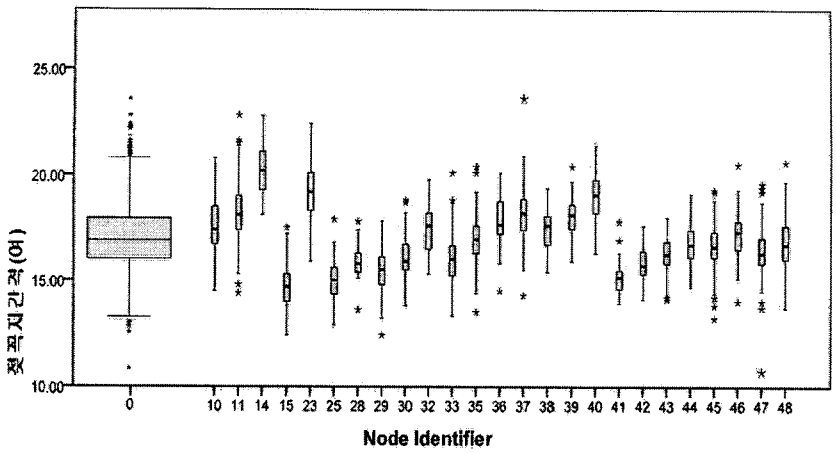

그림 4.3. 인체측정자료 끝마디의 상자도표

표 4.3. 인체측정자료 끝마디의 상자도표

마디번호	N	최소값	최대값	평균	표준편차
0 (전체)	2493	10.7	23.6	16.9436	1.5677
10	120	14.5	20.8	17.5842	1.2859
11	258	14.4	22.8	18.1628	1.3112
14	55	18.1	22.8	20.1382	1.2471
15	73	12.4	17.5	14.6699	1.0648
23	43	15.9	22.4	19.3279	1.5343
25	43	12.9	17.9	14.9930	1.0051
28	25	13.6	17.5	15.9040	0.8734
29	139	12.4	17.8	15.4827	0.9659
30	94	13.8	18.8	16.0989	1.0327
32	65	15.3	19.8	17.5246	1.0484
33	151	13.3	20.1	15.9960	1.1444
35	144	13.5	20.4	16.9694	1.0468
36	28	14.5	20.1	17.7929	1.2211
37	144	14.3	23.6	18.1194	1.1653
38	56	15.4	19.4	17.4214	0.9871
39	47	15.9	20.4	18.0872	0.9157
40	36	16.3	21.4	18.9417	1.1960
41	28	13.9	17.8	15.1464	0.8975
42	69	14.1	17.6	15.8029	0.8080
43	40	14.1	18.0	16.1925	0.9480
44	57	14.7	19.1	16.7965	0.9558
45	244	13.2	19.3	16.6598	1.0241
46	95	14.0	20.5	17.1579	1.0283
47	192	10.7	19.6	16.3677	1.0470
48	247	13.7	20.6	16.8283	1.2036

비 > 27.150이고 가슴둘레 > 93.550이고 가슴너비 > 30.550인 경우에 젖꼭지간격을 20.1382로 예측한다. 가슴너비가 30.550보다 크고 가슴너비가 27.150보다 크다는 것은 30.550보다 크다는 것을 의미하

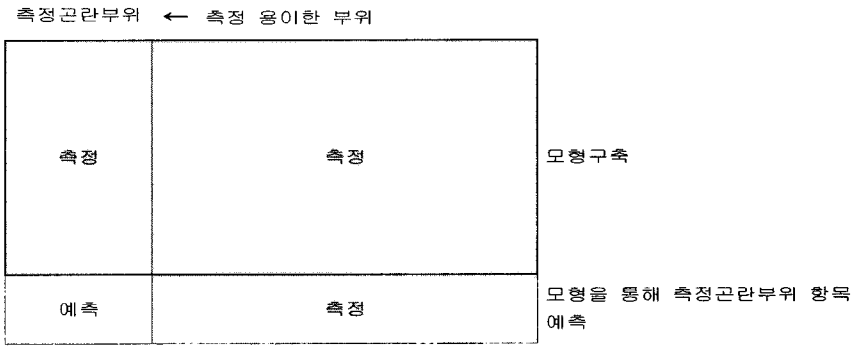

그림 5.1. 인체측정조사의 새로운 조사방법 제안

므로 30.550 < 가슴너비, 93.550 < 가슴둘레를 표 4.2에 나타낸 것이다. 그러나 실제로 관측치를 예측할 때에는 분리순서에 따른다. 그림 4.3은 추천 모형에 기반을 두고 전체 자료와 각 끝마디에 대한 젖꼭지간격의 상자도표(Box Plot)를 작성한 결과이다. 표 4.3은 이에 대한 기술통계량이다. 이를 통해 해당 마디에 따른 젖꼭지간격의 분포를 가시적으로 살펴 볼 수 있다.

5. 토의 및 결론

본 연구는 연속형 목표변수의 경우에 대하여 의사결정나무의 추천 모형 탐지 절차를 연구한 결과이다. 일반적으로 나무가 클수록 위험 추정치는 감소하지만 일반화의 성질을 충족시키기 어렵다는 점에서 교차타당성 평가를 통해 위험 추정치가 작으면서 안정된 후보 모형을 탐지하기 위한 비교실험을 실시한 것이다. 그 결과 분리가 이루어지기 위한 상위 및 하위 마디의 최소 개체수와 나무의 최대 성장 깊이를 결정해 줌으로써 활용된 알고리즘인 CHAID, Exhaustive CHAID, CART에서의 안정된 후보 모형을 선별하고 우수한 예측력을 갖는 추천 모형을 선택할 수 있었다. 알고리즘별로 우수한 모형을 선택하는 데 시각적인 판단이 용이하였고 결과의 효율성도 볼 수 있었다.

그 결과 측정곤란부위인 '젖꼭지간격'을 예측하는데 CART 모형이 좋은 결과를 보였다. CART 모형에 기반하여 구축된 예측을 위한 의사결정나무의 현실 활용면을 검토해 보면 인체측정자료가 활용되는 의류, 가구류 등 산업분야에서 의류 치수제정과 같은 제품표준화와 설계합리화에 분석결과를 활용할 수 있을 것이다. 또한 향후 인체측정조사시 피측정자의 수치심으로 인해 야기될 수 있는 측정곤란부위 항목에 대한 측정오차를 극복하기 위한 방안으로 그림 5.1과 같이 측정곤란 부위 예측을 위한 모형을 구축 적용함으로써 조사의 난점을 줄여나갈 수 있겠다.

부록

표 A.1. 인체측정자료 측정부위별 측정항목

Part I	(1) 키	(13) 옆으로뻗은손끝길이
	(2) 눈높이	(14) 앞으로뻗은손끝길이
	(3) 어깨높이	(15) 양팔뻗은손끝길이
	(4) 목뒷높이	(16) 어깨너비
	(5) 허리높이	(17) 몸통너비
	(6) 팔굽힌팔꿈치높이	(18) 가슴너비

	(7) 엉덩이밑높이	(19) 허리너비
	(8) 손끝높이	(20) 엉덩이너비
	(9) 회음높이	(21) 젖꼭지간격(여)
	(10) 대퇴돌기높이	(22) 가슴두께
	(11) 무릎안쪽높이	(23) 배두께
	(12) 머리위로뻗은Grip높이	(24) 엉덩이두께
Part II	(25) 목둘레	(41) 앉은눈높이
	(26)목밑둘레	(42) 앉은어깨높이
	(27) 진동둘레	(43) 앉은팔꿈치높이
	(28) 윗팔둘레	(44) 대퇴높이
	(29) 아래팔둘레	(45) 앉은무릎높이
	(30) 윗가슴둘레(여)	(46) 앉은오금높이
	(31) 가슴둘레	(47) 엉덩이무릎길이
	(32) 밑가슴둘레(여)	(48) 엉덩이오금길이
	(33) 허리둘레	(49) 앉은엉덩이너비
	(34) 배둘레	(50) 뒤허리발뒤꿈치길이
	(35) 엉덩이둘레	(51) 어깨점팔꿈치길이
	(36) 넓적다리둘레	(52) 팔꿈치손끝길이
	(37) 무릎둘레	(53) 머리길이
	(38) 장딴지둘레	(54) 얼굴길이
	(39) 앉은키	(55) 눈턱끝길이
	(40) 앉아머리위로뻗은Grip높이	
Part III	(56) 머리너비	(71) 앞품
	(57) 머리두께	(72) 앞중심길이
	(58) 귀구슬사이너비	(73) 뒤품
	(59) 귀구슬사이턱밑길이	(74) 등길이
	(60) 귀구슬사이턱끝길이	(75) 둔부길이
	(61) 귀구슬사이머리마루점길이	(76) 소매길이
	(62) 머리둘레	(77) 안소매길이
	(63) 눈동자사이너비	(78) 화장
	(64) 입너비	(79) 밑위앞뒤길이
	(65) 손길이	(80) 발등둘레
	(66) 손바닥길이	(81) 발목둘레
	(67) 손너비	(82) 발길이
	(68) 손두께	(83) 발너비
	(69) 손둘레	(84) 몸무게
	(70) 어깨길이	

표 **A.2.** 인체측정의 추천 모형 나무

마디	평균	표준 오차	N	%	예측 평균	상위 마디	1차 예측변수		
							변수	향상	분할 값
0	16.9436	1.56768	2493	100.0%	16.9436				
1	16.3835	1.29337	1682	67.5%	16.3835	0	가슴너비	.651	≤ 27.150
2	18.1052	1.44462	811	32.5%	18.1052	0	가슴너비	.651	> 27.150
3	15.6658	1.13551	568	22.8%	15.6658	1	가슴둘레	.177	≤ 77.850
4	16.7494	1.21360	1114	44.7%	16.7494	1	가슴둘레	.177	> 77.850

5	17.7978	1.28745	630	25.3%	17.7978	2	가슴둘레	.107	≤ 93.550
6	19.1751	1.45506	181	7.3%	19.1751	2	가슴둘레	.107	> 93.550
7	15.1648	1.05836	213	8.5%	15.1648	3	가슴둘레	.034	≤ 74.050
8	15.9665	1.07363	355	14.2%	15.9665	3	가슴둘레	.034	> 74.050
9	16.6486	1.16544	994	39.9%	16.6486	4	가슴둘레	.038	≤ 86.850
10	17.5842	1.28593	120	4.8%	17.5842	4	가슴둘레	.038	> 86.850
11	18.1628	1.31120	258	10.3%	18.1628	5	뒤품	.023	≤ 37.150
12	17.5446	1.20926	372	14.9%	17.5446	5	뒤품	.023	> 37.150
13	18.7548	1.33815	126	5.1%	18.7548	6	가슴너비	.029	≤ 30.550
14	20.1382	1.24714	55	2.2%	20.1382	6	가슴너비	.029	> 30.550
15	14.6699	1.06480	73	2.9%	14.6699	7	가슴너비	.011	≤ 22.850
16	15.4229	0.96196	140	5.6%	15.4229	7	가슴너비	.011	> 22.850
17	16.4156	1.00132	122	4.9%	16.4156	8	뒤품	.015	≤ 35.050
18	15.7313	1.03649	233	9.3%	15.7313	8	뒤품	.015	> 35.050
19	16.9161	1.08021	404	16.2%	16.9161	9	뒤품	.020	≤ 36.550
20	16.4654	1.18694	590	23.7%	16.4654	9	뒤품	.020	> 36.550
21	17.1035	1.11565	172	6.9%	17.1035	12	가슴둘레	.025	≤ 88.050
22	17.9240	1.15924	200	8.0%	17.9240	12	가슴둘레	.025	> 88.050
23	19.3279	1.53426	43	1.7%	19.3279	13	어깨너비	.009	≤ 35.050
24	18.4578	1.12360	83	3.3%	18.4578	13	어깨너비	.009	> 35.050
25	14.9930	1.00508	43	1.7%	14.9930	16	몸통너비	.005	≤ 35.450
26	15.6134	0.88231	97	3.9%	15.6134	16	몸통너비	.005	> 35.450
27	16.5474	0.99364	97	3.9%	16.5474	17	앉은팔꿈치높이	.003	≤ 25.250
28	15.9040	0.87344	25	1.0%	15.9040	17	앉은팔꿈치높이	.003	> 25.250
29	15.4827	0.96594	139	5.6%	15.4827	18	가슴너비	.009	≤ 24.650
30	16.0989	1.03274	94	3.8%	16.0989	18	가슴너비	.009	> 24.650
31	16.7994	1.04799	339	13.6%	16.7994	19	가슴둘레	.012	≤ 83.250
32	17.5246	1.04837	65	2.6%	17.5246	19	가슴둘레	.012	> 83.250
33	15.9960	1.14437	151	6.1%	15.9960	20	가슴둘레	.018	≤ 80.050
34	16.6269	1.15928	439	17.6%	16.6269	20	가슴둘레	.018	> 80.050
35	16.9694	1.04676	144	5.8%	16.9694	21	엉덩이너비	.006	≤ 33.550
36	17.7929	1.22109	28	1.1%	17.7929	21	엉덩이너비	.006	> 33.550
37	18.1194	1.16531	144	5.8%	18.1194	22	앞품	.008	≤ 35.350
38	17.4214	0.98714	56	2.2%	17.4214	22	앞품	.008	> 35.350
39	18.0872	0.91571	47	1.9%	18.0872	24	가슴너비	.006	≤ 29.650
40	18.9417	1.19604	36	1.4%	18.9417	24	가슴너비	.006	> 29.650
41	15.1464	0.89752	28	1.1%	15.1464	26	가슴너비	.003	≤ 23.550
42	15.8029	0.80804	69	2.8%	15.8029	26	가슴너비	.003	> 23.550
43	16.1925	0.94798	40	1.6%	16.1925	27	몸통너비	.003	≤ 37.650
44	16.7965	0.95580	57	2.3%	16.7965	27	몸통너비	.003	> 37.650
45	16.6598	1.02413	244	9.8%	16.6598	31	엉덩이너비	.007	≤ 32.050
46	17.1579	1.02828	95	3.8%	17.1579	31	엉덩이너비	.007	> 32.050
47	16.3677	1.04701	192	7.7%	16.3677	34	손너비	.009	≤ 8.050
48	16.8283	1.20355	247	9.9%	16.8283	34	손너비	.009	> 8.050

성장방법: CART * 끝마디는 음영표시됨

참고문헌

강현철, 한상태, 최종후, 이성건, 김은석, 엄익현, 김미경 (2006). <고객관계관리(CRM)를 위한 데이터마이닝 방법론>, 자유아카데미.

공업진흥청 (1992). KRISS-92-144-IR, <산업제품의 표준치 설정을 위한 국민표준체위 조사보고서>.

박경수 (1993). <인간공학-작업경제학>, 영지문화사.

최종후, 권기만, 김수택 (2002). <신용평점모형>, 세창출판사.

최종후, 소선하 (2005). <사례로 배우는 데이터마이닝>, 자유아카데미.

Biggs, D., de Ville, B. and Suen, E. (1991). A Method of choosing multiway partitions for classification and decision trees, *Journal of Applied Statistics*, **18**, 49–62.

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). *Classification and Regression Tree*, Chapman & Hall/CRC, New York.

Han, J. and Kamber, M. (2006). *Data Mining: Concepts & Techniques*, 2/e, Elsevier Inc, New York.

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data, *Journal of Applied Statistics*, **29**, 119–127.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection, *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, 1137–1143.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*, Morgan-Kaufmann, California.

A Study on Exploration of the Recommended Model of Decision Tree to Predict a Hard-to-Measure Mesurement in Anthropometric Survey

J. H. Choi[1] · S. K. Kim[2]

[1]Department of Information & Statistics, Korea University; [2]Korea Scoring

(Received July 2009; accepted August 2009)

Abstract

This study aims to explore a recommended model of decision tree to predict a hard-to-measure measurement in anthropometric survey. We carry out an experiment on cross validation study to obtain a recommened model of decision tree. We use three split rules of decision tree, those are CHAID, Exhaustive CHAID, and CART. CART result is the best one in real world data.

Keywords: Decision tree, k-fold cross validation, CHAID, exhaustive CHAID, CART.

[1]Corresponding author: Professor, Department of Information & Statistics, Korea University, Jochiwon-eup, Yeongi-gun, Chungnam, 339-700, Korea. E-mail: jhchoi@korea.ac.kr