Journal of Korean
Data & Information Science Society
2002, Vol. 13, No.1 pp. 97 104

A Study of Combined Splitting Rules in Regression Trees[1)]

Yung-Seop Lee[2)]

Abstract

Regression trees, a technique in data mining, are constructed by splitting function-a independent variable and its threshold. Lee(2002) considered one-sided purity(OSP) and one-sided extreme(OSE) splitting criteria for finding a interesting node as early as possible. But these methods cannot be crossed each other in the same tree. They are just concentrated on OSP or OSE separately in advance.

In this paper, a new splitting method, which is the combination and extension of OSP and OSE, is proposed. By these combined criteria, we can select the nodes by considering both pure and extreme in the same tree. These criteria are not the generalized one of the previous criteria but another option depending on the circumstance.

주제어: 관심노드, 분류기준, 회귀의사결정나무, CART.

1. 서 론

컴퓨터의 발달과 데이터의 축적으로 인한 대용량화 등으로 데이터의 중요성이 증가함에 따라 데이터 저장, 분석방법, 결과해석 및 현업에서의 적용에 관하여 다양한 연구가 활발하게 진행되고 있다. 1990년대 후반부터 CRM에 대한 관심의 증대와 함께 CRM의 중요한 축으로서 데이터 마이닝에 대한 필요성의 증가로 여러 분야, 특히 통계학의 중요성이 증가하고 있다. 특히 분석도구와 결과해석을 위해서는 통계적인 여러 기법들이 필요하게 된다. 기존의 여러 통계기법 중, 로지스틱 회귀분석, 다변량 분석 등의 전통적인 모수적 방법들이 많이 쓰인다. 또한 자료가 방대해짐으로 인하여 모수적인 방법들의 기본 가정을 충족시키지 못한 자료가 빈번해짐에 따라 비모수적인 여러 방법들이 제시되고 있다.

---

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

2. Lecturer, Department of Statistics, Dongguk University, Seoul, 110-715, Korea. E-mail: yung@dongguk.edu

데이터 마이닝의 여러 기법들 중 비모수적 방법인 의사결정나무는 통계학을 전공하지 않아도 결과를 해석하고 이해하기가 쉽기 때문에 널리 이용되고 있다. CART(Breiman et al., 1984)가 통계학에 의사결정 나무를 도입하는데 중대한 역할을 하였다. 의사결정나무는 분류하는 알고리즘에 따라 여러 종류가 있는데 대표적으로 CHAID(Hatigan, 1975), CART(Breiman et al., 1984), C4.5(Quinlan, 1993)가 있다.

CHAID(Chi-Square Automatic Interaction Detection)는 원래 AID(Morgan and Sonquist, 1963) 알고리즘을 수정한 것으로서 나무구조를 형성할 때 다원분류(multiway splits)가 가능하며 가지치기를 하지 않고 미리 정한 제약에 따라 나무구조의 성장을 멈춘다(pre-pruning). 범주형 자료만 입력변수가 될 수 있으며, 연속형 자료의 경우에는 여러 개의 범주로 그룹을 만든 다음 입력변수로 쓸 수 있다. 무엇보다도 다른 알고리즘과의 큰 차이는 χ^2 검정에 기초하여 분류함수를 정하는 것이다.

C4.5는 기계학습(machine learning) 분야에서 개발된 의사결정나무 알고리즘으로서 범주형 입력변수에는 다원분류, 연속형 입력변수에는 이항분류(binary splits)를 하며, 나무구조를 최대한 성장시킨 다음 가지치기(post-pruning)를 하여 최적의 나무구조를 형성한다. 그리고 다른 알고리즘과의 큰 차이는 엔트로피(entropy)에 기초하여 노드의 불순도를 측정하여 분류함수를 구한다.

한편 가장 많이 알려진 의사결정나무 알고리즘은 CART로서 범주형, 연속형 입력변수에 상관없이 이항분류를 하며, C4.5와 마찬가지로 post-pruning을 하여 최적의 나무구조를 형성한다. 이 경우에는 노드의 불순도를 지니인덱스(Gini index)에 의하여 측정하며 분류함수를 구한다.

의사결정나무는 반응변수의 성질에 따라 크게 두 가지로 나눈다. 반응변수가 범주형일 때를 분류의사결정나무(classification trees), 연속형일 때는 회귀의사결정나무(regression trees)라고 한다.

나무구조의 각 단계에서 분류 기준이 되는 분류함수는 불순도를 계산하여 결정한다. $i(O)$를 어미노드에서의 불순도라 하고, $i(L)$를 왼쪽 자식노드의 불순도, $i(R)$를 오른쪽 자식노드의 불순도라고 하자.

n, n_L, n_R을 각각 어미노드, 왼쪽 자식노드, 오른쪽 자식노드에서의 관찰치수라 하고, Δi를 식(1)과 같이 정의하자.

$$\Delta i = i(O) - \left(\frac{n_L}{n} i(L) + \frac{n_R}{n} i(R) \right) \tag{1}$$

나무구조가 성장할 때 각 단계에서의 분류함수는 어미노드와 자식노드간의 불순도가 가장 많이 차이나게 하는 (즉, 가장 큰 Δi) 독립변수와 그의 분계점이다.

CART의 분류방법(splitting method)은 식(1)과 같이 각 단계에서 분류된 왼쪽과 오른쪽 자식노드의 관찰치수를 가중치로 한 각각의 불순도를 평균한 후에 어미노드의 불순도와 차이가 가장 많이 나게 하는 독립변수와 그의 분계점(threshold)을 그 단계에서의 분류함수로 결정한다.

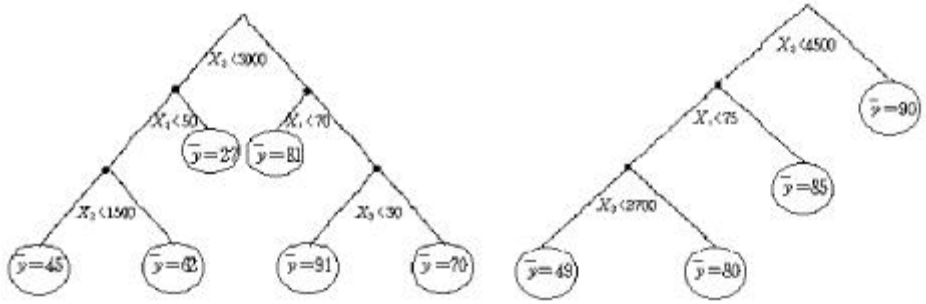
회귀의사결정나무에서 불순도는 잔차제곱합(residual sum of squares)을 기초로 하여 측정되어지고, 분류의사결정나무에서 불순도는 앞서 언급한 바와 같이 알고리즘에 따라 오분류율, 엔트로피, 지니인덱스 등에 의해서 측정된다.

기본적인 나무구조의 형태나 형성구조, 기본적인 의사결정나무에 대한 용어나 설명은 Lee(2001)를 참조하기 바란다.

본 연구는 회귀의사결정나무에서 정확성보다 해석력에 더 관심을 둔 분류함수를 제안한 Lee(2002)방법의 확장으로, Lee(2002)가 제안한 방법들을 보완한 새로운 분류함수를 제시하고자 한다.

2. 관심노드 찾는 분류 기준법

회귀의사결정나무에 이 방법을 적용한 것은 Lee(2002)에 의해서 제안되었다. 일반적으로 나무구조를 평가하는 기준으로는 정확도와 나무구조를 이용한 자료의 해석력의 두 가지가 있다. 대부분의 의사결정나무 알고리즘은 정확도를 높이는데 중점을 두었다. 예를 들면, CART(Breiman et al., 1984), Bagging(Breiman, 1996), AdaBoost (Freund and Schapire, 1996), Bumping(Tibshirani and Knight, 1996)은 전체적인 변동성을 안정시키면서 정확도를 높이는 방법들이다. 그러나 Lee(2002)는 약간의 정확성을 잃더라도 해석력에 더 관심을 가지고 새로운 분류함수를 제안하였다. 이 방법에 의하면 의사결정나무의 해석력을 높였을 때를 그림1에서 알 수 있다. 예를 들면, 어느 회사에서 종업원들의 직업 만족도에 영향을 미치는 요인을 알아보기 위하여 직업만족도(Y,반응변수), 직업적성시험점수(X_1), 연봉(X_2), 통근거리(X_3), 결근횟수(X_4)의 자료를 가지고 종래의 방법(CART)과 Lee(2002)의 단일극한 평균값을 기준으로 한 분류법으로 회귀의사결정나무를 형성하였을 때 가상적으로 그림1과 같은 모양을 얻었다고 하자. 그림1의 왼쪽은 CART에 의한 방법, 오른쪽은 극단값 찾기에 의한 방법(one-sided extremes, OSE)이다. 정확도(R^2)의 측면에서는 CART가 OSE보다 높을 수도 있다. 그러나 만약 정확도보다 직업만족도가 높은 집합을 가능한 한 빨리 간단한 표현으로 찾고 싶은 경우를 고려해 보자.



CART OSE

그림 1: 나무구조의 해석력

그림1에서 CART는 $\bar{y}$가 가장 높은 노드를 찾기 위해서 X_2가 3000이상, X_1이

70이상, X_3이 30이하라는 여러 단계를 거쳤을 때만이 가장 높은 노드($\bar{y}$ = 91)를 찾을 수가 있다. 그러나 OSE의 경우에는 단지 X_2가 4500이상이라는 조건 하나만으로 직업 만족도가 가장 높은 노드($\bar{y}$ = 90)를 찾을 수가 있다. 이와 같이 정확도는 다소 낮을지라도 원하는 부분집단을 가능한 한 빨리, 간단한 몇 개의 조건으로 찾을 수 있기 때문에 해석력이 높다고 말할 수 있다. 이러한 현상은 Lee(2002)의 예제에서 볼 수 있었다. Lee(2002)는 순수도를 기준으로 한 분류법(one-sided purity, OSP)과 앞의 OSE의 두 가지 방법을 제시하였다. 그러나 이러한 방법은 나무 구조를 형성하기 전에 미리 정하여야만 한다. 예를 들면, 순수한 집단만을 가능한 한 빨리 찾고 싶다면 OSP를 미리 정하여야 하고 나무 구조를 형성해 나가면서 어느 단계에서 OSE로 바꿀 수가 없다. 반대의 경우도 마찬가지이다.

3. 제안된 분류기준법

앞서 언급한 관심노드 분류법의 단점에 대한 보완된 방법으로 두 방법을 혼합한 분류방법을 제시하고자 한다.

1) $\min\left(\frac{\hat{\mu}_L}{\hat{\sigma}_L^2}, \frac{\hat{\mu}_R}{\hat{\sigma}_R^2}\right)$ 2) $\min\left(\frac{\hat{\sigma}_L^2}{\hat{\mu}_L}, \frac{\hat{\sigma}_R^2}{\hat{\mu}_R}\right)$

1)과 2)는 각 노드의 각 독립변수(X)와 그의 분계점들에 대한 왼쪽과 오른쪽 자식 노드 각각의 평균과 분산 추정치(여기서는 MLE)의 비율 중 작은 값이라고 정의할 때, 모든 독립 변수와 그의 분계점들에 대해서 1)의 값을 얻을 수 있다. 이 값들 중 가장 작은 값(즉, 최소 값 중 최소 값)이 그 단계에서의 분류 함수가 된다. 2)의 경우도 비율의 분자, 분모만 다를 뿐 같은 방법으로 분류함수를 찾는다. 이 방법들은 순수한 부분집합이나 극단적인 값을 가진 부분집합 중에서 먼저 나타나는 부분집합을 가능한 한 빨리 분류해 나가는 것이다. 이렇게 함으로써 순수한 집단이나 극단적인 집단을 찾는 분류기준을 미리 정할 필요 없이 나무구조를 형성해 나가면서 순수하거나 극단적인 집단을 분리해 나간다. 따라서 하나의 나무구조 안에서 두 가지 방법이 서로 번갈아 가면서 나무구조를 형성해 나간다. 참고로 2)의 방법은 변동계수(coefficient of variation, CV)와 비슷한 형태이다.

4. 예 제

CART(Breiman et al., 1984)에서 회귀의사결정나무의 예제로 쓰인 보스톤 지역의 집 값에 영향을 미치는 요인을 알아보는 자료를 이용하여 본 연구에서 제안된 분류기준과 Lee(2002)의 OSP와 OSE를 비교하여 분석하고자 한다.

보스톤 인근 지역의 506개의 지역(tract)에서 13개 독립변수들과 그 지역의 집 값의 중위수(median)를 반응변수로 하여 얻어진 자료이다. 각 변수와 그에 대한 설명은 표1과 같다.

관심노드를 찾는 방법 중 <그림 4.1>은 Lee(2002)의 <그림3.2>와 같다. 이 그림은

OSP방법, 즉 순수집단을 가능한 한 빨리 찾는 방법으로서 본 연구의 비교를 위하여 인용하였다. <그림 4.2>는 Lee(2002)에서 제안된 OSE 중 $\max(\hat{\mu}_L, \hat{\mu}_R)$로서 극단값 중 고평균을 가진 집단을 가능한 한 빨리 찾는 방법이다. Lee(2002)의 예제에는 이 방법을 이용한 의사결정나무 그림이 수록되지 않아 본 연구와의 비교를 위해 새로 만들어진 의사결정나무이다.

표1. 보스톤 자료의 변수

변 수 명	내용
CRIM	범죄율
ZN	25,000평방 피트 이상의 주거지 비율
INDUS	공업 지역 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)

각 의사결정나무 그림에서 m은 각 노드의 반응변수의 평균이며 sz는 노드 크기로서 전체 관찰치(여기서는 506) 중 해당노드의 관찰치의 비율(%)을 나타낸다. <그림 4.3>과 <그림 4.4>는 본 연구의 분류기준 1)과 2)를 이용하여 만들어진 회귀의사 결정나무들이다.

<그림 4.3>은 $R^2 = 0.22$로서 현저히 낮고 또 이 자료에는 적합치 않은 의사결정나무이다. 그러나 <그럼 4.4>의 R^2는 0.68로써 CART, OSE, OSP보다 다소 떨어졌지만, $m = 17$인 첫 번째 오른쪽 자식노드는 <그림 4.1> OSP의 첫 번째 노드와 같다. 즉 가장 순수한 부분집합이 먼저 분류되었다. $m = 45.10$인 두 번째 자식노드는 <그림 4.2> OSE 의 고평균 의사결정나무의 첫 번째 노드와 같다. 즉 가장 평균값이 높은(집 값이 비싼) 부분집합을 분류했다. 앞서 언급한 바와 같이 OSP와 OSE가 번갈아 가면서 첫 번째와 두 번째 관심노드로 분리되었다. Lee(2002)의 방법은 OSP나 OSE

를 미리 정하고 나무구조 중간에서는 방법을 서로 바꿀 수는 없는 반면 본 논문에서 새로 제시한 <그림 4.4>는 순수한 집단이든 극단적인 값을 갖는 집단이든 먼저 나타나는 부분집합이 분리되어 두 방법이 번갈아 가면서 혼합한 형태의 나무 구조를 형성해나간다. 이와 같이 두 방법을 혼합하여 관심 있는 집단을 분류해 나가는 것이 본 연구의 목적이다.

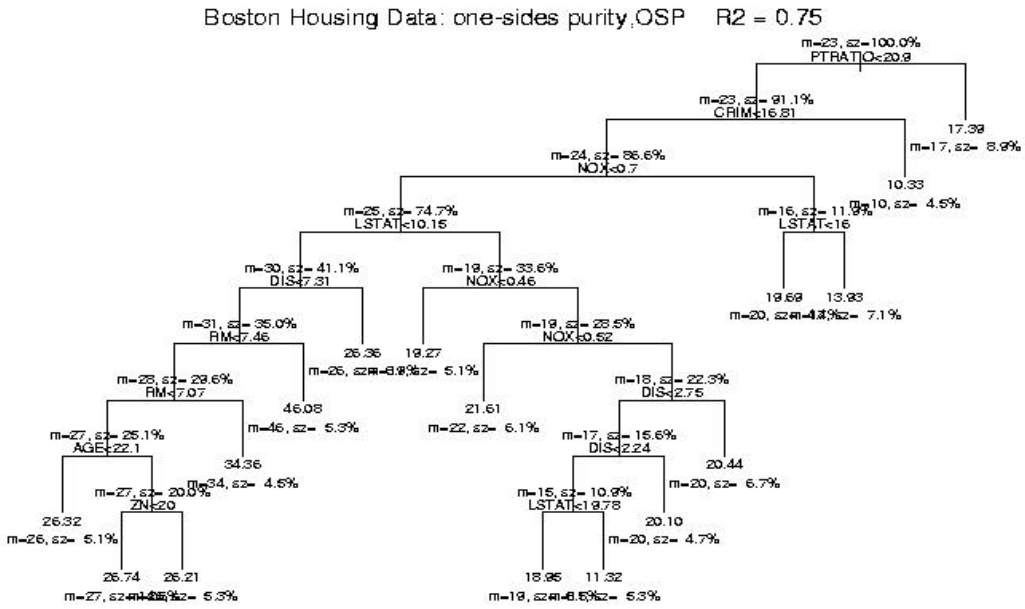
5. 결 론

현대통계학에서 자료분석에 이용되는 통계적 기법은 다양하다. 자료에 따라서 사용되는 기법이 다르고, 모든 자료에 항상 우수한 기법은 없다. 모든 것은 자료에 따라, 또 원하는 목적에 따라서 적재적소의 방법들을 사용하여야 한다. 데이터 마이닝 기법 중 하나인 의사결정나무도 마찬가지이다. 정확도를 높이는 대신 분석목적에 따라서는 간결하면서 해석력이 높은 의사결정나무가 더 필요할 경우도 있다. 본 연구는 회귀의사결정나무의 해석력을 높이고자 시도된 Lee(2002)에서 제시된 분류함수들의 단점을 보완하면서 OSP와 OSE를 혼합한 방법이다. 즉, 상황에 따라서는 OSP와 OSE를 미리 고정하여 의사결정나무를 형성하지 않고, 둘 중에서 먼저 만족하는 기준에 의한 부분집합을 분류해나가고 싶은 경우에 본 논문의 방법을 사용할 수 있다. 형성된 의사결정나무는 하나의 나무구조에서 두 방법이 번갈아 가면서 순수한 집단이나 극단적인 값을 갖는 집단을 분류해 나간다. 물론 여기에서 제안된 방법이 다른 모든 방법보다 항상 우수한 것은 아니지만 앞서 언급한 바와 같이 분석방법이나 목적에 따라 이용할 수 있는 대안을 제시하는데 본 연구의 의의가 있다.

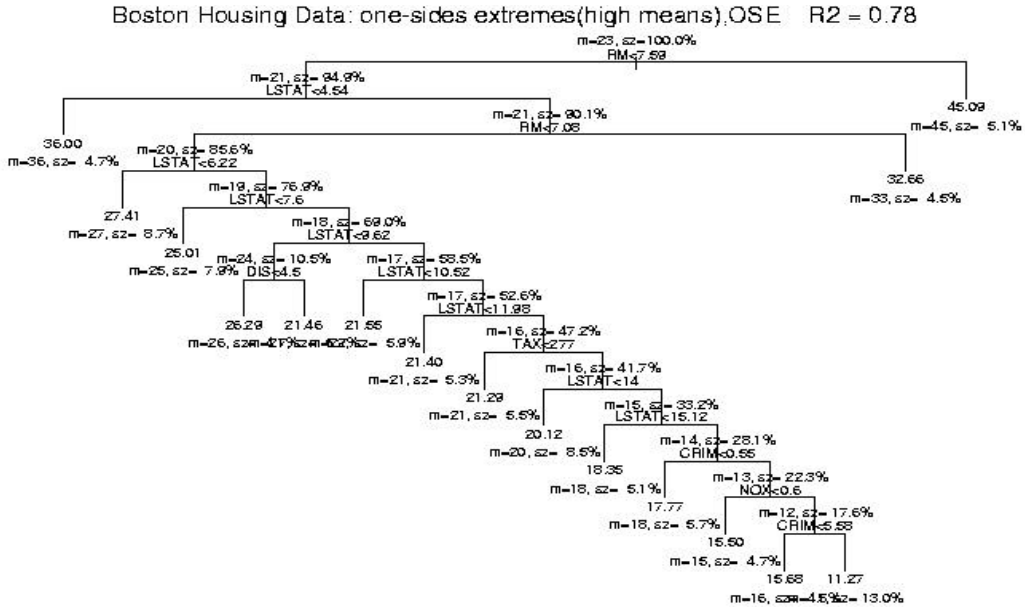
참 고 문 헌

1. Breiman,L. (1996), Bagging predictors, *Machine Learning*, 24, 123-140.
2. Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984), *Classification and Regression Trees*, Pacific Grove, CA: Wadsworth.
3. Freund, Y. and Schapire, R. E. (1996), A Decision-Theoretic Generalization of On-Line Learning and Application to Boosting, *Journal of Computer and System Science*, 55, 119-139.
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*, 8, 885-894.
6. Lee, Y-S. (2002), Interesting Node Finding Criteria for Regression Trees, *The Korean Journal of Applied Statistics* (in process).
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. Quinlan, J. R. (1993), C4.5: *Programs for Machine Learning*, San Mateo, CA: Morgan Kaufmann.
9. Tibshirani, R. and Knight, K. (1996), Model Search and Inference via

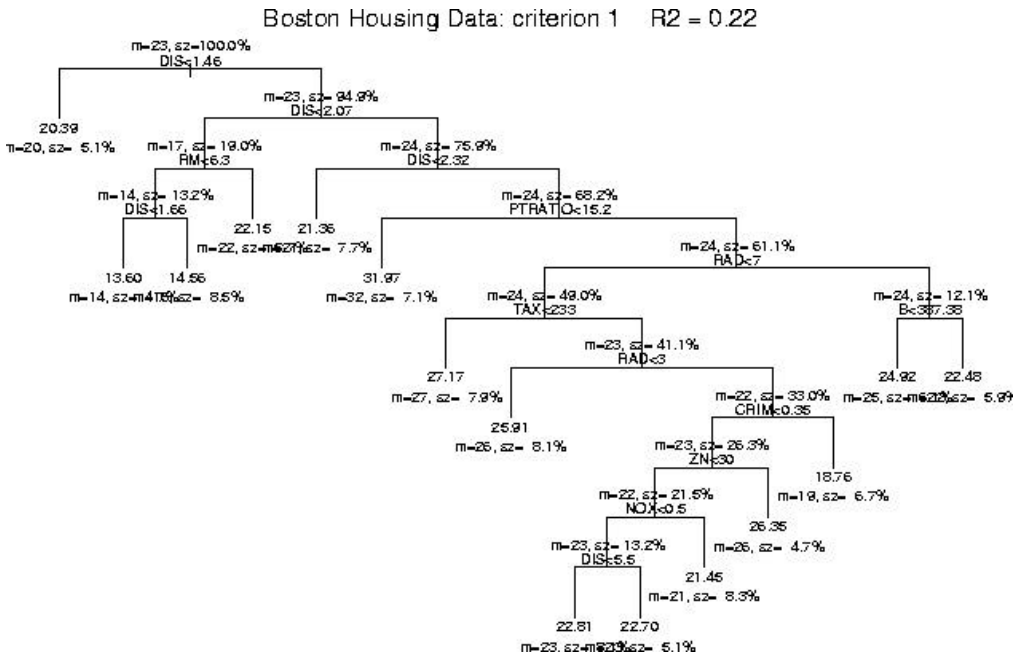
Bootstrap Bumping, Technical Report, Dept. of Statistics, U. of Toronto.



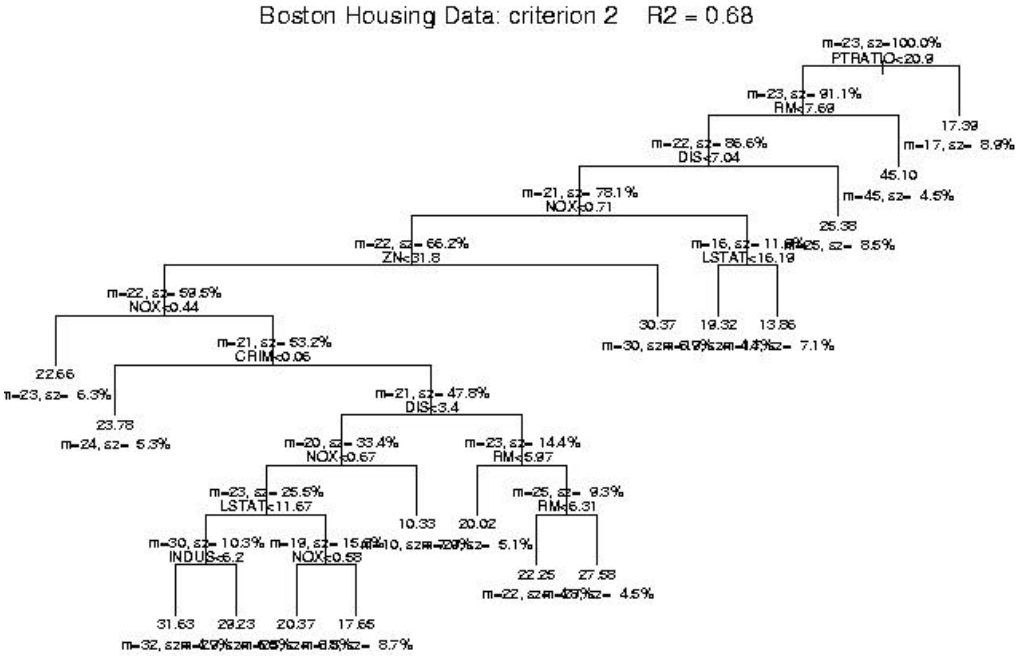
<그림 4.1> 단일 순수노드 (OSP)



<그림 4.2> 극단값(고평균) 노드 (OSE)



<그림 4.3> Criterion 1



<그림 4.4> Criterion 2