The Korean Communications
in Statistics Vol. 9, No. 2, 2002
pp. 411-423

A Split Criterion for Binary Decision Trees

Hyun Jip Choi[1], Myong Rok Oh[2]

Abstract

In this paper, we propose a split criterion for binary decision trees. The proposed criterion selects the optimal split by measuring the prediction success of the candidate splits at a given node. The criterion is shown to have the property of exclusive preference. Examples are given to demonstrate the properties of the criterion.

Keywords : Binary decision tree, Split criterion, Measure of prediction success.

1. 서론

주어진 상황에 따라 차이가 있을 수 있지만 의사결정나무를 얻고자 하는 기본적인 목적은 주어진 입력변수들(input variables)에 의해 목표변수(target variable)의 범주(class)를 정확히 분류하기 위한 분류자(classifier)를 찾거나 혹은 입력변수와 목표변수 사이에 어떠한 예측구조(predictive structure)를 가지고 있는지 밝히는데 있다. 이러한 목적을 수행하기 위하여 제안된 여러 의사결정나무(decision trees) 생성방법들은 대부분 적절한 기준에 의하여 주어진 자료를 분할해 나가면서 나무를 성장 시키는 단계와 최대한 성장된 나무로부터 보다 간결한 구조를 가진 나무를 얻기 위해 가지치기(pruning)를 수행하는 두 단계를 거치게 된다. 따라서 나무를 성장시키기 위하여 주어진 마디(node)에서 고려할 수 있는 여러 분할 중에서 가장 적절한 분할을 선택하기 위한 분할기준(split criterion)은 전체적인 나무의 모양을 결정짓는 가장 중요한 요인이다.

Kass(1980)는 고려할 수 있는 여러 분할 중에서 가장 적절한 분할을 찾기 위한 분할기준으로 피어슨 카이제곱 통계량을 제안하였으며, Clark와 Pregibon(1992) 그리고 Quinlan(1993)등은 엔트로피에 기반을 둔 분할기준을 제안하였다. Breiman, Freidman, Olshen과 Stone(1984)은 분리된 마디의 불순도(impurity) 감소량을 측정하여 분할을 선택하기 위하여 Gini 지수를 분할기준으로 제안하였고, Taylor와 Silverman(1993)은 분할기준이 갖추어야 하는 배타적 선호특성(exclusive preference property)을 지적하고 Gini 지수가 갖추지 못한 이러한 특성을 만족하는 MPI(mean posterior improvement) 기준을 제안하였다. Shih(1999)는 이들 여러 분할기준들이 모두 적절한 가중값에 의한 가중합으로 표현될 수 있다는 사실을 보이고, Read와 Cressi(1988)가 제안한 통계량

1) Assistant professor, Division of Economics, Kyonggi University, Suwon, 442-760, Korea,
E-mail : hjchoi@stat.kyonggi.ac.kr
2) Consultant, StatSoft korea, #44-1(3rd FL), Pyl-Dong 1Ga, Jung-Gu, Seoul, 100-866, Korea,
E-mail : always@statsoft.co.kr

역시 좋은 분할기준이 될 수 있으며 MPI와 같이 배타적 선호특성을 만족하는 것을 보였다. Mola와 Siciliano(1997)는 CART방법에서 Gini 기준에 의한 분할선택을 보다 빠르게 수행하기 위하여 Goodman과 Kruskal의 예측력지수(predictability index)를 분할기준으로 제안하였으며 전병환, 김창수, 송홍엽 그리고 김재희(1997)는 엔트로피에 기반을 둔 새로운 분리기준을 제안하였다.

나무에 속한 주어진 마디에서 가장 적절한 분할을 선택하기 위한 이들 분할기준들은 모두 서로 접근 방법은 상이하지만 고려의 대상이 되는 분할이 얼마나 목표변수의 범주를 가장 잘 분류하는가를 측정하고자 하는 측도들로 볼 수 있다. 다시 말해 가장 적절한 분할을 선택하는 문제는 주어진 마디에서 고려할 수 있는 각 분할들이 얼마나 목표변수의 범주를 잘 분류해내는가를 측정하고, 이들 중에서 목표변수의 범주를 가장 잘 분류하게 하는 분할을 선택하는 문제로 볼 수 있다. 이러한 맥락에서 본 연구에서는 이진의사결정나무의 성장을 위하여 나무에 속한 주어진 마디에서 고려할 수 있는 각 분할들이 잠재적으로 가지고 있는 모든 가능한 규칙들의 예측력을 측정하여 가장 예측력이 우수한 규칙을 가지고 있는 분할을 선택하게 하는 분할기준을 제안하고자 한다.

2절에서는 목표범주가 두 범주만을 가진 경우에, 주어진 마디의 분할을 통해 고려할 수 있는 잠재적인 규칙들의 예측력을 측정하기 위한 예측성공측도(measure of prediction success)와 이를 이용한 분할기준을 제안하였다. 또한 제안된 측도가 Taylor와 Silverman(1993)이 정의한 배타적 선호 특성(exclusive preference property)을 만족하는 것을 보였다. 3절에서는 제안된 측도를 다범주 문제로 확장하고 Breiman, Freidman, Olshen과 Stone(1984)의 Twoing 기준과 유사한 초범주에 의해 분할을 선택하는 방법을 제안하였다. 4절에서는 제안된 측도가 MPI와 같이 배타적 선호 특성을 만족한다는 것을 모의자료를 통해 경험적으로 밝혔으며, 실제 적용 예를 보이기 위하여 목표변수가 두가지 범주를 가진 자료와 다범주의 예를 위하여 세가지 범주를 가진 두 자료의 분석결과를 제시하였다. 마지막으로 결론에서는 본 연구의 결과를 정리하고 제안된 분할기준이 가질 수 있는 문제점과 특징에 관하여 토론하였다.

2. 분할기준 : 예측성공측도

마디 t에서 목표변수가 J개 범주를 가진 경우에 입력변수들로부터 생성 가능한 모든 분할들의 집합을 S 그리고 $s \in S$인 분할 s에 의한 두 하위마디(subnode)를 각각 t_L과 t_R이라고 하자. $J=2$인 경우에 분할 s에 의해 다음과 같은 2×2 분할표를 구성할 수 있다.

	1	2	
L	p_{L1}	p_{L2}	p_L
R	p_{R1}	p_{R2}	p_R
	p_1	p_2	1

<표 1> 분할 s에 의한 마디 t에서의 2×2 분할표

<표 1>에서 p_i, $i \in L, R$, 는 각각 마디 t_L과 t_R의 비율 그리고 p_j, $j=1,2$는 마디 t에서의 목표변수 범주의 비율을 나타낸다. 또한 p_{Lj}는 마디 t_L과 목표변수 범주 j의 비율, p_{Rj}는 마디 t_R과 목표변수 범주 j의 비율을 나타낸다. 이때 $p_L + p_R = 1$, $\sum_{j=1}^{2} p_j = 1$ 이며 $\sum_{j=1}^{2} p_{Lj} = p_L$과 $\sum_{j=1}^{2} p_{Rj} = p_R$ 그리고 $j=1,2$에 대하여 $p_{Lj} + p_{Rj} = p_j$ 이다.

이러한 상황에서 분할 s에 의해 두 하위마디가 생성되었다면 이는 마디 t에서 "자료들이 분할 s에 의해 t_L에 속하면 범주 '1'로 분류 따라서 t_R에 속하면 범주 '2'로 분류" 혹은 "자료들이 분할 s에 의해 t_L에 속하면 범주 '2'로 분류 따라서 t_R에 속하면 범주 '1'로 분류"와 같은 두 분류규칙 중에서 한 분류규칙이 생성되는 것으로 볼 수 있다. 그러므로 만일 전자의 분류규칙에 의해 분류가 일어난다면 <표 1>에서 칸 $(L, 2)$와 $(R, 1)$은 잘 못 분류된 비율을 갖는 칸을 나타내며 이를 오차칸(error cell)이라 부르기로 한다. 또한 이들 오차칸들의 집합은 E로 나타내기로 한다. 즉, $E=\{(L,2), (R,1)\}$. 마찬가지로 후자의 분류규칙에 의해 분류가 일어난다면 오차칸들의 집합 $E=\{(L,1), (R,2)\}$이 될 것이다.

이와 같이 분할 s에 의해 생성되는 두 규칙들이 각기 얼마나 목표변수의 범주를 잘 분류 혹은 예측 해주느냐 하는 것은 Hildebrand, Laing과 Rosenthal(1974a, 1974b)이 이차원 분할표에서 고려할 수 있는 모든 가능한 규칙의 예측력을 측정하기 위하여 제안한 측도를 분할문제에 적용한 다음과 같은 예측성공측도(measure of prediction success)에 의해 측정할 수 있다. 마디 t에서의 분할 s에 의한 두 분류규칙을 R_l, $l=1,2$, 그리고 각 분류규칙에 의한 오차칸들의 집합을 E_l 이라고 하면

$$M_{R_l}(s, t) = 1 - \frac{p_L \sum_{j=1}^{2} \delta_{Lj}^{(l)} p(j|L) + p_R \sum_{j=1}^{2} \delta_{Rj}^{(l)} p(j|R)}{p_L \sum_{j=1}^{2} \delta_{Lj}^{(l)} p_j + p_R \sum_{j=1}^{2} \delta_{Rj}^{(l)} p_j}, \quad l=1,2, \tag{1}$$

여기서

$$\delta_{ij}^{(l)} = \begin{cases} 1, & (i,j) \in E_l, \\ 0, & (i,j) \notin E_l, \end{cases}$$

이며 $i \in \{L, R\}$과 $j=1,2$에 대하여 $p(j|i) = p_{ij}/p_i$이다.

식 (1)에서 만일 분할 s에 의해 생성되는 분류규칙 R_l에 의한 E_l에 속한 칸의 비율이 모두 '0'이면 $M_{R_l}(s, t)=1$이 되어 R_l은 완전한 예측을 수행하는 것으로 판단할 수 있다. 만일 E_l에 속한 칸들에 대하여 $p_j = p(j|i)$, $i \in \{L, R\}$, $j=1,2$, 이라면 $M_{R_l}(s, t)=0$이 된다. 그리고 만일 분할에 의해 모두 같은 목표변수의 범주를 예측하게 하는 규칙을 고려한다고 하여도, 다시 말해

오차칸들의 집합 E_l이 $\{(1,L),(1,R)\}$ 혹은 $\{(2,L),(2,R)\}$이라면 이때의 $M(s,t)=0$인 것에 주의하자. 마지막으로 만일 E_l에 속한 칸들에 대하여 $p_j<p(j|i)$이라면 $M_{R_l}(s,t)<0$이 된다. 그러나 이는 분할에 의한 R_l의 오분류률(misclassification rate)이 원 목표변수 범주의 비율 보다 높은 것을 의미하므로 이때의 R_l은 적절하지 않은 분류규칙이므로 고려의 대상에서 제외하기로 한다. 따라서 p_j는 E_l에 속한 칸들이 가질 수 있는 최대 예측오차를 의미하며, 결국 $M_{R_l}(s,t)$가 가질 수 있는 값의 구간은 $0 \le M_{R_l}(s,t) \le 1$이 된다.

또한 식(1)에서 오른쪽 항의 분자는 분할 s에 의한 하위마디 t_L과 t_R이 주어져 있을 때 오차칸 비율들의 합, 즉 분류규칙 R_l을 적용할 때 발생하는 예측오차들의 가중합을 나타낸다. 반면에 분모는 분할에 의한 오분류들이 가질 수 있는 최대 예측오차인 마디 t에서의 목표변수 범주비율들의 가중합을 나타낸다. 따라서 $M_{R_l}(s,t)$는 분류규칙 R_l을 적용하였을 때의 오차의 감소량을 측정하며 Goodman과 Kruskal(1954)이 제안한 예측력 지수(predictability index) τ와 같이 PRE(proportinate-reduction-in-error) 원칙을 만족한다.

이러한 사실로부터 주어진 마디 t에서 고려할 수 있는 각 분할 s의 분할 적합도(goodness of split)를 측정하기 위한 다음과 같은 분할기준을 제안하기로 한다.

$$M(s,t) = \max\{M_{R_1}(s,t), M_{R_2}(s,t)\} \tag{2}$$

이제 식 (2)에 의하여 마디 t에서 고려 가능한 모든 분할 s에 대한 분할 적합도를 평가할 수 있으므로 결국 주어진 마디 t에서의 최적분할(optimal split)은

$$M(s^*,t) = \max_{s\in S} M(s,t)$$

을 만족하는 s^*로 결정할 수 있다.

식 (2)에 의하면 임의의 분할 s에 대한 분할 적합도를 결정하기 위해서 두 분류규칙 R_1과 R_2에 따른 예측성공측도를 모두 계산하여야 한다. 즉, 두 번의 계산이 수행되어야 한다. 그런데 이들 두 분류규칙에 대하여 식 (1)의 값을 얻기 위해서는 오차칸들의 집합 E_1, E_2가 주어져야 한다. 그러나 오차칸들의 집합이 주어진 것과 분류규칙이 주어진 것은 동등하므로 식 (2)를 만족하는 오차칸들의 집합은 다음을 통하여 결정할 수 있다.

$$E = \begin{cases} \{(L,1),(R,2)\}, & p(1|L) \le p(1|R), \\ \{(L,2),(R,1)\}, & p(1|L) > p(1|R). \end{cases} \tag{3}$$

식 (3)에 의해 만일 $p(1|L) \le p(1|R)$이라면 $E=\{(L,1),(R,2)\}$가 될 것이다. 그런데

$\sum_{j=1}^{2} p(j|L) = \sum_{j=1}^{2} p(j|R) = 1$ 이므로 $p(2|L) \geq p(2|R)$이 된다. p_1과 p_2는 주어진 값이므로 오차칸 E를 갖는 규칙 R은 오차칸 $\{(L,2),(R,1)\}$을 갖는 규칙에 비해 큰 예측성공측도 값을 갖는 것은 너무나 명백하다. 따라서 주어진 분할 s를 위한 분할기준 식 (2)는 식 (3)을 통해 E를 결정할 수 있으므로 한번의 계산만으로 구할 수 있다.

주어진 마디에서 고려할 수 있는 모든 가능한 분할 중에서 서로 배반(mutually exclusive)인 하위마디를 만들어 내는 분할을 선택하게 하는 것은 분할기준이 갖추어야할 좋은 특성 중에 하나이다. Taylor와 Silverman(1993)은 이러한 특성을 충족시키기 위해서는 다음 두 조건을 만족하여야 하며 이를 배타적 선호 특성(exclusive preference property)이라고 정의하였다.

i) 주어진 $p_L p_R$에서 $\sum_j p(j|L)p(j|R) = 0$이면 최대값을 가져야 한다.

ii) 모든 j에 대하여 $p(j|L) = p(j|R) = p_j$에서 최소값을 가져야한다.

제안된 분할기준 (2)는 배타적 선호 특성이 갖추어야 할 두 조건을 모두 만족한다. 이러한 사실을 밝히기 위하여 주어진 마디 t와 분할 s에 대하여 식 (2)는 (3)에 의한 E에 의해 다음과 같이 다시 쓰기로 한다.

$$M(s,t) = 1 - \frac{p_L \sum_{j=1}^{2} \delta_{Lj} p(j|L) + p_R \sum_{j=1}^{2} \delta_{Rj} p(j|R)}{p_L \sum_{j=1}^{2} \delta_{Lj} p_j + p_R \sum_{j=1}^{2} \delta_{Rj} p_j}$$

$$= \frac{p_L \sum_{j=1}^{2} \delta_{Lj} p_j}{p_L \sum_{j=1}^{2} \delta_{Lj} p_j + p_R \sum_{j=1}^{2} \delta_{Rj} p_j} \left(1 - \frac{\sum_{j=1}^{2} \delta_{Lj} p(j|L)}{\sum_{j=1}^{2} \delta_{Lj} p_j} \right) +$$

$$\frac{p_R \sum_{j=1}^{2} \delta_{Rj} p_j}{p_L \sum_{j=1}^{2} \delta_{Lj} p_j + p_R \sum_{j=1}^{2} \delta_{Rj} p_j} \left(1 - \frac{\sum_{j=1}^{2} \delta_{Rj} p(j|R)}{\sum_{j=1}^{2} \delta_{Rj} p_j} \right) \quad (4)$$

목표변수의 범주가 두 개인 경우에 오차칸들의 집합 $E=\{(L,1),(R,2)\}$이거나 혹은 $E=\{(L,2),(R,1)\}$인 것을 상기하자. 이들 오차칸에 속한 비율이 $p(1|L) = p(2|R) = 0$ 혹은 $p(2|L) = p(1|R) = 0$이면 $\sum_{j=1}^{2} p(j|L)p(j|R) = 0$이 되고 식 (4)는 최대값 $M(s,t) = 1$을 갖는다.

만일 $j=1,2$에 대하여 $p(j|L) = p(j|R) = p_j$ 이면 E에 속한 $p(j|L)$과 $p(j|R)$ 역시 p_j와 같은 것을 의미하므로 식 (4)는 $M(s,t) = 0$인 최소값을 갖는다. 즉, 두 번째 조건 역시 만족하며 제안된 분할기준 (2)는 배타적 선호 특성을 만족하는 것을 알 수 있다.

3. 다범주로의 확장

2절에서 정의된 예측성공측도를 $J>2$인 다범주 문제로 확장하기로 한다. 주어진 마디 t와 분할 s에서 목표변수가 범주 $C=\{1,2,...,J\}$를 갖는다고 하면 다음과 같은 $2\times J$ 분할표를 구성할 수 있다.

	1	2	⋯	J	
L	p_{L1}	p_{L2}	⋯	p_{LJ}	p_L
R	p_{R1}	p_{R2}	⋯	p_{RJ}	p_R
	p_1	p_2	⋯	p_J	

<표 2> 분할 s에 의한 마디 t에서의 $2\times J$ 분할표

이러한 상황에서 Breiman, Friedman, Olshen과 Stone(1984)이 제안한 다음과 같은 초범주(super class)를 고려하기로 한다.

$$C_1=\{j_1,j_2,...,j_m\},\ C_2=C-C_1,\ m=1,2,...,J-1\ .$$

여기서 $C=C_1\cup C_2$ 그리고 $C_1\cap C_2=\phi$이다. 이들에 의해 <표 2>는 <표 3>과 같은 2×2 분할표로 재구성할 수 있다.

	C_1	C_2	
L	p_{LC_1}	p_{LC_2}	p_L
R	p_{RC_1}	p_{RC_2}	p_R
	p_{C_1}	p_{C_2}	1

<표 3> 주어진 분할 s에서 두 초범주에 의한 2×2 분할표

<표 3>에서 $i\in\{L,R\}$에 대하여 $p_{iC_k}=\sum_{j\in C_k}p_{ij}$, $\sum_{k=1}^{2}p_{iC_k}=p_i$ 그리고 $j=1,2,...,J$인 각 목표변수들의 범주비율 p_j에 대하여 $p_{C_k}=\sum_{j\in C_k}p_j$이다. <표 3>에서 볼 수 있듯이 초범주가 주어진 경우에 분할 s의 분할 적합도를 구하는 문제는 목표변수의 범주가 C_1과 C_2의 두 범주를 가진 것과 같은 문제가 된다. 그러므로 <표 3>의 각 칸 값은 C_1에 의해 결정되므로 예측성공측도 식 (1)은 다음과 같이 표현할 수 있다.

$$M_{R_l}(s, t, C_1) = 1 - \frac{p_L \sum_{k=1}^{2} \delta_{LC_k} p(C_k|L) + p_R \sum_{k=1}^{2} \delta_{RC_k} p(C_k|R)}{p_L \sum_{k=1}^{2} \delta_{LC_k} p_{C_k} + p_R \sum_{k=1}^{2} \delta_{RC_k} p_{C_k}}, \quad l=1,2, \tag{5}$$

여기서 δ_{iC_k}는 칸 (i, C_k)이 오차칸임을 나타내는 지시변수이다.

주어진 분할 s에서 목표변수가 J개 범주를 가지고 있으므로 범주 집합 C로부터 생성 가능한 서로 다른 원소를 갖는 두 범주집합은 2^{J-1}개가 존재한다. 이러한 사실은 주어진 분할 s의 분할 적합도를 측정하기 위해서는 2^{J-1}개 범주집합에 대하여 식 (5)를 얻어야 하는 것을 의미한다.

이러한 계산문제를 해결하기 위하여 주어진 분할 s에 의해 구성된 <표 2>에 대하여 Breiman, Friedman, Olshen과 Stone(1984)이 Twoing 기준을 위해 제안한 다음과 같은 초범주 결정을 위한 경험적 기준을 고려하기로 한다.

$$C_1^* = \{ j : p(j|L) \leq p(j|R) \} \tag{6}$$

분할 s가 주어진 것은 p_L과 p_R이 주어진 것을 의미하며 또한 목표변수의 비율 p_j들은 결정된 값이므로 식 (6)의 초범주 C_1^*가 결정되면 $p(C_1^*|L) \leq p(C_2^*|R)$인 관계를 만족한다. 더욱이 $\sum_{k=1}^{2} p(C_k^*|L) = \sum_{k=1}^{2} p(C_k^*|R) = 1$ 이므로 $p(C_2^*|L) \geq p(C_2^*|R)$ 이다. 결국 식 (6)에 의해 결정된 초범주 C_1^*에 따른 $E=\{(L, C_1^*),(R, C_2^*)\}$는 식 (3)을 만족하며, E에 의한 규칙 R는 오차칸 $\{(L, C_2^*),(R, C_1^*)\}$을 갖는 규칙에 비해 큰 예측성공측도 값을 갖게 된다.

이러한 사실에 근거하여 $J>2$인 다범주 문제에서 분할 s의 분할 적합도를 평가하기 위한 다음과 같은 분할기준을 제안하기로 한다.

$$M(s, t, C_1^*) = 1 - \frac{p_L \sum_{k=1}^{2} \delta_{LC_k^*} p(C_k^*|L) + p_R \sum_{k=1}^{2} \delta_{RC_k^*} p(C_k^*|R)}{p_L \sum_{k=1}^{2} \delta_{LC_k^*} p_{C_k^*} + p_R \sum_{k=1}^{2} \delta_{RC_k^*} p_{C_k^*}} \tag{7}$$

이제 식 (7)에 의하여 마디 t에서 고려 할 수 있는 모든 가능한 분한 s에 대한 분할 적합도를 평가할 수 있으므로 결국 주어진 마디 t에서 식 (6)에 의해 결정된 초범주 C_1^*에 의한 최적 분할은

$$M(s^*, t, C_1^*) = \max_{s \in S} M(s, t, C_1^*)$$

을 만족하는 s^*으로 결정할 수 있다.

다범주를 위한 제안된 분할기준 식 (7) 역시 배타적 선호 특성을 갖는다. 먼저 E에 속한 칸들에 대하여 $p(C_1^*|L)=p(C_2^*|R)=0$이라면 $\sum_{k=1}^{2} p(C_k^*|L)p(C_k^*|R)=0$이 된다. 이때 최대값 $M(s,t,C_1^*)=1$을 가지므로 첫 번째 조건을 만족한다. 또한 $p(C_k^*|L)=p(C_k^*|R)=p_{C_k}$이면 $M(s,t,C_1^*)=0$인 최소값을 갖고 이는 두 번째 조건을 만족하는 것을 의미한다.

4. 배타적 선호특성 및 적용 예

2절과 3절에서 예측성공측도를 통하여 최적분할을 선택케 하는 제안된 분할기준은 Gini지수가 갖고 있지 못한 배타적 선호특성을 만족하는 것을 보였다. 이 절에서는 먼저 Shih(1999)에서 수행한 것과 같은 가상자료를 통하여 제안된 분할기준이 배타적 선호특성을 만족하는 것을 보이고자 하며, 또한 Gini지수와 Twoing기준과의 비교를 위하여 Shih가 인용한 Merz와 Murphy(1996)에서 발췌한 와인인식자료(wine recognition data)의 분석 결과를 제시하기로 한다.

4.1. 배타적 선호특성

Taylor와 Silverman(1984) 그리고 Shih(1999)는 Gini 기준으로 불리우는 불순도 감소(the decrease in impurity) 기준은 배타적 선호특성을 만족하지 않지만 그들이 제안한 MPI 기준과 Read와 Cressie(1988)의 통계량에 의한 분할기준이 배타적 선호특성을 만족하는 것을 경험적으로 설명하기 위하여 $J=4$인 두 분할 s_1과 s_2에 따른 <표 4>와 같은 가상자료를 제시하였다.

	1	2	3	4
L	40	20	0	0
R	0	0	10	10

(가) 분할 s_1

	1	2	3	4
L	40	3	10	7
R	0	17	0	3

(나) 분할 s_2

<표 4> 두 분할 s_1과 s_2에 따른 가상자료

<표 4>에서 분할 s_1이 분할 s_2에 비해 보다 바람직한 분할이라는 것은 너무도 명확한다. 그러나 이들 두 분할에 대한 여러 분할기준들에서 얻어진 결과를 정리한 <표 5>에서 볼 수 있듯이 Gini 기준은 s_2를 더 적합한 분할로 선택하게 한다. 하지만 MPI 기준과 Shih가 제안한 분할기준의 특수한 경우인 피어슨 X^2 그리고 우도비에 의한 G^2 기준은 분할 s_1을 선택하게 한다. 여기서 G^2 기준은 엔트로피 기준(entropy criterion)으로 불리우기도 한다.

이제 분할 s_1에 대하여 식 (6)에 의한 초범주 $C_1^*=\{3,4\}$가 결정되고 이때 $E=\{(L,C_1^*),(R,C_2^*)\}$이므로 $M(s_1,t,C_1^*)$=1.0이다. 그러나 분할 s_2에 대하여 $C_1^*=\{2,4\}$이므로

	Gini	MPI	X^2	G^2	M
분할 s_1	0.1979	0.1875	80.0000	89.9736	1.0000
분할 s_2	0.2087	0.1293	55.1999	60.8479	0.7143

<표 5> 가상자료에 의한 여러 분할기준의 값

$M(s_2, t, C_1^*)=0.7143$이다. 즉, 제안된 분할기준 (7) 역시 s_1을 선택하게 하므로 배타적 선호특성을 만족하는 것을 확인할 수 있다.

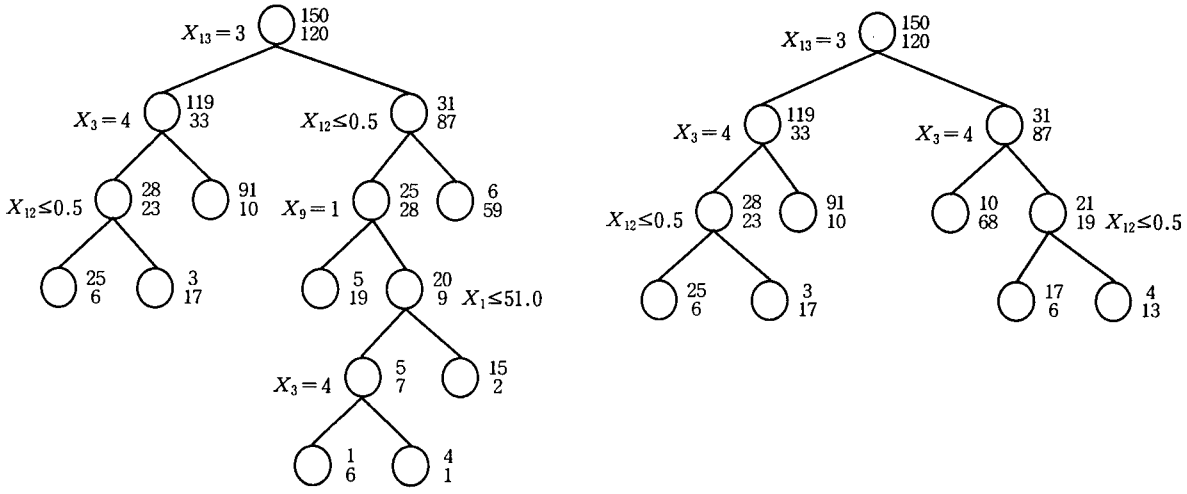
4.2. 심장질환자료

Merz와 Murphy(1996)에서 인용한 심장질환(heart disease) 자료는 심장질환의 유·무를 예측하기 위하여 270명의 환자들로부터 13가지 항목의 검진을 실시한 결과이다. 13개 입력변수는 7개의 연속형 변수와 6개 범주형 변수로 구성되었다.

이 자료를 이용하여 심장질환의 유·무를 예측하기 위하여 Gini 기준과 제안된 분할기준 M에 의한 이진의사결정나무 분석결과가 <표 6>에 주어져 있다. 각 분할기준에 의한 나무는 잎(leaf)에 세개 관찰값만이 속할 때까지 성장시킨 후에 Breiman, Freidman, Olshen과 Stone(1984)이 제안한 cost-complexity 방법을 통하여 가지치기를 수행하였다. 가지치기에 의한 각 나무의 오분류률은 10-folds 교차타당성(cross validation) 방법에 의해 추정하였다. <표 6>에서 $|T_k|$는 가지치기 수행에 의한 나무의 잎의 수를 나타내며 $R^{CV}(T_k)$는 추정 오분류률을 나타낸다. 또한 1-SE 방법에 의해 선택된 나무는 잎의 수에 별 표시가 되어있다. 선택된 최종나무는 <그림 1>에서 볼 수 있다. 그림에서 각 마디 옆의 두 값은 각각 해당마디에서의 목표변수의 범주도수를 나타낸다. 예를 들어 뿌리마디는 목표변수의 범주가 '1'인 도수가 150이며 '2'인 도수가 120인 것을 나타낸다. 이는 전체 270명의 환자중에서 150명의 환자가 심장질환을 앓고 있는 것을 의미한다.

	Gini		M	
k	$\|T_k\|$	$R^{CV}(T_k)$	$\|T_k\|$	$R^{CV}(T_k)$
1	30	0.2333	32	0.2555
2	28	0.2185	29	0.2407
3	23	0.2185	23	0.2222
4	16	0.2000	15	0.2111
5	13	0.2074	9	0.2037
6	11	0.1888	6*	0.2148
7	8*	0.2148	4	0.2407
8	6	0.2852	2	0.3037
9	4	0.3889	1	0.4444
10	2	0.4444		
11	1	0.4444		

<표 6> 심장질환자료의 Gini 기준과 제안된 기준에 의한 분석 결과



(가) Gini 기준에 의한 나무 (나) 제안된 M 에 의한 나무

<그림 1> 심장질환자료에 대한 1-SE 방법에 의해 선택된 나무

우선 <표 6>으로부터 1-SE 방법에 의해 선택된 두 기준에 의한 나무의 오분류률이 같게 추정된 것을 알 수 있다. 그러나 제안된 분할기준 M에 의한 나무가 너 적은 잎을 가진 것을 알 수 있으며 따라서 보다 간결한 구조를 가진 것을 <그림 1>에서 볼 수 있다. 한가지 주목할 점은 Gini 기준에 의한 <그림 1>의 (가)에서 뿌리마디의 첫 번째 오른쪽 하위마디에서의 분할만에 의한 규칙은 majority rule에 의하면 분할을 위해 선택된 변수 X_{12}가 0.5보다 작거나 큰 것에 상관없이 모두 목표변수의 범주 '2'로 예측한다는 사실이다. 그러나 (나)의 나무에서는 Gini 기준과는 달리 변수 X_3을 선택하여 분할에 의한 결과로부터 얻어진 규칙은 왼쪽 마디와 오른쪽 마디에 각각 범주 '2'와 범주 '1'로 예측하도록 하는 것을 알 수 있다.

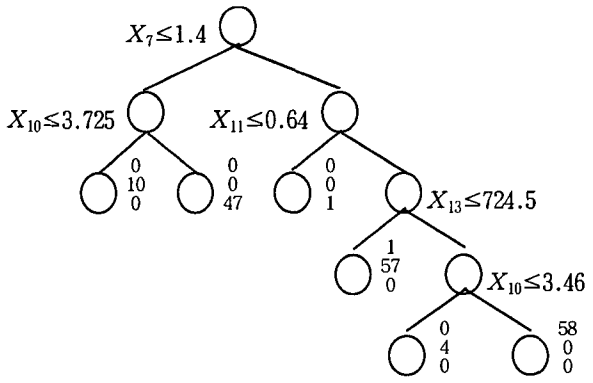
4.3. 와인인식자료

다범주 적용 예를 위해 Merz와 Murphy(1996)에서 인용한 와인인식자료(wine recognition data)는 이탈리아의 동일한 한 지방에서 생산된 세 가지 종류의 와인에 대하여 각 와인이 가지고 있는 특성을 분석하기 위하여 조사된 자료이다. 자료는 13개 연속형 변수에 대하여 와인 '1', 와인 '2' 그리고 와인 '3'에 대하여 각각 59, 71, 48번 측정한 전체 178개 관찰값으로 구성되어 있다.

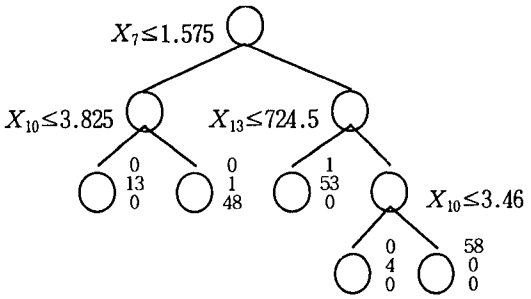
이 자료에 대하여 Gini, MPI, 피어슨 X^2, 우도비에 의한 G^2 그리고 제안된 분할기준 M에 의한 이진의사결정나무 분석결과가 <표 7>에 주어져 있다. 각 분할기준에 의한 나무는 잎에 한개 관찰값만이 속할 때까지 성장시킨 후에 심장질환자료와 마찬가지로 cost-complexity 방법을 통하여 가지치기를 수행하였다. 가지치기에 의한 각 나무의 오분류률 역시 10-folds 교차타당성 방법에 의해 추정하였으며, 1-SE 방법에 의해 선택된 나무는 잎의 수에 별 표시가 되어있다.

k	Gini		MPI		X^2		G^2		M	
	$\|T_k\|$	$R^{CV}(T_k)$	$\|T_k\|$	$R^{CV}(T_k)$	$\|T_k\|$	$R^{CV}(T_k)$	$\|T_k\|$	$R^{CV}(T_k)$	$\|T_k\|$	$R^{CV}(T_k)$
1	14	0.1011	15	0.1067	8	0.0674	8	0.0787	11	0.0618
2	10	0.1011	11	0.1067	6*	0.0618	6	0.0730	7	0.0618
3	8*	0.1124	10*	0.1011	5	0.0899	5*	0.0674	6*	0.0618
4	5	0.1854	7	0.1461	4	0.1348	4	0.1292	5	0.1011
5	4	0.2865	5	0.2247	2	0.3427	3	0.2697	4	0.2978
6	3	0.3371	4	0.2472	1	0.6011	2	0.5056	2	0.5112
7	2	0.4775	3	0.4101			1	0.6011	1	0.6011
8	1	0.6011	2	0.6011						
9			1	0.6011						

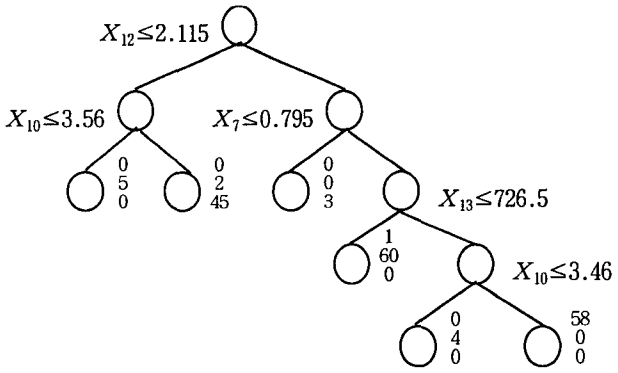
<표 7> 와인인식자료의 여러 분할기준에 의한 분석 결과



(가) X^2에 의한 나무



(나) G^2에 의한 나무



(다) 제안된 M에 의한 나무

<그림 2> 와인인식자료에 대한 X^2, G^2 그리고 제안된 M에 의한 나무

<표 7>에서 1-SE 방법에 의하여 선택된 최종 나무의 교차타당성 방법에 의한 오분류률은 X^2과 제안된 분할기준 M을 통해 성장된 나무가 가장 작은 값을 가진다는 것을 알 수 있다. 이에 반해 G^2은 상대적으로 약간 오분류률이 큰 값을 가지지만 X^2과 M에 의한 나무에 비해 다섯 개의 잎을 가진 보다 간결한 나무구조를 갖는 것을 볼 수 있다. Shih(1999)는 X^2과 G^2에 의한 나무가 Gini 혹은 MPI 기준에 의한 나무에 비해 와인인식자료 분석에는 더욱 적합하다고 설명하고 있으며, 이러한 설명은 제안된 분할기준에 대해서도 동일한 설명이 가능하다는 것을 알 수 있다. X^2, G^2 그리고 제안된 M에 의해 생성된 최종 나무의 모양은 <그림 2>에서 볼 수 있다.

5. 결론

본 연구에서는 이진의사결정나무의 성장을 위하여 주어진 마디에서 가장 적절한 분할을 선택하기 위한 새로운 분할기준을 제안하였다. 제안된 분할기준은 주어진 분할이 목표변수의 범주를 얼마나 잘 예측해주는가를 측정하여 최적분할을 선택하게 한다. 따라서 제안된 분할기준은 목표변수의 범주를 가장 잘 예측하게 하는 예측문제에 유용하게 적용될 수 있을 것이다.

분할기준이 갖추어야 할 바람직한 특성 중에서 제안된 분할기준은 Talor와 Silverman (1993)이 제안한 배타적 선호 특성을 만족한다. 그러나 4절에서 수행한 제안된 분할기준을 통한 분석에서는 전체탐색(exhaustive search)을 통하여 최적분할을 선택하였으므로 기존의 분할기준들이 가지고 있는 변수선택편의(variable selection bias) 문제를 가질 수 있다.

참고문헌

[1] 전병환, 김창수, 송홍엽, 김재희(1997). A new criterion in selection and discretization of attributes for generation of decision trees, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19, 1371-1375.

[2] Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J.(1984). *Classification and Regression Trees*, Chapman & Hall.

[3] Clark, L. A. and Pregibon, D.(1992). Tree-based models, In: J. M. Chambers and T. J. Hastie (eds) *Statistical Models in S*, Wadsworth & Brooks/Cole.

[4] Goodman, L. A. and Kruskal, W. H.(1954). Measures of association for cross-classifications , *Journal of the American Statistical Association*, 49, 732-764.

[5] Hildebrand, D. K., Laing, J. D., and Rosenthal , H. L.(1974a). Prediction logic: A method for empirical evaluation of formal theory, *Journal of Mathematical Sociology*, 3, 163-185.

[6] Hildebrand, D. K., Laing, J. D., and Rosenthal , H. L.(1974b). Prediction logic and quasi-independence in empirical evaluation of formal theory, *Journal of Mathematical Sociology*, 3, 197-209.

[7] Kass, G. V.(1980). An exploratory technique for investigation large quantities of categorical data, *Applied Statistics*, 29, 119-127.

[8] Merz, C. J. and Murphy, P. M.(1996). *UCI Repository of Machine Learning Databases*, Department of Information and Computer Science, University of California, Irvine, CA.

[9] Mola, F. and Siciliano, R.(1997). A fast splitting procedure for classification trees, *Statistics and Computing*, 7, 209-216.

[10] Quinlan, J. R.(1993). *C4.5: Programs for Machine Learning*, Morgan Kaufmann.

[11] Read, T. R. C. and Cressie, N. A. C.(1988). *Goodness-of-Fit Statistics for Discrete Multivariate Data*, Springer-Verlag.

[12] Shih, Y. S.(1999). Families of splitting criteria for classification tree, *Statistics and Computing*, 9, 309-415.

[13] Shih, Y. S.(2001). Selecting the best splits for classification trees with categorical variables, *Statistics & Probability Letters*, 54, 341-345.

[14] Taylor, P. C. and Silverman, B. W.(1993). Block diagrams and splitting criteria for classification trees, *Statistics and Computing*, 3, 147-161.

[2002년 4월 접수, 2002년 6월 채택]