이 성 건 교수지도

석사학위청구논문

# 상사성의 측도를 이용한 다변량 의사결정나무에 관한 연구

2008

성신여자대학교 대학원

통 계 학 과

이 주 현

# 상사성의 측도를 이용한
# 다변량 의사결정나무에 관한 연구

이 성 건 교수지도

이 논문을 석사학위논문으로 제출함

2008년 5월

성신여자대학교 대학원
통 계 학 과
이 주 현

# 인 준 서

**이주현의 석사학위 논문으로 인준함.**

**심사위원 _____ 인**

**심사위원 _____ 인**

**심사위원 _____ 인**

**성신여자대학교 대학원**

# 논 문 개 요

연구대상이 되는 개체들을 어떤 관점에서 분류하거나 예측하는 것은 통계적 연구에서 가장 본원적인 목표들 중의 하나이다. 방대한 양의 자료들이 쏟아지는 요즘, 대용량의 자료들을 관심이 되는 집단으로 분류하거나 예측하고자 하는 방법론들이 제안되고 있다. 그 중 의사결정나무분석(decision tree analysis)은 연구자가 분석과정을 쉽게 이해할 수 있고, 그 해석이 용이하여 흔히 이용되고 있다.

적절한 의사결정나무를 구축하기 위해서는 주어진 자료에 적합한 의사결정나무 알고리즘을 적용하여야 한다. 그러나 대표적인 의사결정나무 알고리즘들은 반응변수가 일변량인 경우에 주로 사용되어, 다변량 반응변수의 다양한 형태에 대한 의사결정나무분석을 수행하는데 한계가 있다.

본 논문은 범주형과 연속형 변수가 혼합된 다변량 반응변수인 경우에 대한 새로운 알고리즘을 제안한다. 제안된 알고리즘을 R 소프트웨어를 이용하여 모의실험한 후, 이 결과를 바탕으로 다변량 의사결정나무가 보다 개선되었는지 비교 연구한다. 또한 새로운 알고리즘이 실제자료에도 잘 적용되고 있음을 보여준다.

# 목 차

# 제1장 서 론

정보화 사회로 진입하면서, 우리는 다양한 정보를 접하면서 살아가고 있다. 이러한 정보들은 여러 가지 형태의 자료들로 수집되고 있다. 이렇게 수집된 많은 양의 자료들을 의미 있는 결과로 도출하기 위해 자료를 분석하는 일은 효과적이고 과학적인 의사결정을 하기 위해서 필수적이다. 의사결정나무분석은 자료를 분류하거나 예측하는 데 나무가지 모양으로 도표화하여 수행하는 분석방법으로, 분석과정의 해석이 쉬워 유용하게 사용되고 있다.

의사결정나무분석은 관심이 되는 집단을 분리기준에 의해서 몇 개의 소집단으로 분류하거나 예측할 때 쓰이고 있다. 그러나 대부분의 의사결정나무는 반응변수가 일변량 또는 동일한 형태의 다변량 반응인 경우에만 적용할 수 있다. 따라서 다양한 종류의 자료들로 넘쳐나는 요즘, 기존의 분리기준만으로는 다양한 형태의 자료를 분석하는 데에 한계가 있다.

본 연구에서는 혼합 반응에 대한 다변량 의사결정나무를 구축하고자 한다. 즉, 연속형과 범주형 변수가 혼합된 다변량 반응변수에 대한 방법론을 제안한다. 첫째로, 범주형 변수를 가변수화한 후 마할라노비스 거리(Mahalanobis distance)를 이용한 분리기준에 대한 의사결정나무를 제안한다. 두 번째로, Gower(1971)가 제시한 상사성 계수(similarity coefficient)를 이용한 의사결정나무를 제안한다.

본 논문의 2장에서는 기존의 알고리즘에 대한 의사결정나무를 소개한다. 3장에서는 마할라노비스 거리와 Gower의 분리기준을 이용한 의사결정나무를 자세히 설명한다. 나아가 4장에서는 구축된 분리기준이 얼마나 잘 분리되는가를 알아보기 위해, R 소프트웨어를 이용하여 분리기준에 대한 모의실

험을 수행한다. 또한, 이 분리기준을 이용하여 실제자료에 적용하고, 끝으로 5장에서는 결론 및 향후 연구방향에 대해서 논의한다.

# 제2장 의사결정나무

의사결정나무분석을 수행하기 위해서는 반응변수들을 잘 분류할 수 있는 분리기준이 필요하다. 일변량 반응변수일 때 대표적으로 사용되는 알고리즘으로는 CHAID와 CART 등이 있다. 다변량 반응변수일 때 사용되는 분리기준으로는 마할라노비스 거리(Mahalanobis distance)와 엔트로피 지수(Generalized Entropy Index)등이 있다.

## 2.1 일변량 의사결정나무

일변량 의사결정나무분석을 수행하기 위한 다양한 분리기준, 정지규칙, 가지치기 방법들이 제안되고 있는데, 그 중에서도 일변량 의사결정나무를 형성하는 알고리즘의 분리기준에 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 알고리즘이다. 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

$$\chi^2 = \sum_{i,j} \frac{(f_{ij} - e_{ij})^2}{e_{ij}} \ , \tag{1}$$

와 같이 정의되고, 우도비 카이제곱 통계량은

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

로 정의된다. 이 때 두 통계량의 자유도(degree of freedom)는 반응변수의 범주 수를 $r$이라 하고, 설명변수의 범주 수를 $c$라 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.

$$e_{ij} = \frac{f_{i.} \times f_{.j}}{f_{..}} \ . \tag{3}$$

식(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 설명변

수가 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 보여진다. 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 | $r-1$ | $SST=\sum_{i=1}^{r} n_i(\overline{y_i}-\overline{y})^2$ | $MST=SST/(r-1)$ | $F=\frac{MST}{MSE}$ |
| 오차 | $n-r$ | $SSE=\sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij}-\overline{y_i})^2$ | $MSE=SSE/(n-r)$ | |
| 전체 | $n-1$ | $TSS=\sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij}-\overline{y})^2$ | | |

$F$ 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 형성

하는 최적분리를 한다.

### 2.1.2 CART 알고리즘

CART(Classification and Regression Trees; Breiman et al., 1984) 알고리즘은 범주형 반응변수 또는 연속형 반응변수의 분류 및 예측을 수행하는 알고리즘이다. 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 $i$번째 범주로부터 추출되었고, 그 개체를 반응변수의 $j$번째 범주에 속한다고 오분류(misclassification)할 확률은 $P(i)P(j)$가 된다. 여기에서 $P(i)$는 각 마디에서 한 개체가 반응변수의 $i$번째 범주에 속할 확률이다. 이러한 오분류 확률을 모두 더하여,

$$G=\sum_{j=1}^{c}\sum_{i\neq j}P(i)P(j)\ , \tag{4}$$

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

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

$$G=\sum_{j=1}^{c}P(j)(1-P(j))=1-\sum_{j=1}^{c}P(j)^2=1-\sum_{j=1}^{c}(n_j/n)^2, \qquad (5)$$

와 같이 표현될 수 있다. 여기에서 $n$은 그 마디에 포함되어 있는 개체수를 말하고, $n_i$는 반응변수의 $i$번째 범주에 속하는 개체수를 말한다. 지니 지수는 $n$개의 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.

$$G=2P(1)P(2)=2(\frac{n_1}{n})(\frac{n_2}{n}). \qquad (6)$$

식(6)은 카이제곱 통계량을 사용하는 것과 같은 결과를 갖는다.

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

$$\Delta G=G-\frac{n_L}{n}G_L-\frac{n_R}{n}G_R\ , \qquad (7)$$

여기서 $n$은 부모마디의 관측치 수를 말하고, $n_R$과 $n_L$는 각각 자식마디의 관측치 수를 의미한다. 즉, 자식마디로 분리되었을 때의 불순도가 가장 작도록

자식마디를 형성하는 것이며, 이는 다음과 같은 자식마디에서의 불순도의 가중합을 최소화하는 것과 동일하다.

$$P(L)\,G_L + P(R)\,G_R = \frac{n_L}{n}\,G_L + \frac{n_R}{n}\,G_R \ . \tag{8}$$

## 2.2 다변량 의사결정나무

다변량 반응변수일 경우, 의사결정나무분석을 수행하기 위해서 사용되는 분리기준으로는 마할라노비스 거리(Mahalanobis distance)와 엔트로피 지수(Generalized Entropy Index)등이 있다.

### 2.2.1 마할라노비스 거리(Mahalanobis distance)

Segal(1992)에 의해 제안된 이 방법은 반응변수의 형태가 연속형인 다변량 의사결정나무를 형성한다. $T\times 1$인 $y'_i = (y_{i1}, y_{i2}, \ldots, y_{iT})$ 반응변수로 어떤 한 마디를 $g$ 라고 할 때, $g$ 마디내의 평균행렬을 $\mu(g)$, 알려지지 않은 모수 $\theta$에 의한 반응변수의 공분산행렬을 $V(\theta, g)$라 한다.

이 때, 불순도(impurity)를 측정하는 값으로

$$SS(g) = \sum_{i \in g} (y_i - \mu(g))' V(\theta, g)^{-1} (y_i - \mu(g)) \ , \tag{9}$$

가 이용된다. 이것은

$$\phi(s,g) = SS(g) - SS(g_L) - SS(g_R) \ , \tag{10}$$

로 표현되며, 부모마디 $g$의 $SS(g)$, 분리된 왼쪽 자식마디의 $SS(g_L)$, 오른쪽 자식마디의 $SS(g_R)$으로 정의된다. 분리되기 전, 부모마디 $g$의 $SS(g)$에서 자식마디의 $SS(g_L)$와 $SS(g_R)$을 빼 준 값을 $\phi(s,g)$라 할 때, $\phi(s,g)$이 최대가 되는 설명변수의 최적분리에 의해서 자식마디를 형성한다. 이는 $SS(g_L)$와 $SS(g_R)$의 합이 최소가 되는 설명변수의 최적분리에 의해 자식마디를 형성하는 것과 동일하다.

### 2.2.2 엔트로피 지수(Generalized Entropy Index)

Zhang(1998)에 의해 제안된 엔트로피 지수(Generalized Entropy Index)는 다변량 이항 반응변수에 대한 동질성(homogeneity)을 측정하기 위한 값이다.

먼저, 로그 선형 모형을 사용하는 일변량 엔트로피 기준을 일반화하고, 다변량 반응변수 $Y$의 결합 분포가 1차 항과 다른 성분들의 곱인 2차 항의 합으로 이루어짐을 가정한다. 다시 말해서, $Y$의 결합 확률 분포는 다음과 같다.

$$F(y:\psi,\theta) = \exp(\psi' y + \theta w - A(\psi,\theta)). \tag{11}$$

여기서, $w = \sum_{i<j} y_i y_j$, 을 나타낸다.

어떤 한 마디를 $g$ 라고 할 때, 마디 $g$의 동질성은 식(11)의 분포로부터 유

도된 로그 우도의 최대값으로 표현되며 아래와 같다.

$$h(g) = \sum \hat{\Psi} y_i + \hat{\theta} w_i - A(\hat{\Psi}, \hat{\theta}). \tag{12}$$

여기서, $\hat{\Psi}$와 $\hat{\theta}$는 각각 $\Psi$와 $\theta$의 최대우도추정치이다. 마디 $g$의 불순도는 $-h(g)$로 나타낼 수 있다.

지금까지 2장에서 소개된 방법들은 일변량 또는 동일한 형태를 가지는 반응들의 다변량 의사결정나무를 생성하는데 유용하게 쓰일 수 있을 것이다.

# 제3장 혼합반응에 대한 다변량 의사결정나무

여러 알고리즘이 의사결정나무분석을 수행하는데 사용되고 있으나, 반응변수의 형태가 범주형과 연속형으로 혼합된 경우의 의사결정나무를 형성하는 데는 한계가 있다. 본 연구에서는 범주형과 연속형의 혼합된 반응변수들의 의사결정나무를 형성하기 위해서, 첫 번째 방법으로 범주형 반응변수를 가변수화한 후, Segal(1992)의 마할라노비스 거리(Mahalanobis distance)를 이용하고, 두 번째 방법으로 Gower(1971)의 상사성 계수(similarity coefficient)를 이용하여 다변량 의사결정나무를 생성한다.

## 3.1 마할라노비스 거리를 이용한 다변량 의사결정나무

일반적으로 반응변수가 연속형일 경우, 마할라노비스 거리를 이용하여 다변량 의사결정나무를 형성하게 된다. 혼합반응에 대한 다변량 의사결정나무를 형성하는 경우, 마할라노비스 거리를 이용하기 위해서, 범주형 반응변수를 0과 1의 값을 가지는 가변수로 변환하여 연속형 반응변수와 함께 마할라노비스 거리를 계산한다.

어떤 한 마디를 $g$ 라고 할 때, 반응변수행렬 $T\times 1$인 $y'_i=(y_{i1}, y_{i2}, \cdots, y_{iT})$, $g$ 마디내의 평균행렬을 $\mu(g)$, 알려지지 않은 모수 $\theta$에 의한 반응변수의 공

분산행렬을 $V(\theta, g)$라 한다.

이 때, 불순도(impurity)를 측정하는 값으로

$$SS(g) = \sum_{i \in g} (y_i - \mu(g))' V(\theta, g)^{-1} (y_i - \mu(g)), \tag{13}$$

와 같이 사용된다. 이것은

$$\phi(s, g) = SS(g) - SS(g_L) - SS(g_R), \tag{14}$$

로 표현되며, 부모마디 $g$의 $SS(g)$, 분리된 왼쪽 자식마디의 $SS(g_L)$, 오른쪽 자식마디의 $SS(g_R)$으로 정의된다. 분리되기 전, 부모마디 $g$의 $SS(g)$에서 자식마디의 $SS(g_L)$와 $SS(g_R)$을 빼 준 값을 $\phi(s, g)$라 할 때, $\phi(s, g)$이 최대가 되는 설명변수에 의해 자식마디를 형성하도록 하는 분리기준이다. $\phi(s, g)$값이 최대가 되는 설명변수는 동질한 개체들의 자식마디로 분리된다고 볼 수 있으므로, 이 값이 최대가 되는 설명변수를 찾는 것이 목적이다. 그러나 위 함수는 각 마디의 모수 추정치 $\hat{\theta}$, $\widehat{\theta_L}$, $\widehat{\theta_R}$가 각기 다르므로 $g$, $g_L$, $g_R$의 공분산행렬이 다르게 된다. 또한 식(14)의 $\phi(s, g)$가 음수 값이 될 수 있기 때문에, $\phi(s, g)$값을 최대로 하는 설명변수가 반드시 동질성(homogeneity)를 증진시키지 않는다. 그러므로 $\phi(s,g) \geq 0$로 유지하기 위해, 부모마디 $g$의 공분산행렬과 분리된 자식마디 $g_L$, $g_R$의 공분산행렬을 동일하게 다음의 식과 같이 가정한다.

$$V(\theta, g) = V(\theta_L, g_L) = V(\theta_R, g_R). \tag{15}$$

## 3.2 Gower의 상사성 계수를 이용한 다변량 의사결정나무

Gower(1971)의 상사성 계수(similarity coefficient)는 혼합반응에 이용되는 가장 대표적인 측정값이다. 계산과정은 다음과 같다.

어떤 한 변수 $k$의 두 개체 $i$와 $j$를 비교한다. 일반적으로 개체 중에 결측치가 있다면 비교를 할 수 없다. 이항변수일 때를 제외하고 $\delta_{ijk}$은 개체들의 비교가 가능한 지를 나타내는 값으로, 변수 $k$의 두 개체 $i$와 $j$의 비교가 가능하면 1의 값을, 비교가 가능하지 않다면 0의 값을 나타낸다. 개체 $i$와 $j$사이의 상사성은 다음과 같이 모든 가능한 개체들 비교의 평균점으로 정의된다.

$$S_{ij} = \sum_{k=1}^{v} s_{ijk} / \sum_{k=1}^{v} \delta_{ijk} \ . \tag{16}$$

가중 평균의 형식인,

$$\widetilde{S_{ij}} = \sum_{k=1}^{v} s_{ijk}\, \delta_{ijk} / \sum_{k=1}^{v} \delta_{ijk} \ , \tag{17}$$

으로도 나타낼 수 있다. 만일 모든 변수에서 $\delta_{ijk} = 0$이라면, $S_{ij}$는 정의될 수 없다. $s_{ijk}$는 변수의 형태가 이항(dichotomous), 질적(qualitative), 양적(quantitative) 변수에 따라서 계산하는 방법이 다르다. 계산방법은 다음과 같다.

▪ 이항(dichotomous)변수일 경우

변수 $k$에서 값이 존재하면 +, 존재하지 않는다면 −으로 나타낸다. $s_{ijk}$와 $\delta_{ijk}$의 값은 아래의 [표3.1]과 같은 점수를 가진다.

[표3.1]

| | 변수 k의 값 | | | |
|---|---|---|---|---|
| 개체 i | + | + | − | − |
| 개체 j | + | − | + | − |
| $s_{ijk}$ | 1 | 0 | 0 | 0 |
| $\delta_{ijk}$ | 1 | 1 | 1 | 0 |

$s_{ijk}$ : 변수의 형태에 따른 점수

$\delta_{ijk}$ : 개체들의 비교 가능 값

▪ 질적(qualitative) 변수일 경우

변수 $k$에서 개체 $i$와 $j$의 값이 같으면, $s_{ijk}=1$이고 개체 $i$와 $j$의 값이 다르면, $s_{ijk}=0$이 된다.

▪ 양적(quantitative) 변수일 경우

$$s_{ijk}=1-|x_i-x_j|/R_k, \tag{18}$$

로 계산된다. $R_k$는 변수 $k$의 범위이다. $x_i=x_j$이면, $s_{ijk}=1$이다. $x_i$과 $x_j$가 범위의 양 끝 값을 가진다면 $s_{ijk}$는 최소값을 가진다.

변수의 형태에 따라 $s_{ijk}$가 계산되면, 식(16) 또는 식(17)에 대입하여 부모마디 $g$의 상사성 $S_{ij}$와 자식마디 $g_L$, $g_R$의 상사성 $S_{ijL}$과 $S_{ijR}$을 각각 산출한다. 이렇게 산출된 값들은 각 그룹들의 동일한 성질을 가지는 상사성을 나

타냄으로써, 부모마디 $S_{ij}$에서 자식마디 $S_{ijL}$과 $S_{ijR}$을 뺀 값이 최소가 되는 설명변수의 최적분리를 수행하게 된다. 식으로 표현하면 다음과 같다.

$$\phi(s,g)=S_{ij}-S_{ijL}-S_{ijR}\ . \tag{19}$$

# 제4장 모의실험 및 적용

3장에서 제안한 분리기준이 혼합된 다변량 반응변수들을 얼마나 잘 분리하는 가를 모의실험을 통해 확인해보고, 실제자료에서 적용해본다.

## 4.1 분리기준에 대한 모의실험

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

**Ⅰ.(분리변수 생성)** 먼저 베르누이 분포 $B(1, 0.5)$로부터 개체가 300개인 임의의 설명변수($X_1$, $X_2$)를 독립적으로 생성한다.

**Ⅱ.(반응변수 생성)** 설명변수 $X_1$의 값에 따라 $X_1 = 0$ 이면, 평균이 $(0, 0)$, 분산이 $(1, 1)$이고 상관계수가 $\rho = 0.5$인 이변량 정규분포 $MN_1(\underline{\mu}, \sum)$, $\underline{\mu} = (0,0)$, $\sum = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$로부터 개체 1개를 생성하고, $X_1 = 1$ 이면 평균이 $(\varepsilon, \varepsilon)$, 분산이 $(1, 1)$이고 상관계수가 $\rho = 0.5$인 이변량 정규분포 $MN_2(\underline{\mu}, \sum)$, $\underline{\mu} = (\varepsilon, \varepsilon)$, $\sum = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$로부터 개체 1개를 생성시킨다. 이런 방법으로 300번 반복 수행한다. 개체수가 300개인 이변량 정규분포의 표본이 생성되면, 이변량 중 하나를 반응변수 $Y_1$, 나머지 하나를 반응변수 $Y_2$라 한다. 설명변수

$X_1$의 값에 따라 두 그룹의 이변량 정규분포의 표본이 생성되므로, 우수한 의사결정나무는 참 분리기준 변수(true split variable)인 $X_1$의 선택확률이 높게 나타날 것이다.

**Ⅲ.(평균벡터)** $X_1 = 0$일 때 생성되는 이변량 정규분포 $MN_1(\underline{\mu}, \sum)$, $\underline{\mu} = (0,0)$, $\sum = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$는 고정시키고, $X_1 = 1$의 값에 따라 생성되는 이변량 정규분포의 평균차이를 0에서부터 0.1만큼 증가시켜 1까지 변화시킨다.

**Ⅳ.(범주화)** 생성된 연속형 반응변수($Y_1$, $Y_2$)에서 반응변수 $Y_2$은 범주 수(2~9범주)에 따라서, 정규분포를 각 범주 수(2~9범주)로 등분한 기준점으로 나누어 범주형 변수 $C$로 변환시킨다. 특히 마할라노비스 거리를 이용한 분리기준을 적용시키기 위해서는 범주형 변수 $C$를 각 범주 수(2~9범주)에 따라 가변수로 변환시킨다.

**Ⅴ.(분리)** 생성된 연속형, 범주형 반응변수를 설명변수 $X_1$, $X_2$에 의해 마할라노비스 거리와 Gower의 상사성 계수를 이용한 분리기준을 적용시킨다. 그리하여 참 분리기준 변수 $X_1$의 선택확률 계산한다.

**Ⅵ.(반복)** 10,000번의 모의실험을 반복한다.

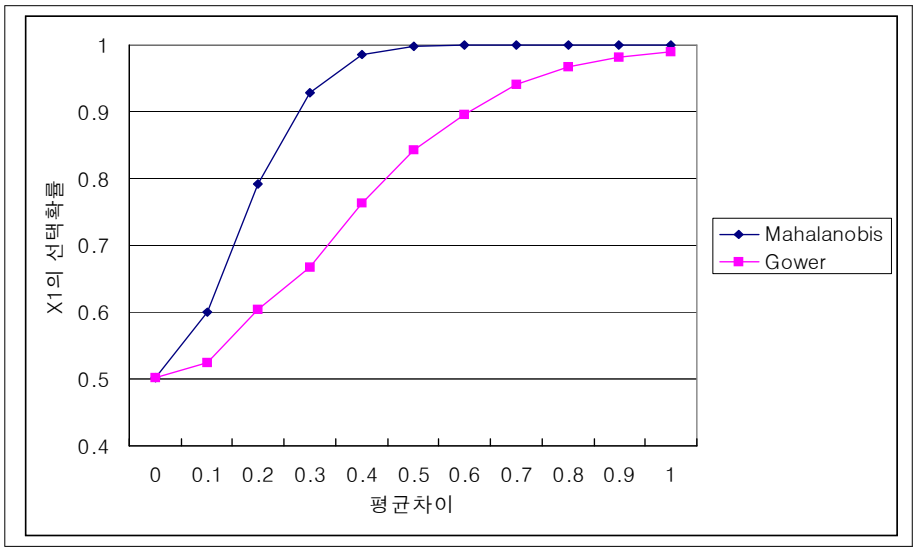
10,000번의 반복수행 결과, 설명변수($X_1$, $X_2$)의 선택확률을 확인할 수 있다. [표4.1]는 연속형 반응변수 $Y_1$와 이항 반응변수 $C$일 때의 제안된 분리기준들에 대한 참 분리변수 $X_1$의 선택확률을 나타낸 것이다.

[표4.1] 연속형 반응변수 $Y_1$와 범주가 이항(binary)인 범주형 반응변수 $C$일 때의 제안된 분리기준들의 설명변수 $X_1$의 선택확률

| 평균 차이 | Mahalanobis | Gower |
|---|---|---|
| 0 | 0.5011 | 0.5013 |
| 0.1 | 0.6003 | 0.5246 |
| 0.2 | 0.7915 | 0.6032 |
| 0.3 | 0.9283 | 0.6681 |
| 0.4 | 0.9863 | 0.7629 |
| 0.5 | 0.9979 | 0.8422 |
| 0.6 | 0.9995 | 0.8964 |
| 0.7 | 0.9999 | 0.9417 |
| 0.8 | 1 | 0.9683 |
| 0.9 | 1 | 0.9824 |
| 1 | 1 | 0.9905 |

분리기준별로 비교했을 때, 제안된 분리기준 모두 이변량 정규분포의 평균의 차이가 클수록 설명변수 $X_1$의 선택확률이 높음을 알 수 있다. 그러나 마할라노비스 거리를 이용한 분리기준이 Gower의 분리기준보다 잘 분리하는 것을 확인할 수 있다. 다음과 같이 [그림4.1]에서 쉽게 비교할 수 있다.

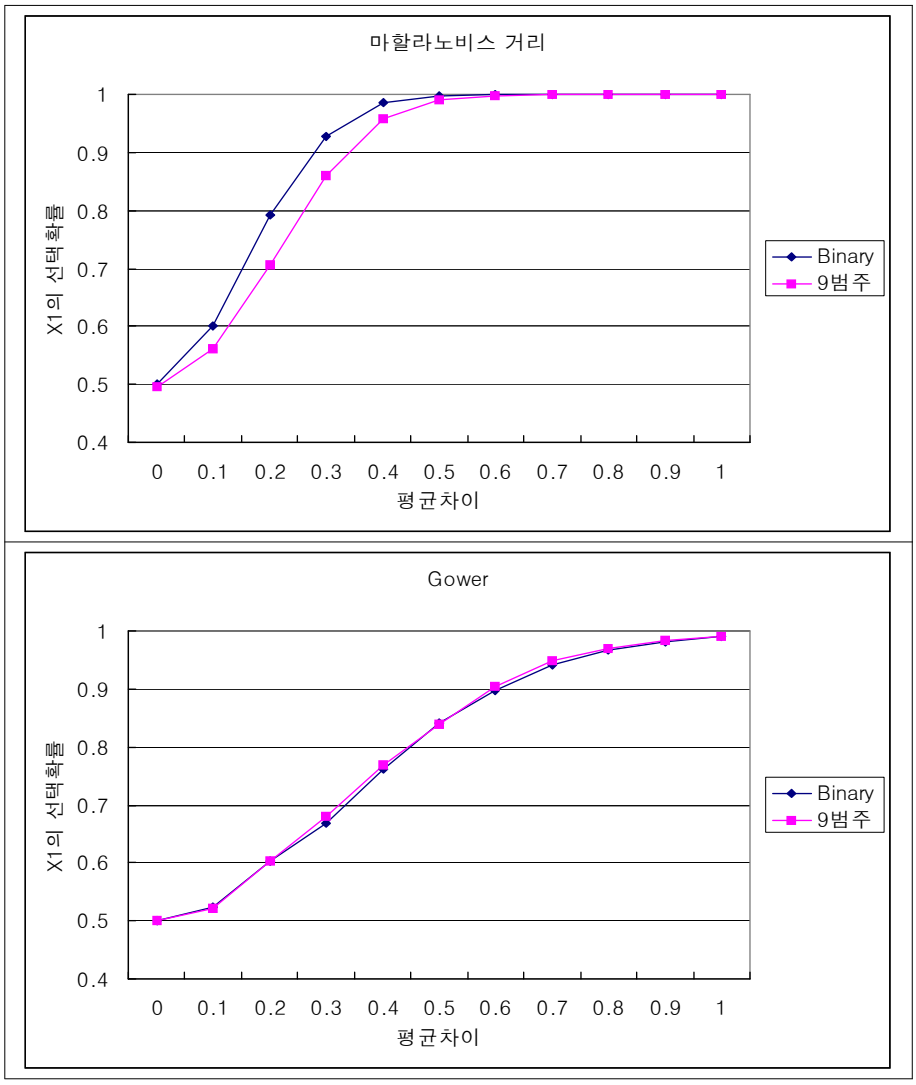
[그림4.1] 마할라노비스 거리와 Gower의 분리기준의 각 평균차이별 비교




[그림4.1]을 보면 두 분리기준 모두 평균차이가 클수록 $X_1$의 선택확률이 높으나, 마할라노비스 거리를 이용한 분리기준이 Gower의 분리기준보다 더 잘 반응변수를 잘 분리함을 알 수 있다.

추가적으로 연속형 반응변수 $Y_1$와 범주가 3~9개인 범주형 반응변수 $C$일 때의 마할라노비스 거리와 Gower의 상사성을 이용한 분리기준으로 $X_1$의 선택확률을 보면, 위의 [표4.1]과 비슷한 양상을 보인다. 그러나 마할라노비스 거리를 이용한 분리기준은 범주의 수가 증가할수록 $X_1$의 선택확률이 낮아짐을 [그림4.2]를 통해 확인할 수 있다.

[그림4.2] 범주수가 2와 9인 각 분리기준별 $X_1$의 선택확률




[그림4.2]를 통해 Gower의 분리기준은 범주수에 관계없이 분리를 하는 반면, 마할라노비스 거리를 이용한 분리기준은 범주수의 증가함에 따라 분리를 잘 수행하지 못함을 알 수 있다. 연속형 반응변수 $Y_1$와 범주가 3~9개인 범주형 반응변수 $C$일 때의 $X_1$의 선택확률은 [표4.3]~[표4.9]을 통해 확인할

수 있다.

[표4.3] 연속형 $Y_1$와 범주수가 3개인 $C$인 반응변수를 가지는 경우, 분리기준별 설명변수 $X_1$의 선택확률

| 평균 차이 | Mahalanobis | Gower |
|---|---|---|
| 0 | 0.5095 | 0.5015 |
| 0.1 | 0.5822 | 0.5312 |
| 0.2 | 0.7799 | 0.6020 |
| 0.3 | 0.9208 | 0.6916 |
| 0.4 | 0.9801 | 0.8067 |
| 0.5 | 0.9976 | 0.8756 |
| 0.6 | 0.9998 | 0.9334 |
| 0.7 | 1 | 0.9710 |
| 0.8 | 1 | 0.9900 |
| 0.9 | 1 | 0.9970 |
| 1 | 1 | 0.9989 |

[표4.4] 연속형 $Y_1$와 범주수가 4개인 $C$인 반응변수를 가지는 경우, 분리기준별 설명변수 $X_1$의 선택확률

| 평균 차이 | Mahalanobis | Gower |
|---|---|---|
| 0 | 0.5020 | 0.5017 |
| 0.1 | 0.5872 | 0.5228 |
| 0.2 | 0.7595 | 0.6008 |
| 0.3 | 0.9028 | 0.6934 |
| 0.4 | 0.9782 | 0.7903 |
| 0.5 | 0.9967 | 0.8628 |
| 0.6 | 0.9998 | 0.9227 |
| 0.7 | 1 | 0.9610 |
| 0.8 | 1 | 0.9833 |
| 0.9 | 1 | 0.9916 |
| 1 | 1 | 0.9986 |

[표4.5] 연속형 $Y_1$와 범주수가 5개인 $C$인 반응변수를 가지는 경우, 분리기준별 설명변수 $X_1$의 선택확률

| 평균 차이 | Mahalanobis | Gower |
|---|---|---|
| 0 | 0.4983 | 0.5014 |
| 0.1 | 0.5692 | 0.5385 |
| 0.2 | 0.7511 | 0.6014 |
| 0.3 | 0.8945 | 0.6918 |
| 0.4 | 0.9688 | 0.7787 |
| 0.5 | 0.9948 | 0.8571 |
| 0.6 | 0.9995 | 0.9123 |
| 0.7 | 0.9999 | 0.9550 |
| 0.8 | 1 | 0.9836 |
| 0.9 | 1 | 0.9917 |
| 1 | 1 | 0.9956 |

[표4.6] 연속형 $Y_1$와 범주수가 6개인 $C$인 반응변수를 가지는 경우, 분리기준별 설명변수 $X_1$의 선택확률

| 평균 차이 | Mahalanobis | Gower |
|---|---|---|
| 0 | 0.5098 | 0.5034 |
| 0.1 | 0.5700 | 0.5167 |
| 0.2 | 0.7317 | 0.6012 |
| 0.3 | 0.8869 | 0.6957 |
| 0.4 | 0.9679 | 0.7731 |
| 0.5 | 0.9937 | 0.8524 |
| 0.6 | 0.9986 | 0.9147 |
| 0.7 | 0.9999 | 0.9565 |
| 0.8 | 1 | 0.9781 |
| 0.9 | 1 | 0.9913 |
| 1 | 1 | 0.9932 |

[표4.7] 연속형 $Y_1$와 범주수가 7개인 $C$인 반응변수를 가지는 경우,

분리기준별 설명변수 $X_1$의 선택확률

| 평균 차이 | Mahalanobis | Gower |
|---|---|---|
| 0 | 0.5046 | 0.4956 |
| 0.1 | 0.5649 | 0.5210 |
| 0.2 | 0.7250 | 0.5949 |
| 0.3 | 0.8751 | 0.6815 |
| 0.4 | 0.9636 | 0.7660 |
| 0.5 | 0.9931 | 0.8404 |
| 0.6 | 0.9990 | 0.9042 |
| 0.7 | 0.9999 | 0.9508 |
| 0.8 | 1 | 0.9747 |
| 0.9 | 1 | 0.9890 |
| 1 | 1 | 0.9956 |

[표4.8] 연속형 $Y_1$와 범주수가 8개인 $C$인 반응변수를 가지는 경우,

분리기준별 설명변수 $X_1$의 선택확률

| 평균 차이 | Mahalanobis | Gower |
|---|---|---|
| 0 | 0.4996 | 0.4985 |
| 0.1 | 0.5638 | 0.5382 |
| 0.2 | 0.7189 | 0.5887 |
| 0.3 | 0.8658 | 0.6831 |
| 0.4 | 0.9585 | 0.7653 |
| 0.5 | 0.9911 | 0.8458 |
| 0.6 | 0.9988 | 0.9112 |
| 0.7 | 0.9998 | 0.9473 |
| 0.8 | 1 | 0.9731 |
| 0.9 | 1 | 0.9849 |
| 1 | 1 | 0.9920 |

[표4.9] 연속형 $Y_1$와 범주수가 9개인 $C$인 반응변수를 가지는 경우,

분리기준별 설명변수 $X_1$의 선택확률

| 평균 차이 | Mahalanobis | Gower |
|---|---|---|
| 0 | 0.4946 | 0.5004 |
| 0.1 | 0.5613 | 0.5220 |
| 0.2 | 0.7048 | 0.6022 |
| 0.3 | 0.8609 | 0.6813 |
| 0.4 | 0.9587 | 0.7686 |
| 0.5 | 0.9902 | 0.8400 |
| 0.6 | 0.9981 | 0.9054 |
| 0.7 | 0.9998 | 0.9484 |
| 0.8 | 1 | 0.9694 |
| 0.9 | 1 | 0.9839 |
| 1 | 1 | 0.9910 |

## 4.2 실제자료의 적용

### 4.2.1 자료 소개

제안된 분리기준을 이용하여 실제자료에 적용해본다.

적용 자료는 고객들의 신용카드사용에 대한 국내 신용카드회사의 자료이다. 적용할 자료에는 현금서비스금액, 현금서비스이용유무, 고객연령대, 고객성별, 신규대출유무, 카드연체유무, 카드할부이용유무, 조회유무 등의 정보가 포함되어 있고, 원 자료에서 5000개의 표본을 단순임의 추출하여 분리기준에 적용해보았다. 자세한 변수설명은 [표4.10]과 같다.

[표4.10] 적용된 자료의 변수 정의

| | 변수 이름 | 변수 설명 |
|---|---|---|
| 혼합 반응변수 | 현금서비스금액 | 2008년 1월부터 3월까지 3개월간 현금서비스를 받은 금액 |
| | 현금서비스이용유무 | 2008년 1월부터 3월까지 3개월간 현금서비스를 받은 경험 |
| 설명변수 | 고객연령대 | 40대 이하와 40대 이상 |
| | 고객성별 | 고객의 성별 |
| | 최근 3개월 내 은행신규대출유무 | 2007년 10월부터 12월까지 3개월간 은행에서 신규대출을 한 경험 |
| | 최근 1년 내 카드할부이용유무 | 2006년 12월부터 2007년 12월까지 1년간 카드할부를 이용한 경험 |
| | 최근 1년 내 조회유무 | 2006년 12월부터 2007년 12월까지 1년간 신용조회를 한 경험 |
| | 최근 1년 내 카드연체유무 | 2006년 12월부터 2007년 12월까지 1년간 카드를 연체한 경험 |

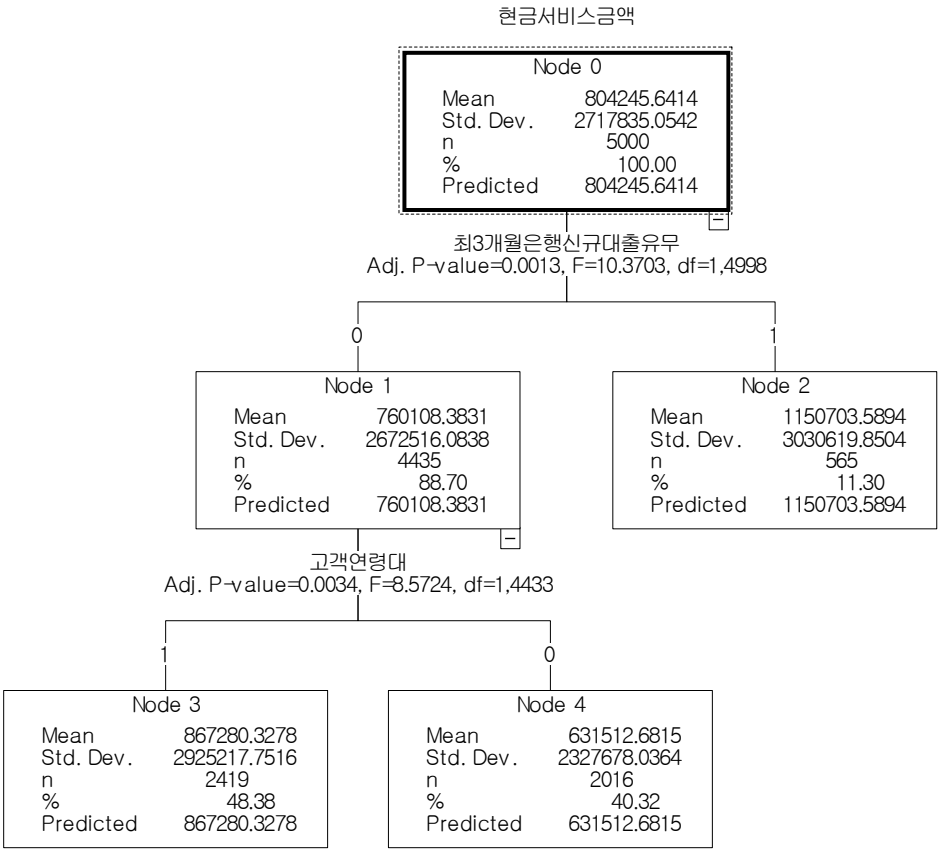
### 4.2.2 일변량 의사결정나무분석 결과

제안된 분리기준에 적용하기 전에, 혼합 반응인 '현금서비스금액'과 '현금서비스이용유무'를 일변량 반응변수의 알고리즘인 CHAID와 CART로 [그림4.4], [그림4.5], [그림4.6], [그림4.7]과 같이 분리해보았다.

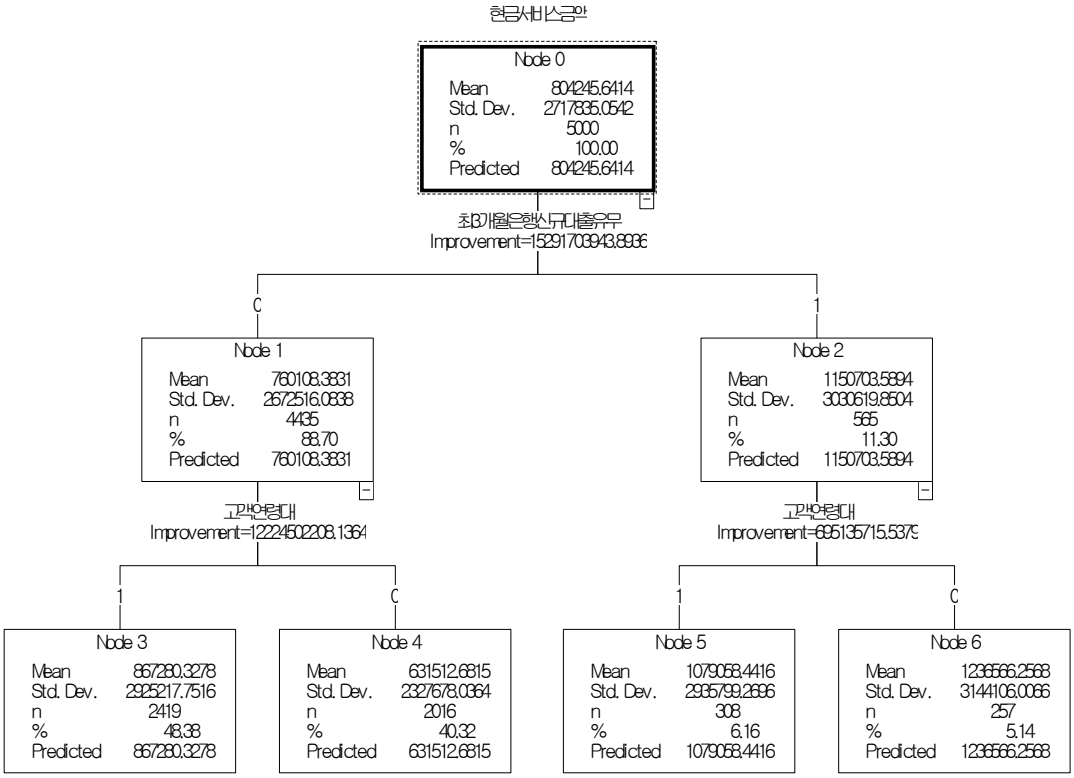
먼저 연속형 반응변수인 '현금서비스금액'을 CHAID 알고리즘으로 의사결정나무를 생성하면 [그림4.4]와 같다. 반응변수 '현금서비스금액'은 설명변수 '최근 3개월 내 은행신규대출유무'에 의해서 최적의 분리가 이루어진다. 최근 3개월 내 은행신규대출이 있는 사람의 현금서비스금액의 평균이 최근 3개월 내 은행신규대출이 없는 사람의 평균보다 높음을 알 수 있다. 여기서 최근 3개월 내 은행신규대출이 없는 그룹은 '고객연령대'인 40대 이하, 40대 이하로 분리되어, 최근 3개월 내 은행신규대출이 없는 사람 중 40대 이상의 현금서비스금액의 평균이 40대 이하보다 높게 나타난다.

CART 알고리즘으로 나타내면 [그림4.5]와 같다. 추가적으로 최근 3개월 내 은행신규대출이 있는 그룹이 '고객연령대'로 분리되는 것을 제외하면, CHAID 알고리즘으로 분리된 것과 같은 의사결정나무를 가진다.

[그림4.4] CHAID 알고리즘에 의한 '현금서비스금액'의 의사결정나무



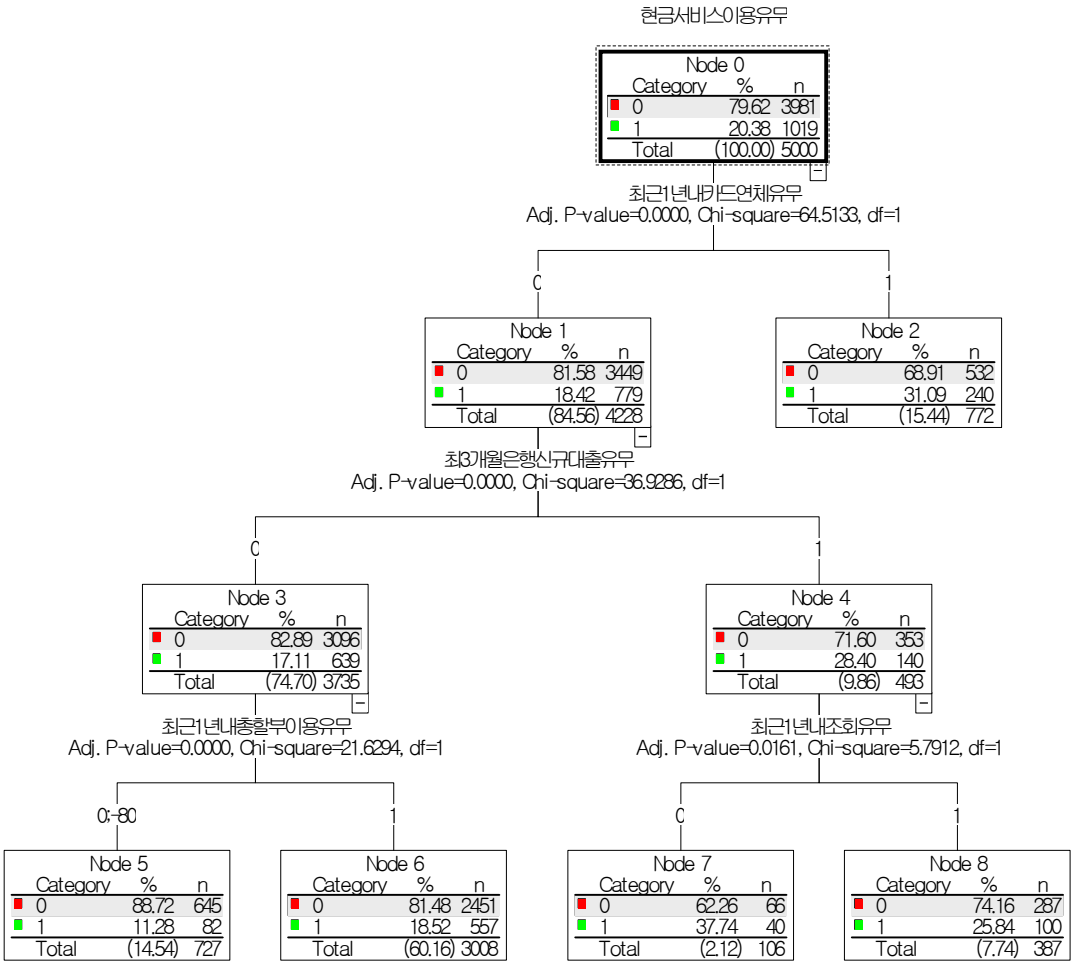
[그림4.5] CART 알고리즘에 의한 '현금서비스금액'의 의사결정나무



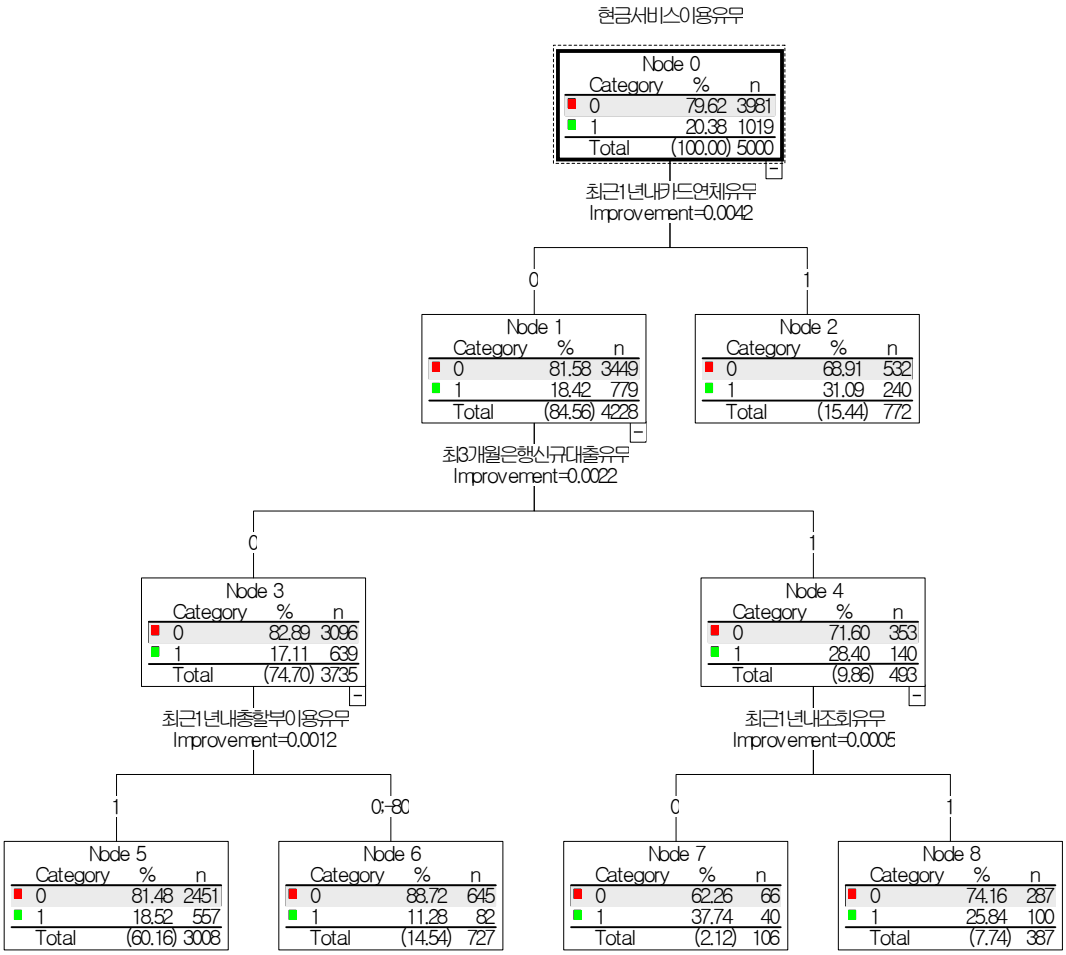

다음으로 범주형 반응변수인 '현금서비스이용유무'를 CHAID와 CART 알고리즘으로 의사결정나무를 [그림4.6]과 [그림4.7]과 같이 형성하였다.

CHAID 알고리즘의 의사결정나무는 설명변수 '최근 1년 내 카드연체유무'에 의해 최적 분리되었다. CART 알고리즘에 의한 의사결정나무 또한 CHAID 알고리즘과 동일한 의사결정나무를 가진다.

[그림4.6] CHAID 알고리즘에 의한 '현금서비스이용유무'의 의사결정나무



[그림4.7] CART 알고리즘에 의한 '현금서비스이용유무'의 의사결정나무



### 4.2.3 다변량 의사결정나무분석 결과

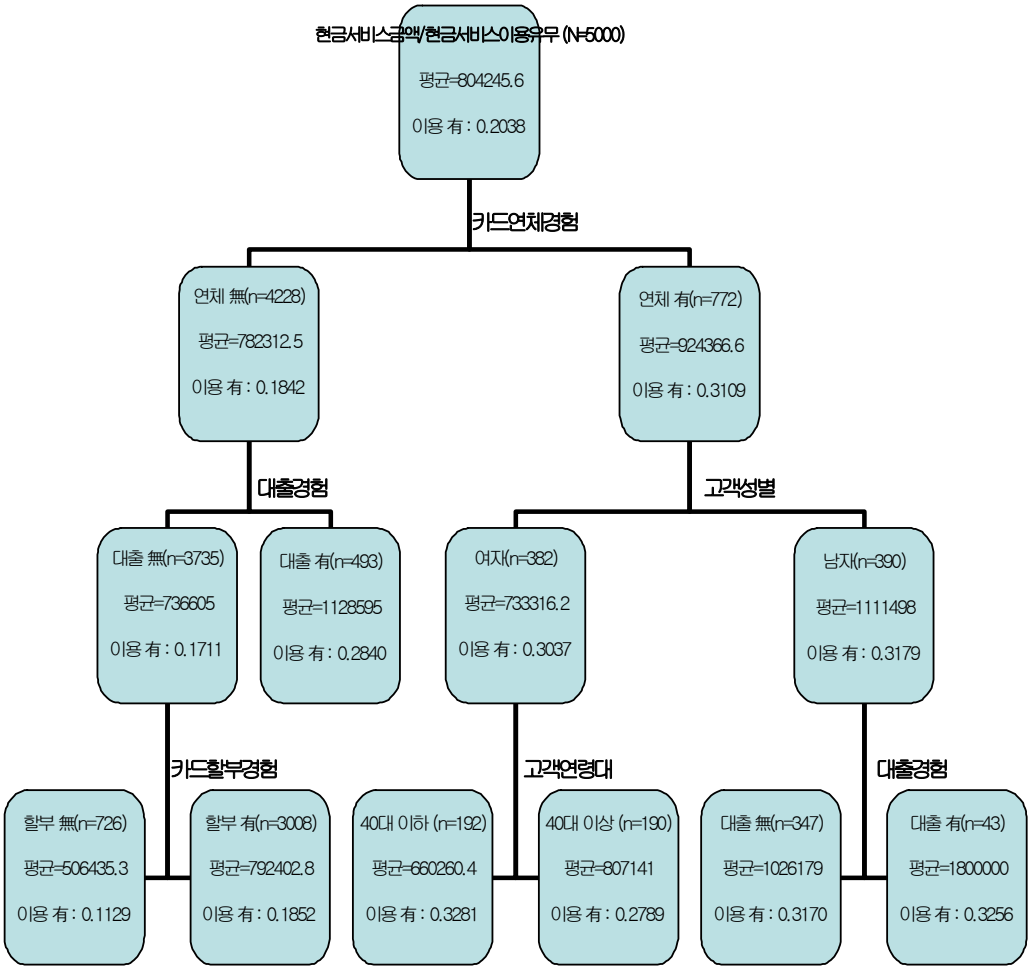
제안된 분리기준으로 자료를 분리한 결과, 마할라노비스 거리를 이용한 분리기준으로는 '최근 1년 내 카드연체유무'가 가장 최적의 분리를 하는 설명변수로 선택되었으며, Gower의 분리기준으로는 '고객연령대'가 선택되었다. 즉, 혼합 반응변수 '현금서비스금액'과 '현금서비스이용유무'는 마할라노비스 거리를 이용한 분리기준에 의해서는 '최근 1년 내 카드연체유무'로 최적의 분리가 되는 반면, Gower의 분리기준에 의해서는 '고객연령대'가 가장 동일한 성질의 개체들의 그룹으로 잘 분리된다고 볼 수 있다.

각 분리기준별로 잘 분리하는 설명변수의 순서를 살펴보면, 마할라노비스 거리를 이용한 분리기준을 이용하여 혼합된 반응변수를 잘 분리하는 설명변수의 순서를 나타내면, '최근 1년 내 카드연체유무', '최근 3개월 내 은행신규대출유무', '최근 1년 내 카드할부이용유무', '고객연령대', '최근 1년 내 조회유무' '고객성별' 순이다.

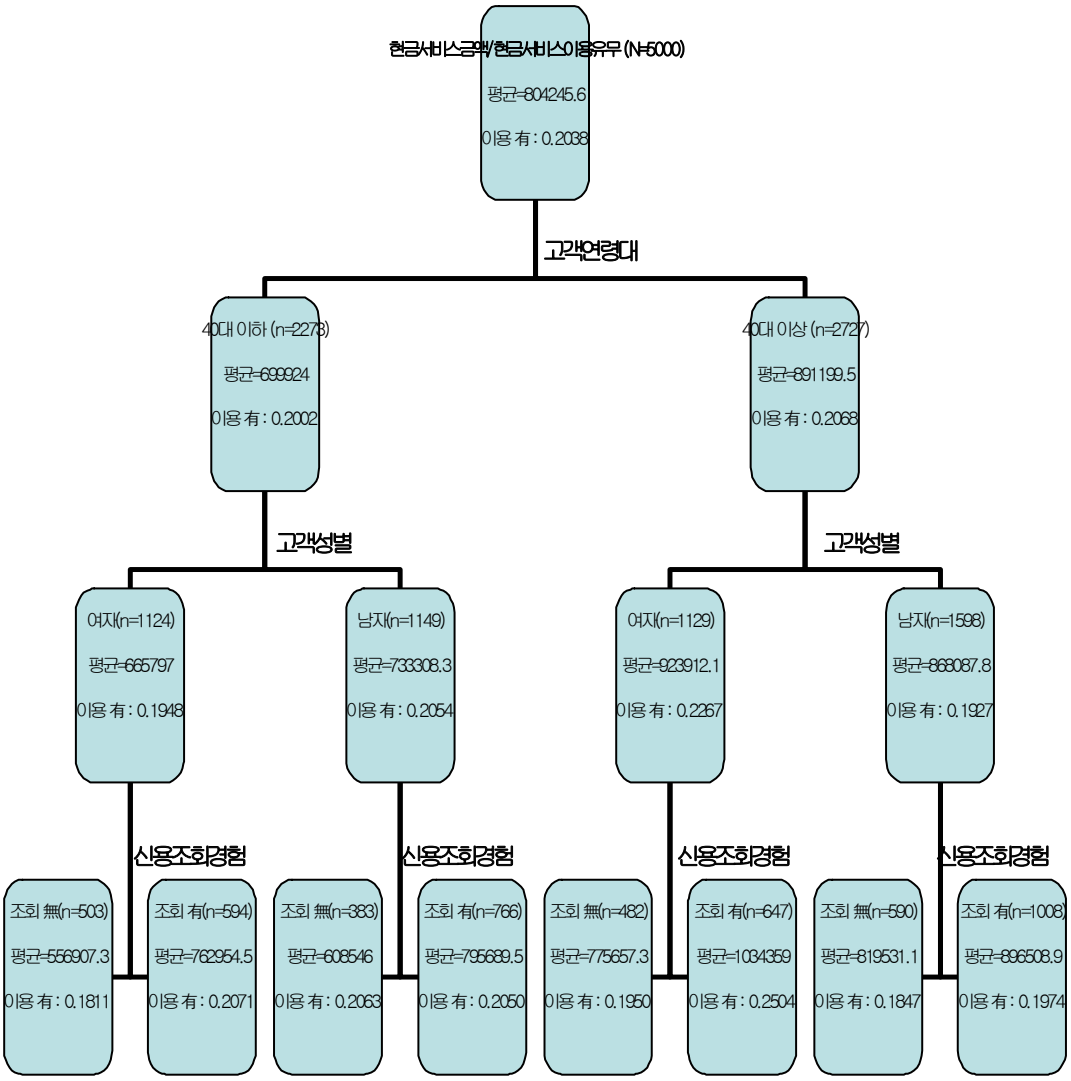
반면, Gower의 분리기준을 이용하여 혼합된 반응변수를 잘 분리하는 설명변수의 순서를 나타내면, '고객연령대', '고객성별', '최근 1년 내 조회유무', '최근 1년 내 카드연체유무', '최근 1년 내 카드할부이용유무', '최근 3개월 내 은행신규대출유무' 순이다.

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

[그림4.8] 마할라노비스 거리를 이용한 의사결정나무



[그림4.9] Gower의 분리기준을 이용한 의사결정나무



각 분리기준별 의사결정나무로 현금서비스금액과 현금서비스이용유무에 따른 카드고객을 분류하면, 4개의 그룹으로 볼 수 있다. 분류한 그룹은 다음의 [표4.4]와 같다.

[표4.4] 각 분리기준별 혼합 반응의 성향에 따른 타겟팅 그룹

| 분리기준 | 혼합 반응변수의 성향 | 타겟팅 그룹 |
|---|---|---|
| 마할라노비스 거리 | 현금서비스금액 ↑<br>현금서비스이용 ↑ | 최근 1년 내 카드연체경험이 있는 남자 중 최근 3개월 내 은행신규대출이 있는 고객 |
| | 현금서비스금액 ↑<br>현금서비스이용 ↓ | X |
| | 현금서비스금액 ↓<br>현금서비스이용 ↑ | 최근 1년 내 카드연체경험이 있는 여자 중 40대 이하인 고객 |
| | 현금서비스금액 ↓<br>현금서비스이용 ↓ | 최근 1년 내 카드연체경험, 최근 3개월 내 은행신규대출, 최근 1년 내 카드할부이용경험이 없는 고객 |
| Gower | 현금서비스금액 ↑<br>현금서비스이용 ↑ | 40대 이상인 여자 중 최근 1년 내 신용조회경험이 있는 고객 |
| | 현금서비스금액 ↑<br>현금서비스이용 ↓ | 40대 이상인 남자 고객 |
| | 현금서비스금액 ↓<br>현금서비스이용 ↑ | 40대 이하인 남자 고객 |
| | 현금서비스금액 ↓<br>현금서비스이용 ↓ | 40대 이하인 여자 중 최근 1년 내 신용조회경험이 없는 고객 |

[표4.4]를 보면, 마할라노비스 거리를 이용한 분리기준에 의해서 최근 1년 내 카드연체경험이 있는 남자 중 최근 3개월 내 은행신규대출경험이 있는 고객의 현금서비스금액과 현금서비스이용률 모두 높다는 것을 알 수 있고

최근 1년 내 카드연체경험이 있는 여자 중 40대 이하인 고객의 현금서비스 이용률은 높지만 현금서비스금액은 낮다는 것을 알 수 있다. 또한 최근 1년 내 카드연체경험, 최근 3개월 내 은행신규대출경험, 최근 1년 내 카드할부이용이 없는 고객에서 현금서비스금액과 이용률 모두 낮았다.

반면, Gower의 분리기준에 의해서는 40대 이상의 여자 중 최근 1년 내 신용조회경험이 있는 고객에서 현금서비스금액과 현금서비스이용률이 높다고 볼 수 있었다. 또한 40대 이상 남자고객의 현금서비스금액은 높지만 이용률은 낮았고 40대 이하 남자고객의 현금서비스금액은 낮지만 이용률은 높음을 확인할 수 있다. 그리고 40대 이하 여자 중 최근 1년 내 신용조회경험이 없는 고객에서 현금서비스금액과 이용률 모두 낮음을 알 수 있다.

그러므로 위의 결과에 따라 각 분리기준별로 분류된 타겟팅 고객에게 적절한 상품광고나 서비스규제 등을 할 수 있을 것이다.

# 제5장 결론 및 향후 연구방향

본 연구에서는 혼합반응에 대한 다변량 의사결정나무를 형성하는 분리기준을 제안하였다. 혼합반응에 적합하게 제안된 분리기준을 모의실험을 통해 검증한 후, 실제 카드회사자료를 적용하여 구축된 분리기준으로 혼합된 다변량 반응변수를 분리시켜보았다.

마할라노비스 거리와 Gower의 상사성 계수를 이용한 의사결정나무 모두 혼합 반응변수간의 평균차이에 따라 분리결과에 영향을 미치는 것을 확인할 수 있었다.

또한 마할라노비스 거리를 이용한 분리기준은 Gower의 분리기준과 대조적으로 범주수가 증가함에 따라 분리를 잘 수행하지 못함을 알 수 있었으나, 전반적인 결과를 보면 마할라노비스 거리의 기준이 Gower의 방법보다 더 좋게 나타남을 알 수 있었다. 그러므로 실제로 어떤 자료의 혼합 반응을 잘 분리시키기 위해서는 Gower의 분리기준보다 마할라노비스 거리를 이용한 기준을 적용하는 것이 좋을 것이다.

본 논문은 기존 일변량 의사결정나무와 동일한 종류의 다변량 반응변수에만 국한되었던 다변량 의사결정나무분석에서 연속형과 범주형 자료가 혼합된 반응에 대한 다변량 의사결정나무를 제안하였다는 데에 그 의의가 있다. 나아가 추후, 의사결정나무의 구조를 형성하는 정지규칙(stoping rule)과 가지치기(pruning)에 관한 알고리즘을 보완하여 보다 향상된 의사결정나무를 구현할 수 있을 것이다.

# 참 고 문 헌


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

[2] Breiman, L., Friedman, J., Olshen, R. & Stone, C. (1984). *Classification and regression trees.*Wadsworth, Belmont. CA.

[3] Gower, J. C. (1971). A General Coefficient of Similarity and Some of Its Properties. *Biometrics*, Vol.27, No.4, 857-871.

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

[5] Lee, S. K. (2006). On Classification and Regression Trees for Multiple Responses and Its Application. *Journal of Classification,* Vol.23, 123-141.

[6] Segal, M. R. (1992). Tree-Structured Methods for Longitudinal Data. *Journal of the American Statistical Association*, Vol.87, No.418, 407-418.

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

[8] The R Development Core Team. (2008). *R Version 2.6.2. User's Manual.*

[9] Zhang, H. P. (1998). Classification Tree for Multiple Binary Responses. *Journal of the American Statistical Associatio,.*Vol.93, 180-193.

# ABSTRACT

## A Study on Multivariate Decision Tree using Similarity Measures

Ju-hyun Lee
Department of Statistics
The Graduate School
Sungshin Women's University

In many fields, decision trees can be used to classify and predict the responses of data. There are many tree methods for univariate responses, such as CHAID, CART and QUSET. Recently, researches based on these methods like Mahalanobis distance and Generalized Entropy Index for multivariate responses have also been studied. But these algorithms are limited to be applied to mixed multivariate responses.

In this thesis, we modify previous multivariate decision trees and suggest new splitting criterions using Mahalanobis distance and Gower's Similarity coefficient for any type of mixed responses. To compare the performance of proposed trees, simulation studies on two new splitting criterions using Mahalanobis distance and Gower's Similarity coefficient are performed. Finally, an application using the data of Korean credit card company is illustrated.

# 부록 - R 프로그램

1. 반응변수가 연속형 $Y_1$과 범주가 이항(binary)인 $Y_2$의 경우, 마할라노비스 거리를 이용한 분리기준에 대한 모의실험

```
for (j in seq(0.0,1,0.1)){
count=0
for (i in 1:10000){

X<-array(0, c(300,5))
Sigma<-matrix(c(1,0.5,0.5,1),2,2)

for (i in 1:300){
X[i,1]<-rbinom(1,1,0.5) #독립변수 x1생성
X[i,2]<-rbinom(1,1,0.5) #독립변수 x2생성

ifelse                                                    ((X[i,1]<1),
(MN<-mvrnorm(1,rep(0,2),Sigma)),(MN<-mvrnorm(1,rep(j,2),Sigma)))

X[i,3]<-MN[1] #연속형종속변수생성
X[i,4]<-MN[2]
ifelse ((X[i,4]>0),(X[i,5]<-1),(X[i,5]<-0))#범주형종속변수생성
}
cov_g<-cov(X[,c(3,5)]) #부모노드공분산
m_g<-colMeans(X[,c(3,5)]) #부모노드평균
SS_p<-sum(mahalanobis(X[,c(3,5)],m_g,cov_g)) #부모노드의SS
#예측변수 x1일 때의 노드평균
m_l1<-colMeans(X[X[,1]<1,c(3,5)])
m_g1<-colMeans(X[X[,1]>0,c(3,5)])
#예측변수 x1일 때의 SS
SS_l<-sum(mahalanobis(X[X[,1]<1,c(3,5)],m_l1,cov_g))
SS_g<-sum(mahalanobis(X[X[,1]>0,c(3,5)],m_g1,cov_g))

#예측변수 x1일 때, 분리함수값
a<-SS_p-SS_l-SS_g

#예측변수 x2일 때의 노드평균
m_l2<-colMeans(X[X[,2]<1,c(3,5)])
```

```
m_g2<-colMeans(X[X[,2]>0,c(3,5)])
#예측변수 x2일때의 SS
SS_l2<-sum(mahalanobis(X[X[,2]<1,c(3,5)],m_l2,cov_g))
SS_g2<-sum(mahalanobis(X[X[,2]>0,c(3,5)],m_g2,cov_g))

#예측변수 x2일 때, 분리함수값
b<-SS_p-SS_l2-SS_g2

#분리
if (a-b>0)(count=count+1)
}
if (j>2) break
cat(j,":", count/10000, "\n")
}
```

2. 반응변수가 연속형 $Y_1$과 범주가 이항(binary)인 $Y_2$의 경우, Gower의 분리기준에 대한 모의실험

```
for (j in seq(0.0,1,0.1)){
count=0
for (i in 1:10000){

X<-array(0, c(300,5))
Sigma<-matrix(c(1,0.5,0.5,1),2,2)

for (i in 1:300){
X[i,1]<-rbinom(1,1,0.5) #독립변수 x1생성
X[i,2]<-rbinom(1,1,0.5) #독립변수 x2생성

ifelse                                                    ((X[i,1]<1),
(MN<-mvrnorm(1,rep(0,2),Sigma)),(MN<-mvrnorm(1,rep(j,2),Sigma)))

X[i,3]<-MN[1] #연속형종속변수생성
X[i,4]<-MN[2]
ifelse ((X[i,4]>0),(X[i,5]<-1),(X[i,5]<-0))#범주형종속변수생성
}

#부모노드 dissimilarity
DS_p<-sum(daisy(X[,c(3,5)], metric="gower", type=list(asymm=2)))

#예측변수 x1일때의 dissimilarity
x1_l<-(X[X[,1]==0,c(3,5)])
```

```
x1_g<-(X[X[,1]==1,c(3,5)])

DS_l<-sum(daisy(x1_l, metric="gower", type=list(asymm=2)))
DS_g<-sum(daisy(x1_g, metric="gower", type=list(asymm=2)))

#예측변수 x1일 때, 분리함수값
a<-DS_p-DS_l-DS_g

#예측변수 x2일 때의 dissimilarity
x2_l<-(X[X[,2]==0,c(3,5)])
x2_g<-(X[X[,2]==1,c(3,5)])

DS_l2<-sum(daisy(x2_l, metric="gower", type=list(asymm=2)))
DS_g2<-sum(daisy(x2_g, metric="gower", type=list(asymm=2)))

#예측변수 x1일 때, 분리함수값
b<-DS_p-DS_l2-DS_g2

#분리
if (a-b>0)(count=count+1)
}
if (j>2) break
cat(j, ": ", count/10000,"\n")
}
```