분류나무와 결측치

곽 영 훈*, 김 지 현**

Missing Values in Classification Trees

Young-Hoon Kwak*, Ji-Hyun Kim**

Abstract

There are many statistical methods for classification or discriminant analysis. The classification tree is one of them. There are several algorithms for classification tree. In this paper a comparison study is done of the performance of those algorithms for classification tree when we have missing values in predictors at hand. With the misclassification error rate as a performance measure, different algorithms are compared for different patterns of missing values.

접수일자: 2001년 12월

* (주)부뜰정보시스템 시스템개발연구소 대리

** 정보통계학과 부교수

Ⅰ. 서 론

판별분석에 적용되는 통계적 방법론 중 하나인 분류나무(Classification Tree)는 해석상 편리하다는 장점 때문에 흔히 쓰이며, 이를 위한 소프트웨어가 많이 개발되어 있다. 우리가 일상적으로 접하는 자료들은 많은 경우 결측치를 포함하고 있는데, 결측치가 있는 경우 분류나무에서 처리하는 방법들이 소프트웨어마다 다르다. 결측치는 통계적 특성에 따라 완전한 랜덤 결측(Missing Completely At Random, MCAR)인 경우와 랜덤 결측(Missing At Random, MAR)인 경우, 그리고 무시할 수 없는 결측(Non-ignorable)인 경우 등 세 가지 유형으로 흔히 구분한다.

본 논문에서는 세 가지 결측의 경우에 대하여 분류나무의 각 소프트웨어들이 어떻게 처리하며 그 분류 결과에 얼마나 차이가 있는지에 대해 연구하였다. Ⅱ장에서 S-PLUS, QUEST, SAS, SPSS 등의 분류나무 프로그램 중 결측치 처리 알고리즘(algorithm)에 대해 소개하고, Ⅲ장에서는 모의실험(simulation) 방법과 결과를 정리하였다.

Ⅱ. 분류나무의 결측치 처리 알고리즘

결측치를 처리하는 알고리즘은 두 가지의 목적 때문에 필요하다. 첫째는 학습 단계(training stage)에서 완전한 자료이든 아니든 최대한의 자료를 사용하기 위함이고, 둘째는 일부 변수에 결측치가 있는 관측치를 나무모형에 의해 분류하기 위함이다.

Breiman et al. (1984)이 제안한 CART (Classification And Regression Tree) 알고리즘에서는 결측치가 있을 때 대리분리(surrogate splits)를 사용한다.

S-PLUS에서의 결측치 처리 방법

S-PLUS에서는 결측치를 처리하는 '*na.tree.replace()*' 라는 함수가 있다. 이 함수는 결측치가 있을 때 결측치를 새로운 범주로 간주하여 변수의 값을 하나 더 생성하여 결측치를 처리한다. 만일 변수가 연속형일 경우에는 연속형 변수를 범주형으로 변환을 하여 같은 방법으로 처리를 한다.

예를 들면 범주형 자료가 a, b, c 세 개의 범주를 갖는 변수라면, 만일 이 변수에 결측치가 있을 때 a, b, c, NA 이렇게 네 개의 범주를 갖는 변수로 결측치를 처리하게 된다. 연속형 변수에 결측치가 있으면 범주형으로 처리를 한다. 예를 들어 1, 1, 2, 3, 3, 4, 5 와 같은 값을 갖는 연속형 변수가 결측치를 포함한다면, 모든 자료를 범주형으로 간주하여 1, 2, 3, 4, 5 다섯 개의 범주와 결측치 NA 한 개, 모두 여섯 개의 범주를 갖는 범주형 변수로 간주한다. 연속형 자료인 경우에는 1은 2보다 1만큼 크다는 것이 의미가 있으나, 범주형으로 처리를 하게 되면 이러한 연속형 자료의 중요한 정보를 모두 잃어버리게 된다. 따라서, 설명변수가 모두 연속형 자료라면 S-PLUS의 성능은 다른 방법보다 나쁜 결과를 보일 것으로 예상된다.

QUEST에서의 결측치 처리 방법

QUEST에 관한 자세한 내용은 Loh et al. (1997)을 참고하기 바란다.

QUEST에서 결측치가 있을 때 마디 평균(node mean)이나 최빈값(node mode)을 이용하여 결측치를 추정하는 방법을 사용하고 있다.

예를 들어 어느 마디에서 분리기준으로 a, b, c 세 개의 범주를 갖는 변수가 선택되었을 때, 이 변수에 결측치가 있으면 마디에서 각 범주의 빈도를 계산하여 가장 많은 빈도를 나타내는 범주로 결측치를 추정하여 사용한다. 연속형 자료인 경우에는 마디의 평균치로 결측치를 대체하여 사용한다.

SAS에서의 결측치 처리 방법

SAS에서는 결측치 처리 방법으로 두 가지를 제공한다. 하나는 CART에서와 같은 대리분리(surrogate rule)이고 다른 한 방법은 결측치를 하나의 새로운 값으로 인식을 하는 방법인데, 두 방법 중 하나를 선택적으로 사용할 수 있다. 본 논문에서는 대리분리를 사용하였다.

SPSS 에서의 결측치 처리 방법

SPSS에서는 C5.0 이라는 알고리즘을 사용하고 있다. C5.0에 대한 자세한 내용은 Michael et al.(1997)을 참고하기 바란다. 본 논문에서는 CART 알고리즘과의 차이점에 대해 간략히 언급한다.

CART에서는 분리의 기본은 이진 분리(binary split)이다 C5.0에서는 각 마디에서 여러 개의 자식마디로 분리가 가능하다. C5.0에서 설명변수가 연속형 변수인 경우에는 CART에서와 매우 유사하다. 그러나, 범주형 변수인 경우에는 상당히 다르다. C5.0에서는 범주형 변수가 분리 변수로 선택이 되면, 기본적으로 범주의 수만큼의 분리가 발생한다. 예를 들어 '색상' 이라는 변수가 분리 기준으로 선택이 되었고 범주로는 '빨강', '노랑', '파랑', '검정', '흰색' 등 다섯 개의 값을 갖는 변수라면, 각 범주에 대응하는 다섯 개의 자식마디로 분리가 발생하게 된다. 본 논문의 관심인 결측치 처리 방법도 다른 프로그램과 상당히 다르다. C5.0 에서는 나무모형을 생성할 분리 기준으로 한 변수를 선택한 후, 그 변수에 결측치가 있는 경우에는 결측치가 없는 값으로만 분리 기준을 만든다. 학습 자료에 대해서는 분리기준으로 사용된 변수에 결측치가 포함되어 있다면 그 관측치는 분리를 하지 않는다.

Ⅲ. 모 의 실 험

현재 상당히 많은 나무모형 알고리즘이 구현되어 있으나, 본 논문에서는 S-PLUS, QUEST, SAS와 SPSS에서 제공하는 알고리즘을 사용하였다. 모의 실험을 위해 3가지 자료로 비교연구를 하였는데, 지면 관계상 파도모양(waveform) 자료에 대한 결과만 보고한다. 독일의 은행자료와 미국의 피임방법 자료에 대한 모의실험은 곽영훈(2000)을 참고하기 바란다.

파도모양 자료(Breiman et al. 1984)는 $h_1(t)$, $h_2(t)$, $h_3(t)$ 형태에 기초를 하여 세 개의 집단으로 되어있다. 각 집단은 $h_1(t)$, $h_2(t)$, $h_3(t)$ 중에서 두 개씩의 확률적 조합으로 이루어져있고, 설명변수는 모두 21개로 되어있다. 집단 1의 변수들을 만들려면 독립적인 균일분포를 따르는 난수

u 를 생성하고, 21개의 표준정규분포를 따르는 난수 $\varepsilon_1,\cdots,\varepsilon_{21}$ 를 생성한다. 그 다음은 아래와 같은 함수를 이용하여 변수 x를 생성한다.

$$x_m = uh_1(m) + (1-u)h_2(m) + \varepsilon_m$$

집단 2의 변수들을 생성하려면 다음과 같은 함수를 이용하여 생성한다.

$$x_m = uh_1(m) + (1-u)h_3(m) + \varepsilon_m$$

집단 3의 변수들을 생성하려면 다음과 같은 함수를 이용하면 된다.

$$x_m = uh_2(m) + (1-u)h_3(m) + \varepsilon_m$$

위와 같은 방법으로 각각의 집단을 200개씩 600개로 학습 자료를 생성하였다. 실험의 반복은 10회로 하였다. 또한 나무모형의 성능을 테스트하기 위하여 3000개의 독립적인 테스트 표본을 만들었다. 즉, 600개의 학습 표본으로 나무모형을 생성하고 3000개의 테스트 표본으로 정확성을 측정하였다. 시험의 정확성을 위하여 10회 반복 시험을 하였다.

결측치 생성 방법

결측치의 종류에는 여러 가지가 있으나 본 논문에서는 완전한 랜덤(Missing Completely At Random, MCAR)인 경우와 랜덤(Missing At Random, MAR)인 경우 그리고 무시할 수 없는 경우(Non-ignorable) 등 세 가지의 경우에 대하여 실험을 실시하였다. 세 가지 유형의 결측치가 분류나무에 어떠한 영향을 미치는지 살펴보고자 한다.

1. MCAR : 전체 x변수의 값 중 $h_1(t)$, $h_2(t)$, $h_3(t)$가 가장 큰 값을 갖는 점인 x_7, x_{11}, x_{15}를 랜덤하게 결측으로 처리하는 방법을 택하였다. 또한 대체처리(surrogate)와 다른 방법과의 공평성을 유지하기 위하여 선택된 점 양쪽의 점도 결측치로 처리하였다. 각각 10, 20, 30, 40, 50%의 비율로 결측치를 줌으로써 영향을 살펴보고자 한다.
2. MAR 인 경우 : 파도모양 자료는 세 개의 봉우리를 갖는 각각의 h(t) 함수 중 두 개의 결합 형태이다. 가중치로서 각각 균일분포(0,1)의 난수를 생성하여 주었으므로 어느 한 봉우리가 높으면 다른 쪽은 낮은 형태로 나타날 것으로 예상하여, 집단 1의 경우, h_1 이 최고점인 변수 x_7의 값이 2이상이면 x_{15}의 값과 양옆의 값을 결측치로 만들었고, 집단 2의 경우 h_3 이 최고점인 변수 x_{11}의 값이 2이상이면 x_7의 값과 양옆의 값을 결측치로 만들었으며, 끝으로 집단 3의 경우 h_2 가 최고점인 변수 x_{15}의 값이 2이상이면, x_{11}의 값과 양옆의 값을 결측치로 만들었다. 주어진 조건을 만족하는 관측치에 대하여 각각 20%, 40%, 60%, 80%, 100% 의 비율로 결측치를 생성하여 실험을 실시하였다.
3. 무시할 수 없는 경우 (Non-Ignorable case) : 집단 1의 경우 변수 x_{15}가 2이상이면 x_{15}자신과 양옆의 값을 결측치로 만들었고, 집단 2의 경우 변수x_7이 2이상이면 x_7 자신과 양옆의 값을 결측치로 만들었으며, 집단 3의 경우 변수x_{11}이 2이상이면 x_{11}자신과 양옆의 값을 결측치로 만들었다. 이 경우에도 주어진 조건을 만족하는 관측치에 대하여 각각 20%, 40%, 60%, 80%, 100%의 비율로 결측치를 생성하였다.

모의실험 결과

1. S-PLUS 결과

<표 1>에서 <표 3>까지는 결측치의 유형과 정도에 따라 시험 자료에 대한 모의실험을 10회 반복하여 오분류 오차율의 평균과 표준편차를 나타낸 것이다. MCAR 인 경우는 결측치가 발생하였을 때 오분류 오차율이 증가하기는 하나 결측치가 많아지면서 변화가 심하지 않다. 주어진 조건에서 20 50%의 결측치가 있더라도 오분류 오차율은 크게 증가하지 않는다. MAR인 경우와 무시할 수 없는 경우에는 결측치의 비율이 증가함에 따라 오분류 오차율이 크게 증가하는 추세를 보이고 있다.

S-PLUS에서 사용되어지는 결측치 처리 방법이 앞에서 언급을 하였듯이 설명변수가 연속형 자료로만 이루어진 파도모양 자료에는 다른 방법보다 오분류 비율이 나쁘게 나타난다. 그 이유는 연속형 자료의 중요한 정보를 잃어버리기 때문이다.

<표 1> S-PLUS결과 (MCAR인 경우)

결측치	전체	10%	20%	30%	40%	50%
오차율	0.2887	0.4465	0.3696	0.3410	0.3489	0.3534
표준편차	0.0247	0.0610	0.0281	0.0293	0.0255	0.0256

<표 2> S-PLUS 결과 (MAR인 경우)

결측치	전체	20%	40%	60%	80%	100%
오차율	0.2887	0.4783	0.4990	0.5124	0.5758	0.6232
표준편차	0.0247	0.0186	0.0392	0.0355	0.0559	0.0219

<표 3> S-PLUS 결과 (NI인 경우)

결측치	전체	20%	40%	60%	80%	100%
오차율	0.2887	0.4619	0.4933	0.5467	0.5866	0.6045
표준편차	0.0247	0.0583	0.0322	0.0523	0.0400	0.0472

2. QUEST 결과

QUEST에서는 MCAR일 때 오분류 오차율이 증가하기는 하나 그 정도가 S-PLUS에서 보다 상당히 작게 나타나고 그리 크지 않다. MAR이나 무시할 수 없는 경우, 조건을 만족하는 값 모두가 결측치가 되는 극단적인 경우에는 오분류 오차율이 크게 증가를 한다. 그러나 그 외에는 오차율이 증가하기는 하나 비교적 안정적인 모습을 보이고 있다.

<표 4> QUEST 결과 (MCAR인 경우)

결측치	전체	10%	20%	30%	40%	50%
오분류	0.2717	0.2781	0.2852	0.2960	0.2984	0.3063
표준편차	0.0174	0.0217	0.0223	0.0261	0.0156	0.0256

<표 5> QUEST 결과 (MAR인 경우)

결측치	전체	20%	40%	60%	80%	100%
오분류	0.2717	0.2724	0.2895	0.3080	0.3237	0.4427
표준편차	0.0174	0.0138	0.0151	0.0227	0.0241	0.0232

<표 6> QUEST 결과 (NI인 경우)

결측치	전체	20%	40%	60%	80%	100%
오분류	0.2717	0.28242	0.2921	0.3018	0.3112	0.4912
표준편차	0.0174	0.0124	0.0197	0.0147	0.0118	0.1018

3. SAS 결과

SAS에서는 QUEST의 결과와 매우 유사하게 나타났다. MCAR인 경우에는 오분류 오차율이 증가하기는 하나 그 정도가 S-PLUS보다 상당히 작게 나타나고 QUEST와 비슷하다. 또한 MAR인 경우와 무시할 수 없는 경우, 조건을 만족하는 값 모두가 결측치가 되는 극단적인 경우에는 상당히 나쁘게 나타난다. 그 외에는 증가하는 추세가 있기는 하지만 대체적으로 안정적이라고 할 수 있다.

<표 7> SAS 결과 (MCAR인 경우)

결측치	전체	10%	20%	30%	40%	50%
오분류	0.2784	0.2802	0.2823	0.2935	0.2922	0.3059
표준편차	0.0162	0.0210	0.0195	0.0227	0.0176	0.0272

<표 8> SAS 결과 (MAR인 경우)

결측치	전체	20%	40%	60%	80%	100%
오분류	0.2784	0.2673	0.2731	0.3095	0.3097	0.3486
표준편차	0.0162	0.0231	0.0109	0.0245	0.0215	0.0762

<표 9> SAS 결과 (NI인 경우)

결측치	전체	20%	40%	60%	80%	100%
오분류	0.2784	0.2821	0.2855	0.3122	0.3357	0.4199
표준편차	0.0162	0.0168	0.0164	0.0260	0.0201	0.0557

4. SPSS 결과

MCAR인 경우와 MAR인 경우에 모두 증가를 하기는 하지만 그 정도가 상당히 작다. 무시할 수 없는 경우에는 증가폭이 앞의 두 경우보다는 뚜렷하다. 이는 앞의 다른 프로그램에서의 결과와 매우 유사하다. 그러나, SPSS에서 결측치를 처리하는 방법이 앞에서 언급하였던 결측치를 처리하는 알고리즘이 필요한 이유 두 가지 모두에 부합하지 않은 방법이므로 결측치 처리를 위한 알고리즘으로 적절하지 않다.

<표 10> SPSS 결과 (MCAR인 경우)

결측치	전체	10%	20%	30%	40%	50%
오분류	0.2688	0.2747	0.2799	0.2890	0.2989	0.2919
표준편차	0.0124	0.0068	0.0119	0.0136	0.0105	0.0089

<표 11> SPSS 결과 (MAR인 경우)

결측치	전체	20%	40%	60%	80%	100%
오분류	0.2688	0.2730	0.2800	0.2856	0.2934	0.3106
표준편차	0.0124	0.0092	0.0124	0.0104	0.0148	0.0188

<표 12> SAS 결과 (NI인 경우)

결측치	전체	20%	40%	60%	80%	100%
오분류	0.2784	0.2821	0.2855	0.3122	0.3357	0.4199
표준편차	0.0124	0.0135	0.0102	0.0126	0.0128	0.0358

Ⅳ. 결론

우리의 주변에서 얻어지는 자료는 흔히 결측치를 포함하고 있다. 결측치의 유형에 따른 S-PLUS, QUEST, SAS, SPSS의 분류나무의 영향을 살펴보았다. 비록 한 인위적 자료에 대해서이지만 각 결측치의 유형과 정도에 따라 분류나무의 오분류 오차율이 변하는 모습을 보았다.

결측치 처리에 대한 효과를 알고리즘별로 정리를 하면, S-PLUS에서는 자료가 연속형 변수인 경우에는 그 성능이 많이 떨어진다는 것을 확인하였으며, SPSS C5.0에서는 결측치를 처리하는 목적 두 가지에 모두 적절하지 않은 방법이다. 결측치가 많은 경우에는 분리되지 않는 많은 관측치가 남게 되어 예측에 있어서 정확성이 많이 떨어진다는 것을 확인하였다. SAS에서 사용되는 대리분리나 QUEST에서 사용되는 마디의 평균 혹은 마디의 최빈값으로 결측치를 처리할 때 MCAR인 경우에는 결측치가 어느 정도 있더라도 오분류 오차율에는 많은 영향을 주지 못한다. MAR인 경우와 무시할 수 없는 경우에는 결측치의 비율이 많아지면 나빠지는 것을 보았다. 그러나 극단적인 경우를 제

외하면 대체로 안정적인 모습을 보여주고 있음을 확인할 수 있었다.

참 고 문 헌

[1] Breiman, L., Friedman, J., Olshen, R, and Stone, C. (1984). *Classification and Regresstion Trees*. Champman and Hall, New York, NY.

[2] Loh, W.-Y., and Shih, Y.-S. (2000). A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms, *Machine Learning Journal*.

[3] Loh, W.-Y. and Shih, Y.-S. (1997). Split selection methods for classification trees. *Statistica Sinica*.

[4] Venables, W. N. and Ripley, B. D. (1997). *Modern Applied Statistics with S-Plus*. Springer, New York, NY, 2nd edition.

[5] Chambers, J. M. and Hastie, T. J. (1990). *Statistical Models in S*. Wadsworth.

[6] Berry, M. J. A, Linoff, Gordon (1997). *Data Mining Techniques For Marketing, Sales, and Customer Support*, John Wiley & Sons, Inc.

[7] **곽영훈** (2000). CART**에서의 결측치 처리**, **숭실대학교 석사 논문**.