The Korean Communications
in Statistics Vol. 8, No. 3, 2001
pp. 885-894

New Splitting Criteria for Classification Trees[1)]

Yung-Seop Lee[2)]

Abstract

Decision tree methods is the one of data mining techniques. Classification trees are used to predict a class label. When a tree grows, the conventional splitting criteria use the weighted average of the left and the right child nodes for measuring the node impurity. In this paper, new splitting criteria for classification trees are proposed which improve the interpretablity of trees comparing to the conventional methods.

The criteria search only for interesting subsets of the data, as opposed to modeling all of the data equally well. As a result, the tree is very unbalanced but extremely interpretable.

Keywords : Classification trees, The Gini Index, binary split, CART

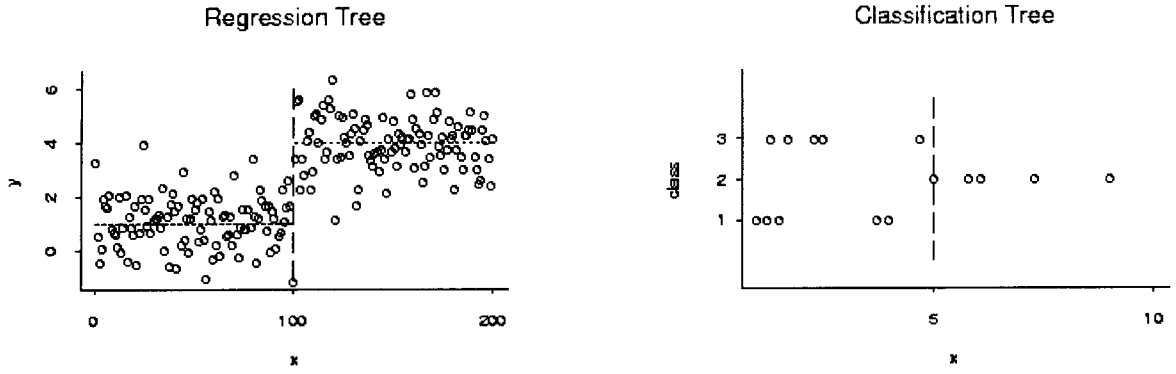
1. 서 론

데이터 마이닝에서 통계적인 분류(classification)법은 로지스틱 회귀분석(logistic regression), 신경망 분석(neural networks) 그리고 의사결정나무(decision tree)방법이 주로 사용된다. 그 중에서 의사결정나무는 어떤 계급(class)이나 값(value)을 분류하기 위하여 여러 개의 if-then문으로써 표현하고 있다. 이것을 rules라고도 한다. 의사결정나무에서는 반응변수(response variable)의 특성에 따라 두 가지로 나누어지는데, 그것은 분류나무(classification trees)와 회귀나무(regression trees)이다(Breiman et al., 1984, and Kang et al., 2000). 분류나무는 반응변수가 범주형(class) 변수로써 각 관찰치를 의사결정나무에 의하여 계급을 예측하는 것이며, 회귀나무는 반응변수가 연속적인 값으로써 회귀분석에서와 같이 반응값을 예측하는 것이 목적이다. <그림 1.1>은 하나의 독립변수로 회귀나무와 분류나무에서 어떻게 분할(partition)되는가를 보여주고 있다.

왼쪽의 회귀나무는 반응변수(y)의 평균을 최대한 잘 분류하는 독립변수(x)의 부분집합을 찾는 것이 목적이다. 반면 오른쪽의 분류나무는 반응변수의 계급을 최대한 잘 분류하는 독립변수의 부분집합을 찾는 것이 목적이다.

---

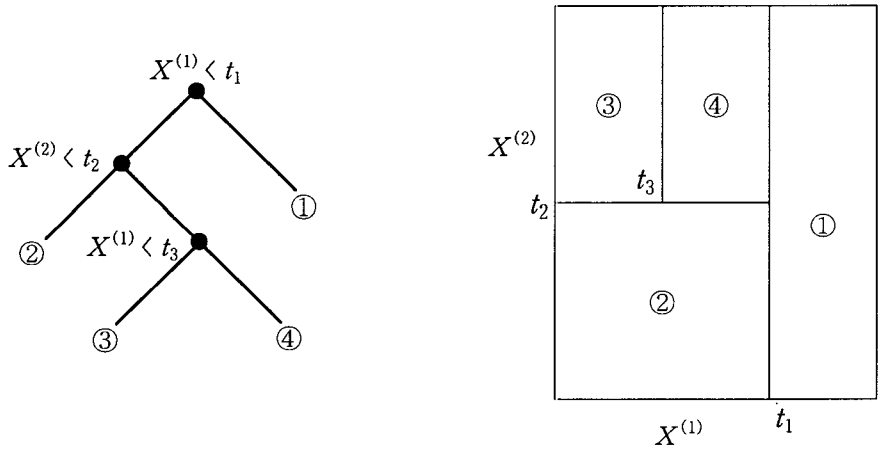
1) This work was supported by Dongguk University Research Fund in 2001
2) Lecturer, Department of Statistics, Dongguk University, Seoul, 110-715, Korea.
E-mail : yung@dongguk.edu



<그림 1.1> 회귀 나무와 분류 나무에서의 분할

1.1. 의사결정나무의 형성 구조

의사결정나무의 형성 구조를 보면, 데이터가 이항 분류(binary splits)에 의해서 더 작은 부분집합으로 반복적으로 분할된다. <그림 1.2>에서처럼 각각의 나무구조에서 어미노드(parent node)로부터 이항 분류에 의하여 두 자식노드(child node)로 나뉘어 진다. 이 때 분류함수는 하나의 독립변수와 그의 분계점(threshold)으로 이루어진다. 예를 들면, 뿌리 노드(root node)에서 $X^{(1)} < t_1$이면 왼쪽으로, $X_0^{(1)} \geq t_2$는 오른쪽으로 가고, 다음 단계에서는 $X^{(1)} < t_1$중에서 $X^{(2)} < t_2$이면 다시 왼쪽으로, 그렇지 않으면 오른쪽으로 가게 된다. 첫 번째 분류 변수인 $X^{(1)}$은 아래 단계에서 다시 나타날 수 있다. 이렇게 반복적인 부분집합으로 분류하는 것이 의사결정나무의 구조이며 이러한 분할을 기하학적인 모형으로 나타낸 것이 <그림 1.2>의 (b)이다.



(a) (b)

<그림 1.2> 나무 구조 형성 과정

1.2. 분류함수 결정 기준

각 단계에서 분류함수를 결정하는 기준은 불순도(measure of impurity)에 의해서 결정된다. 즉, 자식노드들의 불순도와 어미노드의 불순도의 차이를 비교하여 결정하는데, 전통적인 분류 기준(conventional splitting criteria)은 오른쪽과 왼쪽의 자식노드의 관찰치 수를 고려한 불순도를 평균한 후에 어미노드의 불순도와의 차이가 가장 많이 나는 분류함수를 찾는다. 불순도의 차이가 많이 난다는 것은 곧 그 만큼 주어진 분류 함수를 사용함으로써 분류 효과가 크다는 의미이다.

그러나, 본 논문에서는 모든 데이터를 똑같이 취급하는 것이 아니라, 원하는 부분 집합을 가능한 한 빨리 분류해 내는 것이 목적이다. 따라서 자식노드의 불순도를 측정하는데 있어서 둘의 평균 불순도를 측정하는 것이 아니라 하나의 자식노드가 불순도가 작다면 단지 그것으로 자식노드의 불순도로써 측정하는 것이다. 즉 자식노드 중 한 쪽 노드의 불순도에만 관심을 가진다.

2. 의사결정나무의 사용

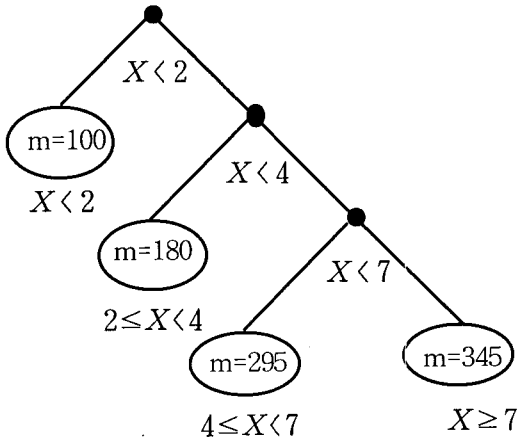
의사결정나무를 형성하는데 있어서 두 가지 측면에서 만족하여야 한다. 첫번째는 예측에 있어서의 정확성(accuracy), 두번째는 해석력(interpretability)이다. 이 두 가지는 양립할 수 없고 하나를 다소 희생하여야 한다. 일반적으로 정확성에 더 많은 관심이 있고 이 정확성을 높이고자 여러 가지 알고리즘들이 개발되고 있다. 그러나 종종 해석력에 관심이 더 많은 경우도 있다. 즉, 정확성을 많이 잃어버리지 않는 범위 내에서 유용하고 설명력이 있는 부분집합을 갖는 노드를 찾고자 할 경우가 있다. 이것을 해석력이 있는 의사결정나무(interpretable trees)라고 한다. 이러한 나무구조에서는 상위단계는 단지 몇 개의 rules로써 표현되기 때문에 아주 가치가 있는 부분이며 또한 하위단계와는 달리 조금 다른 표본에 대해서도 상당히 안정적이다. 즉, 상위단계의 변수는 커다란 변화가 없이 여러 나무 구조에서 비슷하다.

3. 의사결정나무의 단순성(simplicity)과 해석력

본 논문에서는 간단하면서 설명력있는 의사결정나무를 구축하고자 한다. 즉, 나무구조의 상위단계에서 가능한 빨리 의미가 있고 설명 가능한 노드를 찾는 방법을 찾고자 한다. 상위부분에서의 설명력을 강조하고 하위부분에서는 큰 관심이 없기 때문에 예측에 대한 정확성은 다소 떨어질 수는 있다. 한편 기존의 나무구조는 균형적인 구조를 선호하고 있지만 설명력있는 나무구조를 구축하는데 있어서는 아주 불균형한 나무모양이 될 수도 있다.

<그림 3.1>에서 볼 수 있듯이 구조는 아주 불균형적이지만 모든 노드는 노드의 단계에 상관없이 하나 또는 두 개의 조건(rule)으로 간단하게 설명된다. 또한 <그림 3.1>에서처럼 하나의 독립변수가 반복적으로 나타날 때 독립변수와 반응변수와의 관계가 단조증가 또는 감소(monotone increasing or decreasing)가 있다는 것을 알 수가 있다. 뒤에 나타나는 본 논문의 예제에서는 이러한 현상을 볼 수 있다. 본 논문에서 제안하는 분류방법은 기존의 방법과 비교할 때 불균형적이고 정확도가 떨어질 수도 있다. 그러나 분석자가 설명력이 있고 이해하기가 쉬우며 원하는 노드를

빨리 찾고자 할 때는 본 논문의 제안된 분류방법이 훨씬 효과적이다.



<그림 3.1> 불균형적이지만 간단하면서 설명력이 있는 나무 구조

4. 분류의사결정나무(Classification Trees)에서의 일반적인 분류 기준법

앞에서도 언급하였듯이 분류의사결정나무는 반응변수가 범주형 자료일 경우로, 본 논문에서는 반응 변수가 두 계급만을 가질 때를 고려하겠다. 계급값은 0과 1로 하고, 주어진 단계에서 왼쪽과 오른쪽 자식노드를 분류된 확률로 나타낼 때, $P_L^0 + P_L^1 = 1$, $P_R^0 + P_R^1 = 1$이 성립한다. 여기에서 ${P_L}^0$, ${P_L}^1$은 왼쪽 자식노드에서의 계급 0과 1의 비율, ${P_R}^0$, ${P_R}^1$은 오른쪽 자식노드에서의 계급 0과 1의 비율을 말한다.
일반적으로 주어진 노드에서의 불순도의 대표적인 측정방법은 아래와 같다. (편의상 왼쪽 자식노드만 한정함.)

1) 오분류율 (Misclassification rate) : $\min(P_L^0, P_L^1)$
주어진 노드를 어떤 계급으로 결정할 때 다수의 법칙(majority vote)에 의해서 노드계급을 결정하고, 이 때 나머지 계급의 비율로써 오분류율을 결정한다.

2) 엔트로피 (Entropy(Information)) : $-P_L^0 \log P_L^0 - P_L^1 \log P_L^1$
주어진 노드에서의 각 계급의 확률은 상대적인 비율로 할 때, 엔트로피는 다항분포 (multinomial distribution)에서 $E_L[\min[-\log likelihood/N_L]]$과 같다.
이 측정도는 C4.5(Quinlan, 1993)와 S-Plus(Statsci, 1995)에서 사용한다.

3) 지니 계수(The Gini index) : $P_L^0 P_L^1$
이것은 CART(Breiman et al., 1984)에서 제안된 것으로 반응변수가 0과 1일 때 평균 제곱 오차(mean square error)와 같다. 즉,

$$MSE_L = E_L(Y - P_L^1)^2 = P_L^0 P_L^1$$

위에서 구한 노드 불순도를 각 노드의 관찰치 수를 가중치로 하여 평균한 것이 일반적인 분류기준법이다. 즉,

1)′ $P_L \min(P_L^0, P_L^1) + P_R \min(P_R^0, P_R^1)$

2)′ $P_L(-P_L^0 \log P_L^0 - P_L^1 \log P_L^1) + P_R(-P_R^0 \log P_R^0 - P_R^1 \log P_R^1)$

3)′ $P_L P_L^0 P_L^1 + P_R P_R^0 P_R^1$

여기에서 P_L, P_R은 어미노드에서 왼쪽, 오른쪽의 자식노드로 분류될 관찰치들의 주변확률이며 $P_L + P_R = 1$이다.

위의 세 측정도는 CART(Breiman et al., 1984, p32)에서 불순도 측정함수의 성질을 정의한 조건을 모두 만족한다. 즉, 각 노드에서 계급 비율이 같을 때 최대이며, 오직 하나의 계급만을 가질 때(순수노드일 때) 최소값을 가진다. 그리고 대칭함수이다. 이 중에서 오분류율에 의한 측정도는 Breiman et al.(1984, p96)에서 지적하였듯이 P_L^1이나 P_R^1이 0이나 1에 가까울 경우에 올바른 측정도가 되지 못한다.

일반적으로 자식노드 중 하나가 순수할수록 우리는 그러한 분류함수를 더 선호하게 되는데, 오분류율을 사용하면 그것의 선형성 때문에 이러한 경우를 특별히 선호하는 경향이 없이 단지 두 자식노드의 오분류율을 더하여 계산하기 때문에 다른 분류함수들과 별 차이가 없다. 이러한 이유 때문에 오목(concave)함수인 엔트로피나 지니 계수를 사용한다. 즉 어떤 노드의 특정 계급의 비율이 0이나 1에 가까울 때(순수노드일 때) 선형성보다 더 빨리 불순도가 떨어질 측정도이다. 이 두 측정도에는 큰 차이가 없으나 다계급(multi-class)의 반응변수인 경우는 약간의 차이가 있다(Breiman, 1996). 지니 계수의 경우는 두 자식노드들의 크기가 차이가 있을 지라도 되도록이면 한쪽 자식노드를 순수하게 만들려고 하는 경향이 있는 반면, 엔트로피의 경우에는 되도록이면 두 자식노드들의 크기가 균형적으로 분류하려고 한다. 일반적으로 이항분류일 경우는 지니 계수가, 다계급분류일 경우는 엔트로피가 더 좋다고 알려져 있다.

5. 관심노드 분류 기준법

본 논문에서는 반응변수가 이항적(two class)일 때만 고려하겠다. 앞에서도 언급하였듯이, 본 논문의 목적은 특정계급의 확률(P^0 또는 P^1)을 왼쪽이나 오른쪽 자식노드 둘 다 고려하지 않고, 오직 하나의 노드에서 아주 작거나 큰 부분집합을 찾는 것이다. 또 다른 목적은 두 노드 중 계급1(또는 계급0)의 분산이 아주 작은 부분 집합(순수 집합)을 찾는 것이다 . 예를 들면, 보건 통계 자료에서 치사율이 높은 집단 또는 약의 효과가 높은 집단을 가능한 한 빨리 찾고 싶을 때 사용할 수 있는 분류기준법을 제안하고자 한다.

5.1. 순수도를 기준으로 한 분류법(One-sided purity)

반응변수 계급(0 또는 1)에 상관없이 하나의 순수한 노드를 가능한 한 빨리 찾고자 할 때, 왼쪽

과 오른쪽의 가중평균 합 대신 아래와 같은 분류법을 제안한다.

$$crit_{LR} = \min(P_L^0, P_L^1, P_R^0, P_R^1)$$

또는

$$crit_{LR} = \min(P_L^0 P_L^1, P_R^0 P_R^1)$$

또는

$$crit_{LR} = \max(P_L^0, P_L^1, P_R^0, P_R^1)$$

위의 세 기준법은 같은 결과를 가져온다. 왜냐하면 첫 번째와 두 번째에서 $\min(P_L^0, P_L^1)$은 $P_L^0 P_L^1$의 단조 변환이기 때문이며, 첫 번째와 세 번째에서는 P_L^0 나 P_L^1 중 하나가 최대이면 다른 하나는 최소이기 때문이다. 세 번째 기준법이 순수집합을 찾는데 더 이해하기 쉽고 직관적이다.

5.2. 관심노드의 비율을 기준으로 한 분류법(One-sided extremes)

관심있는 계급이 1이라고 한다면, L(왼쪽 자식노드)과 R(오른쪽 자식노드)중에서 계급1의 순수노드를 찾는 방법이다.

$$crit_{LR} = \max(P_L^1, P_R^1)$$

또는

$$crit_{LR} = \min(P_L^0, P_R^0)$$

이 두 기준법은 동일하다. 왜냐하면 $P_L^0 = 1 - P_L^1$과 $P_R^0 = 1 - P_R^1$이기 때문이다.

이와 같이 정해진 기준법에 의하여 모든 가능한 분류변수 및 분계점에서 기준값을 구한 다음 가장 작은 값(불순도에 있어서 어미노드와 가장 차이가 많이 나는 값)을 그 단계에서의 분류함수로 정한다. 위에서 제안된 기준법은 자식노드 중 순수한 하나의 노드에만 관심이 있고 다른 노드에는 관심이 없기 때문에 순수한 노드는 더 이상 분리되지 않는다. 따라서 나무구조 형태가 매우 불균형 적일 수가 있다. 그러나 우리가 찾고자 하는 부분집합을 빨리 찾을 수 있고 따라서 간단히 몇 개의 조건으로 표현할 수 있기 때문에 설명력이 있다.

6. 실증적 자료 분석

캘리포니아 주립대학(UC Irvine)에 있는 데이터 저장소(http:// www.ics.uci.edu/~mlearn/MLRepository.html)에서 피마 어메리칸 인디안 족의 자료를 가지고 비교 분석하고자 한다. 이 자료는 21세 이상의 피마족 여성 768명을 대상으로 그들의 당뇨병 여부를 반응변수로 하고, 당뇨병에 영향을 미칠 것이라고 생각되는 8개의 독립변수를 조사하였다. 768명 중 268명이 당뇨병(반응

계급 1)환자이며, 나머지는(반응계급 0) 비당뇨병이다.
8개의 변수와 반응변수에 대한 설명은 아래와 같다.
· PRGN : 임신 횟수
· PLASMA : 혈당량
· BP : 확장기 혈압(최저 혈압) (mmHg)
· THICK : 삼두근 두께(mm)
· INSULIN : 인슐린
· BODY : 몸무게를 키의 제곱으로 나눈 것(bmi)
· PEDIGREE : 당뇨병 혈통관계(유전적 요인)
· AGE : 나이
· RESPONSE : 반응변수(당뇨병이면 1, 아니면 0)

비교의 편의상 노드 중 최소 관찰치 수는 표본집단의 5%인 35로 미리 정하였으며 정확도보다 해석력에 더 관심이 있기 때문에 가지치기(pruning)는 하지 않았다. 각각의 노드에서 p는 계급 0과 계급1의 비율을 말하며 sz는 노드크기(%)이다.

6.1. CART($N_L P_L^0 P_L^1 + N_R P_R^0 P_R^1$) (<그림 6.1>) - 일반적인 방법

6단계까지 내려간 균형적인 나무구조를 하고 있다. 가장 강력한 변수는 PLASMA로써 첫 번째 분류함수를 포함하여 5번이나 나타났다. 그 다음으로 BODY(3번), PEDIGREE(3번) 순이며 전체적으로는 큰 무리가 없는 나무 구조이나 설명력이 떨어진다.

6.2. 순수도를 기준으로 한 분류법(One-sided purity, min($P_L^0, P_L^1, P_R^0, P_R^1$)) (<그림 6.2>)

12단계까지 내려가는 아주 불균형적인 나무 구조이나 전체적인 형태는 간결하다. 오른쪽으로 내려감에 따라 계급0의 부분집합들이 계속해서 분리되었다가 다시 왼쪽으로 가면서 계급1의 부분집합들이 반복적으로 분리되어진다. BODY와 PLASMA가 상위 단계에서, AGE와 PEDIGREE가 하위 단계에서 가장 많이 나타나는 변수들이다.

6.3. 관심있는 계급(0) 분류법(One-sided high class 0, max(P_L^0, P_R^0)) (<그림6.3>)

불균형적인 나무 구조이나 구조는 아주 간결하다. 계급0의 비율이 높은 노드만 찾기 때문에 구조 형태가 오른쪽을 불균형이다. BODY와 PLASMA가 각각 반복적으로 교호작용을 일으키면서 계급0의 비율이 높은 부분집합을 분리해 나가고 있다. 즉, 반응변수에 이 두 변수가 결합하여 단조 의존적인(monotone dependence) 관계를 가지고 있음을 알 수 있다. 자료를 간단하면서도 보다 설득력있게 설명하고 있다.

이상의 여러 나무 구조를 볼 때, PLASMA가 가장 영향력 있는 변수이고 그 다음으로 BODY이다. 특히 세 번째(계급0) 나무구조는 아주 성공적이다.

7. 결론 및 향후 과제

앞서 예측한대로 일반적인 분류 기준법에 의한 나무구조는 균형적이면서 전체적으로는 큰 무리가 없으며 정확도도 높다. 그러나 우리가 원하는 부분집합을 찾고자 한다거나 전체적인 자료구조를 알고자 할 때는 다소 무리가 있는 경우가 있다. 이와 반면, 본 논문에서 제안된 방법은 불균형적인 나무 구조를 가지지만 이것 때문에 오히려 더 설명력이 있고 우리가 원하는 부분집합을 찾기엔 더 빠르고 간결하여 효과적이다. 정확도가 일반적인 방법에 비해 다소 떨어지지만 유의할 만한 수준은 아니다.

우리가 원하는 목적에 따라 다양한 방법들을 사용하여 그 목적에 가장 적합한 모형을 찾는 것이 현대 통계학의 흐름임을 감안할 때, 본 논문에서 제안된 방법은 보다 효과적이라 할 수 있다. 앞으로는 회귀 의사결정나무(Regression Trees)에도 이와 비슷한 방법을 적용하여 보다 광범위하게 사용하고자 한다.

References

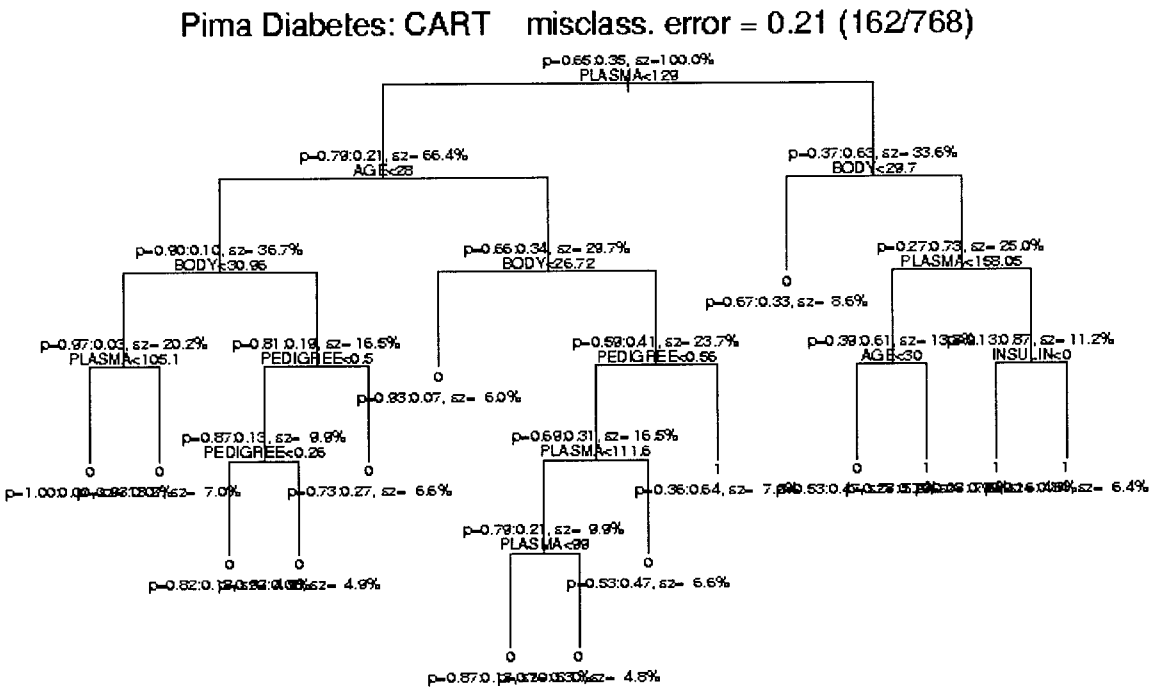
[1] Breiman, L. (1996), Technical Note: Some Properties of Splitting Criteria, *Machine Learning*, 24, 41-47.

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

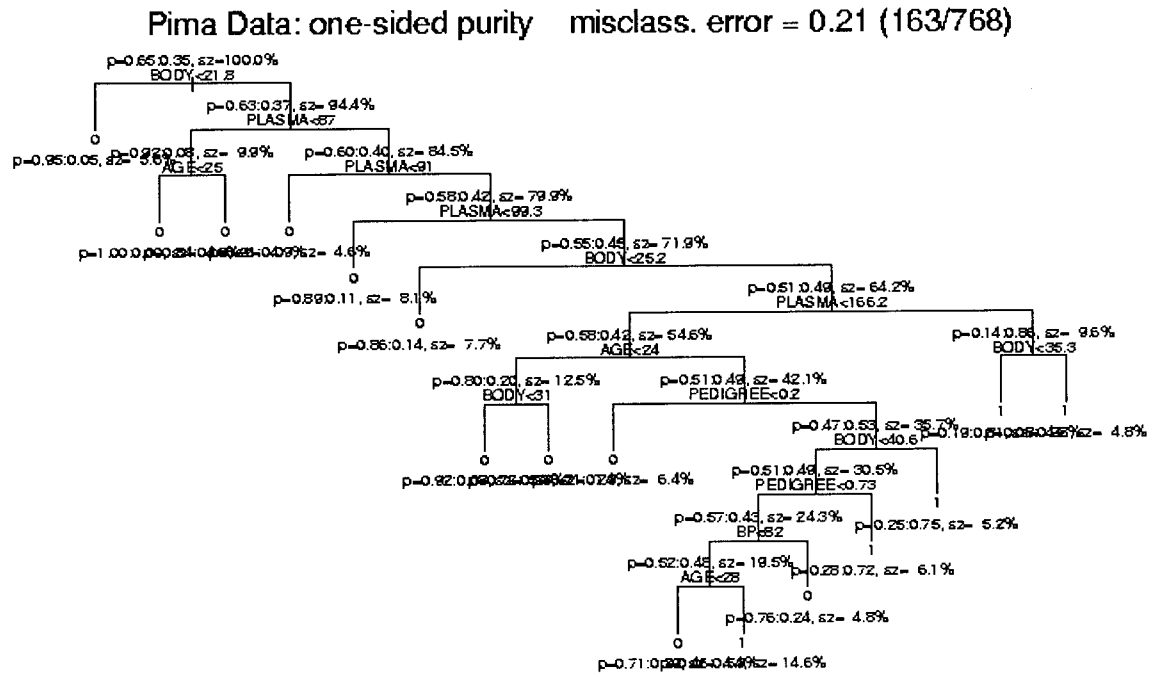
[3] Kang, Hyuncheol, Han, Sang-Tae, and Choi, Jong-Hoo (2000), Interpretation of Data Mining Prediction Model Using Decision Tree, *The Korean communications in Statistics*, Vol. 7, 937-943.

[4] Quinlan, J. R. (1993), C4.5: *Programs for Machine Learning*, San Mateo, CA: Morgan kaufmann.

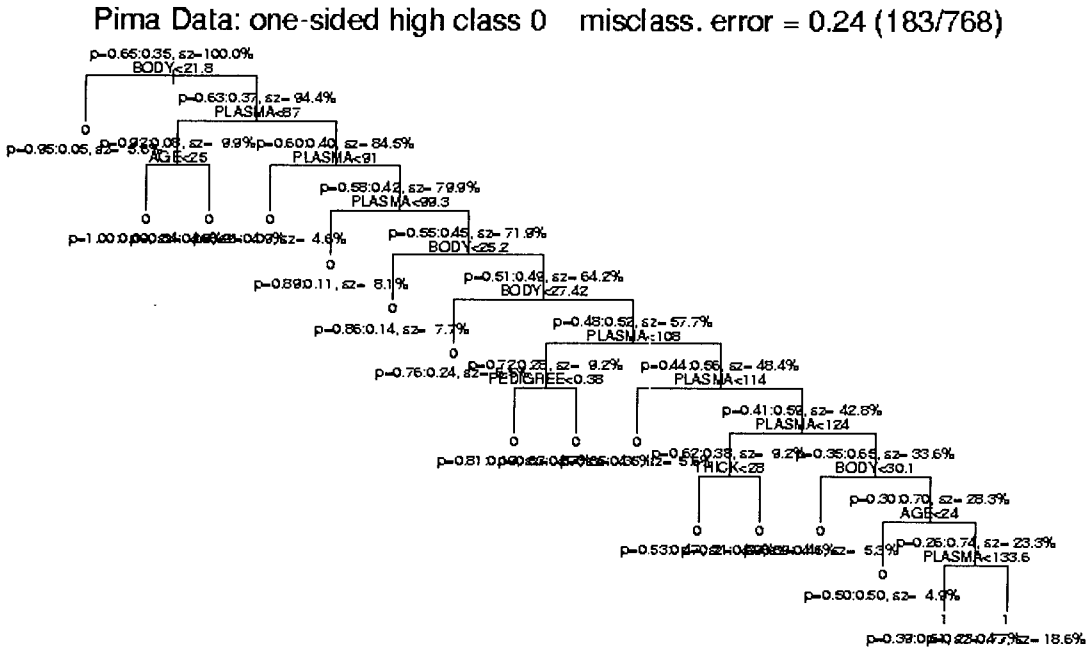
[5] StatSci (1995), *S-PLUS Guide to Statistical and Mathematical Analysis,* Version 3.3, Seattle: MathSoft. Inc..



<그림 6.1> 일반적인 방법 (CART)



<그림 6.2> 순수도를 기준으로 한 분류법 (One-Sided Purity)



<그림 6.3> 관심있는 계급(0) 분류법 (One-Sided High Class 0)