Proceedings of the Autumn Conference in 2001
and the 30th Anniversary of the Korean Statistical Society

불완비 데이터에서 분류 나무의 구축

우 주 성[1] , 김 규 성[2]

요 약

본 논문에서는 결측치가 있는 불완비 데이터에서 분류나루를 구축하는 방법을 고찰하였다. 기존의 결측치 처리 방법인 대리 분리 방법의 대안으로 대체 방법으로 결측치를 처리한 후 분류나무를 구축하는 방법을 제안하였다.

주요용어 : 결측치, 대리 분리, 대체 방법, 오분류 오차율.

1. 서론

분류 나무(Classification Tree)는 범주형 목표변수가 취하는 값들이 동질적이 되도록 예측변수의 공간을 배타적으로 연속 분할하여 순차적인 하위마디를 생성함으로써 생성하며, 최종 마디에 의해서 목표변수의 값을 분류 혹은 예측하게 된다. 이러한 분류 나무는 예측의 과정이 나무 구조에 의해 표현되기 때문에 분석자가 그 과정을 쉽게 이해하고 설명할 수 있다는 장점이 있으며, 탐색과 모형 구축이라는 두 가지 기능을 동시에 지닌다. 분류 나무는 CART(Breiman 외 3인, 1984), QUEST(Loh와 Shih, 1997)와 같은 이진 분류(binary split) 나무와 FACT(Loh, Vanichetakul, 1988), C4.5(Quinlan, 1993) CHAID(Kass, 1980)와 같은 다지 분류(nonbinary; multiway) 나무로 구분된다. 또한 CART, C4.5와 같은 비 모수적인 모형 구축 방법과 FACT, Quest와 같은 모수적 모형 구축 방법으로 구분되기도 한다.

이러한 분류 나무들은 모두 완비 자료에 근거하여 개발되었다. 그런데 최근에 분석의 대상이 되는 대용량 관측 데이터(observational data)는 결측치를 포함하는 경우가 많으며, 결측치의 비율이 작다하더라도 결측치의 수는 무시하기 힘든 경우가 많다. 따라서 대용량 데이터를 분석할 때 결측치를 적절히 처리하여 분석 방법이 요구된다.

본 논문에서는 결측치를 포함하는 관측 데이터가 주어진 상황에서 분류 나무를 구축하는 방법을 고찰한다. 기존의 분류 나무 구축 방법들은 결측치가 있는 경우에 대리 분리(surrogate split) 방법을 이용하여 분류나무를 구축하는 것이 일반적이다. 시중에서 판매되고 있는 AnswerTree, E-Miner, S-plus 등에서 사용하는 CART, Quest, C4.5 등에서는 분류 나무를 구축할 때 대리 분리 방법을 이용한다. 대리 분리를 이용하여 결측치를 처리하는 방법에 대안으로 본 논문에서는 대체(imputation) 방법을 고려하여 결측치를 대체한 후 분류나무를 구축하는 방법을 제안한다.

제 2장에서 결측치가 있는 데이터에서 분류나무를 구축하는 기존의 방법들을 고찰하고, 제 3장에서는 결측치 대체 방법을 이용하여 처리한 후 분류 나무 모형의 구축 방법을 제안한다. 마지막으로 제 4장에서는 향 후 연구과제를 언급한다.

1) (130-743) 서울특별시 동대문구 전농동 90, 서울시립대학교 컴퓨터 · 통계학과, 대학원
2) (130-743) 서울특별시 동대문구 전농동 90, 서울시립대학교 컴퓨터 · 통계학과, 부교수
E-mail : kskim@uoscc.uos.ac.kr

2. 분류 나무 모형에서 결측치 처리 방법

대부분의 분류 나무는 나무의 형성과 가지치기 그리고 타당성 평가와 해석 및 예측의 단계를 거친다. 각 단계에서 정지 규칙, 분리기준, 평가기준 등을 어떻게 적용하느냐에 따라 상이한 나무가 생성될 수 있다. 분류 나무의 가지치기(split)를 하는 과정에서 가지를 치는 분리 규칙은 자료의 순수도(purity)가 가장 크게 증가하도록 하는 것이 일반적인데, 여기서 순수도의 증가란 분할 후 각 마디에 속하는 자료의 구성이 서로 동질적이 될 비율이 높다는 의미이다. 즉, 분류 나무는 순수도가 최대 증가하도록, 또는 불순도(impurity) 감소 폭이 최대가 되도록 분리를 시행하게 된다. 그런데 만약 결측치가 존재한다면 결측치가 나타나는 변수에서는 분리를 할 수 없으므로 적절한 처리 방법이 요구되어진다.

CART에서는 결측치가 있는 경우에 다음에 설명하는 대리분리(surrogate split) 방법을 이용하여 가치를 치고 분류 나무를 만들어 나간다. 먼저 마디 t에서 s^*를 t_L과 t_R로 분리하는 최적의 분리 기준이라 하자. S_m을 변수 x_m에서 모든 분리의 집합이라 하고, $\overline{S}_m$을 S_m의 분리에 대한 여집합이라 하자. 임의의 분리기준 $\hat{s}_m \in S_m \bigcup \overline{S}_m$이 식(2.1)의 조건을 만족 할 때 대리 분리(surrogate split)라 정의한다.

$$p(s^*, \hat{s}_m) = \max_{s_m} p(s^*, s_m) \tag{2.1}$$

여기서, $p(s^*, s_m) = p_{LL}(s^*, s_m) + p_{RR}(s^*, s_m)$ 은 s_m이 s^*를 정확히 예측할 확률의 추정치이다. 일반적으로 대리분리 $\hat{s}_m$은 유일하며 변수 x_m에서 최적의 분리 기준 s^*에 의해 분리되는 과정을 가장 정확하게 예측하는 분리기준으로 해석이 가능하다. 만일 변수 x_m에서 결측치가 발생하면 x_m이 값을 가지는 모든 값에서 분리를 수행하여 최적의 분리를 수행하며, 분리 후 불순도 감소의 폭이 최대가 되는 방향으로 결측치를 할당한다.

S-plus의 CART는 이탈도(deviance)함수를 불순도 함수로 정의하여 분리를 수행하며 CART에서와 같은 비용복합 가지치기(cost-complexity pruning)에 의한 최적의 분류나무를 선택한다. 그러나 결측치를 처리하는 방법은 디폴트로써 결측치를 갖는 개체를 나무생성에서 제외시키거나 또는 이산형 예측변수에 대해서는 "NA"라 부르는 새로운 범주로 취급하여 나무를 생성한다. 그러나 연속형 예측변수의 경우, 사분위수에 근거한 범주화를 실시하여 결측치를 새로운 범주로 취급한다. 따라서 연속형 변수가 가지는 중요한 속성을 상실하게 된다. CHAID는 이산형 목표변수에 한하여 피어슨의 카이제곱 검정을 이용하여 다지분리를 수행하는 모형으로 각 마디에서 최선의 분리를 결정하기 위하여 목표변수에 대하여 유의하지 않은 두 예측 변수의 범주를 병합하는 과정을 더 이상 유의하지 않은 두 범주가 발견되지 않을 때까지 반복한다. 또한 결측치가 있는 경우 결측치를 하나의 유동 범주(floating category)로 취급한다. QUEST는 명목형 목표변수에 한하여 CART 또는 CHAID와는 달리 분리 변수 및 규칙을 구분하여 실시하며 CART와 같은 전체탐색(exhaustive search) 알고리즘에 비하여 분리변수 선택에 있어서의 비편향 성질과 분리 수행 시간이 짧다는 특성을 갖는다. 그리고 FACT의 단점을 개선, 발전시킨 모형으로 판별분석을 적용하기 이전에 2-평균 군집 분석(two-means clustering)을 적용하여

2개의 상위범주(superclass)로 나눔으로써 CART와 같은 이진 분리가 수행되도록 개선하였다. 또한 상이한 범주내의 분산(unequal class variance)이 무시되는 단점에 대하여 2개의 상위범주에 대해서 2차 판별분석(QDA)을 적용한다. 그러나 예측 변수가 범주형인 경우 변수의 범주값을 0또는 1인 가벡터(dummy vector)로 치환하여 순서형으로 변환하여 분리를 수행한다.

3. 결측치 대체 후 분류 나무 모형 구축

결측치를 포함하고 있는 학습자료로부터 분류 나무를 생성하고자 한다. 분리 변수로 x_m이 선택되었다고 하자. 그러면, 변수 x_m에 대해 결측치를 갖는 개체는 두 개의 하위 마디 중 어느 쪽으로 할당해야 하는가? CHAID와 S-plus의 방법처럼 결측치를 또 다른 범주로 취급할 수도 있고, 대리 분리 방법을 채택 할 수도 있다. 그러나 여기서는 결측치를 대체하는 방법을 고려하기로 하자. 만일 결측치가 적절한 방법으로 대체되면 하위 마디로의 할당이 분명해질 것이다.

분리 변수로 x_m의 i번째 개체를 x_{mi}로 표현하고 이 값이 결측치라고 하자. 또한 i번째 개체의 목표변수 y_i가 C 개의 값을 갖는 범주형 변수라고 하고 x^*_{mi}를 결측치 x_{mi}에 대한 대체값이라 하자. 결측치 대체의 효과를 높이기 위해서는 적절한 대체 층을 만드는 일이 중요한데, 이를 위해서는 적절한 보조변수를 선택해야 한다. 분류 나무 생성을 위한 대체 층 형성에는 목표변수 Y를 이용할 수 있다. 따라서 목표변수 Y를 이용하여 C개의 대체 층을 구성한 후 대체 층 내에서 결측치 대체를 하여 완비 자료를 만든다.

평균 대체 방법을 C개의 대체 층에 적용하면 결측치 x_{mi}의 i번째 개체가 속하는 대체 층 내에서 식(3.1)와 같은 동일 예측 변수내의 평균값으로 대체함으로써 최초 분리 이전의 뿌리 마디를 얻는다.

$$x^*_{mi} = \bar{x}_{mp}, \quad y_i = y_p = b_c \ , \ p = 1, 2, \ldots, C \tag{3.1}$$

만약, x_m이 이산형 변수라 한다면 위 방법에서 대체 층의 평균값 대신 최빈값이 사용된다.

비 대체 방법을 적용하기 위해서는 결측치를 갖고 있는 예측변수와 상관 관계가 높은 보조변수를 선택하는 작업이 선행되어야 한다. 여기서 x_m과 가장 상관 관계가 높은 예측 변수를 x_r이라 할 때, 비 대체 방법에 의한 적합값은 식(3.2)과 같다.

$$x^*_{mi} = \left(\frac{\bar{x}_{mp}}{\bar{x}_{rp}} \right) x_{ri}, \quad y_i = y_p = b_c \ , \ p = 1, 2, \ldots, C \tag{3.2}$$

핫덱 방법을 적용하기 위해서는 i번째 개체가 속한 대체 층에서 랜덤하게 하나의 값을 선택한다. 이를 h번째 개체의 관측 값 x_{mh}라 한다면 핫덱 방법에 의한 결측치 대체 값은 식(3.4)와 같다.

$$x^*_{mi} = x_{mh}, \quad y_i = y_h = b_c \ , \ h = 1, 2, \ldots, C \tag{3.3}$$

위에서 언급한 대체 방법이 적용되는 단계로는 맨 처음 뿌리 노드 단계를 고려할 수 있고 이 후에 순차적으로 가지를 쳐 나가면서 대체 방법을 활용하는 방법을 고려할 수 있다. 이에 대한 성능 평가가 요구된다.

4. 향 후 연구과제

본 논문에서는 결측치를 포함하는 데이터를 이용한 분류 나무 구축 방법에 대하여 고찰하였다. 기존의 방법이 대리 분리를 이용한 처리 방법임에 비하여 본 논문에서는 결측치를 대체한 후 분류 나무를 구축하는 방법을 제안하였다. 향 후 두 방법에 대한 비교 연구가 필요하다.

참 고 문 헌

[1] Breiman, L., Friedman, J., Olshen, R., and Ston, C., (1984). Classification and Regression Trees. Chapman and Hall, New York, N.Y.

[2] Loh, W.Y, Shin, Y.-S. (1997). Split selection methods for classification trees, Statistica Sinica 7 , 815-840

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

[4] Kass, G. (1980), An exploratory technique for investigating large quantities of categorical data, Applied Statistics. 29, 2, 119-127.

[5] Quinlan, J. R. (1993) C4.5 Programs for machine learning. Morgan Kaufmann, San Mateo, CA.