응용통계연구 제 19권 1호, 2006년, pp. 149–161

의사결정나무에서 순서형 분리변수 선택에 관한 연구*

김현중[1)]

요 약

CART로 대표되는 의사결정나무의 알고리즘에서 가장 중요한 요소는 분리변수의 선택방법이다. 대부분의 알고리즘은 변수의 형태가 연속형인지, 혹은 명목형(nominal)인지에 따라 별개의 변수선택방법을 적용한다. 하지만 변수의 형태가 순서형(ordinal)인 경우에는 그 변수를 연속형으로 취급하여 연속형 변수선택방법을 적용하는 것이 대부분이다. 이것은 CART와 같은 Greedy 탐색을 이용하는 방법에는 문제점이 발생하지 않는다. 하지만 Greedy 탐색의 약점을 보완하기 위해 통계이론을 이용하여 개발된 최근의 방법들에는 최선의 대처방법이 아니다. 따라서 본 연구에서는 의사결정 나무에서 분리변수를 선택하는데 있어서 비모수적 접근 방법인 Cramer-von Mises 검정을 이용한 방법을 순서형 변수에 사용하는 것을 제안하고, CART, C4.5, QUEST, CRUISE등 기존 알고리즘과 본 연구에서 제안하는 방법의 순서형 변수 선택력을 비교하였다. 모의실험의 결과, Cramer-von Mises 검정을 이용한 변수선택방법은 순서형 변수의 분류력을 기존 방법들에 비해 더 정확히 예측하는 좋은 성과를 보여주었다.

주요용어: 의사결정나무, 비모수방법, Cramer-von Mises 검정, 순서형 변수, CART

1. 연구 배경과 목적

분류(Classification)의 문제에 있어서 $X = (X_1, \cdots, X_p)'$를 자료공간이라 하고 Y를 그룹을 지칭하는 변수라 하자. 의사결정나무는 분류의 한 기법으로, 전체 데이터공간을 분할규칙(partition rule)을 이용하여 분할하고, 분할된 각 하부공간에 대하여 다시 반복적으로 분할규칙을 사용하는 방법이다. 서로 소(mutually exclusive)인 분할된 자료공간을 X^s라 하면 $X = \cup_s X^s$와 같은 관계가 성립된다. Y와 X는 중간노드들에 의해 분할되어 각각 $Y = (Y^1, \cdots, Y^s)'$와 $X = (X^1, \cdots, X^s)'$로 쓸 수 있다. 분할의 반복과정은 하부공간에 위치한 데이터가 대부분 같은 그룹에 속할 때까지 계속되므로 자료공간의 분할의 개수인 S는 미리 정해지지 않는다. 즉, Y^s내의 구성원이 대부분 같은 그룹인 경우에 분할이 멈춘다. 이를 분류영역에서는 불순도가 낮다고 하는데 흔히 사용하는 불순도의 정의는 다음과 같다.

$$\text{불순도}(s) = 1 - \sum_j p^2(j|s),$$

여기서 $p(j|s)$는 이 하부공간 s에 위치한 관찰값 중 그룹 j에 속한 비율이다. 즉 $p(j|s)$가 어느 특정 그룹 j에서 큰 값을 가지고 나머지 그룹에서 작은 값을 가지면 불순도는 매우 낮아

* 본 연구는 한국학술진흥재단의 지원에 의하여 연구되었음 (KRF-2004-003-C00046).

1) (120-749) 서울시 서대문구 신촌동 134, 연세대학교 상경대학 응용통계학과 조교수
E-mail: hkim@yonsei.ac.kr

지게 된다. 불순도가 충분히 낮은 경우에 반복분할을 중단한다. 물론 하부공간의 규모가 매우 작게 되어 분할의 실효성이 없을 때에도 하부공간에 대한 분할을 중단한다. 이렇게 구축된 데이터 공간의 반복적 분할과정을 나무의 형태로 요약한 것이 의사결정나무라 할 수 있겠다.

반복 분할된 데이터 공간은 과적합(over-fitting)의 가능성이 매우 높으므로 분할된 공간을 서로 합병하는 과정을 거치게 된다. 이러한 과정을 가지치기(pruning)라 하는데 이 가지치기는 의사결정나무의 크기를 줄이는 효과가 있으므로 해석력을 향상시키는데 큰 도움이 된다. 더 나아가 분류예측력도 향상 시키는 것으로 인정되고 있다.

의사결정나무의 구축에 있어서 가장 핵심적인 내용은 분할규칙이라 할 수 있다. 분할규칙은 하나의 변수를 사용하는 방법과 여러개의 변수를 병합하여 사용한 것등 두가지로 나뉜다. 전자는 단변량 분할이라 칭하며 흔히 의사결정나무라 할 때 일컫는 방법이다. 후자는 선형결합분할이라 하여 여러 변수의 선형결합을 이용하여 데이터를 분할하는 기법으로 피셔의 선형판별함수는 이 방법의 특수한 형태라 할 수 있다. (즉 피셔의 선형판별함수는 분할이 한번만 행해지는 단순구조의 의사결정나무이다.)

본 논문에서는 단변량 분할규칙에 관한 논의에 초점을 맞추고자 한다. 단변량 분할규칙의 기본 구조는 연속형 변수인 경우 $X_i \leq c$ 와 $X_i > c$ 인 구조를, 명목형 변수인 경우 $X_j \in A$와 $X_j \in A^c$의 구조를 갖는다. 여기서 c는 연속형 변수인 공간상의 한 점으로, 이점을 기준삼아 하부공간으로 분할했을 때 자료의 불순도가 가장 낮았기 때문에 선택된 점이다. 또한 A는 명목형 변수 X_j의 부분집합 공간중 분할후 불순도를 가장 낮게 하는 부분집합을 의미한다. 여기서 분할에 사용되는 변수의 형태에 따라 분할규칙의 구조가 다름을 쉽게 알 수 있다. 분할변수가 일단 선택되어지고 나면 불순도를 가장 낮게 하는 분할점 c나 분할부분집합 A를 찾는 일은 단순 알고리즘에 의한 계산과정만 거치면 된다.

의사결정나무의 대부분 방법들은 변수의 형태가 연속형 혹은 명목형인 경우의 분할규칙만을 고려한다. 변수의 형태가 순서형(ordinal) 자료인 경우, 대개는 연속형 변수로 취급하고 연속형 변수의 분할규칙에 의해 변수선택 절차를 수행하게 된다. 본 논문은 순서형 변수의 형태에 맞는 새로운 변수선택 방법을 제안하고 연속형 변수용으로 개발된 기존 방법과 비교하여 순서형 자료에 대한 변수 선택력을 측정해 보고자 한다.

2. 의사결정나무에서의 변수선택

흔히 의사결정나무의 단점으로 나무구조의 불안정성을 든다. 이는 적합자료(training data)에 약간의 변형이 가해지면 나무구조가 전체적으로 크게 바뀌게 된다는 사실에서 기인한다. 이는 CART(Breiman, Friedman, Olshen, Stone 1984)와 같은 의사결정나무의 초창기 방법에 더 심각한 문제로 그 후 개발된 방법들에서는 많이 해결된 문제이다(Kim & Loh 2001, Loh & Shih 1997, Kim & Loh 2003). 나무구조의 불안정성은 데이터의 변형에 따라 최선의 분할규칙이 바뀐다는 데에 문제의 근원이 있다. 그 결과는 해석의 관점에서는 심각한 문제이지만 예측력의 관점에서는 그다지 심각하지 않는 문제이다. 예를 들어 그림 2.1의 의사결정나무를 고려해보자(Martin 1997).

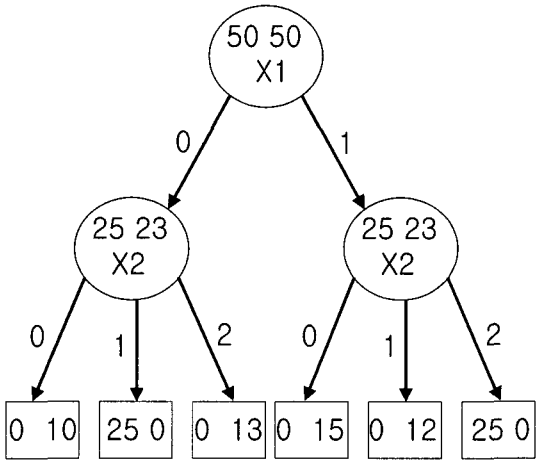

그림 2.1: 의사결정나무의 예

그림 2.1의 의사결정나무는 그룹이 2개인 자료를 분류하는 나무구조이다. 여기서 X_1변수는 0과 1의 값을 갖는 이진변수이고 X_2변수는 0, 1, 2의 값을 갖는다. 각 노드내의 숫자는 두 그룹에 속하는 개체수를 의미한다. 나무구조의 최종노드의 수는 6개이며 나무의 깊이가 2개인 구조를 갖고 있고 훈련자료에 100%의 분류정확도를 가지고 있다. X_1변수를 먼저 분할규칙에 사용하였고 X_2변수를 두 번째 단계에서 사용하였다. 동일 데이터를 이용하여 다른 구조의 의사결정나무를 고려해보자. 예로 X_2를 먼저 분할에 사용하고 X_1변수를 두 번째로 사용해 볼 수도 있다. 또한 Y변수를 분류할 능력이 전혀 없는 Noise 변수 X_3을 고려했을 때 $X_1 - X_2 - X_3$의 순서로 나무구조를 구축하거나, 순서를 바꾸어 $X_1 - X_3 - X_2$ 혹은 $X_3 - X_1 - X_2$의 순서로 구축되는 나무구조를 생각해 볼 수 있다. 표2.1은 여러 가지 조합에 의해 나무구조가 구축됐을 때의 나무의 크기와 깊이 그리고 훈련자료의 분류정확도를 요약한 것이다.

표 2.1에서 알 수 있듯 서로 다른 분할규칙이라 할지라도 분류의 정확도는 영향을 받지 않을 수도 있다. 하지만 분류력이 없는 변수(예로 X_3)가 나무구조의 구축에 사용된다면 불필요하게 나무구조를 복잡하게 만들고 잘못된 해석 (예로 X_3는 유용한 변수라 오해)을 제공할 수도 있는 것이다. 비슷한 맥락으로 Liu & White(1994)는 분류력이 있는 변수와 분류력이 없는 변수를 구분하는 것의 중요성에 대한 논의를 전개한 바 있다. 여기서 분류력이 있는 변수란 목표 변수에 대한 분류정보를 가지고 있는 변수를 일컫는다.

더 나아가 의사결정나무의 중요한 부분인 가지치기의 효과를 고려하면 분류력이 있는 변수를 선택하는 것이 검정자료(test data)에 대한 분류의 정확도를 향상시키는 일임이 명백해진다. 다시 말해서 잘못된 분할규칙은 분류정확도에 악영향을 미친다. 예로 $X_1 - X_3 - X_2$의 나무구조에서 가지치기의 결과 마지막 분할이 가지치기 된다면 $X_1 - X_3$의 나무구조만 남게 된다. 이것은 결국 분류정확도가 현저히 감소하는 결과로 귀결된다. 결론적으로 가지치기의 결과에 따라 비효율적 나무구조는 예측력이 감소하는 위험이 있는 것이다. 따라서 당

표 2.1: 분할의 순서가 나무구조에 미치는 영향

분할순서	최종노드의 수	깊이	훈련자료의 분류정확도
$X_1 - X_2$	6	2	100%
$X_2 - X_1$	5	2	100%
$X_3 - X_1 - X_2$	12	3	100%
$X_3 - X_2 - X_1$	10	3	100%
$X_1 - X_3 - X_2$	12	3	100%
$X_2 - X_3 - X_1$	9	3	100%

장의 분류정확도를 따지기 보다는 더 효율적인 나무구조, 즉 분류력이 있는 변수가 먼저 분할규칙에 사용되는 나무구조를 구축하는 것이 해석력 뿐만 아니라 예측력의 확보에도 필요한 것이다.

분류력이 있는 변수들과 분류력이 없는 변수들이 혼합되어 있는 경우에, 분류력이 있는 변수가 분리 변수로 선택될 확률을 변수 선택력이라 한다. 의사결정나무 알고리즘은 변수 선택력이 커야 나무구조의 효율성을 확보할 수 있다. 그러면 기존 의사결정나무의 변수 선택력에 대해 살펴보도록 하자.

2.1. Greedy 탐색에 의한 변수선택

의사결정나무의 매뉴얼이라 할 수 있는 CART 방법은 Greedy 탐색법으로 요약할 수 있다. 즉 CART방법은 분할이 가능한 모든 분할규칙을 고려한다. 예로, 연속형 변수 X_1은 모든 관찰값이 unique 하다고 하자. 총 관찰치의 수가 n이라 하면 X_1은 $(n-1)$개의 분할이 가능하다. 만약 연속형 변수 X_2의 값은 같은 값이 반복 관찰되어 $n/2$개의 unique한 값이 있다하면 X_2는 $(n/2-1)$개의 분할이 가능하게 된다. 명목형 변수 X_3가 M개의 범주가 있는 변수라 가정하면, 부분집합의 개수를 고려하여 X_3는 모두 $(2^{M-1}-1)$개의 분할이 가능하다. 여기서 한가지 관찰할 수 있는 점은 각 변수별 분할가능한 회수가 서로 다르다는 점이다.

CART는 공간 s상에서 w라는 분할규칙의 분할유효성을 다음과 같이 측정하게 된다.

$$\text{분할유효성}(w,s) = \text{불순도}(s) - P_L \bullet \text{불순도}(s_L) - P_R \bullet \text{불순도}(s_R).$$

공간 s상에 있는 자료들이 s_L공간으로 분리될 비율 P_L과 s_R공간으로 분리될 비율 P_R은

$$P_L = \frac{N(s_L)}{N(s)}, \quad P_R = \frac{N(s_R)}{N(s)}$$

와 같이 정의되며, s공간에서의 총 자료 수는 $N(s)$라 한다. 불순도에 사용되는 $p(j|s)$는

$$p(j|s) = \frac{N_j(s)}{N(s)}$$

으로 정의된다. 여기서 s공간에서 그룹 j에 속하는 자료 수를 $N_j(s)$라 한다.

CART 방법은 각 변수의 모든 분할규칙에 분할유효성을 계산한다. 그리고 각 변수별로 분할유효성을 최대로 만드는 분할규칙을 찾은 후 변수들끼리 비교하여, 가장 큰 분할유효성을 갖는 변수를 분리변수로 채택한다.

통계학 영역에서 CART가 의사결정나무의 대명사화 되었듯, C4.5 (Quinlan, 1993)는 전산과학 영역에서 의사결정나무의 대표적 방법이다. CART가 각 마디에 이원분할을 형성하며 이지분리 나무구조를 만드는데 반하여, C4.5는 연속형 예측 변수에 관해서는 이지 분리를 하지만, 명목형 변수에 관해서는 각 범주가 하나의 마디를 가지는 다지 분리 구조를 갖는 나무를 구성된다. C4.5는 gain ratio라는 통계량을 분할유효성 대신 사용하여 CART방법의 아이디어와 비슷하게 변수를 선택하므로 Greedy 탐색방법중의 하나이다.

Greedy 탐색방법의 특징은 각 변수마다 서로 다른 개수의 분할규칙이 존재하나 이를 고려치 않고 모든 변수에 똑같은 사전확률로 비교한다는 점이다. 이에 대한 부작용으로 분할규칙이 많이 제공되는 변수는 그만큼 분리변수로 선택될 가능성이 높아지고 그렇지 못하면 선택되기 힘들어 진다. (White & Liu, 1994; Loh & Shih, 1997; Kim & Loh, 2001). 중요한 변수를 선택하지 못하는 경우가 생기게 되면 나무구조의 효율성을 저해하는 원인이 된다는 것은 자명하다. 따라서 Greedy 탐색방법은 매우 신중히 사용되어야 한다.

2.2. 통계방법을 이용한 변수선택

CART와 더불어 많이 사용되는 의사결정나무로 CHAID(Kass, 1975; Kass, 1980)라는 방법이 있다. 연속형 변수에 대해서는 CART나 C4.5와 동일한 알고리즘에 분할유효성 대신 카이제곱 검정통계량의 유의확률을 사용하는 점만 다르다. 명목형 변수에 대해서는 C4.5처럼 각 범주가 하나의 마디를 가지는 다지 분리 구조를 갖는 나무로 부터 시작하여 유의성이 낮은 가지끼리 병합하는 과정을 반복하고, 병합한 후의 분할결과와 Y변수의 분할표에서 유의확률을 구한다. 결국 모든 변수에 대한 유의확률을 비교하게 되는데 가장 작은 유의확률을 갖는 변수를 선택한다. 연속형 변수인 경우, CHAID방법은 Greedy 탐색방법이므로 변수선택에 약점을 갖게 된다. 명목형 변수의 경우에는 범주끼리의 병합의 결과에 따라 변수선택이 일관되지 않을 수도 있다. 따라서 CHAID방법은 통계방법을 의사결정나무의 분할규칙에 처음 적용한 의의는 있으나 Greedy 탐색의 틀을 벗어나지 못한 한계가 있다.

예측변수인 X변수와 그룹변수인 Y변수간 유의적인 관계가 있는 지를 확률적으로 판단하여 가장 유의한 변수를 분할규칙에 사용하고자 하는 방법은 White & Liu(1994)에 의해 처음 제안되었다. 이 방법은 명목형 변수인 경우에만 사용되는 방법으로, 연속형 변수는 허용이 안되므로 연속형 변수를 사전에 범주화하여 명목형 변수로 변환한 후 사용하는 약점이 있다. Loh & Shih(1997)는 연속형 변수는 ANOVA의 유의확률 혹은 Levene의 등분산검정 유의확률을 사용하고, 명목형 변수는 카이제곱 분할표검정의 유의확률을 사용한 변수

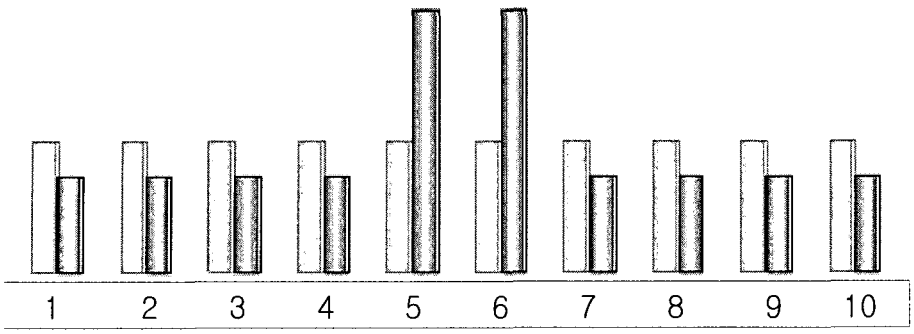

그림 3.1: 10개의 순서형 값을 갖는 한 변수에 대한 두 그룹별 분포 예

선택 방법인 QUEST 알고리즘을 제안하였다. QUEST 방법은 평균과 분산이 동일하나 서로 상이한 분포를 갖는 연속형 변수를 선택하지 못하는 약점이 있다. Kim & Loh(2001)는 Loh & Shih(1997) 방법의 약점을 보완함과 함께 변수들간 상호작용을 변수선택의 기준에 포함시키는 알고리즘인 CRUISE를 제안하였다. 이상의 통계적 유의확률을 이용한 방법들은 기존 Greedy 탐색 방법의 한계였던 변수선택의 편의(bias)를 제거함으로써 의사결정나무의 불안정성 문제도 대부분 해결하는 장점이 있다.

이상 열거된 변수선택방법들은 연속형 혹은 명목형인 변수만을 대상으로 개발된 것이다. 하지만 데이터에 순서형 변수가 포함된 경우에는 이에 적절한 방법이 개발되어 있지 않다. 대부분의 경우에는 순서형 변수를 연속형 변수처럼 취급하고 연속형 변수의 선택방법을 적용한다. 자료의 형태에 적절한 통계방법을 사용하지 않으면 검정력(Power)의 감소로 인하여 선택하여야 할 변수를 선택하지 못하는 경우가 있을 수 있다. 예를 들어, 정수값을 갖는 자료에 대한 ANOVA 모형의 유의확률은 부정확할 수 있는 것과 같다.

본 논문은 의사결정나무를 구축함에 있어서 연속형 변수, 명목형 변수, 그리고 순서형 변수에 대한 선택방법을 구분하고, 그중 순서형 변수에 더 적절한 변수선택 방법을 제안하고자 한다. 구체적 변수선택의 절차로, 연속형 변수와 명목형 변수에 대하여는 QUEST나 CRUISE같은 방법을 통하여 유의확률을 구하고 순서형 변수에서는 본 논문에서 제안된 방법을 통해서 유의확률을 구한 후 모든 변수의 유의확률값중 가장 작은 값을 가지는 변수를 분리변수로 선택하는 것이다.

본 논문에서는 그룹의 개수가 2개인 경우로만 논의를 국한하고자 한다. 그룹이 세 개 이상인 경우에는 Loh & Shih(1997) 방법처럼 2-means clustering 과정을 거쳐서 유사한 그룹을 한 군집으로 묶어줌으로써 두 개의 그룹으로 만든 후 본 논문에서 제안된 방법을 사용할 수도 있다.

3. 순서형 자료의 변수선택방법

이해를 돕기위해 그림 3.1와 같은 10개의 순서형 값을 갖는 변수에 대한 두 그룹별 분포를 고려해보자.

그림 3.1에서 보듯 두 그룹의 분포는 약간의 차이를 보이고 있다. 이 순서형 변수에 대해 2장에 사용된 통계적 방법을 적용해 볼 수 있을 것이다. 먼저 이 변수를 연속형으로 취급하고 ANOVA와 같은 방법을 사용할 수 있다. 혹은 이 변수를 명목형으로 취급하고 카이제곱 분할표 검정과 같은 방법을 사용해 볼 수도 있다. 하지만 이 두가지 경우 모두 정확한 결과를 얻는다는 보장이 없다.

본 논문에서는 순서형 관찰값별로 두 분포의 차이를 측정하여 누적하는 방법을 제안한다. 이는 누적확률분포를 비교하는 방법이라 할 수 있다. 이는 카이제곱 분할표 검정방법과 유사한 측면이 있다. 하지만 카이제곱 분할표 검정방법은 각 변수의 값에 따른 차이에 관심이 있다기 보다는 전체적인 그룹간 비율이 각 변수값별로 얼마나 잘 유지되는지에 관심이 있다고 할 수 있으며 또한 분할표에서 값의 순서에 영향을 받지 않는다.

3.1. 경험누적확률분포의 비교

모집단의 분포가 정확히 알려져 있지 않으므로 주어진 적합데이터를 이용하여 두 그룹의 누적확률분포가 차이가 있는지를 검정하기 위해서는 경험누적확률분포를 활용하여야 한다. 임의의 한 변수에 대하여 그룹1에 속하는 X변수의 관찰치를 $X_{11}, \cdots, X_{1n}$이라 하자. 마찬가지로 그룹2에 속하는 X변수의 관찰치를 $X_{21}, \cdots, X_{2m}$이라 하자. 여기서 n은 그룹1에 속하는 관찰치의 수, m은 그룹2에 속하는 관찰치의 수이다. 그리고 각각의 경험누적확률분포를 $S_1(x)$과 $S_2(x)$라 하였을 때

$$\sum_x |S_1(x) - S_2(x)|$$

를 이용하여 두 누적확률분포의 일치성을 측정해 볼 수 있다. 여기서 x는 그룹을 무시한 이 변수의 관찰값의 범위를 지칭한다. (참고로 Kolmogorov-Smirnov라는 통계량은 $\sup|S_1(x) - S_2(x)|$으로 두 누적확률분포 차의 최대치를 추정한 값이 된다.) 하지만 변수에 따라 unique한 값의 개수가 다르므로 위 공식은 unique한 값의 개수가 적은 변수에 대해서는 불리할 수도 있다.

또한 연속형 혹은 명목형 변수들과의 비교시 유의확률을 사용하고자 하므로 위의 통계량은 비교목적으로 사용되기도 어렵다.

3.2. Cramer-von Mises 방법

Fisz(1960)은 두 누적확률분포의 일치성을 검정하는 통계량을 다음과 같이 제안하였다.

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

이것은 흔히 Cramer-von Mises 통계량으로 알려져 있고 그것의 표본분포표는 Burr(1964)에 의해 제공되었다. 위 통계량에 대한 유의확률은 변수간 unique한 값의 개수가 다르더라도 공정한 비교 도구가 될 것으로 생각된다. 따라서 본 논문에서는 Cramer-von Mises 통계량

표 4.1: 대칭형 순서형 분포

분포기호	내용
U_k	정수 $1, 2, \cdots, k$를 취하는 순서형 균일분포, $Pr(X_1 = i) = 1/k$
A_5	정수 $1, 2, 3, 4, 5$를 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 $1, 2, \cdots, 10$을 취하는 순서형 분포, $Pr(X_1 = 5) = Pr(X_1 = 6) = .2, Pr(X_1 = \text{other than 5 or 6}) = .075$
A_{20}	정수 $1, 2, \cdots, 20$을 취하는 순서형 분포, $Pr(X_1 = 10) = Pr(X_1 = 11) = .14, Pr(X_1 = \text{other than 10 or 11}) = .04$
A_{30}	정수 $1, 2, \cdots, 30$을 취하는 순서형 분포, $Pr(X_1 = 15) = Pr(X_1 = 16) = .136, Pr(X_1 = \text{other than 15 or 16}) = .026$
A_{50}	정수 $1, 2, \cdots, 50$을 취하는 순서형 분포, $Pr(X_1 = 25) = Pr(X_1 = 26) = .116, Pr(X_1 = \text{other than 25 or 26}) = .016$

에 의한 유의확률을 순서형 변수의 분류력을 측정하는 도구로 사용할 것을 제안한다. 이 방법은 변수의 형태가 순서형 뿐만 아니라 연속형인 경우에도 우수한 변수선택력을 보일 것으로 기대하나 연속형 변수인 경우 연산시간이 과다한 약점이 있다.

4. 모의실험

본 장에서는 여러개의 순서형 변수를 생성시킨 모의실험을 통하여 3장에서 제안된 Cramer-von Mises 방법을 사용한 순서형 변수선택방법의 효과를 판단하고자 한다.

먼저 X_1, X_2, X_3, X_4, X_5등 5개의 예측 변수를 생성시켜, 500개의 관찰치를 500번 반복실험 하였다. 변수 Y는 0과 1 두 개의 그룹을 포함하는 그룹 변수로 250개의 관찰치가 각 그룹에 속하도록 하였다. 변수의 선택력을 비교하는 것이 목적이므로 나무모형을 완성할 필요없이 한 노드에서 분리변수로 선택되는 변수만을 관찰하였다. X_1은 순서형이지만 나머지 예측변수 X_2, X_3, X_4, X_5는 $N(0, 1)$로부터 생성된 연속형 변수들이다. 마지막으로 X_1변수만 분류력이 있고, 나머지 변수들은 분류력이 없도록 데이터를 생성하였으므로 X_1에 대한 선택력이 높은 변수선택 방법이 우수한 것으로 판단할 수 있다.

4.1. 대칭형 자료

먼저 표 4.1와 같은 대칭형 순서형 분포를 가정하자. 이 표는 X_1변수의 분포를 의미한다. X_1변수값중 그룹이 0인 값들은 U_k의 분포를 갖고 그룹이 1인 값들은 A_k의 분포를 갖게 되므로 X_1변수만 분류력이 있는 변수이다

표 4.1의 데이터에 기존의 의사결정나무 방법들인 CART, C4.5, QUEST, CRUISE와 본 논문에서 제안하고 있는 Cramer-von Mises방법을 비교한 결과는 표 4.2에 요약되어 있다.

표 4.2: 대칭형 자료를 이용한 변수선택력 비교

X_1		$X_2, \cdots, X_5$	$P(X_1$을 분리변수로 선택)				
그룹 0	그룹 1		CART	C4.5	QUEST	CRUISE	Cramer
U_5	A_5	Noise 변수	0.76	0.75	0.27	0.98	0.94
U_{10}	A_{10}	Noise 변수	0.75	0.73	0.21	0.91	0.91
U_{20}	A_{20}	Noise 변수	0.67	0.71	0.22	0.82	0.85
U_{30}	A_{30}	Noise 변수	0.79	0.80	0.21	0.87	0.91
U_{50}	A_{50}	Noise 변수	0.81	0.78	0.22	0.88	0.93

기존의 의사결정나무 방법들은 순서형 자료에 대한 특별한 방법이 고안되어 있지 않으므로 X_1변수를 연속형으로 취급하여 분석한다.

표 4.2에서 '$P(X_1$을 분리변수로 선택)'은 X_1을 분리변수로 선택하게 되는 확률추정값을 의미한다. X_1변수는 분류력을 가지도록 생성되었으므로 높은 확률추정값을 보일수록 좋은 방법이라 할 수 있다. 표 4.2에 의하면 Cramer-von Mises 방법은 X_1변수를 항상 높은 비율로 선택하는 것을 알 수 있다. 한가지 흥미로운 사실은 순서형 변수의 범주가 적을 때에는 CRUISE 방법이 Cramer-von Mises 방법보다 우수한 경우가 있다는 것이다. 특히 CRUISE 방법은 모든 경우에 있어서 Cramer-von Mises방법과 유사한 변수 선택력을 보였다. 반면 QUEST 방법은 가장 문제가 많았다. 그 이유로는 두 그룹의 평균이 같으므로 평균의 차이를 검정하는 ANOVA 방법이 효과가 없었기 때문이다.

4.2. 비대칭형 자료

표 4.3은 모의실험에서 사용된 X_1변수의 분포로서 비대칭형 순서형 분포를 정의하고 있다. X_1변수값중 그룹이 0인 값들은 B_k의 분포를 갖고 그룹이 1인 값들은 C_k의 분포를 갖게 되므로 X_1변수만 분류력이 있는 변수이다.

표 4.4에는 기존 의사결정나무 방법들과 Cramer-von Mises방법을 비교한 결과가 요약되어 있다. Cramer-von Mises 방법은 비대칭형 순서형 자료에 대해서도 매우 만족할 만한 결과를 보이고 있다. 기존 방법들 중에서는 CRUISE 방법이 전반적으로 좋은 결과를 보인다.

4.3. 혼합형 자료

이상의 모의실험은 순서형 변수만이 분류력이 있다고 가정하였다. 하지만 현실에서 모든 변수는 어느 정도의 분류력을 갖고 있다. 따라서 분류력이 다른 변수보다 우수한 변수를 잘 선택하는지 여부가 좋은 알고리즘인지에 대한 판단의 기준이 된다. 이에 대한 실험으로 다음과 같은 모형을 설정한다.

$$p = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon)}$$

표 4.3: 비대칭형 순서형 분포

분포기호	내용
B_5	정수 $1,2,3,4,5$를 취하는 순서형 분포, $Pr(X_1=4)=.33, Pr(X_1=\text{other than } 4)=.1675$
B_{10}	정수 $1,2,\cdots,10$을 취하는 순서형 분포, $Pr(X_1=7)=.208, Pr(X_1=\text{other than } 7)=.088$
B_{20}	정수 $1,2,\cdots,20$을 취하는 순서형 분포, $Pr(X_1=14)=.202, Pr(X_1=\text{other than } 14)=.042$
B_{30}	정수 $1,2,\cdots,30$을 취하는 순서형 분포, $Pr(X_1=20)=.188, Pr(X_1=\text{other than } 20)=.028$
B_{50}	정수 $1,2,\cdots,50$을 취하는 순서형 분포, $Pr(X_1=35)=.167, Pr(X_1=\text{other than } 35)=.017$
C_5	정수 $1,2,3,4,5$를 취하는 순서형 분포, $Pr(X_1=5)=.28, Pr(X_1=\text{other than } 5)=.18$
C_{10}	정수 $1,2,\cdots,10$을 취하는 순서형 분포, $Pr(X_1=8)=.154, Pr(X_1=\text{other than } 8)=.094$
C_{20}	정수 $1,2,\cdots,20$을 취하는 순서형 분포, $Pr(X_1=16)=.107, Pr(X_1=\text{other than } 16)=.047$
C_{30}	정수 $1,2,\cdots,30$을 취하는 순서형 분포, $Pr(X_1=23)=.101, Pr(X_1=\text{other than } 23)=.031$
C_{50}	정수 $1,2,\cdots,50$을 취하는 순서형 분포, $Pr(X_1=38)=.069, Pr(X_1=\text{other than } 38)=.019$

표 4.4: 비대칭형 자료를 이용한 변수선택력 비교

X_1		$X_2,\cdots,X_5$	$P(X_1$을 분리변수로 선택)				
그룹 0	그룹 1		CART	C4.5	QUEST	CRUISE	Cramer
B_5	C_5	Noise 변수	0.75	0.62	0.50	0.93	0.94
B_{10}	C_{10}	Noise 변수	0.35	0.36	0.21	0.48	0.61
B_{20}	C_{20}	Noise 변수	0.59	0.62	0.30	0.73	0.80
B_{30}	C_{30}	Noise 변수	0.59	0.51	0.27	0.64	0.81
B_{50}	C_{50}	Noise 변수	0.55	0.61	0.31	0.70	0.76

표 4.5: 로지스틱 모형의 계수

계수	β_0	β_1	β_2	β_3	β_4	β_5
사용값	.24*m	.11	.01	.01	.01	.10

표 4.6: 혼합형 자료를 이용한 변수선택력. 아래줄은 각각 $P(X_1)$과 $P(X_5)$임.

X_1의 분포	$P(X_1$ 혹은 X_5를 분리변수로 선택) $(P(X_1), P(X_5))$				
	CART	C4.5	QUEST	CRUISE	Cramer
U_5	0.71 (0.24, 0.47)	0.78 (0.35, 0.43)	0.85 (0.44, 0.41)	0.76 (0.39, 0.37)	0.86 (0.51, 0.35)
U_{10}	0.74 (0.32, 0.42)	0.75 (0.36, 0.39)	0.87 (0.49, 0.38)	0.78 (0.42, 0.36)	0.89 (0.57, 0.32)
U_{20}	0.75 (0.34, 0.41)	0.76 (0.34, 0.42)	0.87 (0.48, 0.39)	0.75 (0.39, 0.36)	0.89 (0.56, 0.33)
U_{30}	0.75 (0.36, 0.39)	0.74 (0.36, 0.38)	0.87 (0.48, 0.39)	0.76 (0.39, 0.37)	0.89 (0.56, 0.33)
U_{50}	0.73 (0.27, 0.46)	0.78 (0.35, 0.43)	0.90 (0.54, 0.36)	0.80 (0.48, 0.34)	0.91 (0.63, 0.28)

여기서 X_1과 X_2는 U_k분포를 따른다($k = 5, 10, 20, 30, 50$). U_k분포의 평균과 표준편차를 m과 s라 할 때, X_3는 지수분포(s)를, X_4과 ϵ은 정규분포$(0, s^2)$를, 그리고 X_5는 균등분포$(0, \sqrt{12}s)$를 갖는다. 단, X_3, X_4, X_5는 평균값이 m이 되도록 상수를 더해준다. 결론적으로 모든 변수들은 동일한 평균과 분산을 갖는다. 만약 p값이 0.5보다 크면 Y는 1이라하고 작으면 0이라고 하여 두 개의 그룹을 생성한다. 마지막으로 위 모형의 계수로서 표 4.5와 같은 값을 부여한다. 즉, X_1과 X_5의 분류력이 다른 변수에 비해 좋으며 X_1의 분류력은 X_5보다 약간 우수하다.

모의실험의 결과는 표 4.6에 나와 있다. 비교에 사용된 모든 방법들에 있어서 X_1과 X_5의 변수선택확률이 다른 변수에 비해 높았다. 그중 Cramer-von Mises 방법이 X_1변수를 가장 잘 선택하였으며, X_1 혹은 X_5를 선택하는 비율도 가장 높았다. 결론적으로 본 연구에서 제안하는 Cramer-von Mises 방법은 순서형 변수가 분류력이 있는 경우에 기존 방법에 비해 훨씬 더 정확하게 해당 변수를 선택하는 것을 알 수 있다.

5. 맺음말

본 연구에서는 의사결정 나무에서 분리 변수를 선택하는데 있어서 비모수적 접근 방법인 Cramer-von Mises 검정을 이용한 방법을 순서형 변수에 사용하는 것을 제안하고, CART, C4.5, QUEST, CRUISE등 기존 알고리즘과 본 연구에서 제안하는 방법의 변수 선택력을 비교하였다. 실험 결과 CART나 C4.5 알고리즘은 전반적으로 변수 선택력이 낮게 나타났으며, QUEST 알고리즘은 순서형 변수에는 매우 낮은 변수 선택력을 보였으나 혼합형 자료에 대한 변수 선택력은 우수한 결과를 보여주었다. CRUISE 방법은 기존의 방법중에서 순서형 변수에 대한 변수선택력이 대체로 우수하였다. 본 연구에서 제안하는 Cramer-von Mises 검정을 이용한 방법은 순서형 변수의 분류력을 기존 방법에 비해 더 정확히 평가하는 우수한 방법인 것으로 판단된다.

참고문헌

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

Burr, E. J. (1964). Small-sample distributions of the two-sample Cramer-von Mises' W2 and Watson's U2. *The Annals of Mathematical Statistics*, **35**: 1091-1098.

Fisz, M. (1960). On a result by M. Rosenblatt concerning the von Mises-Smirnov test. *The Annals of Mathematical Statistics*, **31**: 427-429.

Kass, G. V. (1975). Significance testing in automatic interaction detection (A.I.D), *Journal of Applied Statistics*, **24**: 178-189.

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data, *Journal of Applied Statistics*, **29**: 119-127.

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

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

Liu,W. Z. and White, A. P. (1994). The importance of attribute-selection measures in decision tree induction, *Machine Learning*, **15**: 25-41.

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

Martin, J. K. (1997). An exact probability metric for decision tree splitting and stopping, *Machine Learning*, **28**: 257 - 297.

Quinlan, J. (1993). *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo.

White, A. P. and Liu, W. Z. (1994). Bias in information-based measures in decision tree induction, *Machine Learning*, **15**: 321-329.

[2005년 8월 접수, 2005년 12월 채택]

Ordinal Variable Selection in Decision Trees*

Hyunjoong Kim[1)]

ABSTRACT

The most important component in decision tree algorithm is the rule for split variable selection. Many earlier algorithms such as CART and C4.5 use greedy search algorithm for variable selection. Recently, many methods were developed to cope with the weakness of greedy search algorithm. Most algorithms have different selection criteria depending on the type of variables: continuous or nominal. However, ordinal type variables are usually treated as continuous ones. This approach did not cause any trouble for the methods using greedy search algorithm. However, it may cause problems for the newer algorithms because they use statistical methods valid for continuous or nominal types only. In this paper, we propose a ordinal variable selection method that uses Cramer-von Mises testing procedure. We performed comparisons among CART, C4.5, QUEST, CRUISE, and the new method. It was shown that the new method has a good variable selection power for ordinal type variables.

Keywords: Decision Trees, Nonparametric statistics, Cramer-von Mises test, Ordinal variable, CART

* This work was supported by Korea Research Foundation Grant (KRF-2004-003-C00046).

1) Assistant Professor, Department of Applied Statistics, Yonsei University, Shinchon-Dong 134, Seodaemun-Gu, Seoul 120-749, Korea.
E-mail: hkim@yonsei.ac.kr