응용통계연구 제 17권 3호, 2004년, pp. 459-473

변수선택 편향이 없는 회귀나무를 만들기 위한 알고리즘

김진흠 [1] 김민호 [2]

요 약

본 논문에서는 Breiman 등(1984)의 전체탐색법이 갖고 있는 변수선택 편향을 극복할 수 있는 알고리즘을 제안하였다. 제안한 알고리즘은 노드의 분리변수를 선택하는 단계와 그 선택된 변수에 대해서만 이진분리를 위한 분리점을 찾는 단계로 나뉘어져 있다. 예측변수가 연속형 일 때는 스피어만의 순위상관계수에 의한 검정을 수행하고, 범주형 일 때는 크루스칼-왈리스의 통계량에 의한 검정을 수행하여 통계적으로 가장 유의한 변수를 분리변수로 선택하였고 Breiman 등(1984)의 전체탐색법을 그 변수에만 적용하여 노드의 분리기준을 정하였다. 모의실험 연구를 통해 Breiman 등(1984)의 CART와 제안한 알고리즘을 변수선택 편의, 변수선택력과 평균제곱오차 측면에서 서로 비교하였다. 아울러 두 알고리즘을 실제 자료에 적용하여 효율을 서로 비교하였다.

주요용어: 변수선택력; 변수선택 편의; 스피어만의 순위상관계수; 크루스칼-왈리스 검정; CART

1. 서론

Breiman 등(1984)의 CART(Classification And Regression Tree)에서 사용하는 전체탐색법(exhaustive search method)은 변수선택 편향(bias in variable selection)이 심각한 것으로 알려져 있다 (송문섭, 윤영주, 2001; Loh와 Shih, 1997; Lee와 Song, 2002). 다시 말해 전체탐색법은 많은 범주 수를 가진 범주형 변수가 예측변수(predictor variables)에 포함되어 있는 경우 이 범주형 변수가 목표변수(target variable)와 연관이 적어도 분리변수(split variable)로 선택될 가능성이 높은 문제점을 가지고 있다. 이와 같은 편향이 없는 분류나무(classification trees)를 만들기 위해 여러 연구자들은 먼저 노드의 분리변수로 목표변수와 통계적으로 가장 유의하게 연관된 예측변수를 선택하고 그 변수에만 의존하는 최적의 분리점(split point)을 찾아 노드의 분리기준(split rule)을 정하는 방법을 제안하였다 (Loh와 Vanichsetakul, 1988; Loh와 Shih, 1997; Kim과 Loh, 2001; Lee와 Song, 2002). 한편 변수선택 편향이 없는 회귀나무(regression trees)를 만들기 위해 Loh(2002)는 'GUIDE(Generalized, Unbiased Interaction Detection and Estimation)'로 불리우는 알고리즘을 제안하였다.

본 논문에서는 목표변수가 연속형 일 때 변수선택 편향이 없는 회귀나무를 만들 수 있는 알고리즘을 제안하고자 한다. 모의실험 연구를 통해 Breiman 등(1984)의 CART 알고리

1) (445-743) 경기도 화성시 봉담읍 와우리 산 2-2호, 수원대학교 자연과학대학 통계정보학과, 부교수
E-mail: jinhkim@suwon.ac.kr
2) (445-743) 경기도 화성시 봉담읍 와우리 산 2-2호, 수원대학교 자연과학대학 통계정보학과, 석사과정
E-mail: sapire@suwon.ac.kr

즘과 제안한 알고리즘을 변수선택 편향, 변수선택력(variable selection power)과 평균제곱오차(mean squared error; MSE) 측면에서 서로 비교하고자 한다. 아울러 두 알고리즘을 실제 자료에 적용하여 효율을 서로 비교하고자 한다.

2. 제안 알고라즘

노드의 분리기준을 정하기 위해 Breiman 등(1984)이 사용하는 전체탐색법은 분리변수 선택과 분리점 결정을 동시에 하기 때문에 연속형 변수 중에서는 서로 다른 값을 많이 갖는 변수가 분리기준으로 선택될 가능성이 높고, 범주형 변수와 연속형 변수 중에서는 범주수에 따라 가능한 분리가 지수적으로 증가하는 범주형 변수가 분리기준으로 선택될 가능성이 높다. 노드의 분리변수 선택에서 발생하는 이러한 편향과 모든 변수의 탐색으로 발생하는 계산의 복잡성(computational complexity)을 동시에 극복하기 위한 한 방법으로 본 논문에서는 통계적 방법에 의해 분리변수를 선택하는 단계와 선택된 변수에만 전체탐색법을 적용하여 분리점를 찾는 단계로 나누어 노드의 분리기준를 정하는 알고리즘을 제안하고자 한다.

K개의 예측변수 중에서 $X_1, \ldots, X_{K_1}$은 연속형이고 $X_{K_1+1}, \ldots, X_K$는 범주형이라고 하자. 총 개체 수는 N이라고 하자. $y_i (i = 1, \ldots, N)$는 i번째 개체의 목표변수 Y의 관찰 값, $x_{ij} (i = 1, \ldots, N; j = 1, \ldots, K)$는 i번째 개체의 예측변수 X_j의 관찰 값을 나타낸다. 또한, $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iK_1}, x_{iK_1+1}, \ldots, x_{iK})$는 i번째 개체의 예측변수들의 관측 값들로 이루어진 벡터이다. 목표변수와 예측변수의 쌍으로 이루어진 서로 독립인 N 개의 자료는 다음과 같다.

$$(\boldsymbol{x}_i, y_i),\ i = 1, \ldots, N.$$

그리고 노드 t에서 $M_k(t)$는 범주형 예측변수 $X_k (k = K_1 + 1, \ldots, K)$의 범주 수를 나타내고, $N(t)$는 노드 내 개체 수를 나타낸다고 하자.

2.1. 분리변수 선택

두 변수 사이의 독립성을 검정하기 위해 널리 쓰이고 있는 여러 검정법 가운데서 두 변수 모두가 연속형 일 때는 스피어만(Spearman)의 순위상관계수에 기초한 검정, 두 변수 모두가 범주형 일 때는 피어슨(Pearson)의 카이제곱 검정, 하나는 연속형이고 다른 하나는 범주형 일 때는 크루스칼-왈리스(Kruscal-Wallis)의 통계량에 기초한 검정을 생각할 수 있다(Randles과 Wolfe, 1979). Eubank 등(1987), 이승천, 허문열(2003)은 비록 이 세 가지 검정법들이 서로 다른 유형의 변수에 대한 연관성을 측정하고 있지만 같은 기원을 갖고 있기 때문에 각 검정법의 유의확률 값(p값)은 연관성 측도로서 일관성을 갖는다는 점을 밝혔다.

따라서 본 논문에서는 예측변수가 연속형 일 때는 스피어만의 순위상관계수에 의한 검정을 수행하고, 범주형 일 때는 크루스칼-왈리스의 통계량에 의한 검정을 수행하여 통계적으로 가장 유의하게 목표변수와 연관된 예측변수를 노드의 분리변수로 선택하는 규칙을 따르고자 한다. 각 노드에서 분리변수를 선택하는 절차를 구체적으로 설명하면 아래와 같다.

Step 1: [연속형인 경우] 노드 t에서 예측변수 X_k $(k = 1, \ldots, K_1)$와 목표변수 Y 사이의 스피어만의 순위상관계수에 기초한 독립성 검정을 수행하여 p값 즉, $\hat{\alpha}(k)$를 구하고 다음 조건을 만족하는 k의 값을 k_1으로 정의한다.

$$\hat{\alpha}(k_1) = \min_{1 \le k \le K_1} \hat{\alpha}(k).$$

Step 2: [범주형인 경우] 노드 t에서 예측변수 X_k $(k = K_1 + 1, \ldots, K)$와 목표변수 Y 사이의 크루스칼-왈리스의 통계량에 기초한 독립성 검정을 수행하여 p값 즉, $\hat{\alpha}(k)$를 구하고 다음 조건을 만족하는 k의 값을 k_2으로 정의한다.

$$\hat{\alpha}(k_2) = \min_{K_1+1 \le k \le K} \hat{\alpha}(k).$$

Step 3: 예측변수 $X_{k'}$를 분리변수로 선택한다. 단,

$$k' = \begin{cases} k_1, & \text{if } \hat{\alpha}(k_1) \le \hat{\alpha}(k_2) \\ k_2, & \text{if } \hat{\alpha}(k_1) > \hat{\alpha}(k_2). \end{cases}$$

2.2. 분리점 선택

노드 t 내 개체들의 목표변수의 평균을 $\bar{y}(t) = \sum_{(\boldsymbol{x}_i, y_i) \in t} y_i / N(t)$로 정의하고 노드 내 MSE를 다음과 같이 정의하자.

$$s^2(t) = \frac{1}{N(t)} \sum_{(\boldsymbol{x}_i, y_i) \in t} \{y_i - \bar{y}(t)\}^2.$$

또한, 노드 t에서 $M_k(t)$ 개의 범주를 갖는 범주형 예측변수 X_k의 범주 $l = 1, \ldots, M_k(t)$에 대해 그 범주에 속한 개체 수를 $N_l(t) = \sum_{(\boldsymbol{x}_i, y_i) \in t} I(x_{ik} = l)$로 정의하고 목표변수의 평균을 $\bar{y}_l(t) = \sum_{\{(\boldsymbol{x}_i, y_i) \in t, x_{ik} = l\}} y_i / N_l(t)$로 정의하자. 단, $I(\cdot)$은 지시함수이다.

노드의 최적 분리기준은 2.1절의 절차에 따라 선택된 분리변수에만 Breiman 등(1984)의 전체탐색법을 적용하여 정하고자 한다. 2.1절에서 선택된 분리변수 $X_{k'}$의 유형에 따라 분리점를 찾는 방법을 구체적으로 설명하면 아래와 같다.

1. [연속형인 경우]

   Step 1: $X_{k'}$의 관측 값 $x_{1k'}, \ldots, x_{N(t)k'}$를 $x_{(1k')} \le \cdots \le x_{(N(t)k')}$와 같이 순서화 한다.

   Step 2: 각각의 $l = 1, \ldots, N(t) - 1$에 대해 $x_{(lk')} \ne x_{(l+1,k')}$ 일 때 $x_{(1k')}, \ldots, x_{(lk')}$ 값을 갖는 개체는 왼쪽 노드(t_L)로 보내고, $x_{(l+1,k')}, \ldots, x_{(N(t)k')}$ 값을 갖는 개체는 오른쪽 노드(t_R)로 보내어 다음과 같이 MSE의 가중평균을 구한다.

   $$s_l^2(t) = p_L s^2(t_L) + p_R s^2(t_R).$$

   단, $p_L(p_R)$은 $N(t)$ 개의 개체 중 노드 $t_L(t_R)$에 포함된 개체 수의 비율을 나타내다.

Step 3: $\tilde{s}^2(t) = \min_{\{l;\ x_{(lk')} \neq x_{(l+1,k')}\}} s_l^2(t)$와 같이 정의할 때 예측변수 $X_{k'}$의 분리점은 $\tilde{s}^2(t)$에 대응하는 l의 값 즉, $x_{(lk')}$으로 결정한다. 따라서 노드 t의 분리기준은

$$`X_{k'} \leq x_{(lk')}'$$

으로 주어진다.

2. [범주형인 경우]

Step 1: $X_{k'}$의 범주 $1, \ldots, M_{k'}(t)$에 대해 목표변수의 평균 $\bar{y}_1(t), \ldots, \bar{y}_{M_{k'}(t)}(t)$ 를 구하고 이를 $\bar{y}_{(1)}(t) \leq \cdots \leq \bar{y}_{(M_{k'}(t))}(t)$와 같이 순서화 한다.

Step 2: 각각의 $l = 1, \ldots, M_{k'}(t) - 1$에 대해 $\bar{y}_{(l)}(t) \neq \bar{y}_{(l+1)}(t)$ 일 때 범주 $(1), \ldots, (l)$에 속하는 개체는 왼쪽 노드(t_L)로 보내고, 범주 $(l+1), \ldots, (M_{k'}(t))$에 속하는 개체는 오른쪽 노드$(t_R)$로 보내어 다음과 같이 MSE의 가중평균을 구한다.

$$s_l^2(t) = p_L s^2(t_L) + p_R s^2(t_R).$$

Step 3. $\tilde{s}^2(t) = \min_{\{l;\ \bar{y}_{(l)}(t) \neq \bar{y}_{(l+1)}(t)\}} s_l^2(t)$와 같이 정의할 때 예측변수 $X_{k'}$의 분리집합은 $\tilde{s}^2(t)$에 대응하는 $\{(1), \ldots, (l)\}$으로 결정한다. 따라서 노드 t의 분리기준은

$$`X_{k'} \in \{(1), \ldots, (l)\}'$$

으로 주어진다.

2.1절~2.2절의 절차를 따라 노드를 분리해나가는 과정은 노드 내 개체 수가 미리 지정해 놓은 값보다 큰 값을 갖는 한 계속 진행되다가 모든 노드에서 이 정지규칙(stopping rule)이 만족될 때 비로소 끝나게 된다. 이후로는 제안한 알고리즘을 'SKES(Spearman or Kruscal-Wallis test and Exhaustive Search)'으로 약칭하겠다.

3. 두 알고리즘 CART와 SKES의 비교

본 절에서는 모의실험 연구를 통해 변수선택 편향, 변수선택력, MSE 측면에서 CART 알고리즘과 SKES 알고리즘을 서로 비교하고자 한다.

3.1. 변수선택 편향의 비교

변수선택 편향 측면에서 CART와 SKES를 서로 비교하기 위해 모든 예측변수가 목표변수와 서로 독립인 모형을 고려하였다. 목표변수 Y는 표준정규분포로부터 생성하였고, 예측변수군에 포함될 변수 Z, W, U, B, C, BC들은 각각 다음과 같은 분포로부터 생성하였다.

$$Z \sim N(0,1);\ W \sim \text{Exp}(1);\ U \sim \text{Unifrom}\{1,2,3,4\}, \tag{3.1}$$

$$B \sim \text{Unifrom}\{1,2\};\ C \sim \text{Unifrom}\{1,\ldots,M\};\ BC = \begin{cases} 1, & \text{if } C \leq M/2 \\ \text{Uniform}\{1,2\}, & \text{if } C > M/2. \end{cases}$$

단, M은 범주 수를 나타낸다. 변수 B, C, BC는 범주형이고, U는 순서형이다. 위 5개의 변수를 적절히 변환하여 예측변수 $X_1 \sim X_5$를 만들었다. 변환된 형태에 따라 예측변수들 사이의 종속관계가 서로 '독립'인 경우, '약상관'인 경우, '강상관'인 경우로 구분되는 데 구체적인 형태는 표 3.1과 같다. 특히 '약상관'의 경우에는 $\mathrm{Corr}(X_1, X_2) = 0.392$ 이고 '강상관'의 경우에는 $\mathrm{Corr}(X_1, X_2) = 0.995$ 이다. 모의실험에서 고려한 표본의 수는 $N = 200, 500$ 이고, 예측변수 X_5의 범주 수는 $M = 5, 15$ 이다. 예측변수들 사이의 종속관계별로 N, M의 값에 따라 모의실험을 300번 반복수행하였다. 모든 예측변수가 목표변수와 서로 독립이므로 근노드(root node)에서 각 예측변수가 분리변수로 선택될 확률은 이론적으로 0.2이다. 따라서 300번 반복수행으로 추정된 각 변수의 선택확률이 0.2로 부터 크게 벗어난다면 이것은 변수선택 편향을 보여준다고 할 수 있다.

표 3.1: 예측변수 $X_1 \sim X_5$의 구성

예측변수	독립	약상관	강상관
X_1	Z	$U+W+Z$	$W+0.1Z$
X_2	W	W	W
X_3	U	U	U
X_4	B	BC	BC
X_5	C	C	C

표 3.2: 독립모형에서 300번 반복수행으로 추정된 근노드의 변수선택 확률

			독립		약상관		강상관	
N	M	예측변수	CART	SKES	CART	SKES	CART	SKES
200	5	X_1	.380	.147	.377	.210	.400	.167
		X_2	.420	.157	.420	.187	.373	.113
		X_3	.060	.243	.030	.193	.043	.287
		X_4	.017	.243	.013	.216	.030	.213
		X_5	.123	.210	.160	.193	.153	.220
	15	X_1	.103	.187	.093	.167	.097	.160
		X_2	.107	.193	.117	.220	.083	.153
		X_3	.010	.200	.017	.190	.020	.243
		X_4	.003	.200	.007	.213	.003	.197
		X_5	.777	.220	.767	.210	.797	.247
500	5	X_1	.447	.177	.373	.190	.390	.117
		X_2	.377	.243	.420	.187	.363	.207
		X_3	.033	.173	.027	.177	.063	.240
		X_4	.023	.200	.023	.273	.020	.237
		X_5	.120	.207	.157	.173	.163	.200
	15	X_1	.097	.200	.133	.157	.113	.170
		X_2	.127	.207	.137	.220	.083	.143
		X_3	.010	.217	.003	.230	.013	.227
		X_4	.003	.157	.000	.200	.000	.197
		X_5	.763	.220	.727	.193	.790	.263

표 3.2는 두 알고리즘 SKES와 CART의 변수선택 편의를 비교하기 위해 모든 예측변수 $X_1 \sim X_5$가 목표변수와 서로 독립인 모형에 대해 몬테칼로(Monte Carlo) 모의실험을 300번 반복수행하여 추정된 각 예측변수의 선택 비율이다. 표 3.2의 모든 추정 값의 표준오차(se)는 0.023 이다. 표 3.2에서 볼 수 있듯이 CART는 예측변수들 사이의 종속관계에 관계없이 $M = 5$ 일 때는 예측변수 X_1과 X_2를 분리변수로 더 많이 선택하였고, $M = 15$ 일 때는 X_5를 분리변수로 더 많이 선택하였다. 따라서 CART 알고리즘은 노드를 분리할 수 있는 방법의 수가 많은 변수를 분리변수로 선택하려고 하는 변수선택 편향을 지니고 있음을 알 수 있다. 반면에 SKES는 개체 수, 변수 X_5의 범주 수, 예측변수들 사이의 종속관계와 무관하게 모든 예측변수가 분리변수로 선택되는 비율이 몇 가지 경우를 제외하고 0.154~0.246($0.2 \pm 2\text{se}$)의 범위 내에 포함되므로 변수선택 편향이 없는 알고리즘이라 할 수 있다.

3.2. 변수선택력의 비교

변수선택력 측면에서 두 알고리즘 SKES와 CART를 서로 비교하기 위해 모의실험을 수행하였다. 3.1절에서처럼 변수 Z, W, B, U, C, BC들은 각각 (3.1)과 같이 생성하였고 모든 예측변수 $X_1 \sim X_5$는 예측변수들 사이의 종속관계에 따라 표 3.2와 같이 구성하였다. 변수선택력을 비교하기 위해서 모의실험에서는 예측변수 중에서 일부 예측변수만 목표변수와 연관된 모형을 고려하였다. 각 모형에 대해 모의실험에서 고려한 개체 수는 $N = 200,500$ 이고, 예측변수 X_5의 범주 수는 $M = 5, 15$ 이고, 목표변수와 상관된 특정 예측변수의 목표변수와의 상관계수는 $\rho = 0.1, 0.2$ 이다. N, M, ρ의 모든 가능한 8개의 조합 각각에 대해 모의실험을 300번 반복수행하였다. 각 예측변수의 변수선택 확률은 3.1절에서처럼 근노드에서 그 변수가 선택되는 비율로 추정하였고 추정된 각 변수선택 확률의 표준오차는 0.023이다.

3.2.1. 선형모형

목표변수 Y와 예측변수 X_1이 서로 연관된 다음과 같은 선형모형을 고려하였다.

$$Y = cX_1 + \epsilon. \tag{3.2}$$

단, c는 미지의 상수이고 ϵ은 표준정규확률변수이다. 선형모형 (3.2)에서 예측변수들 사이의 종속관계별로 목표변수와 각 예측변수와의 상관계수는 표 3.3와 같다. 표 3.3의 각 칸에 포함된 상수 c의 값은 예측변수들 사이의 종속관계별로 $\rho = \text{Corr}(Y, X_1)$의 값에 따라 결정된다.

표 3.3: 모형 (3.2)에서 목표변수와 각 예측변수와의 상관계수

예측변수	독립	약상관	강상관
X_1	$c/\sqrt{c^2+1}$	$\sqrt{13}c/\sqrt{13c^2+4}$	$\sqrt{101}c/\sqrt{101c^2+100}$
X_2	0	$2c/\sqrt{13c^2+4}$	$10c/\sqrt{101c^2+100}$
X_3	0	$\sqrt{5}c/\sqrt{13c^2+4}$	0
X_4	0	0	0
X_5	0	0	0

표 3.4: 모형 (3.2)에서 300번의 반복수행으로 추정된 근노드의 변수선택 확률

				독립		약상관		강상관	
N	ρ	M	예측변수	CART	SKES	CART	SKES	CART	SKES
200	0.1	5	X_1	.560	.423	.557	.353	.473	.247
			X_2	.290	.153	.310	.193	.363	.247
			X_3	.047	.130	.037	.170	.033	.140
			X_4	.010	.130	.007	.150	.003	.206
			X_5	.093	.163	.090	.133	.127	.160
		15	X_1	.287	.450	.243	.390	.167	.310
			X_2	.077	.117	.097	.163	.143	.240
			X_3	.003	.137	.010	.143	.010	.180
			X_4	.003	.133	.010	.160	.007	.147
			X_5	.630	.163	.640	.143	.673	.123
	0.2	5	X_1	.900	.857	.783	.740	.560	.603
			X_2	.057	.023	.163	.090	.437	.380
			X_3	.010	.037	.033	.110	.000	.003
			X_4	.003	.027	.000	.037	.000	.007
			X_5	.030	.057	.020	.023	.003	.007
		15	X_1	.630	.870	.576	.787	.357	.483
			X_2	.040	.017	.083	.080	.290	.373
			X_3	.003	.040	.030	.077	.003	.043
			X_4	.003	.023	.003	.027	.000	.050
			X_5	.323	.050	.307	.030	.350	.050
500	0.1	5	X_1	.817	.710	.673	.647	.477	.393
			X_2	.120	.060	.250	.110	.453	.330
			X_3	.013	.090	.027	.143	.017	.080
			X_4	.000	.050	.000	.063	.003	.117
			X_5	.050	.090	.050	.037	.050	.080
		15	X_1	.440	.740	.386	.636	.243	.410
			X_2	.060	.053	.113	.116	.263	.340
			X_3	.000	.057	.010	.163	.000	.097
			X_4	.000	.100	.003	.033	.007	.067
			X_5	.500	.050	.487	.053	.487	.087
	0.2	5	X_1	.990	.990	.943	.947	.470	.587
			X_2	.010	.007	.053	.023	.530	.403
			X_3	.000	.000	.003	.030	.000	.003
			X_4	.000	.000	.000	.000	.000	.007
			X_5	.000	.003	.000	.000	.000	.000
		15	X_1	.940	.990	.883	.947	.470	.567
			X_2	.000	.003	.060	.013	.467	.423
			X_3	.000	.003	.003	.040	.000	.010
			X_4	.000	.003	.003	.000	.000	.000
			X_5	.060	.000	.050	.000	.063	.000

표 3.4를 살펴보면 $M = 5, \rho = 0.1$ 일 때 예측변수들 사이의 종속관계에 관계없이 CART가 SKES보다 변수 X_1에 대한 선택력은 더 높은 것으로 나타났다. 그러나 표 3.2에서 살펴보았듯이 CART는 $M = 5$ 일 때 변수 X_1과 X_2으로 변수선택 편향을 띄고 있기 때문에 SKES보다 X_1에 대한 변수선택력이 높다고 쉽게 단정할 수는 없다고 생각된다. $M = 5$ 일 때 ρ가 0.1에서 0.2로 증가하면 예측변수 X_1과 목표변수와의 상관관계가 커지기 때문에 변수 X_1에 대한 선택력이 크게 증가함을 관찰할 수 있다. 한편, $M = 15, \rho = 0.1$ 인 경우 예측변수들 사이의 종속관계에 관계없이 CART는 변수선택 편향에 의한 영향을 그대로 보여주고 있다. 다시 말해 목표변수과 연관된 변수 X_1보다 오히려 큰 범주 수를 갖는 변수 X_5를 근노드의 분리변수로 선택하려는 편향을 보이고 있다. 그러나 이와 같은 편향은 ρ가 0.1에서 0.2로 증가하면 크게 줄어들어 변수 X_1에 대한 선택력이 변수 X_5에 대한 선택력보다 커지게 된다. $M = 15$ 인 경우 SKES는 예측변수들 사이의 종속관계와 ρ의 값에 관계없이 CART보다 항상 변수 X_1에 대한 우수한 선택력을 보여주고 있다. $N = 500$ 일 때 경향은 전체적으로 $N = 200$ 일 때 같지만 $N = 200$ 일 때보다 M과 ρ의 모든 조합에서 예측변수들 사이의 종속관계에 관계없이 변수 X_1에 대한 선택력이 크게 증가함을 알 수 있다.

3.2.2. 이동모형

목표변수 Y와 범주형 예측변수 X_4가 서로 연관된 다음과 같은 이동모형을 고려하였다.

$$Y = cI(X_4 = 2) + \epsilon. \tag{3.3}$$

단, c는 미지의 상수이고 ϵ은 표준정규확률변수이다. 이동모형 (3.3)에서 범주 수 M이 홀수일 때 예측변수들 사이의 종속관계별로 목표변수와 각 예측변수와의 상관계수는 표 3.5와 같다. 표 3.5의 각 칸에 포함된 상수 c의 값은 예측변수들 사이의 종속관계별로 $\rho = \mathrm{Corr}(Y, X_4)$의 값에 따라 결정된다.

표 3.5: 모형 (3.3)에서 각 예측변수와 목표변수와의 상관계수

예측변수	독립	약상관	강상관
X_1	0	0	0
X_2	0	0	0
X_3	0	0	0
X_4	$c/\sqrt{c^2+4}$	$c/\sqrt{c^2+t^{-1}}$†	$c/\sqrt{c^2+t^{-1}}$
X_5	0	$(M^3+M^2-1)c/s$†	$(M^3+M^2-1)c/s$

†$t = (3M^2+2M-1)/(16M^2)$, $s = 16M^2\sqrt{(M^2-1)(tc^2+1)/12}$

표 3.6을 살펴보면 $M = 5, \rho = 0.1$ 일 때 예측변수들 사이의 종속관계에 관계없이 CART는 목표변수와 연관된 변수 X_4보다 오히려 변수 X_1이나 X_2를 근노드의 분리변수로 더 선택하는 것으로 나타났다. 이와 같은 현상은 표 3.2에서 보여준 것처럼 $N = 200, M = 5$ 하에서 변수 X_1, X_2에 대한 변수선택 편향으로부터 비롯되었다고 말할 수 있다. 그러나 이런 편향은 ρ가 0.1에서 0.2로 증가하면 극복되어 변수 X_4에 대한 선택력이 가장 크게 나타났

표 3.6: 모형 (3.3)에서 300번의 반복수행으로 추정된 근노드의 변수선택 확률

				독립		약상관		강상관	
N	ρ	M	예측변수	CART	SKES	CART	SKES	CART	SKES
200	0.1	5	X_1	.300	.120	.327	.073	.373	.047
			X_2	.340	.167	.333	.140	.270	.120
			X_3	.057	.143	.027	.143	.037	.167
			X_4	.150	.460	.153	.507	.153	.517
			X_5	.153	.110	.160	.137	.167	.150
		15	X_1	.117	.137	.077	.127	.063	.067
			X_2	.097	.117	.057	.147	.050	.113
			X_3	.003	.093	.013	.103	.013	.133
			X_4	.060	.520	.047	.490	.047	.530
			X_5	.723	.133	.807	.133	.827	.157
	0.2	5	X_1	.127	.040	.123	.010	.117	.017
			X_2	.163	.030	.133	.023	.117	.007
			X_3	.010	.030	.017	.017	.013	.027
			X_4	.647	.880	.617	.903	.640	.903
			X_5	.053	.020	.110	.047	.113	.047
		15	X_1	.070	.037	.050	.013	.030	.013
			X_2	.047	.033	.030	.030	.027	.023
			X_3	.003	.023	.003	.023	.003	.027
			X_4	.373	.877	.307	.903	.313	.903
			X_5	.507	.030	.610	.030	.627	.033
500	0.1	5	X_1	.310	.070	.267	.050	.243	.043
			X_2	.253	.043	.267	.087	.233	.057
			X_3	.017	.057	.020	.043	.040	.060
			X_4	.330	.750	.317	.717	.333	.730
			X_5	.090	.080	.130	.103	.150	.110
		15	X_1	.090	.063	.080	.037	.057	.037
			X_2	.100	.053	.073	.067	.053	.037
			X_3	.000	.077	.003	.067	.003	.080
			X_4	.183	.733	.143	.770	.153	.787
			X_5	.627	.073	.700	.060	.733	.060
	0.2	5	X_1	.030	.007	.027	.000	.010	.000
			X_2	.007	.000	.017	.000	.017	.000
			X_3	.000	.003	.000	.000	.003	.000
			X_4	.953	.990	.933	.997	.943	.997
			X_5	.010	.000	.023	.003	.027	.003
		15	X_1	.003	.000	.007	.000	.007	.000
			X_2	.010	.000	.007	.000	.003	.000
			X_3	.000	.003	.000	.000	.000	.000
			X_4	.903	.997	.787	.997	.790	.997
			X_5	.083	.000	.200	.003	.200	.003

다. 한편, $M = 15, \rho = 0.1$ 일 때 예측변수들 사이의 종속관계에 관계없이 CART 알고리즘은 목표변수와 상관된 변수 X_4보다 오히려 많은 범주 수를 갖는 변수 X_5를 근노드의 분리변수로 더 선택하고 있다. 이런 편향은 ρ가 0.1에서 0.2로 증가하면 많이 줄어들지만 아직도 변수 X_5에 대한 선택력이 변수 X_4에 대한 선택력보다 크게 유지되고 있다. 반면에 SKES는 M과 ρ의 모든 조합에서 예측변수들 사이의 종속관계에 관계없이 변수 X_4에 대한 선택력이 가장 크게 나타났으며 또한, CART보다 변수 X_4에 대한 선택력이 항상 우수하게 나타났다. $N = 500$ 일 때 경향은 전체적으로 $N = 200$ 일 때 같지만 X_4에 대한 변수선택력이 $N = 200$ 일 때보다 크게 증가하는 것으로 나타났으며 $N = 200$ 일 때보다 CART의 편의현상이 극복되는 경우를 찾아볼 수 있다. 그 중에서 가장 두드러지는 경우는 $M = 15, \rho = 0.2$ 일 때 X_5에 대한 변수선택력이 큰 폭으로 감소하고 상대적으로 X_4에 대한 변수선택력이 큰 폭으로 증가하여 변수 X_4에 대한 선택력이 가장 높게 나타났다.

3.3. MSE의 비교

두 알고리즘 SKES와 CART의 효율을 MSE 측면에서 서로 비교하기 위해 모의실험을 수행하였다. 이를 위해 3.1절에서처럼 변수 Z, W, B, U, C, BC들은 각각 (3.1)과 같이 생성하였고 모든 예측변수는 표 3.2와 동일하게 구성하였다. 목표변수 Y와 예측변수 X_1, X_3, X_4가 서로 연관된 다음과 같은 모형을 고려하였다.

$$Y = 0.2X_1 + 0.2X_3 + 0.4I(X_4 = 2) + \epsilon. \tag{3.4}$$

단, ϵ은 표준정규확률변수이다. 모형 (3.4)에서 예측변수들 사이의 종속관계별 목표변수와 각 예측변수와의 상관계수는 표 3.7과 같다. 모의실험에서 고려한 개체 수는 $N = 200, 500$ 이고 변수 X_5의 범주 수는 $M = 5, 15$ 이다. N, M의 서로 다른 4개의 조합 각각에 대해 모의실험을 100번 반복수행했는데 매번 훈련용 표본(training sample)으로부터 회귀나무를 만들고 동일한 조건으로부터 생성된 평가용 표본(test sample)을 그 회귀나무에 적용하여 MSE를 계산하였다. 회귀나무를 만들 때 정지규칙으로 노드 내 개체 수가 총 개체 수의 5%미만일 때 노드의 분리를 멈추는 방법을 사용하였다.

표 3.7: 모형 (3.4)에서 목표변수와 각 예측변수와의 상관계수

M	예측변수	독립	약상관	강상관
5	X_1	.188	.453	.190
	X_2	0	.182	.189
	X_3	.210	.406	.210
	X_4	.188	.166	.173
	X_5	0	.077	.080
15	X_1	.188	.456	.190
	X_2	0	.183	.190
	X_3	.210	.409	.211
	X_4	.188	.162	.167
	X_5	0	.079	.082

표 3.8: 모형 (3.4)에서 300번의 반복 수행으로 추정된 근노드의 변수선택 확률

			독립		약상관		강상관	
N	M	예측변수	CART	SKES	CART	SKES	CART	SKES
200	5	X_1	.500	.237	.857	.753	.270	.103
		X_2	.013	.003	.000	.000	.253	.107
		X_3	.310	.460	.143	.243	.337	.497
		X_4	.170	.300	.000	.000	.137	.287
		X_5	.007	.000	.000	.003	.003	.007
	15	X_1	.313	.287	.830	.790	.213	.173
		X_2	.030	.006	.000	.000	.193	.097
		X_3	.280	.427	.163	.210	.277	.483
		X_4	.130	.277	.003	.000	.073	.243
		X_5	.247	.003	.003	.000	.243	.003
500	5	X_1	.327	.223	.880	.987	.227	.110
		X_2	.000	.000	.000	.000	.190	.077
		X_3	.350	.503	.120	.013	.387	.633
		X_4	.323	.273	.000	.000	.197	.177
		X_5	.000	.000	.000	.000	.000	.003
	15	X_1	.300	.233	.863	.980	.207	.130
		X_2	.000	.000	.000	.000	.227	.063
		X_3	.423	.450	.137	.020	.403	.640
		X_4	.257	.317	.000	.000	.110	.167
		X_5	.020	.000	.000	.000	.053	.000

표 3.9: 모형 (3.4)에서 100번의 반복 수행으로 추정된 MSE의 표본 통계량

			독립		약상관		강상관	
N	M		CART	SKES	CART	SKES	CART	SKES
200	5	m	1.577	1.511	1.633	1.495	1.545	1.477
		s	0.177	0.184	0.189	0.177	0.189	0.210
		r	0.958		0.915		0.955	
		e	4.185		8.451		4.440	
	15	m	1.630	1.453	1.630	1.478	1.567	1.449
		s	0.220	0.166	0.201	0.164	0.172	0.150
		r	0.891		0.907		0.924	
		e	10.859		9.325		7.530	
500	5	m	1.350	1.264	1.354	1.275	1.310	1.251
		s	0.102	0.090	0.124	0.097	0.099	0.089
		r	0.936		0.942		0.954	
		e	6.370		5.835		4.504	
	15	m	1.336	1.252	1.361	1.244	1.325	1.232
		s	0.120	0.099	0.105	0.089	0.099	0.084
		r	0.937		0.914		0.930	
		e	6.287		8.597		7.019	

표 3.8은 모형 (3.4)를 통해 근노드에서 각 변수가 분리변수로 선택될 확률을 추정한 것이다. SKES는 N, M의 어떤 조합에 대해서도 예측변수들 사이의 종속관계에 관계없이 목표변수와 가장 강하게 연관된 예측변수의 변수선택 비율이 가장 높게 나타났다. 다만 '강상관'의 경우 X_1과 X_2가 X_4보다 목표변수와 더 강하게 연관되어 있음에도 불구하고 변수선택 비율이 더 낮게 나타난 것은 X_1과 X_2의 상관계수가 0.995로서 두 변수 사이의 상관이 크기 때문인 것으로 생각된다. 한편, CART도 몇 가지 경우를 제외하고 SKES와 비슷한 결과를 보여주고 있다. 그 중에서 $N = 200, M = 5$일 때 CART는 X_1에 대한 변수선택 비율이 목표변수와 가장 강하게 연관된 X_3에 대한 선택 비율보다 크게 나타났으며, $N = 200, M = 15$일 때는 목표변수와 서로 독립인 변수 X_5를 더 많이 선택하는 편향을 보여주었다. 그러나 개체의 수가 $N = 500$인 경우 CART 알고리즘의 이런 편향은 극복되는 것으로 나타났다. 3.2절의 결과뿐만 아니라 표 3.8의 결과를 통해서 볼 때 SKES는 CART보다 더 우수한 변수선택력을 지니고 있다고 생각된다.

표 3.9는 100개의 평가용 표본으로부터 얻어진 각 알고리즘의 MSE의 평균(m)과 표준편차(s), 두 알고리즘의 MSE의 평균의 비(r)와 상대평균제곱오차감소량(e)이다. 단, $r = m_S/m_C$ 이고 $e(\%) = (m_C - m_S)/m_C \times 100$ 이다. m_S와 m_C는 각각 SKES와 CART의 MSE의 평균을 나타낸다. 표 3.9를 보면 상대평균제곱오차감소량이 4 ~ 11% 정도로 나타났다. 예측변수들이 서로 독립이고 $N = 200, M = 15$일 때 상대평균제곱오차의 감소량이 가장 크게 나타났는데 그 이유는 비록 예측변수 X_5가 목표변수와 서로 독립일지라도 X_5의 범주수가 많기 때문에 CART는 변수선택 편향을 보여 X_5를 분리변수로 많이 선택하게 되었고 이는 MSE를 크게 한 것으로 생각된다.

4. 실제 자료에의 적용

두 알고리즘 CART와 SKES를 Statlib(http://lib.stat.cmu.edu)에서 수집한 '타자' 자료와 '투수' 자료, UCI Machine Learning Repository에서 수집한 'MPG' 자료와 '보스톤' 자료에 적용하여 효율을 서로 비교하였다. '타자' 자료는 미국 프로야구 타자 263명에 대한 1987년 연봉과 22개의 예측변수로 이루어져 있다. '투수' 자료는 미국 프로야구 투수 176명에 대한 1987년 연봉과 17개의 예측변수로 이루어져 있다. '보스톤' 자료는 보스톤 주변 506개의 지역에 대한 집값의 중앙값과 13개의 예측변수로 이루어져 있다. 'MPG' 자료는 398가지 차종에 따른 1갤론당 주행거리와 8개의 예측변수로 이루어져 있다. 변수 'Hspo'가 결측값을 갖는 6개의 개체는 분석에서 제외하고 392개의 개체만 분석에 포함하였다. 회귀나무를 만들 때 정지규칙은 3.3절에서처럼 노드 내 개체 수가 총 개체 수의 5%미만이 되면 노드의 분리를 멈추는 방법을 사용하였고 가지치기(pruning)는 수행하지 않았다. Breiman 등(1984)의 교차타당성(10-fold Cross Validation; 10-fold CV) 방법을 써서 회귀나무의 MSE와 그 추정량의 표준오차를 추정하였다.

표 4.1에서 볼 수 있듯이 분석한 모든 데이타 셋에서 SKES의 MSE가 CART의 MSE보다 작았으며 표준오차도 작거나 거의 같게 추정되었다. 또한, 상대평균제곱오차의 감소량도 약 12%~51% 정도로 SKES가 CART보다 더 우수한 것으로 나타났다.

표 4.1: 10-fold CV 방법으로 추정된 실제 자료의 MSE 비교

자료	알고리즘	MSE±1se	r	e
타자	CART	0.639 ± 0.088		
			0.865	13.459
	SKES	0.553 ± 0.089		
투수	CART	1.488 ± 0.291		
			0.491	50.941
	SKES	0.730 ± 0.198		
보스톤	CART	0.060 ± 0.006		
			0.883	11.667
	SKES	0.053 ± 0.006		
MPG	CART	14.439 ± 1.617		
			0.853	14.655
	SKES	12.323 ± 1.236		

5. 결론

본 논문에서는 회귀나무를 구성함에 있어 CART의 전체탐색법이 갖고 있는 변수선택 편향 문제를 개선하기 위해 분리기준을 정하는 방법으로 노드를 분리하기 위한 변수를 먼저 선택한 후 선택된 변수에 대해서만 분리점을 찾는 알고리즘을 제안하였다. 모의실험을 통해 제안한 알고리즘과 CART를 서로 비교하였다. 그 결과 SKES는 CART보다 변수선택 편향이 적고 변수선택력 또한 우수한 것으로 나타났으며, 알고리즘의 효율을 비교하는 모의실험에서도 SKES가 CART보다 매우 우수하게 나타났다. 한편, 여러 실제 자료의 분석에서도 SKES가 CART보다 MSE를 더 작게 추정하였다.

실제 자료를 보면 자료값 중에 결측값을 포함하는 경우가 종종 있는데 결측값 처리에 대한 연구가 앞으로 진행되어야 한다고 생각한다. 의사결정나무(decision trees)의 효율을 증가시키기 위한 한 방법으로 Breiman(1996)이 제안한 BAGGING (Bootstrap AGGregatING)이 널리 사용되고 있다. 모의실험 결과를 제시하지는 않았지만 제안한 알고리즘에 BAGGING 방법을 적용해 본 결과 예상했던 것처럼 BAGGING을 하지 않은 SKES 알고리즘보다 더 작은 MSE를 갖는 회귀나무를 구성할 수 있었지만 근노드에서 변수선택 편향이 발견되었다. 따라서 향후 연구에서는 변수선택 편향이 없으며 동시에 BAGGING처럼 추정 오차를 줄일 수 있는 알고리즘의 개발이 필요하다고 생각한다.

감사의 글

CART 알고리즘의 S-plus 프로그램 소스를 제공해 준 이윤모 박사에게 감사드립니다.

참고문헌

송문섭, 윤영주 (2001). 데이터마이닝 패키지에서 변수선택 편의에 관한 연구, <응용통계연구>, **14**, 475-486.

이승천, 허문열 (2003). 혼합자료에서 독립성 검정에 의한 연관성 측정, <응용통계연구>, **16**, 151-167.

Breiman, L. (1996). Bagging predictors, *Machine Learning*, **24**, 123-140.

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

Eubank, R. L., Lariccia, V. N., and Rosenstein, R. B. (1987). Test statistics derived as components of Pearson's Phi-squared distance measure, *Journal of the American Statistical Association*, **82**, 816-825.

Kim, G. V. and Loh, W. (2001). Classification trees with unbised multiway splits, *Journal of the American Statistical Association*, **96**, 589-604.

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

Loh, W. (2002). Regression trees with unbiased variable selection and interaction detection, *Statistica Sinica*, **12**, 361-386.

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

Loh, W. and Vanichsetakul, N. (1988). Tree-structured classification via generalized discriminant analysis (with discussion), *Journal of the American Statistical Association*, **83**, 715-728.

Randles, R. H. and Wolfe, D. A. (1979). *Introduction to The Theory of Nonparametric Statistics*, John Wiley and Sons, New York.

[2004년 1월 접수, 2004년 3월 채택]

Regression Trees with Unbiased Variable Selection

Jinheum Kim [1] Min-Ho Kim [2]

ABSTRACT

It has well known that an exhaustive search algorithm suggested by Breiman et al.(1984) has a trend to select the variable having relatively many possible splits as an splitting rule. We propose an algorithm to overcome this variable selection bias problem and then construct unbiased regression trees based on the algorithm. The proposed algorithm runs two steps of selecting a split variable and determining a split rule for binary split based on the split variable. Simulation studies were performed to compare the proposed algorithm with Breiman et al.(1984)'s CART(Classification and Regression Tree) in terms of degree of variable selction bias, variable selection power, and MSE(Mean Squared Error). Also, we illustrate the proposed algorithm with real data sets.

Keywords: CART; Kruscal-Wallis test; Regression trees; Spearman's rank correlation coefficient; Variable selection bias; Variable selection power

1) Associate Professor, Dept. of Applied Statistics, University of Suwon, Gyeonggi-Do, 445-743, Korea
E-mail:jinhkim@suwon.ac.kr
2) Graduate Student, Dept. of Applied Statistics, University of Suwon, Gyeonggi-Do, 445-743, Korea
E-mail: sapire@suwon.ac.kr