회귀나무에서 변수선택 편의에 관한 연구

김 민 호*[1], 김 진 흠[2]

요약

Breiman, Friedman, Olshen and Stone(1984)의 전체탐색법에 의한 회귀나무는 상대적으로 많은 분리가 가능한 변수로 분리기준이 정해지는 편의 현상을 갖고 있다. 본 연구에서는 이런 문제점을 해결할 수 있는 알고리즘을 제안하여 변수선택편의가 없는 회귀나무를 만들고자 한다. 제안하는 알고리즘은 노드의 분리변수를 선택하는 단계와 그 선택된 변수에 의해 이진분리를 위한 분리점을 찾는 단계로 구성되어 있다. 예측변수 중에서 목표변수와 가장 밀접하게 연관된 예측변수는 예측변수의 자료의 종류에 따라 스피어만의 순위상관계수에 의한 검정 혹은 크루스칼-왈리스의 통계량에 의한 검정을 수행하여 가장 통계적으로 유의한 변수로 선택하였고, 선택된 변수에만 Breiman et al.(1984)의 전체선택법을 적용하여 분리점을 결정하였다. 모의실험을 통해 변수선택편의, 변수선택력, 그리고 평균제곱오차 측면에서 Breiman et al.(1984)의 CART(Classification and Regression Trees)와 제안한 알고리즘을 서로 비교하였다. 또한, 두 알고리즘을 실제 자료에 적용하여 효율을 서로 비교하였다.

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

1. 서론

CART 알고리즘에서 사용하는 전체탐색법은 변수선택 편의가 심각한 것으로 알려져 있다. 다시 말해 범주의 수가 많은 범주형 변수가 예측변수 중에 포함되어 있을 때 이 범주형 변수가 목표변수와 상관이 없어도 분리변수로 선택될 가능성이 높은 문제점이 있다. 이와 같은 편의 현상이 없는 분류나무를 만들기 위해 여러 연구자들은 먼저 노드의 분리변수로 목표변수와 통계적으로 가장 유의하게 연관된 예측변수를 선택하고 그 변수에만 의존하는 최적의 분리점을 찾아 노드의 분리기준을 결정하는 방법을 제안하였다 (Loh and Vanichsetakul, 1988; Loh and Shih, 1997; Kim and Loh, 2001; Lee and Song, 2002). 한편, 변수선택 편의가 없는 회귀나무를 만들기 위해 Loh(2002)는 'GUIDE'로 불려지는 알고리즘을 제안하였다. 본 연구에서는 목표변수가 연속형 일 때 변수선택 편의가 없는 회귀나무를 만들 수 있는 알고리즘을 제안하고자 한다. 모의실험을 통해 변수선택 편의, 변수선택력, 그리고 평균제곱오차 측면에서 CART와 제안한 알고리즘을 서로 비교하고자 한다. 또한, 두 알고리즘을 실제 자료에 적용하여 효율을 서로 비교하고자 한다.

2. SKES (Spearman or Kruskal-Wallis test and Exhaustive Search) 알고리즘

변수선택 편의가 없는 회귀나무를 만들기 위해 제안하는 알고리즘에서는 각 노드의 분리기준에 대한 결정을 분리변수를 선택하는 단계와 선택된 분리변수의 분리점를 찾는 단계로 나누어 수행하고자 한다. K개의 예측변수 중에서 $X_1, \ldots, X_{K_1}$은 연속형이고 $X_{K_1+1}, \ldots, X_K$는 범주형이라고 가정하고 총 표본의 크기는 N이라고 하자. 목표변수와 예측변수의 쌍으로 이루어진 N 개의 서로 독립인 자료가 다음과 같이 주어져 있다고 가정하자.

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

단, y_i는 i번째 개체의 목표변수 Y의 관찰 값이고, $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iK_1}, x_{iK_1+1}, \ldots, x_{iK})$는 i번째 개체의 예측변수들의 관측 값들로 이루어진 벡터이고, x_{ij}는 i번째 개체의 예측변수

[1](445-743) 경기도 화성시 봉담읍 와우리 산 2-2호 수원대학교 통계정보학과 석사과정
[2](445-743) 경기도 화성시 봉담읍 와우리 산 2-2호 수원대학교 통계정보학과 교수

X_j의 관찰 값이다. 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

목표변수와 가장 밀접한 관계를 갖는 예측변수의 선택은 각 예측변수와 목표변수 사이의 독립성 검정을 수행하여 통계적으로 가장 유의한 예측변수를 해당노드의 분리변수로 정하는 규칙을 따르고자 한다. 각 노드에서 분리변수를 선택하는 절차를 구체적으로 설명하면 아래와 같다.

Step 1: (예측변수가 연속형일 때) 노드 t에서 각 연속형 예측변수 X_k $(k = 1, \ldots, K_1)$와 목표변수 Y 사이의 스피어만의 순위상관계수에 기초한 독립성검정을 수행하여 유의확률 값 $\hat{\alpha}(k)$를 구한다 (Randles and Wolfe, 1979). 각 $\hat{\alpha}(k)$ 로부터 다음 조건을 만족하는 k의 값으로 k_1을 정의한다.

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

(예측변수가 범주형일 때) 노드 t에서 각 범주형 변수 X_k $(k = K_1+1, \ldots, K)$와 목표변수 Y 사이의 크루스칼-왈리스의 통계량에 기초한 독립성 검정을 수행하여 유의확률 값 $\hat{\alpha}(k)$를 구한다 (Randles and Wolfe, 1979). 각 $\hat{\alpha}(k)$ 로부터 다음 조건을 만족하는 k의 값으로 k_2를 다음과 같이 정의한다.

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

Step 2: k'을 다음과 같이 정의할 때 $X_{k'}$를 분리변수로 선택한다.

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

2.2 분리점 선택

노드 t에서 최적의 분리점을 찾기 위한 측도로 평균제곱오차(MSE)를 다음과 같이 정의한다.

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

단, $\bar{y}(t) = \sum_{\boldsymbol{x}_i \in t} y_i / N(t)$ 이다. 분리변수선택 단계에서 선택된 변수 $X_{k'}$의 속성에 따라 분리점를 찾는 방법이 서로 다르다.

1. $X_{k'}$이 연속형일 때는 아래와 같은 절차를 따라서 분리점을 찾는다.

   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 $x_{(lk')} \neq x_{(l+1,k')}$ 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 가중분산을 구한다.

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

   Step 3: $\tilde{s}^2(t) = \min_{\{l;\, x_{(lk')} \neq x_{(l+1,k')}\}} s_l^2(t)$ 이면 노드 t 에서 예측변수 $X_{k'}$ 의 분리점은 $\tilde{s}^2(t)$에 대응하는 l의 값 $x_{(lk')}$로 결정한다.

2. $X_{k'}$이 범주형일 때는 아래와 같은 절차를 따라서 분리점을 찾는다.

Step 1: $X_{k'}$의 범주 $c_1, \ldots, c_{M_{k'}(t)}$ 에 따라 목표변수의 평균 $\bar{y}_{c_1}(t), \ldots, \bar{y}_{c_{M_{k'}(t)}}(t)$ 를 구하고 이를 $\bar{y}_{c_{(1)}}(t) \leq \cdots \leq \bar{y}_{c_{(M_{k'}(t))}}(t)$ 와 같이 순서화 한다.

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

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

Step 3. $\tilde{s}^2(t) = \min_{\{l;\ \bar{y}_{c(l)} \neq \bar{y}_{c(l+1)}\}} s_l^2(t)$ 라고 정의할 때 노드 t에서 예측변수 $X_{k'}$의 분리점은 $\tilde{s}^2(t)$에 대응하는 l의 값 $c_{(l)}$로 결정한다.

각 노드를 분리해나가는 과정은 노드의 크기가 미리 지정해 놓은 값보다 작을 때까지 반복한다. 이 정지규칙을 따라 더 이상 어느 노드에서도 분리를 할 수 없을 때 회귀나무를 증가시키는 과정을 멈춘다.

3. 두 알고리즘 CART와 SKES의 비교

모의실험을 통해 변수선택 편의, 변수선택력, MSE 측면에서 CART와 SKES를 서로 비교하고자 한다.

3.1 변수선택 편의 비교

변수선택 편의 측면에서 CART와 SKES를 서로 비교하기 위해 모든 예측변수가 목표변수와 서로 독립인 모형을 고려하였다. 각 예측변수의 변수선택 확률은 근노드에서 해당 변수가 분리변수로 선택되는 비율로 추정하였다. 목표변수 Y는 표준정규분포로부터 생성하였고 예측변수군에 포함될 변수 Z, W, U, B, C, BC 들은 각각 다음과 같은 분포로부터 생성하였다.

$$Z \sim N(0,1);\ W \sim \text{Exp}(1);\ U \sim \text{Unifrom}\{1,2,3,4\},$$
$$B \sim \text{Unifrom}\{1,2\};\ C \sim \text{Unifrom}\{1,\ldots,M\};\ BC = \begin{cases} 1, & \text{if } C \leq M/2 \\ \text{Uniform}\{1,2\}, & \text{if } C > M/2. \end{cases}$$

단, M은 범주 수를 나타낸다. 변수 B, C, BC 는 범주형이고, U는 순서형이다. 위 5개의 변수를 적절히 변환하여 예측변수 $X_1 \sim X_5$ 를 만들었다. 변환한 형태에 따라 예측변수들 사이의 종속관계가 서로 '독립'인 경우, '약상관'인 경우, '강상관'인 경우로 구분되는 데 구체적인 형태는 표 3.1과 같다.

표3.1 예측변수 $X_1 \sim X_5$의 구성

	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

모의실험에서 고려한 표본의 수는 $N = 200$ 이고, 예측변수 X_5의 범주 수는 $M = 5, 15$ 이다. 예측변수들 사이의 종속관계별로 M의 값에 따라 모의실험을 300번 반복수행하였다. 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 분리변수

표3.2 독립모형에서 $N = 200$일 때 300번 반복수행으로 추정된 변수선택 확률

		독립		약상관		강상관	
M	X_i	CART	SKES	CART	SKES	CART	SKES
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

로 더 많이 선택하였다. 따라서, CART는 노드를 분리하는 가지 수가 더 많은 변수를 분리변수로 선택하려고 하는 변수선택 편의 지니고 있다고 할 수 있다. SKES는 변수 X_5의 범주 수, 예측변수들 사이의 종속관계에 관계없이 5개 모든 예측변수가 분리변수로 선택되는 비율이 몇 가지 경우를 제외하고 0.154~0.246($0.2 \pm 2\text{se}$)의 범위 내에 포함되므로 변수선택 편의가 없는 알고리즘이라 할 수 있다.

3.2 변수선택력 비교

CART와 SKES의 변수선택력을 비교하기 위해 목표변수 Y와 예측변수 X_1이 서로 연관된 다음과 같은 모형을 고려하였다.

$$Y = cX_1 + \epsilon. \tag{1}$$

단, c는 미지의 상수이고 ϵ은 표준정규확률변수이다. 모의실험에서 고려한 표본수는 $N = 200$ 이고, 예측변수 X_5의 범주 수는 $M = 5, 15$ 이고, $\rho = \text{Corr}(Y, X_1) = 0.1, 0.2$ 이다. 예측변수들 사이의 종속관계별로 M, ρ의 각 조합에 대해 모의실험을 300번 반복수행하였다. 각 예측변수의 변수선택 확률은 3.1절에서처럼 근노드에서 해당변수가 선택되는 비율로 추정하였다.

표3.3을 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 편의를 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 편의에 의한 영향을 그대로 보여주고 있다. 목표변수과 연관된 변수 X_1보다 오히려 큰 범주 수를 갖는 변수 X_5를 근노드의 분리변수로 선택하려는 편의를 보이고 있다. 그러나, 이와 같은 편의 현상은 ρ가 0.1에서 0.2로 증가하면 크게 줄어들고 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.

3.3 MSE 비교

CART와 SKES의 효율을 평균제곱오차 측면에서 서로 비교하기 위해 목표변수 Y와 예측변수 X_1, X_3, X_4가 서로 상관된 다음과 같은 모형을 고려하였다.

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

표3.3 모형 (??)에서 $N = 200$일 때 300번의 반복수행으로 추정된 변수선택 확률

			독립		약상관		강상관	
ρ	M	X_i	CART	SKES	CART	SKES	CART	SKES
0.1	5	X_1	.560	.423	.557	.353	.473	.247
		X_2	.290	.153	.310	.193	.363	.247
		X_3	.047	.130	.007	.150	.003	.206
		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	.093	.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	.780	.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

모의실험에서 고려한 표본수는 $N = 200$ 이고, 변수 X_5의 범주 수는 $M = 5, 15$ 이다. 예측변수들 사이의 종속관계별로 M의 값에 따라 모의실험을 100번 반복수행했는데 매번 훈련용 표본으로부터 회귀나무를 만들고 동일한 조건으로부터 생성된 평가용 표본을 그 회귀나무에 적용하여 MSE를 계산하였다. 회귀나무를 만들 때 정지규칙으로 노드 내 개체 수가 총 표본 수의 5%미만 일 때 노드의 분리를 멈추는 방법을 사용하였다.

표3.4는 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. 표3.4를 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 범주 수가 많기 때문에 CART는 변수선택 편의현상을 보여 X_5를 분리변수로 많이 선택하게 되었고 이는 MSE를 크게 하였기 때문이다.

4. MPG 자료에의 적용

UCI Machine Learning Repository에서 수집한 'MPG' 자료에 두 알고리즘 CART와 SKES에 적용하여 효율을 서로 비교하였다. 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 표본 수의 5%미만이 되면 노드의 분리를 멈추는 방법을 사용하였고 가지치기는 수행하지 않았다. 교차타당성(10-fold CV) 방법을 써서 회귀나무의 MSE를 추정하였다 (Breiman et al., 1984). 표 3.5에서 볼 수 있듯이 SKES의 MSE가 CART

표3.4 모형 (??)에서 $N = 200$일 때 100번의 반복수행으로 추정된 MSE의 표본통계량

		독립		약상관		강상관	
M		CART	SKES	CART	SKES	CART	SKES
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	

의 MSE보다 작았으며 표준오차도 작게 나타났다. 또한, 상대평균제곱오차의 감소량은 약 15% 정도이므로 제안한 알고리즘 SKES가 CART보다 더 우수한 회귀나무를 만든다고 생각된다.

표3.5 MPG 자료의 10-fold CV 방법으로 추정된 MSE의 표본통계량

	MSE±1se	r	e
CART	14.439 ± 1.617	0.853	14.655
SKES	12.323 ± 1.236		

References

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

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.