의사결정나무에서 순서형 분리 변수 선택에 관한 연구

김현중[1],송주미[2]

ABSTRACT

지금까지 의사결정나무에서 분리 변수의 선택에 관한 연구는 많았으나, 대부분 연속형 변수와 명목형 변수에 국한되어 왔다. 본 연구에서는 순서형 변수에 주목하여 CART, QUEST, CRUISE 등 기존 알고리즘과 본 연구에서 제안하는 비모수적 접근 방법인 K-S test, Cramer-von Mises test 방법의 변수 선택력을 비교하였다. 그 결과 본 연구에서 제안하는 Cramer-von Mises test 방법이 다른 알고리즘에 비하여, 변수 선택력과 안정성에 있어서 좋은 성과를 보였다.

주요용어 : CART, QUEST, CRUISE, Kolmogorov-Smirnov test, Cramer-von Mises test

1. 서론

의사결정나무 분석은 예측과 분류를 위한 보편적이고 강력한 툴이다. 이해하기 쉽다는 점 외에 분류의 근거를 설명할 수 있다는 장점이 있기 때문이다. 의사결정나무에서 사용되는 알고리즘에는 CHAID (Kass 1980), CART (Breiman, Friedman, Olshen, Stone, 1984), C4.5 (Quinlan 1993), QUEST (Loh and Shih, 1997), CRUISE (Kim and Loh, 2001; Kim and Loh 2003) 등 여러 가지가 있다. 이러한 알고리즘들은 설명변수를 연속형 변수와 명목형 변수로 구분하거나, 연속형 변수를 그룹화하여 명목형 변수로 만들어주기도 한다. 하지만 범주형 변수중에서 순서형(ordinal) 변수는 지금까지 고려의 대상이 아니었으며, 이 경우 연속형 변수선택 방법이나 명목형 변수선택 방법이 순서형 변수에 대신 적용되어 왔다.

본 연구에서는 순서형 분리 변수의 선택에 주목하여 비모수적 방법인 Kolmogorov-Smirnov test, Cramer-von Mises test를 이용한 방법을 제안하고, 기존에 제시되었던 알고리즘인 CART, QUEST, CRUISE와 비교하기로 한다.

2. 분리 변수 선택 알고리즘

목표 변수에 대한 정보를 가지고 있는 변수들과 의미 없는 변수들이 혼합되어 있는 경우에 정보를 가지고 있는 예측 변수가 분리 변수로 선택될 확률을 변수 선택력이라 한다. 좋은 알고리즘은 정보를 가지고 있는 예측 변수의 변수 선택력이 커야 할 것이다. 본 절에서는 기존에 제시된 알고리즘과 비모수적 접근을 이용한 방법에 대해 살펴보기로 한다.

2.1 CART의 변수 선택

CART(Classification And Regression Trees) 알고리즘은 의사결정나무를 형성하는데 있어서 가장 보편적인 알고리즘이다. CART 알고리즘은 이진트리구조로 모형을 형성하고, 각 마디

1) 120-749 서울시 서대문구 신촌동 134 연세대학교 응용통계학과 조교수
2) 120-749 서울시 서대문구 신촌동 134 연세대학교 응용통계학과 석사과정

(node)에서 자식마디(child node)로 분리시 불순도 함수(impurity function)의 감소를 최대화하는 분리기준을 선택한다. t 마디에서 지니계수로 관찰된 불순도는 다음과 같이 정의된다.

$$imp(t) = 1 - \sum_j p^2(j|t)$$

여기서 $p(j|t)$는 t 마디로 분류되었는데 class j에 속할 확률이다. 그리고 왼쪽 마디로 구분될 확률 P_L과 오른쪽 마디로 구분될 확률 P_R은 각각 아래와 같이 측정한다.

$$P_L = \frac{N(t_L)}{N(t)},\ P_R = \frac{N(t_R)}{N(t)},$$

여기서 $N(t)$는 t마디에서의 총 자료 수, $N(t_L)$는 t마디에서 왼쪽으로 분류된 자료 수, 그리고 $N(t_R)$는 t마디에서 오른쪽으로 분류된 자료 수이다. 지니 계수로 측정된 감소된 불순도, 즉 goodness of a split인 $g(s,t)$은 다음과 같다.

$$g(s,t) = imp(t) - P_L \cdot imp(t_L) - P_R \cdot imp(t_R).$$

Goodness of a split을 최대로 만드는 분리점을 각 변수에 관하여 찾은 후, 변수들끼리의 goodness of a split을 또 측정하여, 가장 큰 변수를 분리변수로 채택한다.

2.2 QUEST의 변수 선택

QUEST (Quick, Unbiased and Efficient Statistical Tree)는 CART의 범주형 예측변수로의 bias를 보완하고 있다. 변수 선택과 분리 점 선택을 두 단계로 나누어서 실시하기 때문에, 변수 선택에 있어서 bias가 거의 없다.

순서형 변수에 대해서는 ANOVA test를 실시하여 p-값(p-value)를 구하고, 범주형 변수에 대해서는 독립성 검정을 하여 p-값을 구한다. 이렇게 해서 구한 p-값이 가장 작은 변수를 분리변수로 선택한다.

2.3 CRUISE의 변수 선택

CRUISE (Classification Rule with Unbiased Interaction Selection and Estimation) 알고리즘 또한 QUEST와 마찬가지로 분리 변수의 선택과 분리 점의 선택, 두 단계로 나누어서 시행하기 때문에, 변수 선택에 있어서 bias가 거의 없다. 또한 분리 변수를 선택할 때 교호작용을 고려하지 않는 CRUISE 1D와 교호 작용을 고려하는 2D로 나뉘는데, CRUISE 1D는 QUEST의 변수 선택 알고리즘과 같다. 따라서 여기에서는 CRUISE 2D의 변수 선택 알고리즘만을 다루기로 한다. $X_1, X_2, \cdots, X_{K_1}$을 연속형 예측 변수들이라 하고, $X_{K_1+1}, \cdots, X_K$을 명목형 예측 변수라고 하자.

1. 연속형 변수의 marginal test
   (a) 자료의 사분위수로 자료를 4개의 그룹으로 나누고, y의 class를 행으로, 4개의그룹을 열로 하는 $J_t \times 4$ 분할표를 작성한다.
   (b) 자유도, v가 $3(J_t-1)$인 x^2-통계량을 구하고, 다음의 Peizer-Pratt의 z-변환을 시행한다.

$$z = \begin{cases} \dfrac{1}{|W|}\left(W - \dfrac{1}{3}\right)\sqrt{(\mathrm{v}-1)\log\left(\dfrac{\mathrm{v}-1}{x^2}\right) + W}, & \mathrm{v} > 1 \\ \sqrt{x^2}, & \mathrm{v} = 1 \end{cases}$$

여기서 $W = x^2 - \mathrm{v} + 1$

K_1개의 z-값 중 가장 큰 값을 z_n으로 나타낸다.

2. 명목형 변수의 marginal test
   (a) 범주의 개수가 C개 일 때, $J_t \times C$ 분할표를 작성한다.
   (b) 자유도, v가 $(J_t-1)\times(C-1)$ x^2-통계량을 구하고, Peizer-Pratt의 z-변환을 시행한다. 이 때 $(K-K_1)$개의 z-값들 중 가장 큰 값을 z_c로 나타낸다.
3. 연속형 변수 (x_k, x_k)의 교호작용 검정
   (a) (x_k, x_k)의 표본 공간을 각 변수의 median에 의해 4개의 그룹으로 나누고, $J_t \times 4$ 분할표를 작성한다.
   (b) 자유도, v가 $3(J_t-1)$인 x^2-통계량을 구하고, Peizer-Pratt의 z-변환을 시행한다. $\frac{K_1(K_1-1)}{2}$개의 z-값들 중 가장 큰 값을 z_m으로 나타낸다.
4. 명목형 변수의 교호작용 검정
   (a) 범주의 개수가 각각 C_1, C_2일 때 $J_t \times C_1C_2$ 분할표를 작성한다.
   (b) 자유도, v가 $(J_t-1)\times(C_1C_2-1)$인 x^2-통계량을 구하고, Peizer-Pratt의 z-변환을 시행한다. $\frac{(K-K_1)(K-K_1-1)}{2}$의 z-값들 중 가장 큰 값을, z_{cc}으로 나타낸다.
5. 연속형 변수 x_k와 명목형 변수 x_k의 교호작용 검정
   (a) 연속형 변수는 median에 의해 2개의 그룹으로 나누고, $J_t \times 2C$ 분할표를 작성한다.
   (b) x^2-통계량을 구하고, Peizer-Pratt의 z-변환을 시행한다. $K_1(K-K_1)$개의 z-값들 중, 가장 큰 값을 z_{nc}으로 나타낸다.

이렇게 실행해서 나온 z값들로 변수를 선택할 경우 발생하는 명목형 예측 변수로의 bias를 조정하기위해 bootstrap bias correction을 실시해주어야 한다.

1. 복원 추출로 종속변수에서 bootstrap sample을 추출한다. 예측변수는 그대로 사용하므로, 종속변수와 예측변수는 서로 독립이다
2. bootstrap sample에 Algorithm 2.5-1을 적용하여, 5개의 z-값을 얻어낸다
3. 주어진 상수 f(>1)에 대해서, $f\max(z_n, z_m) \geq \max(z_c, z_{cc}, z_{nc})$을 만족하면, 연속형 변수를 선택한다. 그렇지 않으면 명목형 변수를 선택한다.
4. 1단계에서 3단계를, 다른 f값으로 B회 반복한다. $\pi(f)$를 연속형 변수가 선택된 비율이라 한다.
5. $\pi(f^*) = \frac{\text{수치형 예측변수의 개수}}{\text{예측변수의 개수}}$을 만족하도록 보간법으로 f^*를 구한다.

bootstrap bias correction 에 의해 결정된 값이 f^*라면, 이 값을 각각 z_n과 z_m에 곱해준다. $\{f^*z_n, z_c, f^*z_m, z_{cc}, z_{nc}\}$의 값들을 비교하여 가장 큰 z값을 골라낸 후, 위에서와 같은 방법으로 분리변수를 선택한다.

$$z^* = \max\{f^*z_n, z_c, f^*z_{nn}, z_{cc}, z_{nc}\}$$

만약 $f^*z_n = z^*$이면, 이에 해당하는 연속형 예측 변수를 선택한다.

만약 $z_c = z^*$ 이면, 이에 해당하는 명목형 예측 변수를 선택한다.

만약 $f^*z_m = z^*$이면, z-값이 보다 큰 연속형 예측 변수를 선택한다.

만약 $f^*z_\alpha = z^*$이면, z-값이 보다 큰 명목형 예측 변수를 선택한다.

만약 $f^*z_\pi = z^*$이면, 이에 해당하는 명목형 예측 변수를 선택한다.

2.4 Kolmogorov-Smirnov test

Kolmogorov-Smirnov test (이하 K-S test)는 두 변수의 확률분포함수가 같은지 여부를 판별하는 방법으로, 두 변수의 경험 누적 확률 분포간의 수직적 차이 중 가장 큰 값을 통계량으로 한다. $X_1, X_2, X_3, \cdots, X_n$의 경험누적확률분포를 $S_1(x)$라 하고, $Y_1, Y_2, Y_3, \cdots, Y_m$의 경험누적확률분포를 $S_2(x)$라 하였을 때 통계량은 $D = \sup|S_2(x) - S_1(x)|$ 이 된다.

이러한 K-S test를 의사결정나무에 적용할 수 있다. 종속변수 y의 class에 따라 예측 변수의 경험 누적 확률 분포를 비교하여, D가 가장 유의한 결과를 갖는 변수를 분리 변수로 선택하는 것이다. 즉 D가 클수록 y의 class에 따른 예측변수의 분포가 크게 달라지기 때문에, D가 가장 유의한 변수가 y를 잘 설명해주는 변수라고 할 수 있다.

2.5 Cramer-von Mises test

위의 K-S test가 한 점만을 고려한 것이라면, Cramer-von Mises test는 전체분포를 고려한 것이라 할 수 있다. Cramer-von Mises test의 통계량 T는 다음과 같이 계산된다.

$$T = \frac{mn}{(m+n)^2}\left\{\sum_{i=1}^{n}[S_1(X_i) - S_2(X_i)]^2 + \sum_{j=1}^{m}[S_1(Y_j) - S_2(Y_j)]^2\right\}$$

K-S test 방법이 통계량 D가 가장 유의한 변수를 분리 변수로 선택했던 것처럼, Cramer-von Mises test 방법은 통계량 T가 가장 유의한 결과를 보이는 변수를 분리 변수로 선택한다.

3.모의실험

MATLAB에서 x1, x2, x3, x4, x5등 5개의 예측 변수를 생성시켜, 500개의 관찰치를 500번 반복실험 하였다. 변수 y는 0과 1 두개의 그룹을 포함하는 그룹변수로 250개의 관찰치가 각 그룹에 속한다.

3.1 순서형 변수의 선택력 비교

먼저 예측변수 x1은 순서형이지만 예측변수 x2, x3, x4, x5는 연속형 변수인 경우를 가정하자. 모의실험에서는 x1 변수만 예측력이 있고, 나머지 변수들은 예측력이 없도록 데이터를 생성하였다. 순서형 변수 x1의 분포는 표3.1과 같이 정의하기로 한다. 예측변수 x2, x3, x4, x5들은 평균은 모두 0, 분산과 공분산은 각각 1과 0.1인 다변량 정규분포에서 변수를 생성하였다.

표3.2는 순서형 변수가 있는 경우의 모의실험 결과이다. CART 알고리즘의 경우 예측변수 x1을 선택하는 선택력은 높지 않음을 알 수 있다. 그리고 QUEST 알고리즘은 순서형범주의 개수가 2인 경우를 제외하고는 변수 선택력이 매우 낮게 나타나고 있다. 또한 기존 알고리즘 중에서는 CRUISE의 변수 선택력이 가장 우수하게 나타나고 있고, 비모수방법을 이용한 방법중에는 Cramer-von Mises test 방법이 높은 변수 선택력을 보여주고 있다. 순서형범주의 개수가 적을

때는 CRUISE 알고리즘이 Cramer-von Mises test 방법보다 우수하지만, 개수가 많아질수록 Cramer-von Mises test 방법이 더 우수한 것으로 나타났다.

표 3.1: 순서형 변수의 모의실험에 사용된 분포

표시	설명
U_k	정수 1,2, ..., k를 취하는 순서형 균일분포, $Pr(X_1=i)=\frac{1}{k}$, $\forall i$=1, ..., k
A_2	정수 1, 2를 취하는 순서형 분포, $Pr(X_1=1)=.6$, $Pr(X_1=2)=.4$
A_3	정수 1, 2, 3을 취하는 순서형 분포, $Pr(X_1=1)=Pr(X_1=3)=.25$, $Pr(X_1=2)=.5$
A_5	정수 1, 2, 3, 4, 5를 취하는 순서형 분포, $Pr(X_1=1)=Pr(X_1=5)=.12$, $Pr(X_1=2)=Pr(X_1=4)=.18$, $Pr(X_1=3)=.4$
A_{10}	정수 1, 2, ..., 10을 취하는 순서형 분포, $Pr(X_1=5)=Pr(X_1=6)=.2$, $Pr(X_1$=other than 5 or 6)=.075
A_{20}	정수 1, 2, ..., 20을 취하는 순서형 분포, Pr(X=10)=Pr(X=11)=.14, $Pr(X_1$=other than 10 or 11)=.04
A_{30}	정수 1, 2, ..., 30을 취하는 순서형 분포, Pr(X=15)=Pr(X=16)=.136, $Pr(X_1$=other than 15 or 16)=.026
A_{50}	정수 1, 2, ..., 50을 취하는 순서형 분포, Pr(X=25)=Pr(X=26)=.116, $Pr(X_1$=other than 25 or 26)=.016

표 3.2 순서형 변수의 모의실험 결과

X_1		$X_2, ..., X_5$	P(X1을 분리변수로 선택)				
class0	class1		CART	QUEST	CRUISE	K-S test	Cramer
A_2	U_2	Noise변수	.32	**.70**	.51	.40	**.64**
A_3	U_3	Noise변수	.62	.21	**.92**	.66	**.90**
A_5	U_5	Noise변수	.76	.27	**.98**	.80	**.94**
A_{10}	U_{10}	Noise변수	.75	.21	**.91**	.87	**.91**
A_{20}	U_{20}	Noise변수	.67	.22	.82	**.83**	**.85**
A_{30}	U_{30}	Noise변수	.79	.21	.87	**.91**	**.91**
A_{50}	U_{50}	Noise변수	.81	.22	.88	**.93**	**.93**

3.2 연속형 변수의 선택력 비교

본 논문에 제시된 비모수적 방법들을 연속형 변수들에 대해서도 사용가능한지 여부를 알기위한 모의실험을 수행하였다.. 따라서 모든 예측변수들이 연속형인 경우에 알고리즘간 변수 선택력을 비교해야 한다.

y가 0인 경우에는 평균은 모두 0, 분산과 공분산은 각각 1과 0.1인 다변량 정규분포에서 변수를 생성하였다. 그리고 y가 1인 경우에는 x1의 평균을 0.15, 0.2, 0.3, 0.4로 변화시켜가며, x2, x3, x4, x5의 평균은 모두 0.1이고 분산과 공분산은 각각 1과 0.1인 다변량 정규분포에서 변수를 생성하였다. 이 경우에도 예측변수 x1이 가장 예측력이 좋으므로 x1에 대한 선택력이 높은 알고리즘을 밝히고자 한다.

표 3.3: 분산구조는 같고, 평균은 서로 다른 다변량 정규분포

class0의 $\underset{\sim}{\mu}$	class1 $\underset{\sim}{\mu}$	P(X1을 분리변수로 선택)				
$(\mu_1, \mu_2, \mu_3, \mu_4, \mu_5)$	$(\mu_1, \mu_2, \mu_3, \mu_4, \mu_5)$	CART	QUEST	CRUISE	k-s test	cramer
(0,0,0,0,0)	(.15,.1,.1,.1,.1)	.29	.33	.27	**.38**	**.41**
(0,0,0,0,0)	(.2,.1,.1,.1,.1)	.45	.52	.42	**.58**	**.62**
(0,0,0,0,0)	(.3,.1,.1,.1,.1)	.78	**.85**	.73	**.85**	**.88**
(0,0,0,0,0)	(.4,.1,.1,.1,.1)	.93	**.98**	.94	.97	**.98**

표 3.3은 다변량 정규분포를 이용한 모의실험에 의한 변수 선택력의 결과이다. 본 연구에서 제안하는 비모수적 방법인 K-S test와 Cramer-von Mises test를 이용한 방법과 QUEST의 변수 선택력이 높게 나타났다. 따라서 비모수적 방법은 순서형 변수뿐만 아니라 연속형 변수의 경우에도 사용이 가능한 우수한 방법임을 확인할 수 있다.

4. 결론 및 향후 방향

본 연구에서는 의사결정 나무에서 분리 변수를 선택하는데 있어서, 변수의 형태가 순서형인 경우에 주목한다. 비모수적 접근 방법인 K-S test와 Cramer-von Mises test를 이용한 방법을 제안하고, CART, QUEST, CRUISE등의 기존 알고리즘과 변수 선택력을 비교하였다.

실험 결과 CART 알고리즘은 거의 모든 경우에 변수 선택력이 낮게 나타났다. CRUISE 알고리즘은 순서형 변수의 범주의 개수가 적을 경우(예를 들면 10이하)에 변수 선택력이 가장 우수한 반면, 범주의 개수가 많아질수록 변수 선택력이 비모수방법에 비해 약화되었다. QUEST 알고리즘은 순서형 변수에는 매우 낮은 변수 선택력을 보였으나 연속형 변수에 대해서는 우수한 변수 선택력을 보였다. 본 연구에서 제안하는 K-S test와 Cramer-von Mises test를 이용한 방법도 일반적으로 좋은 성과를 보여주고 있다. 특히 Cramer-von Mises test는 변수의 형태에 상관없이 대부분의 경우에 좋은 성과를 보여주고 있어 매우 Robust한 방법인 것으로 판단된다.

이 연구에서는 분리 변수 선택의 문제만을 고려하였으나, 분리 변수를 올바르게 선택하였을지라도 분리점 선택이 잘못되면, 의미 없는 분류가 될 수도 있다. 따라서 향후 분리점 선택의 문제도 다루어져야 할 것이다.

참고문헌

Kass, G. V. (1980), An Exploratory Technique for Investigating Large Quantities of Categorical Data, *journal of applied statistics* **29**: 119-127

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

Quinlan, J. R (1993), *C4.5 : programs for machine learning*, Morgan Kaufmann Publishers

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

Kim, H. and Loh, W.-Y. (2003), Classification trees with bivariate linear discriminant node models, *Journal of Computational and Graphical Statistics* **12**: 512-530

Kim, H. and Loh, W.-Y. (2001), Classification trees with unbiased multiway splits, *Journal of the American Statistical Association* **96**: 589-604