응용통계연구 제 **16**권 1호, 2003년, pp. 45-53

회귀의사결정나무에서의 관심노드 찾는 분류 기준법 *

이영섭[1)]

요 약

의사결정나무 분석 기법 중 하나인 회귀의사결정나무는 연속적인 반응변수를 예측할 때 사용된다. 나무 구조를 형성할 때, 전통적인 분류 기준법은 왼쪽과 오른쪽 자식노드의 불순도를 결합하여 이루어진다.

그러나 본 논문에서 제안하는 새로운 분류 기준법은 관심있는 한 쪽만 선택하고 다른 나머지 자식노드는 큰 관심이 없어 무시함으로써 더 이상 결합하여 구하는 것이 아니다. 따라서 나무 구조는 불균형적일 수 있으나 이해하기가 쉽다. 즉, 관심있는 부분집합을 가능한 한 빨리 찾음으로써 단지 몇 개의 조건으로 쉽게 표현할 수 있으며, 정확도는 다소 떨어지지만 설명력은 아주 높다.

주요용어: 회귀의사결정나무, CART, 나무구조의 해석력

1. 서론

데이터 마이닝이 사회 전반에 걸쳐 CRM(Customer Relationship Management)와 함께 많은 분야에서 관심을 가지고 활발한 연구와 응용이 진행되고 있다. 여러 가지 방법 중에서 의사결정나무(Decision Trees)분석 방법은 비록 통계학을 전공하지 않아도 쉽게 이해할 수 있어서 많이 사용되고 있다. 가장 많이 알려진 참고 서적은 CART(Breiman et al., 1984)이며, S-plus를 이용한 나무구조는 S-plus 매뉴얼(StatSci., 1995) 또는 Venables and Ripley(1997)에서 찾을 수 있다.

의사결정나무 방법을 일반적인 통계기법들과 비교할 때 몇 가지의 장점이 있다.
첫째로 이것은 여러 개의 rules로 되어 있기 때문에 결과를 해석하고 이해하기가 쉽다. 따라서, 통계학자가 아니라도 쉽게 이해할 수 있는 예측 모델이다.
둘째로, 비모수적인 방법으로써 어떤 특정한 함수 형태를 요구하지도 않고, 모수적인 모형처럼 선형성, 정규성 또는 공분산성등의 가정을 필요로 하지 않는다. 따라서 독립 변수의 변환없이 그 자체를 사용할 수 있다.
셋째로, 독립변수 선택을 미리 하지 않아도 된다. 따라서 전진, 후진, 단계적 변수 선택방법 등이 필요하지 않으며 모든 가능한 독립변수를 사용할 수 있다.
넷째, 결측치나 이상치를 가진 자료를 잘 처리할 수 있다. 의사결정나무는 주분류변수 (primary splitting variable)가 결측치를 가진 어떤 자료를 분류할 때, CHAID(Morgan와 Sonquist, 1963, Hatigan, 1975)에서처럼 결측치를 또 다른 범주로 생각하던지 또는 CART(Breiman

* 본 연구는 동국대학교 논문게재 연구비 지원으로 이루어졌음
1) (110-715) 서울 중구 필동 3가 26, 동국대학교 이과대학 통계학과 조교수
E-mail : yung@dongguk.edu

et al., 1984)처럼 다른 분류변수(대리변수, surrogate variable)를 사용하여 결측치를 처리하고 있다. 이상치에 대해서는 의사결정나무의 특성상 하나의 노드로 분류되어진다. 그러나, 이러한 장점에도 불구하고 몇 가지 단점이 있다. 첫째로, 불안정성이다. 조그마한 자료의 변화에도 나무구조는 많이 달라질 수 있다. 이것은 나무구조의 국소최적설계(local optimization) 때문이다. 그러나 일반적으로 상위단계의 구조모형은 크게 다르지 않고, 하위단계에서 조금씩 달라질 뿐이며 전체 나무의 정확도에는 큰 차이가 없다. 둘째로, 분류값의 주변에 대한 엄격한 분류(hard split)이다. 즉, 주분류변수의 분계점(threshold) 주변의 조그마한 값의 변화에도 예측값 ($\hat{y}$)의 커다란 변화를 가져 올 수 있다. 이것은 나무구조가 단계함수(step function)로 되어지는 구조이기 때문이다.

의사결정나무는 반응변수(response variable, target variable)의 성질에 따라 크게 두가지로 나눈다. 반응변수가 범주형 변수(categorical variable)일 때는 분류의사결정나무(Classification Trees)라 하며, 연속형 변수(continuous variable)일 때는 회귀의사결정나무(Regression Trees)라 한다.

본 논문은 회귀의사결정나무를 형성할 때의 새로운 분류 기준법(splitting criteria)을 제안하고자 한다. 전통적인 분류 기준법은 왼쪽과 오른쪽 자식노드의 불순도를 관찰치의 숫자를 가중치로 하여 평균 합으로 하였다. 그러나 본 논문에서는 평균 합 대신 관심있는 한쪽 노드의 불순도를 가지고 나무 구조를 형성하고자 한다.

이러한 방법으로 형성된 나무 구조는 비록 불균형적인 형태를 이루고 있지만 훨씬 간단하고 설명력이 있다. 나무 구조의 형성 과정과 분류함수 결정기준과 해석력에 관한 모든 설명은 Lee(2001)을 참조하기 바란다. Lee(2001)가 분류의사결정나무에 새로운 분류 기준법을 적용한 반면, 본 논문은 회귀의사결정나무에 설명력을 높이는 새로운 분류 기준법을 적용, 응용하고자 한다.

2. 회귀의사결정나무(Regression Trees)

2.1. 전통적인 분류방법

Lee(2001)에서 언급하였듯이 의사결정나무는 어떤 분류기준에 의하여 반복적으로 각 단계에서 분류함수 -분류 변수와 분류 분계점-를 이용하여 나무구조를 구축하고 있다. 이때 전통적인 분류기준법은 왼쪽자식노드와 오른쪽자식노드의 불순도를 결합하여 계산한다. 분류의사결정나무(Classification Trees)는 각 노드의 계급0과 1의 확률에 의해서 불순도가 결정되는데 반하여, 회귀의사결정나무는 분산이나 평균에 의하여 구하여지는데, 왼쪽과 오른쪽 자식노드 관찰치 수로 가중치된 분산평균으로 계산하는 것이 전통적인 방법이다. 먼저 기호를 정리하자면

$$\hat{\mu}_L = \frac{1}{N_L}\sum_{n\in L} y_n \qquad \hat{\sigma}_L^2 = \frac{1}{N_L}\sum_{n\in L}(y_n - \hat{\mu}_L)^2$$

$$\hat{\mu}_R = \frac{1}{N_R}\sum_{n\in R} y_n \qquad \hat{\sigma}_R^2 = \frac{1}{N_R}\sum_{n\in R}(y_n - \hat{\mu}_R)^2$$

여기에서 L,R은 각각 왼쪽, 오른쪽 자식 노드를 나타내며, 추정치들은 각각 왼쪽과 오른쪽 자식노드의 정규분포 하에서 구해진 최우추정치(MLE)들이다. CART(Breiman et al., 1984)의 분류 기준법은 아래와 같다.

$$crit_{LR} = \frac{1}{N_L + N_R}[N_L\hat{\sigma}_L^2 + N_R\hat{\sigma}_R^2]$$
$$= \frac{1}{N_L + N_R}[\sum_{n\in L}(y_n - \hat{\mu}_L)^2 + \sum_{n\in R}(y_n - \hat{\mu}_R)^2]$$

따라서 전통적인 분류기준은 위에서처럼 왼쪽과 오른쪽 노드의 가중치된 평균 분산을 사용한다.

2.2. 새로운 분류 기준법 - 관심 노드 분류법(Interesting node splitting criteria)

전통적인 분류 기준 방법으로는 우리가 관심이 있는 부분집합, 즉, 반응변수의 평균이 아주 높거나 낮은 경우, 또는 분산이 아주 적은 경우(순수집단)만을 가능한 한 빨리 찾고자 할 때는 어려울 경우가 있다. 예를 들면, ⟨ 그림 2.1 ⟩ 과 같은 경우이다.
⟨그림 2.1(a)⟩의 경우는 x축의 오른쪽 부분집합의 분산이 아주 작은 경우이고, ⟨그림 2.1(b)⟩의 경우는 오른쪽에 극단적으로 평균이 높은 부분집합이 있는 경우이다. 이러한 경우에 CART는 올바르게 우리가 원하는 데로 분할하지 못한다. 왜냐하면 (a)의 경우는 등분산을 가정하기 때문에, (b)의 경우는 x=300 부근에서 분할할 것이다. 본 논문에서는 이러한 경우에 우리가 관심이 있는 부분집단 - 순수집단 ($\hat{\sigma}^2$이 작은 경우) 또는 극단집단($\hat{\mu}$이 크거나 작은 경우)을 가능한 한 빨리 찾는 것이 목적이다. 따라서 왼쪽과 오른쪽 노드를 결합하여 불순도를 측정하지 않고 새로운 분류기준법을 제안한다.

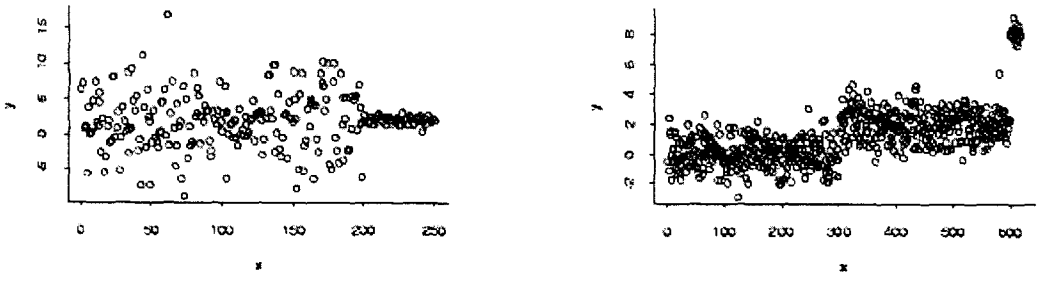

그림 2.1: 순수하거나 극단적인 값을 갖는 부분 집합의 예

1) 단일노드 순수도를 기준으로 한 새로운 분류법(One-sided purity)
이 기준법은 가능한 한 빨리 순수한 노드(pure node)를 찾는 것이다.

$$crit_{LR} = min(\hat{\sigma}_L^2,\ \hat{\sigma}_R^2)$$

이 기준법을 모든 가능한 분류변수 및 분계점에 적용하여 가장 작은 값을 갖을 때를 그 단계에서의 분류함수로 정한다. 즉, 왼쪽이든 오른쪽 자식노드이든 상관없이 가장 작은 분산값을 가지는 부분집합을 가진 노드를 찾는 것이다. 일반적으로 순수한(purity) 노드는 일단 한번 분리되면 더 이상 분리되지 않고, 다른 쪽의 노드만 계속해서 분리되기 때문에 나무구조가 불균형적인 형태를 가진다.

2) 단일 극한 평균값을 기준으로 한 새로운 분류법(One sided extremes with high or low value response)

이 기준법은 반응 변수의 평균값이 큰(또는 작은) 부분집합의 노드를 가능한 한 빨리 찾는 것이 목적이다.

$$crit_{LR} = max(\hat{\mu}_L,\ \hat{\mu}_R)$$

또는

$$crit_{LR} = min(\hat{\mu}_L,\ \hat{\mu}_R)$$

이 기준법을 앞의 1)에서 설명한 것과 같은 방법으로 하여 분류함수를 찾는다. 이 기준법도 평균값이 큰(또는 작은) 부분집합만 관심이 있고 다른 부분집합에는 관심이 없기 때문에 불균형적인 나무 형태를 가질 수 있다. 그러나 우리가 반응변수의 값이 큰 그룹을 빨리 찾고 싶을 때는 제안된 분류법이 아주 유용하며 간단한 몇 가지 조건만으로도 찾고자하는 부분 집합을 설명할 수 있다. 이러한 현상은 뒤의 실제 예에서 더 자세히 설명할 것이다.

3. 예제를 통한 비교

Harrison 와 Rubinfeld(1978)에 의하여 잘 알려진 보스톤 지역의 집 값에 대한 자료를 가지고 비교 분석하였다. 이 자료는 특히 Belsey, Kuh 와 Welch(1980)에 의하여 더욱 더 자세히 알려지면서 많은 연구의 예제로 쓰여지고 있으며 다음의 URP에서 구할 수 있다. (Merz 와 Murphy(1998), http://www.ics.uci.edu/~mlearn /MLRepository.html)

보스톤 지역의 506개 구역(tract)별로 아래 13개 변수들을 조사하고 그 구역에서의 집 값의 중앙값(median)을 반응변수로 하였다. 13개 변수와 반응변수는 다음과 같다.

- CRIM : 범죄율
- ZN : 25,000평방 피트 이상의 주거지 비율
- INDUS :공업지역 비율
- CHAS : 찰스강 인접 여부(인접 : 1, 그 외 : 0)
- NOX : 대기중 일산화질소 (pphm)
- RM : 거주지별 방 수의 평균
- AGE : 1940년 이전에 지어진 건물 비율
- DIS : 보스톤의 5대 고용 중심지와의 거리
- RAD : 고속도로 근접 지수

· TAX : 재산세($ 10,000)
· PTRATIO : 학생과 교사 비율
· B : 흑인 비율 지수
· LATAT : 극빈자 비율 (%)
· MEDV : 주택 가격의 중앙값(반응변수)

비교를 위해서 각각의 나무구조는 16개의 끝노드(terminal node)와 노드당 최소 자료수는 23개로 하고 가지치기는 하지 않았다.
그림에서 m은 반응변수의 평균이며 sz는 각 노드가 전체 중 차지하는 비율(노드크기)이다.

3.1. CART ($N_L\hat{\sigma}_L^2 + N_R\hat{\sigma}_R^2$) (⟨그림 3.1⟩)

⟨그림 3.1⟩에서 볼 수 있듯이 아주 균형적인 나무구조이다. RM이 아주 중요한 변수로서 3번, LSTAT은 6번이나 나타났다. 그러나 아주 작은 집값(m=10.2)을 갖는 노드를 찾기 위해서는 여러 번의 조건을 만족하여야 한다. $R^2 = 0.8$ 이다.

3.2. 단일노드 순수도(One-sided purity, min($\hat{\sigma}_L^2$, $\hat{\sigma}_R^2$)) (⟨그림 3.2⟩)

9단계까지 가는 불균형적인 나무구조이다. PTRATIO가 먼저 나타나고 가능한 한 빨리 순수한 노드를 계속해서 찾아나간다. R^2는 CART보다 조금 떨어진 0.75이다.

3.3. 저평균노드(One-sided low means, min($\hat{\mu}_L$, $\hat{\mu}_R$)) (⟨그림 3.3⟩)

아주 불균형적인 나무 구조이며 RM과 LSTAT가 반복적으로 여러 번 나타난다. 이것은 RM과 LSTAT가 반응변수에 교호작용을 일으키면서 단조함수를 나타내고 있음을 알 수 있다. 즉, B>100.08 에서 반응 변수는 LSTAT과 단조 감소 형태를 이루고 있다가 다시 RM과 단조 증가 형태를 이루고 있다. 한편 단지 CRIM>15.79 로써 저평균노드(m=10.15)를 한번에 찾을 수 있어서 아주 설명력이 있다. 즉 우리가 원하는 노드를 쉽게 그리고 간단히 설명할 수 있다.R^2 는 0.76으로 CART에 비해서 다소 떨어졌지만 우려할 만한 정도는 아니며, 본 논문에서는 앞서 말했듯이 정확성에는 별 관심이 없다.

요약하자면, CART 나무구조는 $R^2 = 0.8$로써 다소 높지만 가장 해석하기 어려운 나무구조를 가지고 있다. 단일 순수노드는 해석하기가 다소 쉽고, 저평균노드 찾기의 나무구조는 가장 해석력이 뛰어나다. 물론 뒤의 두 나무구조는 매우 불균형적이지만, 이 때문에 더욱 더 우리가 찾고자 하는 부분집합은 찾기 쉽다. 해석력이란 면에 있어서는 다음의 두 가지 의미가 있다. 첫째, 저평균 또는 고평균 노드 찾기에서는 그러한 노드들을 단지 몇 개의 조건으로 쉽게 표현할 수 있고 또 설득력이 있다. 둘째, 불균형적인 구조로 인하여 반복적으로 하나 또는 두 개의 독립변수의 부분들을 분리해 나감에 따라 이들과 반응변수들의 단조 증가 또는 감소의 관계가 있음을 알아 낼 수 있다.

이러한 단조 함수의 관계가 발견되면, 의사결정나무 방식 대신 교호작용을 포함한 선형 또는 가법모형(additive model)을 사용하는 것이 더 적당하다. 예를 들면, 위의 예에서 저평균노드 모형(⟨그림 3.3⟩)의 경우에 선형모형으로 나타내자면 아래와 같다.

$$MEDVAL = \beta_{CRIM} * 1_{[CRIM>15.79]} + \beta_B * 1_{[B\leq 100.08]} * 1_{[CRIM\leq 15.79]}$$
$$+ \beta_{LSTAT} * LSTAT * 1_{[LSATA>10.14]} * 1_{[B>100.08]} * 1_{[CRIM\leq 15.79]}$$
$$+ \beta_{RM} * RM * 1_{[LSTAT\leq 10.14]} * 1_{[B>100.08]} * 1_{[CRIM\leq 15.79]} + error$$

4. 결론

이상에서와 같이 전통적인 나무구조(CART)는 다소 정확성은 있지만, 균형적인 구조로써 우리가 원하는 노드(예, m=10.20(저평균))를 찾으려면 여러 번의 단계를 거쳐야 하기 때문에 설명하기가 복잡하다. 그러나 〈그림 3.3〉에서처럼 저평균노드 찾기를 하면 비록 정확성은 다소 떨어지지만 단 한번의 조건(CRIM>15.79)으로 저평균(m=10.15)을 찾을 수 있다. 아주 불균형적인 구조이지만 우리가 관심이 있는 노드만을 찾고 다른 쪽은 관심이 없기 때문에 이러한 현상이 일어난다고 할 수 있다. 데이터 마이닝에서는 어떠한 기법이나 알고리즘도 모든 경우에 다 우수할 수는 없다. 데이터에 따라 또는 분석 목적에 따라 사용하는 기법이나 알고리즘도 달라져야 한다. 이러한 의미에서 본 논문은 하나의 선택으로, 수요자의 요구에 적합한 부분집합을 빨리 찾아 간결하게 설명하는 기법으로써 유용하다.

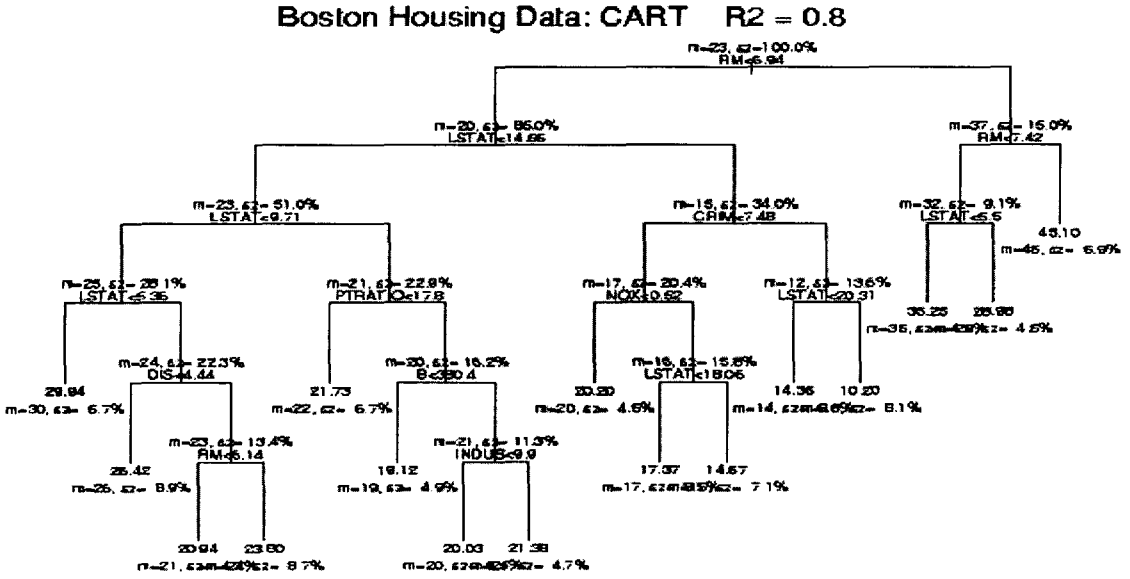

그림 3.1 전통적인 방법 (CART)

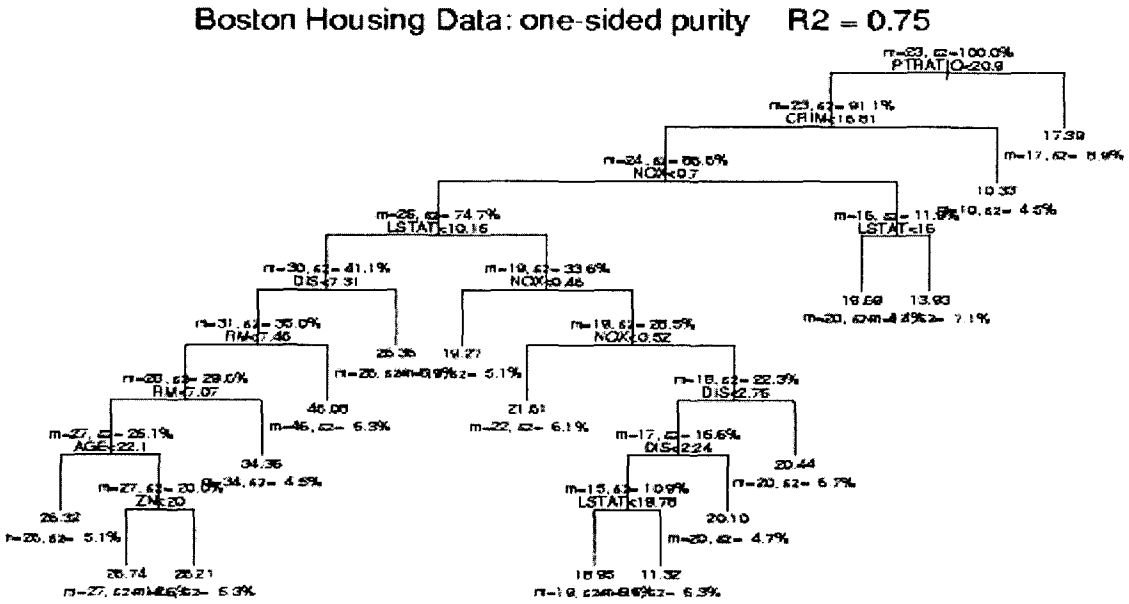

그림 3.2 단일 순수 노드

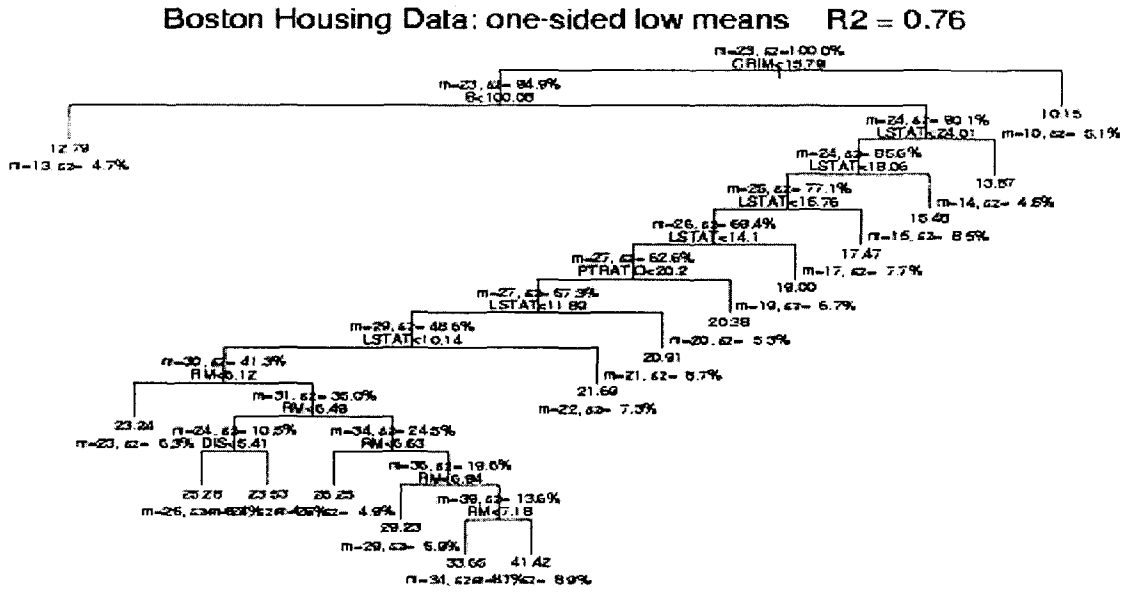

그림 3.3 저평균 노드

참고문헌

[1] Belsley, D.A., Kuh, E., and Welsch, R. E. (1980), Regression Diagnostic, New York, NY: John Wiley & Son, Inc..

[2] Breiman, L., Friedman, J.H., Olshen, R. A., and Stone, C. J. (1984), Classification and Regression Trees, Pacific Grove, CA: Wadsworth.

[3] Harrison, R. J., and Rubinfeld, D. L. (1978), Hedonic Prices and the Demand for Clean Air, Journal of Environmental Economics and Management, 5, 81-102.

[4] Hartigan, J. A. (1975), Clustering Algorithms, New York, NY: John Wiley & Sons, Inc..

[5] Lee, Y-S. (2001), New Splitting Criteria for Classification Trees, The Korean Communications in Statistics. vol. 8, 885-894.

[6] Merz, C. J., and Murphy, P. M. (1998), UCI repository of machine learning data bases (http://www.ics.uci.edu/~ mlearn/MLRepository.html).

[7] Morgan, J. N., and Sonquist, J. A. (1963), Problems in the Analysis of Survey Data, and a Proposal, Journal of the American Statistical Association, 58, 415-434.

[8] StatSci (1995), S-PLUS Guide to Statistical and Mathematical Analysis, Version 3.3, Seattle: MathSoft, Inc..

[9] Venables, W. N., and Ripley, B. D. (1997), Modern Applied Statistics with S-Plus, New York, NY: Springer-Verlag.

[2001년 9월 접수, 2002년 8월 채택]

Interesting Node Finding Criteria for Regression Trees*

Yung-Seop Lee [1)]

ABSTRACT

One of decision tree method is regression trees which are used to predict a continuous response. The general splitting criteria in tree growing are based on a compromise in the impurity between the left and the right child node. By picking up the more interesting subsets and ignoring the other, the proposed new splitting criteria in this paper do not split based on a compromise of child nodes anymore. The tree structure by the new criteria might be unbalanced but plausible. It can find a interesting subset as early as possible and express it by a simple clause. As a result, it is very interpretable by sacrificing a little bit of accuracy.

Keywords: Regression trees, CART, Interpretability of trees.

* This work is supported by the Dongguk University research fund.

1) Assistant Professor, Department of Statistics, Dongguk University, Seoul, 110-715, Korea.
E-mail: yung@dongguk.edu