이 성 건 교수지도

석사학위청구논문

독립변수의 선형결합을 이용한 의사결정나무 분리기준에 관한 연구

2010

성신여자대학교 대학원

통 계 학 과

이 경 혜

독립변수의 선형결합을 이용한 의사결정나무 분리기준에 관한 연구

이 성 건 교수지도

이 논문을 석사학위논문으로 제출함

2009년 11월

성신여자대학교 대학원

통 계 학 과

이 경 혜

인 준 서

이경혜의 석사학위 논문으로 인준함.

심사위원 _____ 인

심사위원 _____ 인

심사위원 _____ 인

성신여자대학교 대학원

논 문 개 요

현실세계에서 발생하는 많은 데이터들로부터 우리가 행하고자 하는 일에 도움이 되는 정보(information)나 지식(knowledge)을 찾아내는 것은 매우 중요하다. 특히 대량의 데이터에서 의미 있는 정보를 추출해내는 데 널리 응용되고 있는 의사결정나무분석(decision tree analysis)은 예측과 분류를 하기 위해서 사용되는 보편적이고 강력한 도구이다. 분석과정이 나무구조에 의한 분할규칙(split rule)에 의해서 표현되기 때문에 다른 방법들에 비해 연구자가 그 과정을 쉽게 이해할 수 있고 두 개 이상의 변수가 결합하여 목표변수에 어떻게 영향을 주는지를 쉽게 이해할 수 있는 장점을 가진다.

분류나무 형성 시 중요한 것은 목표변수를 잘 분리해주는 변수를 선택하는 것이다. 이는 분리의 측도인 분리기준(split criterion)에 영향을 크게 받기 때문에 적절한 분리기준에 의해 의사결정나무를 구축해야 한다. 그러나 대부분의 의사결정나무 알고리즘들은 분리변수 선택에 있어 독립변수가 일변량인 경우만 고려한다.

본 논문에서는 독립변수의 선형결합에 대한 분리기준을 제시함으로서 예측의 정확성을 높이는 분리기준에 대한 연구를 하고자 한다. CHIAD, CART, C4.5, QUEST 등 기존 알고리즘과 본 연구에서 제안하는 방법의 오분류율을 R 소프트웨어를 이용하여 모의실험을 통해 비교하였다. 모의실험 결과 제안된 방법인 선형결합 분리기준과 황금분할을 이용한 분리기준의 오분류율이 낮았다. 또한 새로운 알고리즘이 실제자료에도 잘 적용되고 있음을 보여준다.

목 차

제1장 서 론

지난 수십 년간 여러 가지 형태로 저장되어 있는 데이터의 양은 기하급수적으로 증가되어 왔으며 이러한 데이터의 무제한적인 증가는 우리가 원하는 정보를 찾아내는 일을 보다 어렵게 만들고 있는 것이 현실이다. 이러한 현실 속에서 실제적으로 의미 있는 자료를 찾아내는 과정은 필수적이다.

의사결정나무 분석은 자료를 분류하거나 예측하는데 나무구조로 표현하는 분석방법으로, 모형을 쉽게 이해 할 수 있는 장점을 가지고 있다. 또한, 해석이 용이하며, 나무구조로부터 어떤 입력변수가 목표변수를 설명하는데 있어서 더 중요한지를 쉽게 파악할 수 있기 때문에 유용하게 사용되고 있다.

그러나 분석과정에서 얼마나 잘 분류하거나 예측하느냐가 문제가 되기 때문에 의사결정나무 분석은 제일 먼저 많은 독립변수 중 자료를 가장 잘 분리할 수 있는 분리기준을 찾는 것을 시작으로 한다.

CART로 대표되는 의사결정나무의 알고리즘에서 가장 중요한 요소는 분리변수의 선택방법이다. 대부분의 알고리즘은 분리 변수로 일변량 변수선택 방법을 적용하는 것이 대부분이다. 이것은 불필요하게 나무구조를 복잡하게 만들고 더 나아가 의사결정나무의 중요한 부분인 가지치기를 수행 할 경우 지나치게 간단한 나무구조만 남게 되는 결과를 제공한다.

이러한 약점을 보완하는 해결책으로 본 논문에서는 분석에 필요한 변수를 찾아내고 측정분리기준에 대한 방안을 제시하고자 한다. 즉, 어떤 독립변수를 이용하여 어떻게 분리하는 것이 반응변수를 가장 잘 구별해 주는지를 파악하기 위해 일변량 분리기준(single splits) 대신 이변량 선형결합 분리기준

(bivariate linear combination splits)을 제안한다. 첫째로 일대일 선형결합 분리기준에 대한 의사결정나무를 제안한다. 두 번째로, 변수결합방법으로 황금분할을 이용한 의사결정나무를 제안한다. 일대일 선형결합이란 2개의 독립변수에 대해서 가중치를 동일하게 부여해 결합하는 방법을 말하며 황금분할 이용한 선형결합은 2개의 독립변수에 대해 5:3의 비율을 이용해 결합하는 방법을 말한다.

본 논문의 제2장에서는 기존의 알고리즘에 대한 의사결정나무를 소개한다. 제3장에서는 일대일 선형결합과 황금분할을 이용한 선형결합 분리기준을 이용한 의사결정나무를 설명한다. 제4장에서는 R 소프트웨어를 이용한 모의실험을 통하여 CHIAD, CART, C4.5, QUEST와 제안된 알고리즘의 예측의 정확성을 비교하였다. 또한, 실제데이터를 본 논문의 알고리즘에 적용하였다. 제5장에서는 결론 및 향후 연구방향에 대해서 논의한다.

제2장 의사결정나무

의사결정나무의 구축에 있어서 가장 핵심적인 내용은 분할규칙(split rule)이라 할 수 있다. 분할규칙은 하나의 변수를 사용하는 방법과 여러 개의 변수를 병합하여 사용한 것 등 두 가지로 나뉜다. 전자는 일변량 분할(single split)이라 칭하며 흔히 의사결정나무라 할 때 일컫는 방법이다. 후자는 선형결합분할이라 하여 여러 변수의 선형결합을 이용하여 데이터를 분할하는 기법이다.

본 장에서는 일변량 분할규칙에 관한 논의에 초점을 맞추고자 한다. 의사결정나무를 형성하기 위한 다양한 알고리즘이 제안되어 있는데, 대표적인 알고리즘으로는 CHAID, CART, C4.5, QUEST 등이 있다.

2.1 일변량 의사결정나무

일변량 의사결정나무분석을 수행하기 위한 다양한 분리기준, 정지규칙, 가지치기 방법들이 제안되고 있는데, 그 중에서도 일변량 의사결정나무를 형성하는 알고리즘에 대해서 알아본다.

2.1.1 CHAID 알고리즘

CHAID(Chi-squared Automatic Interaction Detection; Kass, 1980) 알고리즘은 범주형 반응변수 또는 연속형 반응변수의 분류 및 예측을 수행하는 알

고리즘이다. 반응변수가 범주형일 때, CHAID 알고리즘은 분할표에 기초한 피어슨(Pearson) 카이제곱 또는 우도비(likelihood ratio) 카이제곱 통계량을 분리기준으로 사용한다. 반면, 반응변수가 연속형일 때는 두 개 이상의 그룹에 대해서 평균치 차를 검정하는 분산분석표의 F 통계량을 분리기준으로 이용한다.

반응변수가 범주형인 경우, 이용되는 카이제곱 통계량은 관측도수(f_{ij})로 이루어진 분할표(contingency table)로부터 계산된다. Pearson의 카이제곱 통계량은

$$\chi^2 = \sum_{i,j} \frac{(f_{ij} - e_{ij})^2}{e_{ij}} \quad , \tag{2.1}$$

와 같이 정의되고, 우도비 카이제곱 통계량은

$$\chi^2 = 2\sum_{i,j} f_{ij} \times \log_e \left(\frac{f_{ij}}{e_{ij}}\right) \quad , \tag{2.2}$$

로 정의된다. 이 때 두 통계량의 자유도(degree of freedom)는 반응변수의 범주 수를 r이라 하고, 설명변수의 범주 수를 c라 할 때 $(r-1)(c-1)$으로 계산된다. e_{ij}는 분포의 동일성 또는 독립성의 가설 하에서 계산된 기대도수(expected frequency)를 말하며 다음과 같이 계산된다.

$$e_{ij} = \frac{f_{i.} \times f_{.j}}{f_{..}} \quad , \tag{2.3}$$

식 (2.3)에서 $f_{i.}$는 행에 대한 도수의 합, $f_{.j}$는 열에 대한 도수의 합이며, $f_{..}$는 전체도수의 합을 말한다.

카이제곱 통계량이 자유도에 비해서 매우 작다는 것은 설명변수의 각 범주에 따른 반응변수의 분포가 서로 동일하다는 것을 의미하며, 따라서 설명변수가 반응변수의 분류에 영향을 주지 않는다고 결론지을 수 있다. 자유도에 대한 카이제곱 통계량 값의 크고 작음은 p-값으로 표현될 수 있는데, 카이제곱 통계량 값이 자유도에 비해서 작으면 p-값은 커지게 된다. 그러므로 p-값이 큰 값을 가지면, 설명변수의 범주에 의해 분리된 반응변수의 분포가 동질하다고 보여 진다. 결국, 분리기준을 카이제곱 통계량 값으로 한다는 것은 p-값이 가장 작은 설명변수와 그 때의 최적분리에 의해서 자식마디를 형성시킨다는 것을 의미한다.

반면, 반응변수가 연속형인 경우, 이용되는 F 통계량은 자유도 $(r-1, n-r)$인 F-분포를 따르고, [표2.1]과 같이 계산된다.

[표2.1] 분산분석표

요인	자유도	평방합	평균평방	분산비
설명변수	$r-1$	$SST=\sum_{i=1}^{r} n_i(\bar{y}_i-\bar{y})^2$	$MST=SST/(r-1)$	$F=\frac{MST}{MSE}$
오차	$n-r$	$SSE=\sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2$	$MSE=SSE/(n-r)$	
전체	$n-1$	$TSS=\sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2$		

F 통계량이 자유도에 비해서 매우 작다는 것은 설명변수의 각 범주에 따른 반응변수의 평균치 차가 존재하지 않다는 것을 의미하며, 따라서 설명변수가 반응변수의 분류나 예측에 영향을 주지 않는다고 결론지을 수 있다.

자유도에 대한 F 통계량의 크고 작음은 p-값으로 표현될 수 있는데 F 통계량이 자유도에 비해서 작으면 p-값은 커지게 된다. 분리기준을 F 통계량 값으로 할 경우, 가장 작은 p-값을 가지는 설명변수에 의해 자식마디를 형성하는 최적분리를 한다.

2.1.2 CART 알고리즘

CART(Classification and Regression Trees; Breiman et al., 1984) 알고리즘은 설명변수들의 특성에 다라 자료들을 이지 분리하며, 2개의 하위노드를 생성하는 과정을 반복하여 반응변수의 값이 유사한 부분집합으로 만드는 방법이다. 반응 변수로 범주형 변수뿐만 아니라 연속형 반응변수도 사용이 가능 하다. 또 반응변수에 가장 유의적인 설명변수를 찾는 것이 아니라, 노드의 불순도(impurity)와 다양성(diversity)을 가장 많이 줄여주는 설명변수를 선택하는 방식을 택하고 있다.

반응변수가 범주형일 때, CART 알고리즘은 지니 지수(Gini index)를 분리기준으로 사용한다. 반면, 반응변수가 연속형일 때는 분산(variance)의 감소량을 분리기준으로 이용한다.

지니 지수는 카이제곱 통계량과 마찬가지로 불순도(impurity)를 측정하는 하나의 지수이다. 여기서 불순도(impurity)란 반응변수의 특정 범주에 해당 마디의 개체들이 집중되어 있지 않고 섞여있는 정도를 의미한다.

먼저 각 마디에 속하는 개체를 그 마디에서 도수가 가장 많은 반응변수의 한 범주에만 모두 할당하는 분류규칙을 고려한다. 임의의 한 개체가 반응변수의 i번째 범주로부터 추출되었고, 그 개체를 반응변수의 j번째 범주에 속한다고 오분류(misclassification)할 확률은 $P(i)P(j)$가 된다. 여기에서 $P(i)$

는 각 마디에서 한 개체가 반응변수의 i번째 범주에 속할 확률이다. 이러한 오분류 확률을 모두 더하여,

$$G=\sum_{j=1}^{c}\sum_{i\neq j}P(i)\,P(j)\ , \tag{2.4}$$

를 얻을 수 있고, 이는 앞에서와 같은 분류규칙 하에서 오분류 확률의 추정치가 된다. 여기서 c는 반응변수의 범주수를 말한다.

지니 지수는 각 마디에서의 불순도 또는 다양도(diversity)를 재는 측도 중의 하나로써,

$$G=\sum_{j=1}^{c}P(j)\,(1-P(j))=1-\sum_{j=1}^{c}P(j)^2=1-\sum_{j=1}^{c}(n_j/n)^2\ , \tag{2.5}$$

와 같이 표현될 수 있다. 여기에서 n은 그 마디에 포함되어 있는 개체수를 말하고, n_i는 반응변수의 i번째 범주에 속하는 개체수를 말한다. 지니 지수는 n개의 원소 중에서 임의로 2개를 추출하였을 때, 추출된 2개가 서로 다른 그룹에 속해있을 확률을 의미하며, Simpson(1949)의 다양도 지수(diversity index)로도 알려져 있다. 반응변수의 범주가 2개인 경우에는 지니 지수는 다음과 같이 표현된다.

$$G=2P(1)\,P(2)=2(\frac{n_1}{n})(\frac{n_2}{n})\ , \tag{2.6}$$

식 (2.6)은 카이제곱 통계량을 사용하는 것과 같은 결과를 갖는다.

CART는 이 지니 지수를 가장 감소시켜주는 설명변수와 그 변수의 최적 분리를 자식마디로 선택하는데, 지니 지수의 감소량은 다음과 같이 계산된다.

$$\Delta G = G - \frac{n_L}{n} G_L - \frac{n_R}{n} G_R \ , \tag{2.7}$$

여기서 n은 부모마디의 관측치 수를 말하고, n_R과 n_L는 각각 자식마디의 관측치 수를 의미한다. 즉, 자식마디로 분리되었을 때의 불순도가 가장 작도록 자식마디를 형성하는 것이며, 이는 다음과 같은 자식마디에서의 불순도의 가중합을 최소화하는 것과 동일하다.

$$P(L)\, G_L + P(R)\, G_R = \frac{n_L}{n} G_L + \frac{n_R}{n} G_R \ . \tag{2.8}$$

2.1.3 C4.5 알고리즘

C4.5 알고리즘은 Quinlan(1993)에 의해 수정 발전된 의사결정 알고리즘이다. 이것의 초기버전인 ID3 알고리즘은 기계학습(machine learning) 분야에 많은 영향을 주었다. CART가 각 마디에 이원분할을 형성하며 이지분리 나무구조를 만드는데 반하여 C4.5는 연속형 예측 변수에 관해서는 이지분리를 하지만, 명목형 예측변수에 관해서는 각 범주가 하나의 마디를 가지는 다지 분리 구조를 갖는 나무로 구성된다. C4.5에서 의사결정나무를 형성하기 위하여 처음 수행하는 작업은 분할 정복(divide and conquer)이다. 입력되는 훈련 집합(training set)이 성공적으로 분할되도록 모든 하부집합에 하나의 집단(class)

이 속하는 경우들로 구성될 때까지 나무를 형성한다. C4.5는 정보(information)라는 개념을 사용한다. p가 메시지(message)의 확률일 때, 이 메시지로 전달되는 정보는 $-\log_2 p$로 측정한다. 예를 들어 8개의 동일한 확률을 갖는 메시지(equally probable message)가 있을 경우, 한 메시지의 정보는 $-\log_2 \frac{1}{8} = 3$이 된다. 이는 작은 확률로 일어나는 메시지일수록 이를 알기 위해서는 보다 많은 정보가 필요하다는 뜻이다. 개체(case)들의 집합인 S에서 무작위로 한 개체를 선택 할 때, 이 개체가 C_j에 속할 확률은 다음과 같다.

$$\frac{freq(C_j,S)}{|S|} \quad , \tag{2.9}$$

여기서, $|S|$는 S에 속하는 모든 개체의 개수이고, $freq(C_j,S)$는 집합 S에서 C_j에 속하는 개체들의 개수이다. 따라서 이 개체가 전달하는 정보(information)는 다음과 같다.

$$-\log_2\left(\frac{freq(C_j,S)}{|S|}\right) \quad , \tag{2.10}$$

집합 S에서 기대 정보(expected information)를 구하기 위해선, 각 개체가 전달하는 정보를 가중평균하면 된다.

$$Info(S) = -\sum_j^k \left(\frac{freq(C_j,S)}{|S|} \times \log_2\left(\frac{freq(C_j,S)}{|S|}\right) \right) \quad , \tag{2.11}$$

위의 $Info(S)$와 비슷한 개념으로, T가 X에 의해 n개로 분할 된 후의 기대

정보(expected information)를 구하려면 식 (2.12)를 이용할 수 있다.

$$Info\ x(T) = \sum_{i}^{n}\left(\frac{|T_i|}{T} \times Info(T_i)\right) , \tag{2.12}$$

X에 의한 분할로 얻어진 정보(information)는 다음 식 (2.13)에 의해 얻을 수 있다.

$$Gain(X) = Info(T) - Info\ X(T) , \tag{2.13}$$

기존 알고리즘인 ID3에서는 이 Gain을 최대로 하는 테스트를 선택했었다. 그러나 이 경우에는 범주의 수가 많은 변수로의 심각한 편의(bias)가 생기는 문제점이 있다. 예를 들어, 각 최종마디(terminal node)에 한 개체만을 포함하며, 모든 개체들이 1의 확률로 배정되는 분리 변수가 있다고 하자. 이 경우에는 $Info\ X(T) = 0$일 것이다. 따라서 어떤 변수를 사용하는 것보다 정보이득(information gain)이 최대가 될 것이다. 그러나 이러한 분리는 전혀 의미를 갖지 못한다. 그래서 T에 있는 한 개체가 속하는 하부집합(집단 대신)을 정의하는데 필요한 평균 정보의 양(split info)으로 정규화(normalize)시켜줄 필요가 있다.

$$Split\ Info(X) = \sum_{i=1}^{n}\left(\frac{|T_i|}{|T|} \times \log_2(\frac{|T|}{|T|})\right) , \tag{2.14}$$

정보의 양(split info)은 T가 n개의 하부집합으로 분할됨에 따라 발생하는 정보(information)의 양이다. Gain을 정보의 양으로 나누어 주면, 분할에 의해 생

성된 유용한 정보의 비율인 정보 비율(gain ratio)이 된다.

$$Gain\,Ratio(X) = \frac{Gain(X)}{Split\ \ Info(X)} \quad . \tag{2.15}$$

변수별로 정보 비율을 최대화 시켜주는 분리지점(split point)을 찾고, 이를 각 변수 별로 실시하여 변수끼리 최대 정보 비율도 비교하여, 그중에서 가장 큰 변수를 선택하면 된다. 즉 C4.5도 변수의 선택과 분리지점(split point)의 선택이 동시에 이루어진다.

2.1.4 QUEST 알고리즘

QUEST(Quick, Unbiased, Efficient, Statistical Tree; Loh & Shin, 1997) 알고리즘은 CART에서와 같이 이지분리(binary split)를 수행하는 알고리즘이다.

QUEST에서 변수선택 알고리즘을 간단히 요약하면 다음과 같다.

1. 순서형 예측변수에 대해서는 F검정의 p-값을 계산한다.
2. 범주형 예측변수에 대해서는 예측변수와 목표변수의 분할표로부터 카이제곱 검정의 p-값을 계산한다.
2. 1, 2 단계에서 가장 작은 p-값이 Bonfferroni 수정된 임계값 보다 작으면 그에 대응되는 변수를 분리변수로 선택한다.
4. 그렇지 않으면, 순서형 예측변수에 대하여 F검정의 p-값하고 Bonferroni 수정된 임계값과 비교한다. p-값에 대응되는 변수를 분리변수로 선택한다.

QUEST는 명목형 목표변수에 대해서만 분석을 수행할 수 있으며 예측변수의 측도에 따라서 서로 다른 분리규칙을 사용한다. 예측변수가 순서형 또는 연속형인 경우에는 분리규칙으로 앞에서 설명한 ANOVA F-검정 또는 Levene의 검정(Levene's robust test of homogeneity of variance)을 사용하며, 예측변수가 명목형인 경우에는 Pearson의 카이제곱 검정을 사용한다.

목표변수의 범주가 3개 이상인 경우에는 CART의 Twoing 기준에서와 유사하게 2-평균 군집분석(two-means clustering)을 수행하여 두 개의 그룹을 만든 후 분석을 수행한다.

각 마디에서 목표변수의 범주의 집합을 $C=\{1, 2,\cdots, c\}$라 할 때, 이를 임의의 두 집합 C_1과 $C_2 = C - C_1$로 그룹화 하는 경우, 이러한 임의의 그룹화에 대해서 목표변수의 범주수를 2라 가정한 후 $P(1)p(2)$의 감소량을 계산하여 이를 $\Delta G(C_1)$이라고 표현한다. 이 때 모든 가능한 C_1 에 대해서 $\Delta G(C_1)$을 최대화 하는 것은 결국 다음과 같은 Twoing 지수 T를 최대화하는 것과 같게 된다.

$$T = \frac{P_L P_R}{4}\left[\sum_{i=1}^{c}|P_L(i) - P_r(i)|\right] = \frac{1}{4}\left(\frac{n_L}{n}\right)\left(\frac{n_R}{n}\right)\left[\sum_{i=1}^{c}\left|\frac{n_{iL}}{n_L} - \frac{n_{iR}}{n_R}\right|\right]^2 . \qquad (2.16)$$

또한 각 예측변수의 최적분리를 찾기 위하여 2차 판별분석(quadratic discriminant analysis)을 수행하고, 목표변수를 가장 잘 분류하는 예측변수의 최적분리를 이용하여 자식마디를 형성한다.

일반적으로 CART는 자식마디를 형성할 때 보다 많은 이산값을 가지는 예측 변수를 선택하는 경향이 있기 때문에, 계산시간이 다소 많이 걸리고 분류 또는 예측오차가 커질 가능성이 있다. QUEST는 위와 같은 알고리즘을 이용하여 변수선택 편의(bias)나 계산시간을 줄이고자 하는 방법이다.

QUEST는 관측치의 수가 많거나 복잡한 자료에 대해서는 효율적이지만, 기존의 방법과 비교하여 분류 정확도나 나무 크기 면에서는 우세하다고 할 수 없다.

2.2 선형결합 의사결정나무

단일 분리기준 의사결정나무는 예측변수 한 개로 분리가 일어나기 때문에 변수가 가지고 있는 측도에 따라서 자료의 분리 되는 병합에 영향을 받게 된다. 또한, 독립변수의 모든 선형결합의 조합을 고려하는 것에 비해 2개인 경우는 2차원 플롯(plot)을 통해 시각화(visualization)하기도 쉬울 뿐 아니라 예측력도 뛰어나며 알고리즘 계산시간도 효율적이기 때문에 본 논문에서는 독립변수의 개수가 2개인 경우로만 논의를 국한하고자 한다.

2.2.1 CART 선형결합 알고리즘

CART 선형결합은 단계적으로 $\sum a_i x_i \leq c$의 해 상수 c와 계수 $\{a_i\}$을 찾음으로써 분리가 이루어진다. 이 방법은 모든 상수 c에 대해 계수 $\{a_i\}$와 최적 분리를 찾기 위한 부분집합(subset) A를 고려하는데, 이는 각 마디에서의 불순도(impurity)를 최소화하는 것이다.

선형결합분리기준은 선형판별분석(linear discriminant analysis)에 의해 일반화 되며 계수행렬에 대한 주성분분석(principal component analysis)이 각 마디(node) 마다 수행된다. 선형판별함수(linear discriminant function)는 β번째 최대 고유값(the largest eigenvalue)을 초과하는 고유값의 주성분으로부터 얻는다. 분리변수는 다음의 판별함수에 의해 선택된다.

$$d_j(y) = \hat{u_j}' \widehat{\sum}^{-1} y - \frac{1}{2} \hat{u_j}' \widehat{\sum}^{-1} \hat{\mu_j} + \ln\{p(j|t)\} \ , \qquad (2.17)$$

여기서, y는 최대주성분 공간상의 벡터이며, $\hat{\mu_j}$는 j번째 개체의 표본평균벡터(sample mean vector), $\widehat{\sum}$은 마디에서의 공분산행렬 합동추정값(pooled estimate of the covariance matrix)이다. 각 마디는 J개의 부마디(subnode)로 분할되며 기대 오분류 비용(expected misclassification cost)은 최소화 되도록 추정되며 다음과 같이 표현된다.

$$\sum_{j=1}^{J} C(i|j) \exp\{d_j(y)\} = \min \sum_{j=1} C(m|j) \exp\{d_j(y)\} \ , \qquad (2.18)$$

여기서, $i = j$이면 $C(i|j) = 0$이다. $i \neq j$일 때, $C(i|j)$가 1의 값을 가지는 경우를 제외하고, 오분류 비용은 사전확률(prior)을 적절히 이용하여 수정되는데, 사전확률은 다음과 같이 계산된다.

$$\pi'(j) = \frac{C(j)\pi(j)}{\sum_i C(i)\pi(i)} \ , \qquad (2.19)$$

여기서, $C(j) = \sum_i C(i|j)$이다.

2.2.2 CRUISE 2D 선형결합 알고리즘

Kim & Loh(2001)는 Loh & Shin(1997) 방법의 약점을 보완함과 함께 변수들 간 상호작용을 변수선택의 기준에 포함시키는 알고리즘인 CRUISE를 제안하였다. 기본 의사결정나무 방법의 한계였던 변수선택의 편의(bias)를 제거함으로써 의사결정나무의 불안정성 문제도 대부분 해결하는 장점이 있다.

또한 이것을 발전시켜 Kim & Loh(2009)는 선형결합의 최적분리를 찾기 위하여 선형판별분석(LDA)에 적용하여 나무구조를 형성하는 방법을 제안하였다.

먼저 선형판별모형에 적합한 2개의 변수를 선택하기 위해 2가지 방법을 고려한다. 첫 번째 방법은 'C'(for cost) 방법이라 한다. 모든 예측변수를 적합시켜 오분류 비용을 추정하고, 오분류율을 최소화하는 변수의 조합을 선택한다. 결측값이 존재하는 변수는 제외하고 모형을 적합하며 결측값의 경우 연속형 변수인 경우 평균(mean)으로 대체되며 범주형 변수인 경우 중앙값(mode)로 대체된다.

두 번째 방법은 'M'(for MANOVA)방법으로 'C' 방법이 결측값이 존재하는 변수를 분리변수로 선택하는 것을 피하기 위해 다변량 분산분석(multivariate analysis of variance)을 수행하여 p-값이 가장 작은 변수의 조합을 선택한다. 어떤 한 마디를 t라고 할 때, 독립변수의 조합 $x=(x_l,\ x_m)'$, t 마디에서 j집단(class)의 수를 n_j, j 집단(class)에서 i번째 관측치를 x_{ij}, n_j로부터의 표본평균벡터(sample mean vector)를 $\overline{x_j}$, 전체표본 평균벡터를 $\overline{x}$라 한다.

이 때, 그룹간 제곱합 행렬(between group sum of squares matrices)로

$$B=\sum_{j} n_j(\overline{x_j}-\overline{x})(\overline{x_j}-\overline{x})', \tag{2.20}$$

와 그룹내 제곱합 행렬(within group sum of squares matrices)로

$$W=\sum_{j}\sum_{i=1}^{n_j}(x_{ji}-\overline{x_j})(x_{ji}-\overline{x_j})', \tag{2.21}$$

이 같이 사용된다.

집단평균 벡터의 동질성(equality)을 테스트 하기위한 MANOVA Wilks' Λ는 다음과 같다.

$$\Lambda=|W|/|B+W| \ , \tag{2.22}$$

p-값을 측정하기 위해 Bartlett's(1938) 근사에 의한

$$-\{n-1-(2+J_t)/2\}\log\Lambda \sim \chi^2_{2(J_{t-1})} \ . \tag{2.23}$$

이 이용된다.

여기서, $n=\sum_{j}^{n_j}$이고, J_t는 t에서의 집단의 수를 의미한다.

제3장 이변량 선형결합 분리기준을 이용한 의사결정나무

이변량 선형결합 분리기준은 일변량(single split) 분리기준에 비해 예측의 정확성(prediction accuracy)이 뛰어나며 의사결정나무의 크기를 줄이는 효과가 있다. 오직 2개의 독립변수만을 이용하기 때문에 2개의 변수가 결합하여 어떻게 분리되었는지에 대해 2차원 플롯(plot)을 통해 시각화하기 쉬운 장점을 가진다. 변수가 3개 이상인 경우에는 복잡하여 마디(node)에 대해 판별모형을 적용시킬 수가 없다. 따라서 본 논문에서는 변수가 2개인 경우로만 논의를 국한하고자 한다.

의사결정나무를 형성하기 위해서, 첫 번째 방법으로 2개의 독립변수에 대해 동일하게 가중치를 부여한 분리기준을 이용하고, 두 번째 방법으로 황금분할 5:3으로 가중치를 부여한 분리기준을 이용하여 의사결정나무를 형성한다.

3.1 일대일 선형결합 의사결정나무

연속형 변수인 경우 선형결합 분할 규칙의 기본 구조는 다음과 같다.

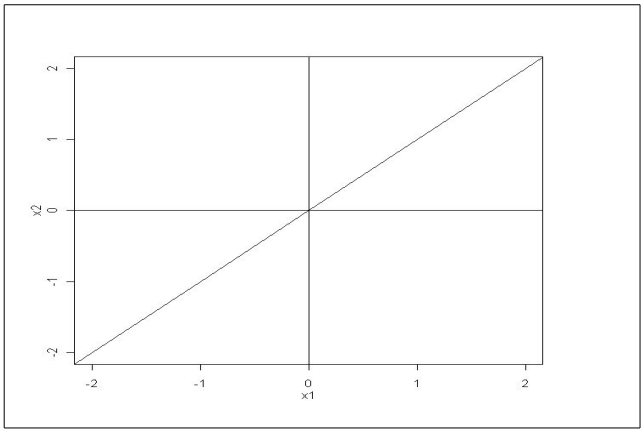
$$\sum_{k=1}^{K} a_k x_k \leqq c \ , \tag{3.1}$$

선형결합 분리기준을 찾기 위해서 단계적으로 $\sum_{k=1}^{K} a_k x_k \leq c$의 해 상수 c와 계수 $\{a_i\}$을 구하게 된다. 여기서 c는 연속형 변수인 공간상의 한 점으로, 이 점을 기준삼아 하부공간(subset)으로 분할했을 때 자료의 불순도가 가장 낮았기 때문에 선택된 점이다. 불순도를 가장 낮게 하는 분할점 c를 찾는 일은 나무구조 형성에 있어서 예측력의 관점에 있어서는 뛰어나지만 연산시간이 과다한 약점이 있다.

따라서 분류정확도를 따지기 보다는 더 효율적인 나무구조를 구축하기 위해 독립변수에 동일한 가중치를 두어 결합하여 분리기준으로 제시하고자 한다. 예로 $X_1 + X_2, X_1 - X_2$와 같은 변수를 분할에 사용하여 나무구조를 구축하는 것이다.

즉, 일대일 선형결합 의사결정 나무란 [그림3.1]과 같이 $\angle A$와 $\angle B$가 45°가 되도록 X_1과 X_2에 동일한 가중치로 결합하여 분리기준으로 사용하는 것이다.

이와 같은 분할규칙(partition rule)을 이용하여 분할하고, 분할된 각 하부공간에 대하여 다시 반복적으로 분할규칙을 사용하여 불순도가 충분히 낮은 경우에 반복분할을 중단한다.



[그림3.1] 일대일 결합 분리변수 생성방법

제안된 방법으로 변환된 변수를 이용하여 모형구축을 위한 분리변수 선택과정은 다음과 같다. 불순도는 지니지수(Gini index)와 엔트로피(entropy)로 측정하게 되는데 선형결합을 이용한 분류나무는 분리변수 선택 과정에서 CART와 마찬가지로 지니지수(Gini index)를 사용한다. 지니지수를 중심으로 분리 알고리즘을 살펴보기로 한다.

s마디에서의 총 자료 수를 $N(s)$ 라고 하고, s마디에서 그룹 j에 속하는 자료의 수를 $N_j(s)$라고 한다. 또한 각 그룹의 사전 확률을 $\pi(j)$라고 하는데, 주로 $\frac{N_j}{N}$로 측정한다.

이 때 s 마디에서 그룹 j의 확률 $p(j,s)$은 다음과 같다.

$$p(j,s) = \pi(j) \cdot \frac{N_{j(s)}}{N_j} = \frac{N_j(s)}{N} \quad , \tag{3.2}$$

또한 s마디로 분류될 확률 $p(s)$는 그 마디에서의 각 그룹의 합과 같다.

$$p(s) = \sum_j P(j,s) \quad , \tag{3.3}$$

s 마디로 분류되었는데 그룹 j에 속할 확률, $p(j|s)$는 다음과 같다.

$$p(j|s) = \frac{p(j,s)}{p(s)} = \frac{N_j(s)}{N(s)} \quad , \tag{3.4}$$

여기서 $p(j|s)$는 이 하부공간 s에 위치한 관찰값 중 그룹 j에 속한 비율이다. 즉 $p(j|s)$가 어느 특정 그룹 j에서 큰 값을 가지고 나머지 그룹에서 작

은 값을 가지면 불순도는 매우 낮아지게 되며, 다음 식의 성립해야 한다.

$$\sum_{j} p(j|s) = 1 \ , \tag{3.5}$$

s 마디에서 지니지수로 관찰된 불순도의 정의는 다음과 같다.

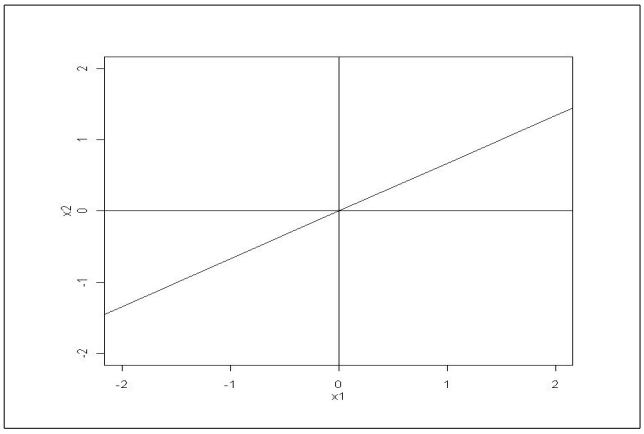
$$\text{불순도}(s) = 1 - \sum_{j} p^2(j|s) \ . \tag{3.6}$$

3.2 황금분할을 이용한 의사결정나무

고대 그리스의 수학자 Eudoxos(BC408~BC355)는 어떤 대상이 62 : 38로 양분되었을 때 그러한 분할비율을 황금분할(golden section)이라 부르기 시작했다. 혹자는 이를 줄여서 5 : 3황금률 이라 일컫기도 한다. 황금분할은 건축, 조각 등 조형예술 분야에서 하나의 원리로 활용되어 왔으며, 일상생활에서도 널리 적용되고 있다. 이점을 활용하여 2개의 독립변수 선형결합 가중치를 위해 사용하고자 한다.

황금분할 의사결정 나무란 [그림3.2]와 같이 X_1축과 X_2축을 5 : 3의 비율로 분할하여 이 가중치를 통해 선형결합을 행하여 분리기준으로 사용하는 것을 말한다.

결국 황금분할 의사결정나무는 [그림3.2]와 같이 $\angle A$가 33.75° (56.25°)이고 $\angle B$가 56.25° (33.75°)가 되도록 X_1과 X_2에 가중치를 갖도록 결합하여 분리기준으로 사용하는 것이다.



[그림3.2] 황금분할을 이용한 분리기준

즉, $\tan 33.75 = 0.6682 \simeq 0.67$ 이 성립하기 때문에 다음과 같은 선형결합 식으로 이루어진다.

$$\begin{cases} X_1 + 0.67X_2 \\ X_1 - 0.67X_2 \end{cases}, \quad \begin{cases} 0.67X_1 + X_2 \\ 0.67X_1 - X_2 \end{cases}, \tag{3.7}$$

황금분할 의사결정나무 역시 분할된 각 하부공간에 대하여 다시 반복적으로 분할규칙을 사용하여 불순도가 충분히 낮은 경우에 반복분할을 중단하며 불순도를 측정하기위해 지니지수(Gini index)를 사용한다.

제4장 모의실험 및 적용

3장에서 제안한 선형결합을 이용한 분리기준이 반응변수들을 얼마나 잘 분리하는 가를 모의실험을 통해 확인해보고, 실제자료에 적용해본다.

4.1 분리기준에 대한 모의실험

제안한 분리기준이 잘 수행되는 지를 알아보기 위해서, R 소프트웨어를 이용하여 선형결합을 이용한 분리기준에 대한 모의실험을 수행한다. 모의실험의 과정은 다음과 같다.

- 상관계수를 변화시킨 모의실험이다.

Ⅰ.(반응변수 생성) 먼저 베르누이 분포 $B(1, 0.5)$로부터 크기가 100인 반응변수 벡터 Y를 생성한다.

Ⅱ.(분리변수 생성) 반응변수 Y의 값에 따라 $Y=0$ 이면, 평균이 $(1, 0)$, 분산이 $(1, 1)$이고 상관계수가 ρ인 이변량 정규분포 $MN_1(\underline{\mu}, \sum)$, $\underline{\mu}=(1,0)$, $\sum=\begin{pmatrix}1 & \rho \\ \rho & 1\end{pmatrix}$로부터 생성하고, $Y=1$ 이면 평균이 $(0,1)$, 분산이 $(1, 1)$이고 상관계수가 ρ인 이변량 정규분포 $MN_2(\underline{\mu}, \sum)$, $\underline{\mu}=(0,1)$, $\sum=\begin{pmatrix}1 & \rho \\ \rho & 1\end{pmatrix}$로부터 생성시킨다. 이런 방법으로 100번 반복 수행한다. 크기가 100인 이변량 정규분포의

표본이 생성되면, 이변량 중 하나를 분리변수 X_1, 나머지 하나를 분리변수 X_2라 둔다. 또한 위에서 생성된 분리변수 X_1과 X_2의 선형결합인 X_1+X_2, X_1-X_2 와 같은 선형결합의 조합 또한 분리변수로 사용한다.

Ⅲ.(상관계수) -0.9에서부터 0.1만큼 증가시켜 0.9까지 변화시킨다.

Ⅳ.(분리) 생성된 이항 반응변수를 설명변수 X_1, X_2에 의해 선형결합을 이용한 분리기준을 적용시킨다. 그리하여 분리된 의사결정나무의 오분류율(missclassification rate)을 계산한다.

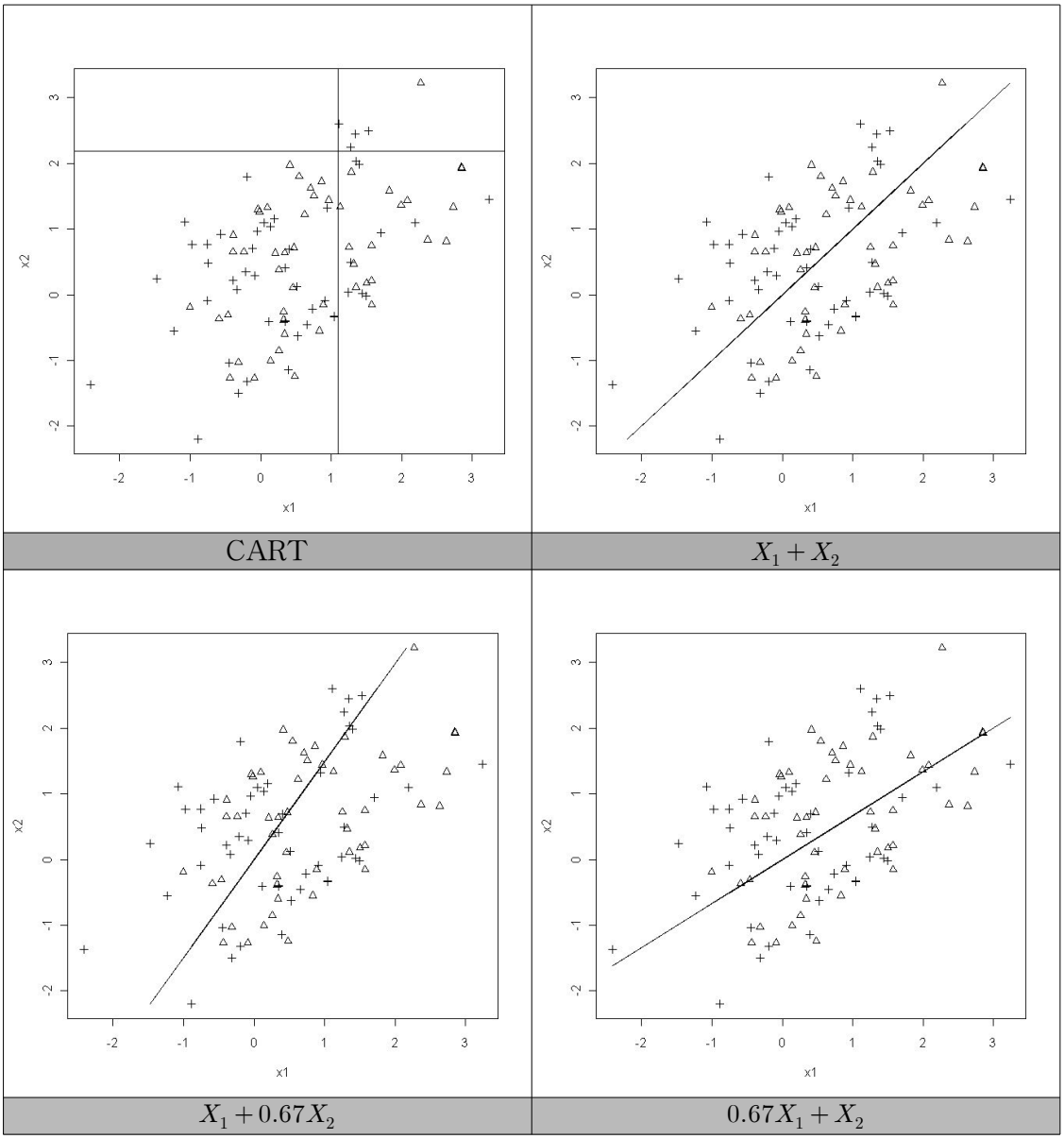
Ⅴ.(반복) 10,000번의 모의실험을 반복한다.

10,000번의 반복수행 결과, 각 분리기준의 오분류율을 확인할 수 있다. [표 4.1]은 이항 반응변수 Y일 때의 CART와 제안된 분리기준의 오분류율을 나타낸 것이다.

[표4.1] $\sum=\begin{pmatrix}1 & 0.5\\ 0.5 & 1\end{pmatrix}$인 경우 분리기준별 반응변수 Y의 오분류율

상관계수	CART	X_1+X_2 X_1-X_2	$X_1+0.67X_2$ $X_1-0.67X_2$	$0.67X_1+X_2$ $0.67X_1-X_2$
-0.9	0.266	0.262	0.261	0.261
-0.8	0.262	0.257	0.257	0.257
-0.7	0.259	0.252	0.253	0.253
-0.6	0.256	0.249	0.248	0.248
-0.5	0.252	0.244	0.245	0.244
-0.4	0.248	0.238	0.239	0.239
-0.3	0.244	0.232	0.233	0.233
-0.2	0.249	0.226	0.226	0.226
-0.1	0.232	0.219	0.220	0.219
0	0.226	0.209	0.211	0.211
0.1	0.217	0.199	0.201	0.202
0.2	0.208	0.187	0.190	0.190
0.3	0.198	0.172	0.178	0.177
0.4	0.156	0.156	0.164	0.163
0.5	0.170	0.135	0.145	0.146
0.6	0.154	0.110	0.125	0.125
0.7	0.135	0.079	0.099	0.099
0.8	0.112	0.043	0.068	0.068
0.9	0.086	0.064	0.033	0.032

분리기준별로 비교했을 때, 제안된 분리기준 모두 양의 상관계수의 값이 클수록 오분류 확률이 작음을 알 수 있다. 또한 [표4.1]에서 상관계수가 0일 때 일대일 선형결합 분리기준의 오분류율은 0.209으로 가장 작았으며, 전체적으로 다른 분리기준에 비해 오분류율이 작았다. 다음과 같이 [그림4.1]에서 쉽게 비교할 수 있다.



[그림4.1] $\rho = 0.9$인 경우 각 분리기준의 비교

• 다음은 생성된 반응변수 Y에 따른 분리변수 (X_1, X_2)의 이변량 정규분포의 평균을 변화시킨 모의실험이다.

Ⅰ.(반응변수 생성) 먼저 베르누이 분포 $B(1, 0.5)$로부터 크기가 100인 반응변수 벡터 Y를 생성한다.

Ⅱ.(분리변수 생성) 반응변수 Y의 값에 따라 $Y=0$ 이면, 평균이 $\underline{\mu_1}=(\mu,\mu)$, 분산이 $(1,1)$이고 상관계수가 $\rho=0.5$인 이변량 정규분포 $MN_1(\underline{\mu}, \sum)$, $\underline{\mu_1}=(\mu,\mu)$, $\sum=\begin{pmatrix}1 & 0.5\\ 0.5 & 1\end{pmatrix}$로부터 생성하고, $Y=1$ 이면 평균이 $\underline{\mu_2}=(\mu,\mu)$, 분산이 $(1,1)$이고 상관계수가 $\rho=0.5$인 이변량 정규분포 $MN_2(\underline{\mu}, \sum)$, $\underline{\mu}=(\mu,\mu)$, $\sum=\begin{pmatrix}1 & 0.5\\ 0.5 & 1\end{pmatrix}$로부터 생성시킨다. 이런 방법으로 100번 반복 수행한다. 크기가 100인 이변량 정규분포의 표본이 생성되면, 이변량 중 하나를 분리변수 X_1, 나머지 하나를 분리변수 X_2라 한다. 또한 위에서 생성된 불리변수 X_1과 X_2의 선형결합인 X_1+X_2, X_1-X_2 또한 분리변수로 사용한다.

Ⅲ.(평균벡터) 이변량 정규분포의 평균을 $\begin{cases}\underline{\mu_1}=(1,0)\\ \underline{\mu_2}=(0,1)\end{cases}$, $\begin{cases}\underline{\mu_1}=(-1,-1)\\ \underline{\mu_2}=(1,1)\end{cases}$, $\begin{cases}\underline{\mu_1}=(2,0)\\ \underline{\mu_2}=(0,2)\end{cases}$, $\begin{cases}\underline{\mu_1}=(-2,-2)\\ \underline{\mu_2}=(2,2)\end{cases}$ 으로 각각 변화 시킨다.

Ⅳ.(공분산행렬) $\sum=\begin{pmatrix}1 & 0.5\\ 0.5 & 1\end{pmatrix}$, $\sum=\begin{pmatrix}2 & 0.5\\ 0.5 & 2\end{pmatrix}$ 2가지 경우를 고려한다.

Ⅴ.(분리) 생성된 이항 반응변수를 설명변수 X_1, X_2에 의해 선형결합을 이용한 분리기준을 적용시킨다. 그리하여 분리된 의사결정나무의 오분류 확률을 계산한다.

Ⅵ.(분리) 10,000번의 모의실험을 반복한다.

10,000번의 반복수행 결과, 선형결합 분할의 오분류 확률을 확인 할 수 있다. [표4.2]와 [표4.3]은 이항 반응변수 Y일 때의 CART와 제안된 분리기준의 오분류 확률을 나타낸 것이다.

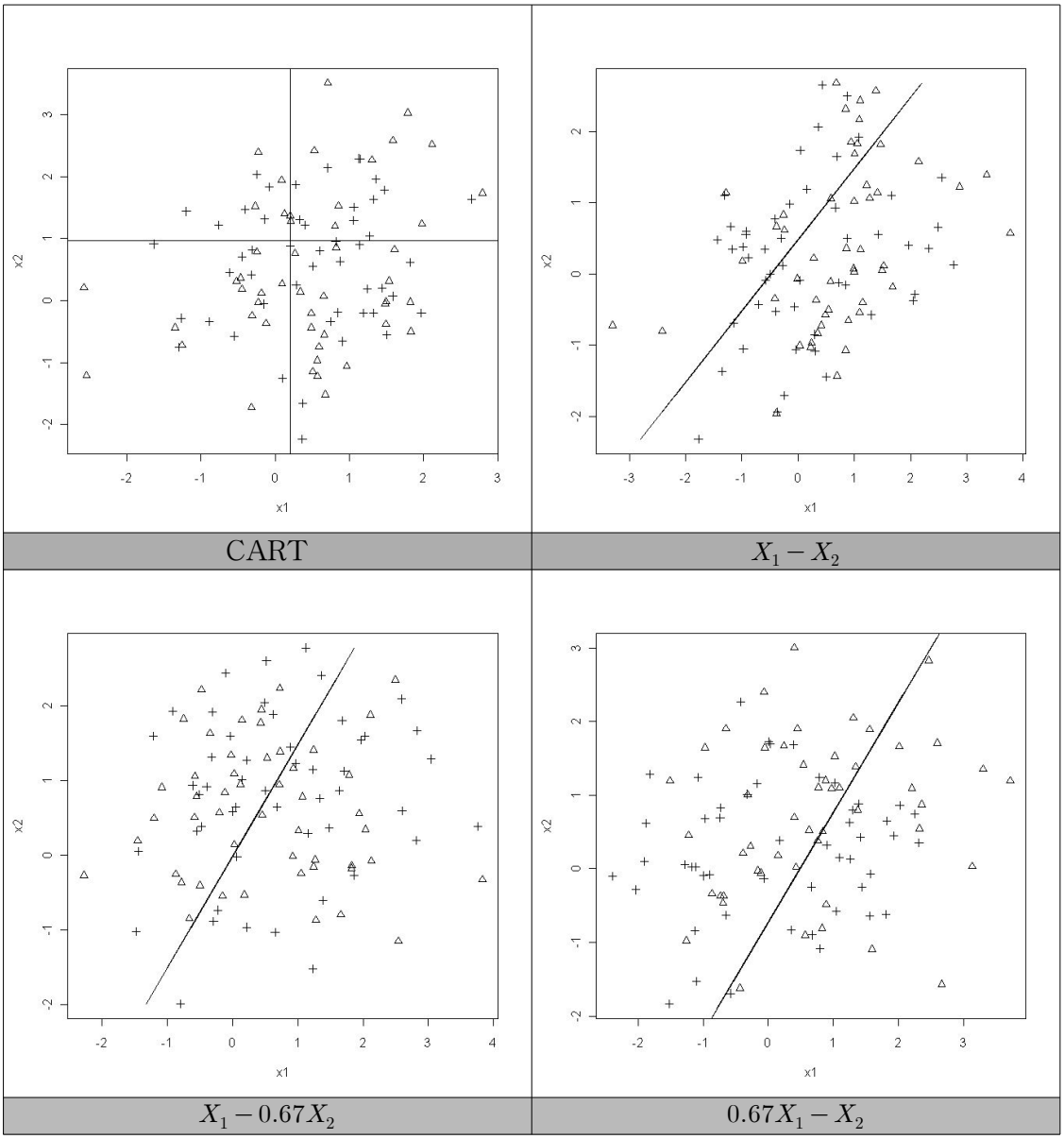
[표4.2] $\sum = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$인 경우 분리기준별 반응변수 Y의 오분류율

이변량 정규분포의 평균	CART	$X_1 + X_2$ $X_1 - X_2$	$X_1 + 0.67X_2$ $X_1 - 0.67X_2$	$0.67X_1 + X_2$ $0.67X_1 - X_2$
$\underline{\mu_1} = (1,0)$ $\underline{\mu_2} = (0,1)$	0.138	0.123	0.125	0.123
$\underline{\mu_1} = (-1,-1)$ $\underline{\mu_2} = (0,1)$	0.105	0.096	0.096	0.096
$\underline{\mu_1} = (2,0)$ $\underline{\mu_2} = (0,2)$	0.051	0.014	0.018	0.020
$\underline{\mu_1} = (-2,-2)$ $\underline{\mu_2} = (2,2)$	0.009	0.046	0.004	0.004

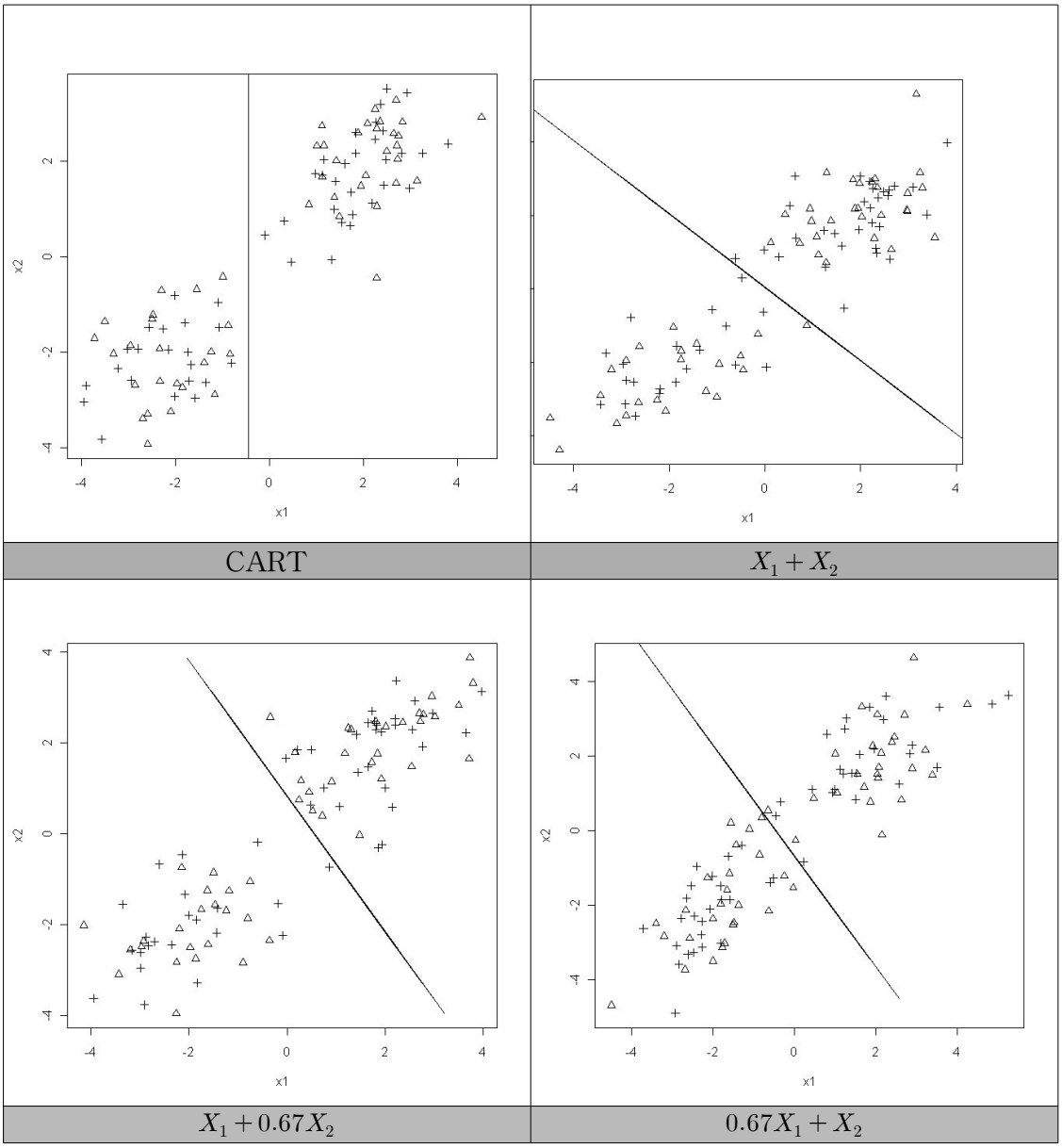
[표4.3] $\sum = \begin{pmatrix} 2 & 0.5 \\ 0.5 & 2 \end{pmatrix}$인 경우 분리기준별 반응변수 Y의 오분류율

이변량 정규분포의 평균	CART	$X_1 + X_2$ $X_1 - X_2$	$X_1 + 0.67X_2$ $X_1 - 0.67X_2$	$0.67X_1 + X_2$ $0.67X_1 - X_2$
$\underline{\mu_1} = (1,0)$ $\underline{\mu_2} = (0,1)$	0.212	0.199	0.197	0.198
$\underline{\mu_1} = (-1,-1)$ $\underline{\mu_2} = (0,1)$	0.150	0.139	0.140	0.141
$\underline{\mu_1} = (2,0)$ $\underline{\mu_2} = (0,2)$	0.114	0.097	0.101	0.101
$\underline{\mu_1} = (-2,-2)$ $\underline{\mu_2} = (2,2)$	0.045	0.025	0.026	0.026

[표4.2]와 [표4.3]에서는 각 알고리즘에 대하여 평균의 변화에 따른 오분류율을 계산하였다. 분리기준별로 비교했을 때, 제안된 분리기준이 CART분리기준보다 오분류율이 작았다. 일대일 선형결합 분리기준이 대체로 우수한 결과를 보여주었다. 다음과 같이 [그림4.2]와 [그림4.3]에서 쉽게 비교할 수 있다.



[그림4.2] $\sum = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$, $\underline{\mu_1} = (1,0)$, $\underline{\mu_2} = (0,1)$인 경우 각 분리기준 비교



[그림4.3] $\sum = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$, $\underline{\mu_1} = (-2,-2)$, $\underline{\mu_2} = (2,2)$인 경우 각 분리기준 비교

4.2 실제자료의 적용

4.2.1 자료 소개

제안된 분리기준을 이용하여 실제자료에 적용해본다.

적용 자료는 고객들의 신용가치에 대한 독일 은행의 자료이다. 적용할 자료에는 고객나이, 대출금액, 보증인수, 대출기간, 보유계좌 수, 부채비율, 거주기간 등의 정보가 포함되어 있고, 원 자료의 수는 1,000개이다. 자세한 변수설명은 [표4.4]와 같다.

[표4.4] 적용된 자료의 변수 정의

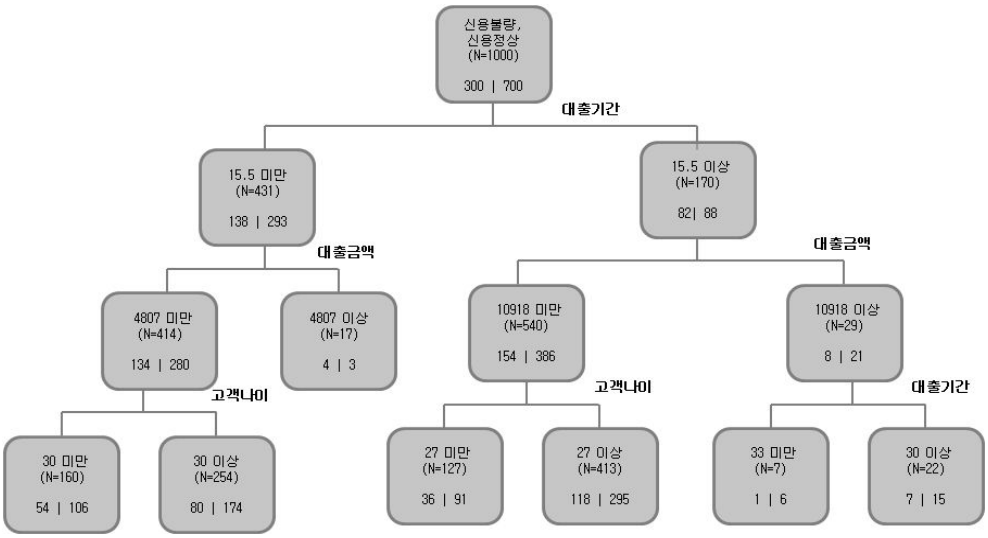
	변수 이름	변수 설명
반응변수	신용비율 상태	1 : 신용정상 2 : 신용불량
설명변수	고객나이	고객의 나이, 단위 : 년
	대출금액	대출한 금액
	보증인수	채무를 대신 이행할 보증인의 수
	대출기간	대출 존속기간, 단위 : 월
	보유계좌 수	현재은행에서 보유하고 있는 계좌의 수
	부채비율	가처분 소득 중 부채가 차지하고 있는 비율, 단위 : %
	거주기간	현주거지에서 거주한 기간, 단위 : 년

4.2.2 일변량 의사결정나무 분석결과

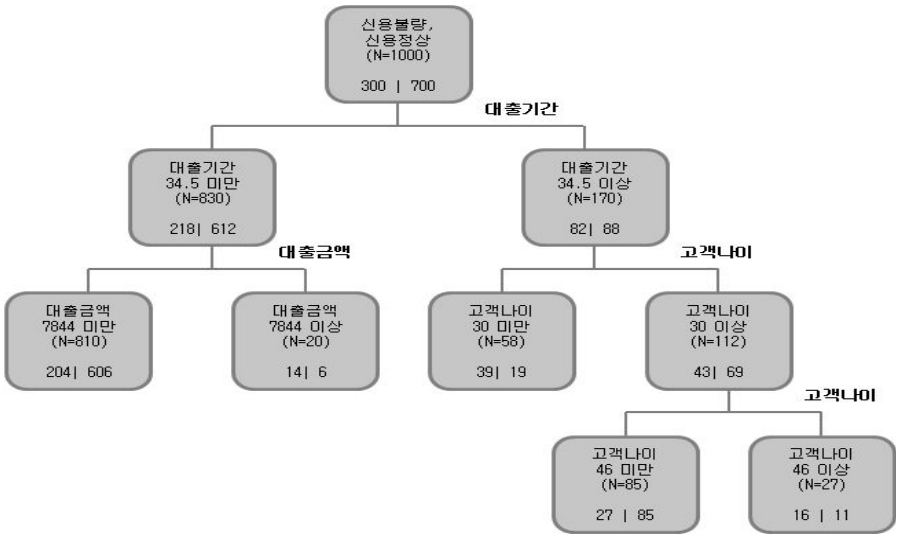
제안된 분리기준에 적용하기 전에, 이항 반응인 '신용비율 상태'를 일변량 반응변수의 알고리즘인 CHAID와 CART, C4.5로 [그림4.4], [그림4.5], [그림4.6]과 같이 분리해보았다.

먼저 이항 반응변수인 '신용비율 상태'를 CHAID 알고리즘으로 의사결정나무를 생성하면 [그림4.4]와 같다. 반응변수 '신용상태'는 설명변수 '대출기간'과 '대출금액', '고객나이'에 의해서 최적의 분리가 이루어진다.

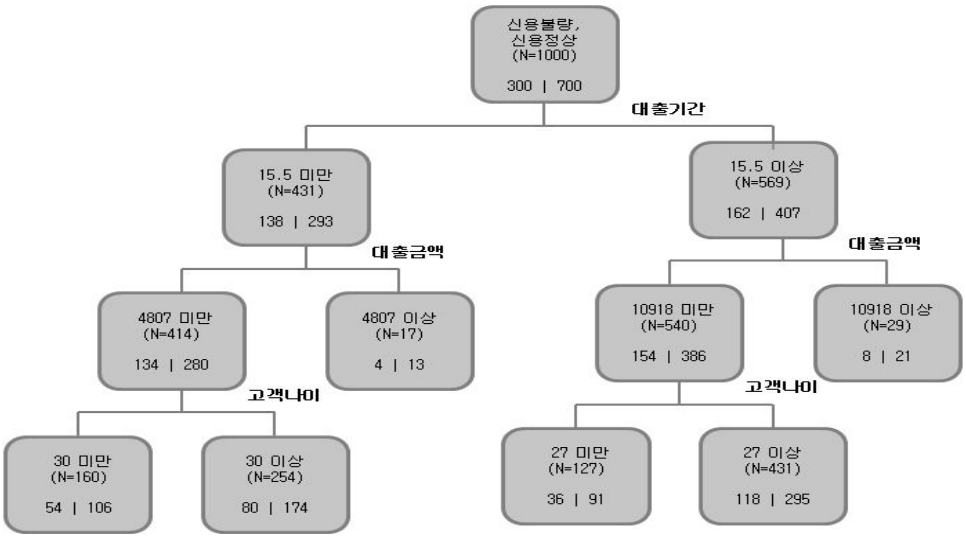
CART 알고리즘으로 나타내면 [그림4.5]와 같다. '대출금액' 대신 '고객나이'로 분리 되었고 '대출금액'에 대한 추가적인 분리가 이루어 지지 않았다. C4.5 알고리즘으로 나타내면 [그림4.6]과 같다. CHAID 알고리즘에서 대출금액이 10918이상인 그룹이 분리되지 않을 것을 제외하면, CHAID 알고리즘으로 분리된 것과 같은 의사결정나무를 가진다.



[그림4.4] CHAID 알고리즘에 의한 '신용비율 상태'의 의사결정나무



[그림4.5] CART 알고리즘에 의한 '신용비율 상태'의 의사결정나무



[그림4.6] C4.5 알고리즘에 의한 '신용비율 상태'의 의사결정나무

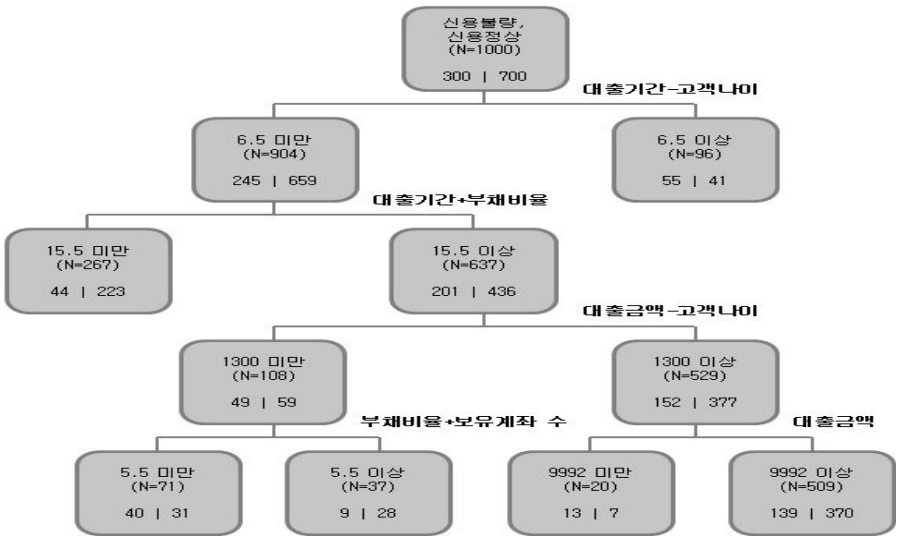
4.2.3 선형결합 의사결정나무 분석결과

제안된 분리기준으로 자료를 분리한 결과, 일대일 선형결합을 이용한 분리기준으로는 '대출기간'과 고객나이'의 선형결합이 가장 최적의 분리를 하는 설명변수로 선택되었으며, 황금분할 선형결합 분리기준으로는 '대출기간과 고객나이'가 선택되었다. 즉, 반응변수 '신용비율상태'는 선형결합을 이용한 분리기준과 황금분할 선형결합을 이용한 분리기준 모두 '대출기간'과 '고객나이' 의 선형결합이 가장 동일한 성질의 개체들의 그룹으로 잘 분리된다고 볼 수 있다.

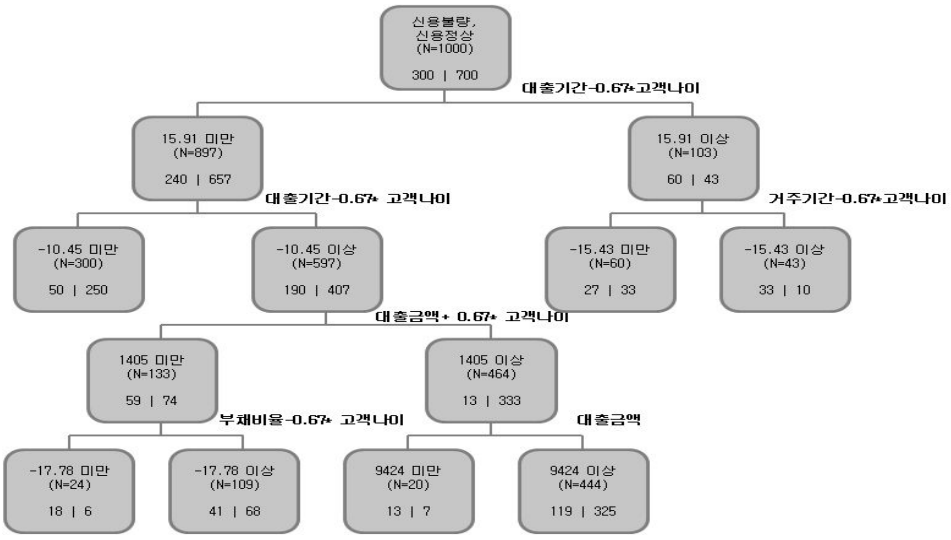
각 분리기준별로 잘 분리하는 설명변수의 순서를 살펴보면, 일대일 선형결합을 이용한 분리기준을 이용하여 반응변수를 잘 분리하는 설명변수의 순서를 나타내면, '대출기간과 고객나이', '대출기간과 부채비율', '대출금액과 고객나이', '부채비율과 보유계좌 수', '대출금액'순이다.

반면, 황금분할 선형결합을 이용하여 반응변수를 잘 분리하는 설명변수의 순서를 나타내면, '대출기간과 고객나이', '대출기간과 부채비율', '대출금액과 고객나이', '부채비율과 고객나이', '대출금액' 순이다.

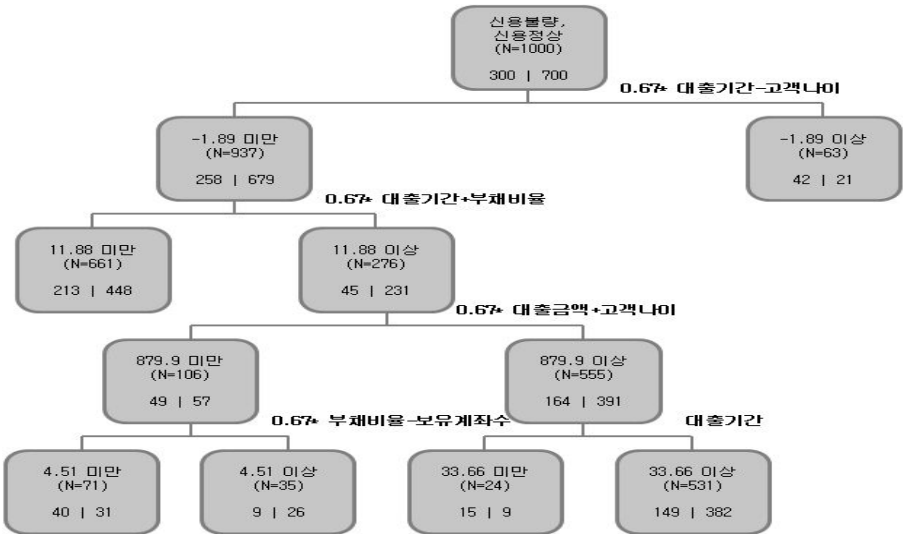
깊이가 4인 이지분리의 의사결정나무로 나타내면, [그림4.7], [그림4.8], [그림4.9]와 같다.



[그림4.7] 일대일 선형결합에 의한 '신용비율 상태'의 의사결정나무



[그림4.8] 황금분할에 의한 '신용비율 상태'의 의사결정나무



[그림4.9] 황금분할에 의한 '신용비율 상태'의 의사결정나무

각 알고리즘의 예측의 정확성 판정기준으로 일변량 및 제안한 분리기준 오분류율을 살펴보면 다음의 [표4.5], [표4.6], [표4.7]과 같다.

[표4.5], [표4.6], [표4.7]에서는 각 알고리즘에 대하여 분리기준별 오분류율과 나무의 크기, 끝마디 수, 나무 깊이 등을 나타내고 있다.

[표4.5] CHAID 알고리즘의 각 분리기준별 오분류율

	오분류율	최종나무크기 (마디수)	끝마디수	나무깊이
single split	0.249	13	7	3단계
$X_1 + X_2$ $X_1 - X_2$	0.249	7	4	3단계
$X_1 + 0.67X_2$ $X_1 - 0.67X_2$	0.249	11	6	3단계
$0.67X_1 + X_2$ $0.67X_1 - X_2$	0.254	7	4	3단계

[표4.6] CART 알고리즘의 각 분리기준별 오분류율

	오분류율	최종나무크기 (마디수)	끝마디수	나무깊이
single split	0.267	9	5	3단계
$X_1 + X_2$ $X_1 - X_2$	0.271	11	6	4단계
$X_1 + 0.67X_2$ $X_1 - 0.67X_2$	0.264	13	7	4단계
$0.67X_1 + X_2$ $0.67X_1 - X_2$	0.261	11	6	4단계

[표4.7] C4.5 알고리즘의 각 분리기준별 오분류율

	오분류율	최종나무크기 (마디수)	끝마디수	나무깊이
single split	0.256	11	6	3단계
$X_1 + X_2$ $X_1 - X_2$	0.221	15	8	3단계
$X_1 + 0.67X_2$ $X_1 - 0.67X_2$	0.279	7	4	3단계
$0.67X_1 + X_2$ $0.67X_1 - X_2$	0.233	13	7	3단계

[표4.5], [표4.6], [표4.7]를 살펴보면, C4.5의 일대일 선형결합 분리기준의 오분류율은 0.221로 가장 작았다. 하지만 가장 큰 나무를 형성하는 단점이 있다. CHAID알고리즘에서는 일변량 분리기준과 제안된 분리기준은 거의 비슷한 오분류율을 보이고 있다. CART, CHAID, C4.5 방법 모두 제안된 분리기준의 오분류율이 대체로 낮으며, 나무크기, 끝마디 수, 나무 깊이 세 가지 측면에서도 가장 작은 나무를 형성한다.

제5장 결론 및 향후 연구방향

의사결정 나무에서 분리 변수를 선택하는데 있어서 일변량 분리기준은 나무크기가 증가하고 가지치기 효과로 인해 나무구조가 간단해지는 문제점이 있으므로 이를 해결하기 위한 노력이 필요하다. 본 논문에서 제안한 선형결합 분리기준은 예측의 정확성을 증가시키는 효과가 있었다.

CHIAD, CART, C4.5 등 기존 알고리즘과 본 연구에서 제안하는 방법의 오분류율을 비교하였다. 모의실험 검증결과와, 실제 독일은행 신용자료에 대해서 구축된 분리기준으로 반응변수를 분리시켜본 결과 일대일 선형결합 분리기준과 황금분할을 이용한 분리기준의 오분류율이 낮은 결과를 보여주었다. 그러나 3가지 제안된 분리기준 중에서 어떠한 방법이 전적으로 우세하다고는 말할 수는 없다.

본 연구에서 제안하는 분리기준을 이용하여 의사결정나무를 형성하는 방법은 기존 일변량 변수만을 이용하여 나무구조를 형성하는 방법에 비해 더 정확한 방법인 것으로 판단된다.

나아가 추후, 의사결정나무의 구조를 형성하는 정지규칙(stoping rule)과 가지치기(pruning)에 관한 알고리즘을 보완하여 보다 향상된 의사결정나무를 구현할 수 있을 것이다. 아울러 이 방법은 독립변수가 연속형인 경우에만 적용할 수 있어 향후 보다 체계적인 연구가 필요할 것으로 판단되며 나무구조의 해석을 어렵게 하는 문제점이 있으므로 이들에 대한 연구는 앞으로의 과제가 될 수 있다.

참 고 문 헌

[1] 김현중. (2006). 의사결정나무에서 순서형 분리변수 선택에 관한 연구, 한국통계학회지, Vol.19, 149-161

[2] 최종후, 한상태, 강현철, 김은석, 김미경, 이성건 (2002). *AnswerTree 3.0을 이용한 데이터마이닝 예측 및 활용*. SPSS아카데미, 서울.

[3] Atkinson, B. & Theneau, T. M. (2009). r*part : Recursive Partitioning. R package. Version 3.1-45*. original at : http://mayoresearch.mayo.edu/mayo/research/biostat/splusfunctions.cfm

[4] Bartlett, M. S. (1938). *Further Aspects of the Theory of Multiple Regression,* in Proceedings of the Cambridge Philosophical Society, Vol.34, 33-40.

[5] Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. (1984). *Classification and Regression Trees*. Chapman & Hall, New York.

[6] Kass, G. V. (1980). An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Applied Statistics,* Vol.29, 119-127.

[7] Kim, H. & Loh, W. Y. (2001). Classification trees with unbiased multiway splits, *Journal of the American Statistical Association*, Vol.96, 589-604.

[8] Kim, H. & Loh, W. Y. (2003). Classification Trees with Bivariate Linear Discriminant Node Models. *Journal of the Computational and Graphical Statistics,* Vol.12, 512-530.

[9] Liu, W. Z. & White, A. P. (1994). The importance of attribute-selection measures in decision tree induction, *Machine Learning*, Vol.15, 24-41.

[10] Loh, W. Y. & Shih, Y. S. (1997). Split Selection Methods for

Classification Trees. *Statistica Sinica*, Vol.7, 815-840.

[11] Loh, W. Y. & Vanichsetakul. N. (1988). Tree Structured Classification Via Generalized Discriminant Analysis. *Journal of the American Statistical Association*, Vol.83, 715-725.

[12] Simpson, E. H. (1949). Measurement of Diversity. *Nature,* Vol.163, 688.

[13] The R Development Core Team. (2009). *R Version 2.9.1. User's Manual.*

[14] Quinlan, J. R. (1993). *C4.5 Programs for Machine Learning*, Morgan Kaufmann, San Mateo.

ABSTRACT

A Study on Decision Tree using Linear Combination Splits

Kyung-hye Lee
Department of Statistics
The Graduate School
Sungshin Women's University

A classification tree is a rule for predicting the class of an object from the values for its predictor variables. The common goal in CART, CHAID, C4.5 and QUEST is to obtain such that in each terminal node is quite pure and simple tree. Occasionally this cannot be achieved with standard algorithms can produce large tree structure because they use only single splits.

This thesis introduce a classification tree split criterion that can improve class prediction. We accomplish this by bivariate linear combination splits. Our splits are general linear combination split and using golden section splits. This splits are carried out misclassification costs of the models.

Furthermore, some simulation and real data experiments are performed to demonstrate the performance of the proposed approach. Our split criterion has better prediction power and lower misclassification rate than CART single algorithms.

감사의 글

새로운 마음으로 대학원에 들어온 지 얼마 지나지 않은 것 같은데 어느덧 졸업을 앞두고 대학원 석사논문을 마무리하게 되었습니다. 처음 의지와는 달리 나약한 모습을 보이며 조금은 버거워 했던 제 자신이 새삼 부끄럽습니다. 스스로 배워나가며 훈련 되어진 2년이란 시간이 제 인생에 있어 소중한 부분이란 생각이듭니다.

논문을 완성할 수 있도록 부족한 제게 많은 가르침을 주시고 세심한 관심과 지도를 베풀어 주신 이성건 지도교수님께 존경과 감사의 뜻을 전합니다. 항상 따뜻한 격려를 해주신 이해용 교수님, 너그러운 마음으로 많은 배려를 해주신 송일성 교수님, 언제나 웃음을 잃지 않도록 해주신 이우선 교수님, 또한 사랑과 격려로 따끔한 충고와 공부 이외 것에서도 많은 가르침을 주신 이종협 교수님께 가슴깊이 감사를 드립니다.

바쁜 와중에도 후배에게 관심을 가져준 향선언니, 애란언니, 가영언니, 힘들고 어려울 때 마다 저에게 많은 조언과 용기를 준 영은언니, 희라언니, 주현언니, 언제나 저에게 의지 할 수 있도록 많은 힘이 되어주고 세심한 부분까지 신경써준 희원언니, 인경언니에게도 감사의 마음을 전합니다. 대학원생활을 함께 해준 정윤, 보미, 하얀, 영아, 명희, 따뜻한 격려와 많은 도움을 주신 광렬오빠에게도 고마움을 전합니다. 언제나 한결같은 마음으로 변하지 않는 나의 오래된 친구들 주영, 주희, 보아, 수연, 혜윰, 항상 저를 챙겨주고 걱정해주는 둘도 없는 친구인 준희에게도 고마움을 전합니다.

마지막으로 가장 가까이에서 저를 지켜 봐주며 못난 투정에도 언제나 사랑으로 대해준 소중한 사람 민식에게도 사랑과 고마움을 전합니다. 묵묵히 지원해주시고 항상 믿어주시는 사랑하는 부모님, 동생 같은 언니의 어리광도 잘 받아주는 동생 신혜, 설희에게도 사랑과 고마움을 전합니다.

그 외에도 많은 분들께 감사의 뜻을 전합니다.

부모님, 교수님, 그리고 저의 소중한 사람들에게 부끄럽지 않은 사람이 되도록 항상 노력하겠습니다.