The Korean Communications
in Statistics Vol. 11 No. 3, 2004
pp. 553-562

A Study on the Bias Reduction in Split Variable Selection in CART[1)]

Hyo-Im Song[2)], Eun-Tae Song[3)], and Moon Sup Song[4)]

Abstract

In this short communication we discuss the bias problems of CART in split variable selection and suggest a method to reduce the variable selection bias. Penalties proportional to the number of categories or distinct values are applied to the splitting criteria of CART. The results of empirical comparisons show that the proposed modification of CART reduces the bias in variable selection.

Keywords : classification tree, variable selection bias, impurity measure, CART

1. 서론

분류나무(classification tree)는 훈련용 데이터를 이용하여 나무구조 형식의 예측모형을 만들고, 이 모형으로부터 새로운 개체(case)에 대한 예측변수를 이용하여 목표변수를 예측한다. 데이터 마이닝 분야에서 유용하게 사용되는 대표적인 알고리즘은 CART(Breiman, Friedman, Olshen and Stone, 1984)와 C4.5(Quinlan, 1993)라고 할 수 있다.

CART와 C4.5는 주어진 분리기준(splitting criteria)에 따라 훈련용 데이터를 반복적으로 분리하여 나무의 노드(node)를 형성해 나간다. 예를 들어 CART는 Gini 지수와 같은 불순도(impurity) 측도를 이용하며, 불순도의 감소를 최대로 하는 분리방법을 찾는다. 그러나 CART와 C4.5는 모두 변수선택 과정에서 편의가 발생하는 문제가 있다. 변수선택에서 편의란, 목표변수와 동일한 연관성을 갖는 설명변수들은 같은 확률로 분리변수로 선택되어야 함에도 불구하고 변수가 지닌 다른 특성으로 인하여 선택되는 확률이 달라지는 것을 의미한다.

CART에서는 예측변수가 연속형 또는 순서형 변수인 경우에는 별개값(distinct values)이 많은 쪽으로, 범주형인 경우에는 범주의 개수가 많은 쪽으로 편의가 발생하게 된다. 이와 같은 편의는 분류나무의 오분류율에는 큰 영향을 미치지 않으나 분류나무를 이용한 예측모형의 해석을 오해시

---

1) This research was supported in part by the Brain Korea 21 Project.
2) Analyst, KB data system, Seoul 110-070, Korea
Email : small-gakduki@hanmail.net
3) Analyst, Dongbu Insurance Co., Ltd., Seoul 135-280, Korea
Email : songeuntae@hanmail.net
4) Professor, Department of Statistics, Seoul National University, Seoul 151-742, Korea
Email : songms@plaza.snu.ac.kr (Corresponding Author)

키거나 변수의 중요도를 왜곡시키는 문제가 발생할 수 있다. 즉, 설명력이 떨어지는 변수가 설명력이 더 있는 변수에 우선하여 선택된다면 변수의 중요도 해석에 오해가 있을 수 있다.

변수선택 편의에 관한 연구는 많은 저자들에 의해 발표되었다. 예를 들어 Quinlan(1996), Loh and Shih(1997), Kim and Loh(2001), 송문섭 · 윤영주(2001), Dobra and Gehrke(2001), Lee and Song(2002), Shin, Jeong and Song(2003), 정성석 · 김순영 · 임한필(2004) 등에서 분류나무의 편의 문제를 다루었다. Loh and Shih(1997)와 Kim and Loh(2001), Lee and Song(2002), 정성석 · 김순영 · 임한필(2004)에서는 불편성의 성질을 갖게 하기 위하여 분리변수 선택과 분리점 선택을 두 단계로 나누는 방법을 제안하였다. Dobra and Gehrke(2001)는 분리변수 선택에서 편의를 수량화하여 정의하고 분리기준의 p-값을 이용한 분리변수 선택법을 제안하였으나 이 방법은 분리기준의 분포를 이용해야 하므로 실제 사용에서는 근사분포를 구해야 하는 문제점이 있다. Shin, Jeong and Song(2003)에서는 C4.5에서 범주형 변수의 편의문제를 해결하기 위하여 이득비율(gain ratio) 측도에 범주의 개수에 비례하는 벌점(penalty)을 부여하는 방안을 제안하였다.

본 논문에서는 CART에서의 편의문제를 해결하기 위하여 불순도의 감소량에 별개값 또는 범주의 개수에 비례하는 벌점을 부여하는 방안을 제안한다. 본 논문에서 제안된 방법은 Quinlan(1996)과 Shin, Jeong and Song(2003)에서 C4.5의 편의문제를 해결하기 위해 제안된 방법을 CART에 적용시킨 것으로 이해할 수 있다.

2절에서는 CART의 변수선택 편의 문제를 간단히 살펴보고, 3절에서는 본 논문에서 제안된 방법인 수정된 분리기준을 설명한다. 4절에서는 제안된 방법을 기존의 방법과 비교한다.

2. CART의 변수선택 편의 문제

2.1 CART의 전체탐색법

CART(Classification and Regression Tree)는 전체탐색법(exhaustive search method)에 의해 최적의 이진분리법(binary splitting rule)을 찾는다. 즉, 각 노드에서 최적의 이진분리법을 찾기 위하여 연속형 또는 순서형 예측변수에 대하여는 모든 "$x \leq c$?"(c는 연속된 두 데이터의 중앙값) 형태의 분리법에 대한 불순도를 계산하고, 범주형 예측변수에 대하여는 모든 "$x \in A$?"(A는 x의 범주의 부분집합) 형태의 분리법에 대한 불순도를 계산하여, 불순도의 감소가 최대가 되는 분리법을 선택한다.

노드 t에서 불순도를 $i(t)$로 표현하면, 이진분리에 의한 불순도의 감소량은 다음과 같다.

$$\Delta i(s,t) = i(t) - p_R i(t_R) - p_L i(t_L). \tag{2.1}$$

여기에서 분리법 s는 노드 t에 있는 개체(case)를 p_R의 확률로 오른쪽 하위노드인 t_R로 보내고 p_L의 확률로 왼쪽 하위노드인 t_L로 보낸다. CART에서 흔히 사용되는 불순도의 측도는 Gini 지수이다. 노드 t에서 한 개체가 class j에 속할 확률의 재대입 추정치를 $p(j|t)$라고 할 때, Gini 지수

는 $\sum_{i \neq j} p(i|t)p(j|t)$로 정의된다. 따라서 Gini 지수를 이용한 불순도는

$$i(t) = \sum_{i \neq j} p(i|t)p(j|t) = 1 - \sum_{j} p^2(j|t)$$

이며, 불순도 감소량이 최대가 되는 예측변수와 그때의 분리법에 의해 하위노드를 형성한다.

2.2 전체탐색법에서 변수선택 편의

CART의 전체탐색법은 모든 가능한 분리법에 대해 불순도를 계산하고, 이 가운데 최적의 이진 분리를 찾는다. 그러나 분리변수를 선택하는 과정에 있어서 범주의 개수 또는 별개값의 개수가 많은 변수로의 편의가 발생한다. d개의 별개값을 갖는 순서형 예측변수에는 모두 $(d-1)$개의 가능한 분리법이 있고, k개의 범주를 갖는 범주형 예측변수에는 $(2^{k-1}-1)$개의 가능한 분리법이 있다. 만약 한 예측변수는 별개값 개수가 많은 순서형이고 다른 예측변수는 범주의 개수가 적은 범주형이면 순서형 변수에 대하여 더 많은 횟수로 불순도를 계산하므로, 두 변수가 목표변수와 같은 연관성을 갖는다면 순서형 변수가 선택될 기회가 많아진다. 반면에 범주형 변수의 범주 개수가 많아지면 모든 가능한 분리법의 개수는 지수적으로 증가하기 때문에 범주형 변수가 선택될 기회가 더 많아지게 된다. 예를 들어, x_1이 200개의 별개값을 갖는 순서형 변수이고 x_2는 15개의 범주를 갖는 범주형 변수라 하면, x_1에는 $200-1=199$개의 분리방법, x_2에는 $2^{15-1}-1=16,383$개의 분리방법이 존재하며, 각 분리법들에 대한 불순도를 계산한다. 두 변수가 동일한 조건이라면 분리방법이 많은 x_2가 분리변수로 선택될 가능성이 더 크게 되며, 따라서 CART의 전체탐색법은 범주의 개수가 많은 범주형 변수 x_2로의 편의가 존재하게 된다.

이러한 편의의 문제점을 살펴보기 위해 간단한 모의실험을 실시하였다. 본 논문에서 사용한 프로그램은 모두 R로 작성되었다.

목표변수는 $\{1,2\}$의 값을 갖는 이진변수이고, 2개의 설명변수 $\{x_1, x_2\}$는 목표변수와 서로 독립이며, 두 변수의 별개값 개수 또는 범주의 개수에 차이가 나도록 2가지 모형을 생성하였다. 각 경우에 200개의 개체들이 생성되고 목표변수는 1과 2를 임의로 100개씩 배정하였다. 이러한 과정을 500회 반복하여 각 변수가 선택되는 횟수를 구하였다. 따라서 각각의 예측변수들이 대략 50%의 비율로 선택되어야 편의가 없다고 할 수 있다.

첫 번째 모형에서는 x_1은 200개의 별개값을 갖는 연속형 변수이며, x_2는 범주형 변수로 범주의 개수가 5, 10, 15개인 경우에 대하여 모의실험을 실시하였다. 두 번째 모형에서는 x_1과 x_2는 모두 범주형 변수인 경우로서, x_1은 범주의 개수가 5개이고 x_2는 범주의 개수가 각각 5, 10, 15개인 경우에 대하여 모의실험을 실시하였다. 본 논문에서 연속형 변수는 모두 $N(1,3^2)$에서 랜덤하게 생성되었다. 500회의 반복실험에서 각 변수가 선택된 비율은 <표 2.1>과 같다. x_1이 200개의 별개값을 갖고 x_2가 5개의 범주를 갖는 경우에는 x_1으로의 편의가 있었으나, x_1이 5개의 범주를 갖고 x_2가 15개의 범주를 갖는 경우에는 x_2로의 편의가 매우 심각함을 확인할 수 있다.

<표 2.1> 각 변수가 선택된 비율(%)

x_2의 범주의 개수	x_1의 별개값 개수 : 200개		x_1의 범주의 개수 : 5개	
	x_1	x_2	x_1	x_2
5개	70.0	30.0	49.4	50.6
10개	29.0	71.0	20.0	80.0
15개	10.2	89.8	5.6	94.4

3. 수정된 분리기준

CART에서 지금까지 연구된 편의문제 개선안은 전체탐색법 대신에 분리기준을 변수선택과 분리점 선택의 2단계로 나누는 방법이 주류를 이루었다. 그러나 2단계 방법들은 전체탐색법에 비하여 변수선택 편의를 개선하고 계산속도가 빠른 장점이 있으나(Lee and Song, 2002), 효율이 떨어질 수 있는 단점이 있다. 예를 들어 Siciliano and Mola(1998, 2002)에 의하면 목표변수가 이진형일 때, CART에서 불순도의 감소를 최대로 하는 최적 분리(best split)에 사용된 변수가 2단계 방법에서 상위 3개 변수에 속할 확률은 1에 가깝지만 상위 2개 변수에 속할 확률은 0.95까지, 최상위 변수로 선택될 확률은 0.80까지도 떨어짐을 보였다. 따라서 본 논문에서는 전체탐색법을 사용하되 불순도 측도에 벌점(penalty)을 부여하는 방법을 생각하기로 한다.

Quinlan(1996)과 Shin, Jeong and Song(2003)에서는 C4.5의 편의문제를 해결하기 위하여 이득(gain)함수에 별개값 또는 범주의 개수에 비례하는 벌점을 주어 편의를 완화하였다. 이와 같은 개념을 CART에도 적용시킬 수 있을 것이다. 즉, CART에서 분리기준으로 사용하는 불순도 감소량에 별개값 또는 범주의 개수에 비례하는 벌점을 부여하는 방법을 제안한다.

k개의 범주를 갖는 범주형 예측변수에는 $(2^{k-1}-1)$개의 가능한 분리법이 있다. 범주의 개수가 많아짐에 따라 분리법의 개수가 지수적으로 증가하므로 log 변환을 통해 벌점의 크기를 조정하기로 한다. 다만 모의실험에 의하면 밑이 e인 경우보다 10인 경우가 더 잘 적합되므로 $\log_{10}$ 변환을 사용하였다. 또한 개체 하나당 증가하는 불순도를 정의하기 위해 노드 t의 개체수인 N_t로 나누어 사용하기로 한다. 따라서 (2.1)식으로 주어진 노드 t에서의 불순도 감소량은 다음과 같이 조정한다.

$$\Delta i(s,t) = i(t) - \left\{ p_R i(t_R) + p_L i(t_L) + \frac{\log_{10}(2^{k-1}-1)}{N_t} \right\}. \tag{3.1}$$

노드 t에서 d개의 별개값을 갖는 연속형 또는 순서형 예측변수에는 모두 $(d-1)$개의 가능한 분리법이 있고 별개값의 개수가 증가함에 따라 가능한 분리법의 개수는 선형적으로 증가한다. 따라서 범주형 변수와 마찬가지로 연속형 변수에 대해서도 동일한 원리로 불이익을 줄 수 있다. 그러나 실제 모의실험을 하는 과정에서 $(\log_{10}(d-1))/N_t$ 항은 d가 증가함에 따라 지나치게 큰 불이익을 주게 되어 오히려 별개값의 개수가 적은 연속형 변수로의 편의가 발생하였다. 이를 보완

하기 위해 보정을 실시하였다. 즉, 벌점의 양을 $m \times (\log_{10}(d-1))/N_t$로 놓고 m의 값을 1.0부터 0.1씩 감소시키면서 모의실험을 시행하였다. 서로 독립인 두 변수 x_1과 x_2의 범주의 개수가 각각 $(d_1, d_2) = (2, 200)$, $(10, 200)$, $(50, 200)$인 경우에 두 변수가 루트(root) 노드에서 선택된 비율을 구하였으며, 결과는 <표 3.1>과 같다. 각 경우에 노드 크기는 200이고 반복 횟수는 500이다.

<표 3.1> m의 값에 따라 각 변수가 선택된 비율(%)

m	$d_1 = 2$	$d_2 = 200$	$d_1 = 10$	$d_2 = 200$	$d_1 = 50$	$d_2 = 200$
	x_1	x_2	x_1	x_2	x_1	x_2
1.0	67.4	32.6	63.8	36.2	60.8	39.2
0.9	63.0	37.0	60.2	39.8	58.2	41.8
0.8	54.8	45.2	56.4	43.6	56.6	43.4
0.7	**47.0**	**53.0**	**51.0**	**49.0**	**54.8**	**45.2**
0.6	36.6	63.4	46.0	54.0	53.0	47.0

범주의 개수가 변하면서 편의의 정도도 변하고 있지만 $m = 0.7$인 경우에 편의의 개선 상태가 가장 좋은 것으로 판단되어, 연속형 변수에 대한 불순도 감소량은 다음과 같이 조정하기로 한다.

$$\Delta i(s,t) = i(t) - \left\{ p_R i(t_R) + p_L i(t_L) + \frac{\log_{10}(d-1)}{N_t} \times 0.7 \right\}. \tag{3.2}$$

위의 (3.1)식 및 (3.2)식으로 수정된 분리기준을 본 논문에서는 P_CART라고 부르기로 한다. 다음 절에서는 (2.1)식에 의한 CART와 본 논문에서 제안된 P_CART를 비교하였다. 다만 불순도의 계산은 Gini 지수를 사용하였다.

4. 비교연구

3절에서 제안된 P_CART를 원래의 CART와 비교하기 위하여 모의실험을 실시하고, 실제 데이터에서 변수가 선택되는 과정을 비교하였다. 제안된 방법을 2단계 방법과도 비교하기 위해 Lee and Song(2002)에서 제안한 CHITES를 비교연구에 포함시켰다.

CHITES는 변수선택 단계에서는 예측변수와 목표변수 사이의 연관성 검정을 χ^2-검정으로 시행하여 가장 유의성이 있는 예측변수를 분리변수로 선택하고, 분리점 선택 단계에서는 선택된 변수를 이용하여 전체탐색법을 적용하는 방법이다.

4.1 모의실험

첫 번째 모의실험은 편의의 정도를 비교하는 실험이다. 모든 경우에 데이터의 개수는 200개로 하였다. 목표변수는 이진형으로서 예측변수와 무관하게 50%씩 배당되었으며, 예측변수는 3가지

경우를 고려하였다. 첫째는 예측변수 x_1과 x_2가 모두 범주형으로서 범주의 수가 각각 5개, 15개인 경우이고, 둘째는 두 예측변수가 각각 순서형으로서 별개값의 수가 각각 10개, 200개인 경우이다. 셋째는 x_1은 별개값의 수가 200개인 순서형이고 x_2는 범주의 수가 15개인 범주형인 경우이다. 각 경우를 500회씩 반복실험 하였으며, 루트 노드에서 x_1과 x_2가 분리변수로 선택된 횟수를 정리한 결과는 <표 4.1>과 같다. 모든 경우에 CART는 심각한 편의 현상을 보이고 있는 반면에 P_CART는 편의가 완화되어 안정된 결과를 보이고 있다. 편의의 측면에서는 CHITES가 가장 안정적인 결과를 보이고 있으며, 이는 2단계 방법의 장점이 된다.

<표 4.1> 각 변수가 선택된 비율(%)

	x_1 $k=5$	x_2 $k=15$	x_1 $N=10$	x_2 $N=200$	x_1 $N=200$	x_2 $k=15$
CART	6.6	93.4	22.4	77.6	10.4	89.6
P_CART	49.8	50.2	46.8	53.2	58.2	41.8
CHITES	51.4	48.6	49.0	51.0	48.8	51.2

$k=5$와 $k=15$는 범주의 수가 각각 5개, 15개인 범주형 변수를 나타내며,
$N=10$과 $N=200$은 별개값의 수가 각각 10개, 200개인 순서형 변수이다.

두 번째 모의실험은 검정력(또는 변수 선택력)을 비교하는 실험이다. 목표변수는 이진형으로서 예측변수 x_1과는 연관성이 있으며 x_2와는 연관성이 없도록 생성되었다. 따라서 x_1이 선택될 확률(비율)이 클수록 검정력이 높은 방법이다. 예측변수는 3가지 경우를 고려하였다. 첫째는 예측변수 x_1과 x_2가 모두 범주형으로서 범주의 수가 각각 2개, 15개인 경우이고, 둘째는 두 예측변수가 각각 순서형으로서 별개값의 수가 각각 10개, 200개인 경우이다. 셋째는 x_1은 범주의 수가 2개인 범주형이고 x_2는 별개값의 수가 200개인 순서형인 경우이다. 각 경우를 500회씩 반복실험하고, 루트 노드에서 x_1과 x_2가 분리변수로 선택된 횟수를 정리한 결과는 <표 4.2>와 같다. 각 경우에 범주 또는 별개값의 개수가 많은 변수로의 편의 때문에 CART의 검정력이 P_CART보다 떨어짐을 보이고 있다. CHITES와 P_CART는 비슷한 검정력을 보이고 있다.

<표 4.2> 각 변수가 선택된 비율(%)

	x_1 $k=2$	x_2 $k=15$	x_1 $N=10$	x_2 $N=200$	x_1 $k=2$	x_2 $N=200$
CART	39.2	60.8	69.0	31.0	71.4	28.6
P_CART	90.6	9.4	85.0	15.0	90.2	9.8
CHITES	94.4	5.6	81.0	19.0	93.8	6.2

$k=2$와 $k=15$는 범주의 개수가 각각 2개, 15개인 범주형 변수를 나타내며,
$N=10$과 $N=200$은 별개값의 개수가 각각 10개, 200개인 순서형 변수이다.

4.2 실증적 비교

이 절에서는 UCI Repository(Blake and Merz, 1998)에 있는 Pima Indians Diabetes(Pima Indians) 데이터와 Credit Approval(CRX) 데이터를 이용하여 CART와 P_CART에 의해 생성된 분류나무를 비교하고자 한다. 비교를 위하여 참고로 CHITES도 포함시켰다. 데이터의 구성은 <표 4.3>과 같다.

Pima Indians 데이터는 세계건강기구의 기준에 의해 환자가 당뇨병 징후를 보이는지를 조사한 자료이다. 전체 개체수는 768개이며, 예측변수(A1~A8)는 8개의 순서형만으로 이루어져 있고 목표변수는 이진형이다. 그러나 별개값의 개수들이 많이 다르지는 않으므로 CART와 P_CART에 의해 생성된 분류나무는 비슷할 것으로 기대한다. CRX 데이터는 신용카드의 발급 여부를 나타내는 자료로서 개인정보의 비밀을 위해 모든 변수이름은 가변수로 나타내었다. 개체수는 모두 690개이나 결측이 있는 37개를 제외한 653개를 사용하였다. 예측변수(V1~V13)는 연속형과 범주형이 섞여 있고, 특히 V6는 범주의 수가 14개인 범주형인 것이 특징이다. 목표변수는 Pima Indians와 마찬가지로 이진형이다.

<표 4.3> 데이터의 구성

자료명	개체수	예측 변수			
		연속형	별개값 개수	범주형	범주 수
Pima Indians	768	Number of times pregnant(A1)	17		
		Plasma glucose(A2)	136		
		Diastolic Blood pressure(A3)	47		
		Triceps skin fold thickness(A4)	51		
		2-Hour serum insulin(A5)	186		
		Body mass index(A6)	248		
		Diabetes pedigree function(A7)	517		
		Age(A8)	52		
Credit Approval (CRX)	653			V1	2
				V4	3
		V2	340	V5	3
		V3	213	V6	14
		V8	131	V7	9
		V11	23	V9	2
		V14	164	V10	2
		V15	229	V12	2
				V13	3

두 자료에서 CART, P_CART 및 CHITES에 의해 선택되는 변수들을 살펴보기 위해 정지규칙

을 이용한 분류나무를 생성하였다. 정지규칙으로 분류나무의 최대 깊이를 5로 하고, 한 노드에서 분리를 시행하기 위한 최소 개체수는 20개로 하였다. Pima Indians와 CRX의 각각에서 전체 데이터를 랜덤하게 10개의 그룹으로 나누고, 차례로 1개 그룹씩 제외한 9개의 그룹을 훈련용 데이터로 사용하여 분류나무를 만드는 작업을 10회 반복하였다. 이렇게 만들어진 10개의 분류나무에서 각 변수들이 분리변수로 선택된 평균 횟수를 구한 결과 <표 4.4> 및 <표 4.5>와 같았다.

Pima Indians 데이터에서는 전체적으로 각 변수들이 비슷한 비율로 선택되었다. 변수 A1은 다른 변수에 비하여 별개값 개수가 적고 A7은 많은 편이다. 따라서 A1은 P_CART에서 더 많이 선택되었고 A7은 CART에서 더 많이 선택되었음을 확인할 수 있다. CHITES는 A5의 선택에서 CART나 P_CART와 다른 양상을 보이고 있다. 즉, CART나 P_CART에서 최적분리로 선택되지 않은 변수가 CHITES에서는 분리변수로 선택되었을 가능성이 있다. 그러나 전체적으로는 많은 차이가 나지는 않는다.

<표 4.4> Pima Indians 데이터에서 분리변수로 선택된 횟수

변수	A1	A2	A3	A4	A5	A6	A7	A8
CART	1.2	4.9	1.7	0.5	1.8	4.9	3.6	3.3
P_CART	2.1	4.9	2.0	0.4	1.5	4.8	2.6	3.4
CHITES	2.4	3.7	1.3	0.9	3.6	4.3	3.4	2.9

CRX 데이터에서는 범주의 개수가 많은 V6의 선택 횟수가 많이 다른 것으로 나타났다. CART에서는 평균 4.1회 선택되었으나 P_CART에서는 0.7회밖에 선택되지 않았다. 범주의 개수가 3개인 V4는 CART에서는 1.1회 선택되었으나 P_CART에서는 3.6회 선택되었다. 이와 같이 CART와 P_CART에서 사용된 분리변수는 범주의 개수에 따라 다를 수 있음을 확인할 수 있다.

<표 4.5> CRX 데이터에서 분리변수로 선택된 횟수

변수	V1	V2	V3	V4	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15
CART	0	1.5	2.3	1.1	4.1	1.0	1.0	1.0	0.9	0.1	0	0	2.4	2.1
P_CART	0.8	0.2	1.5	3.6	0.7	1.2	1.0	1.0	2.2	0.6	0.3	0	2.5	2.4
DHITES	0.5	0.7	1.6	1.6	1.7	1.6	0.5	1.0	1.2	0.1	0	0.7	3.1	3.2

Pima Indians 데이터와 CRX 데이터에서 전체 데이터를 10개 그룹으로 나눈 것은 실제로 10-겹(10-fold) 교차타당성의 기법을 사용한 것이다. 따라서 9개의 그룹으로 만들어진 분류나무에 나머지 한 그룹을 적용시켜 오분류율을 구하고, 이와 같이 구한 10개 오분류율의 평균은 교차타당성에 의한 오분류율의 추정값이 된다. 10-겹 교차타당성으로 구한 오분류율의 추정값은 <표 4.6>과 같다. CRX 데이터에서는 P_CART가 CART보다 오분류율이 적으나, 전체적으로 크게 차이가 나지는 않는다.

<표 4.6> 교차타당성에 의한 오분류율 추정값

	Pima Indians	CRX
CART	0.255	0.168
P_CART	0.253	0.138
CHITES	0.260	0.139

5. 결론

본 논문에서는 CART에서 변수선택 편의에 관한 문제점을 살펴보고, 한 개선책으로 별개값 또는 범주의 개수에 비례하는 벌점을 부여하는 P_CART를 제안하였다. 새로 제안된 방법은 변수선택 편의를 현저하게 개선할 수 있었다. 그러나 오분류율의 측면에서는 CART와 P_CART가 비슷한 결과를 보이고 있었다. 이는 상위 노드에서 연관성이 약한 변수가 선택되어도 하위 노드에서 다시 연관성이 강한 변수가 사용되어 전체적으로 오분류율에는 큰 영향을 주지 않기 때문이다. 그러나 분류나무의 해석이나 변수의 중요도 등에서는 변수선택 편의가 왜곡된 정보를 제공할 수 있으므로 이에 대한 개선은 바람직한 결과라고 할 수 있다. 그러나 본 논문에서 제안된 방법이 최종적인 해결책은 아니며, 이 분야에 관한 하나의 방법을 제안한 것이다. 따라서 앞으로도 변수선택 편의를 줄이면서 오분류율도 동시에 개선할 수 있는 방법은 계속 연구되어야 할 분야이다.

감사의 글

두 심사위원의 지적과 제안에 의해 처음에 제출했던 논문이 많이 개선될 수 있었음에 감사드립니다.

참고문헌

[1] 송문섭, 윤영주 (2001), 데이터마이닝 패키지에서 변수선택 편의에 관한 연구, 「응용통계연구」, 제14권, 475-486.

[2] 정성석, 김순영, 임한필 (2004), 의사결정나무에서 분리 변수 선택에 관한 연구, 「응용통계연구」, 제17권, 347-357.

[3] Blake, C.L. and Merz, C.J. (1998), UCI repository of machine learning databases (http://www.ics.uci.edu/~mlearn/MLRepository.html), University of California, Department of Information and Computer Science, Irvine, CA.

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

[5] Dobra, A. and Gehrke, J. (2001), Bias correction in classification tree construction, *Proceedings of the Seventeenth International Conference on Machine Learning*, 90-97.

[6] Kim, H. and Loh, W.Y. (2001), Classification trees with unbiased multiway splits, *Journal*

of the American Statistical Association, Vol. 96, 589-604.

[7] Lee, Y.M. and Song, M.S. (2002), A study on unbiased methods in constructing classification trees, *The Korean Communications in Statistics*, Vol. 9, 809-824.

[8] Loh, W.Y. and Shih, Y.S. (1997), Split selection methods for classification trees, *Statistica Sinica,* Vol. 7, 815-840.

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

[10] Quinlan, J.R. (1996), Improved use of continuous attributes in C4.5, *Journal of Artificial Intelligence Research*, Vol. 4, 77-90.

[11] Shin, S.C., Jeong, Y.J. and Song, M.S. (2003), Bias reduction in split variable selection in C4.5, *The Korean Communications in Statistics*, Vol. 10, 627-635.

[12] Siciliano, R. and Mola, F. (1998), On the behaviour of splitting criteria for classification trees, In: Hayashi, C. et al. (Eds.), *Data Science, Classification, and Related Methods*, Splinger, Tokyo, 191-198.

[13] Siciliano, R. and Mola, F. (2000), Multivariate data analysis and modeling through classification and regression trees, *Computational Statistics & Data Analysis*, Vol. 32, 285-301.

[2004년 8월 접수, 2004년 9월 채택]