이 성 건 교수지도
석사학위청구논문

영과잉 이산형 자료를 위한 의사결정나무 연구

2011

성신여자대학교 대학원

통 계 학 과

최 보 미

영과잉 이산형 자료를 위한 의사결정나무 연구

이 성 건 교수지도

이 논문을 석사학위논문으로 제출함

2010년 11월

성신여자대학교 대학원

통 계 학 과

최 보 미

인 준 서

최보미의 석사학위 논문으로 인준함.

심사위원 _____ 인

심사위원 _____ 인

심사위원 _____ 인

성신여자대학교 대학원

논 문 개 요

영과잉 포아송 분포는 정상적인 포아송 확률분포보다 영의 값이 과잉 관측 된 경우에 적용할 수 있는 확률분포이다. 이러한 영과잉 포아송 자료는 최근 경제, 금융, 의학 등에서 나타나고 있으며 이를 분석하기 위한 통계적 모형들도 연구되고 있다. 그러나 의사결정나무분석에서는 영과잉 자료를 위한 분리기준이 드물며, 실제 전통적으로 많이 쓰이는 의사결정나무로는 분리가 되지 않을 때가 있어 영과잉 이산형 자료에 적합한 의사결정나무를 구축할 필요가 있다.

본 연구에서는 영과잉 이산형 자료를 위한 의사결정나무를 구축해 보고자 한다. 이는 Chaudhuri 등(1995)이 제안한 일반화 회귀나무의 분리기준을 바탕으로 영과잉 회귀나무를 구축한다. 일반화 회귀나무의 분리점이 그룹의 관찰치 수와 분산에 좌우되는 것을 보완하기 위하여 가중평균을 분리점으로 사용한 영과잉 회귀나무를 제안한다. 모의실험을 통하여 CART의 분리기준과 제안한 방법론의 효율성을 비교하며 가중평균 분리점을 고려한 영과잉 회귀나무가 경쟁력 있는 의사결정나무임을 보이고, 이를 실제자료에 적용해 본다.

목 차

제1장 서론

의사결정나무분석을 수행하기 위한 다양한 분리기준, 정지규칙, 가지치기 방법들이 제안되어 있으며, 이들을 어떻게 결합하느냐에 따라서 서로 다른 의사결정나무 형성 방법이 만들어진다. 또한 정확하고 빠르게 의사결정나무를 형성하기 위해서 다양한 알고리즘이 제안되어 있는데 CHAID, CART, C4.5 등이 이에 속한다. 의사결정나무는 목표변수의 성질에 따라 크게 두 가지로 나눈다. 목표변수가 범주형 변수일 때는 분류나무라 하며, 연속형 변수일 때는 회귀나무라 한다. 분류나무와 회귀나무에 대해 개선된 알고리즘들이 계속 연구되어 발표되고 있다. 분류나무에서는 CHITES(Chi-square test and exhaustive search) 방법과 F&CHITES(F&Chi-square test and exhaustive search) 방법이 등이 제안되었으며, 회귀나무에서는 Chaudhuri 등 (1995)에 의해 일반화 회귀나무가 제안되었다. 그리고 이영섭(2003)에 의해 회귀나무에서의 관심노드를 찾는 분류가 연구되었다.

본 논문의 목적은 목표변수가 영과잉 이산형인 경우에 자료를 영과잉 회귀모형에 적합 시켜 추정한 후 회귀나무를 이용하여 분류하는 것이다. 이는 기존에 영과잉 이산형 자료에 적합한 의사결정나무 연구가 거의 없어 Chaudhuri 등이 제안한 일반화 회귀나무를 적용하여 이를 개선하였다. 분리변수와 분리점의 선택과정에서 가중평균 분리점을 고려하여 회귀나무를 형성하였다.

본 연구 2장에서는 전통적인 기존의 의사결정나무와 일반화 회귀나무에 대해 소개한다. 3장에서는 영과잉 모형과 이를 분류하기 위한 새로운 의사결정나무를 제안한다. 4장에서는 제안한 영과잉 회귀나무의 효율성을 살펴보기 위하여 모의실험을 수행하며, 실제 자료를 이용하여 적용해 본다. 5장에서는 결론 및 향후 연구 방향에 대해 논의한다.

제2장 의사결정나무

의사결정나무는 의사결정규칙(decision rule)을 도표화하여 관심대상이 되는 집단을 몇 개의 소집단으로 분류(classification)하거나 예측(prediction)을 수행하는 분석 방법이다. 의사결정나무는 목표변수의 특성에 따라 분류나무(classification trees)와 회귀나무(regression trees)로 구분된다. 분류나무는 목표변수가 이산형으로써 목표변수의 각 범주에 속하는 계급을 최대한 잘 분류하는 예측변수의 부분집합을 찾는 방법이다. 회귀나무는 목표변수가 연속적인 값으로써 회귀분석에서와 같이 반응값을 예측하는 것이 목적으로 목표변수의 평균을 최대한 잘 분류하는 예측변수의 부분집합을 찾는 방법이다.

의사결정나무는 하나의 나무구조를 이루고 있으며, 자료를 이진분리(binary splits)하여 부모마디(parent node)로부터 두 개의 자식마디(child node)를 반복적으로 분할하므로 나무구조를 형성한다. 분리기준(split criterion)은 하나의 부모마디로부터 자식마디가 형성 될 때 예측변수의 선택과 범주의 병합이 이루어질 기준을 의미한다. 분리기준에 의해 어떤 예측변수를 이용하여 어떻게 분리하는 것이 목표변수의 분포를 가장 잘 구별해 주는지를 파악하여 자식마디를 형성한다. 여기서 목표변수의 분포를 구별하는 정도는 순수도(purity) 또는 불순도(impurity) 등에 의해서 측정하는데, 순수도란 목표변수의 특정 범주에 해당 마디의 개체들이 집중되어 있는 정도를 의미한다. 따라서 분리기준은 부모마디에 비해서 자식마디들의 순수도가 증가하는 정도를 수치화한 것이라 할 수 있다. 이러한 분리기준에 의해 의사결정나무는 부모마디의 순수도에 비해서 자식마디들의 순수도가 증가하도록 자식마디를 형성해 나간다.

본 장에서는 분리기준에 초점을 맞추어 전통적인 의사결정나무 CART,

CHAID, C4.5와 일반화 회귀나무에 대해 알아본다.

2.1 전통적인 의사결정나무

2.1.1 CHAID

CHAID(Chi-squared Automatic Interaction Detection)는 1977년 Kass에 의해 소개된 알고리즘으로, 카이제곱 검정 혹은 F-검정을 이용하여 분리(split)와 병합(merge)을 반복하면서 두 개 이상의 분리를 이루어 내는 다지분리(multiway split)를 수행한다. CHAID 알고리즘은 AID(Automatic Interaction Detection)에서 기원된 것으로 변수들 간의 통계적 관계를 찾는 것이 목적이다. 변수들 간의 관계는 다시 의사결정나무를 통해 표현 할 수 있으므로 분류기법으로 사용한다(Thearling,1995).

CHAID 알고리즘의 분리기준은 목표변수가 범주형 일 때 분할표에 기초한 피어슨(Pearson) 카이제곱 또는 우도비(likelihood) 카이제곱 통계량을 사용하며, 목표변수가 연속형 일 때는 두 개 이상의 그룹에 대해 평균 차이를 검정하는 분산분석의 F 통계량을 사용한다. 카이제곱 통계량은 관측도수 f_{ij}로 이루어진 $r \times c$ 분할표로부터 계산 된다. 분할표 구조는 <표 2.1>과 같다.

<표 2.1> 분할표의 구조

목표변수 / 예측변수	범주1	범주2	…	범주c	합계
범주1	f_{11}	f_{12}	…	f_{1c}	$f_{1.}$
범주2	f_{21}	f_{22}	…	f_{2c}	$f_{2.}$
…	…	…	…	…	…
범주r	f_{r1}	f_{r2}	…	f_{rc}	$f_{r.}$
합계	$f_{.1}$	$f_{.2}$	…	$f_{.c}$	$f_{..}$

<표 2.1>의 분할표로부터, 피어슨의 카이제곱 통계량은

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

와 같이 정의되며, 우도비 카이제곱 통계량은

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

로 정의된다. 이 때 목표변수의 범주 수는 c, 예측변수의 범주 수가 r 이므로 두 통계량의 자유도(degree of freedom)는 $(r-1)(c-1)$로 동일하다. e_{ij}는 분포의 동일성 또는 독립성의 가설 하에서 계산된 기대도수(expected frequency)이다. 기대도수는 다음과 같이 계산된다.

$$e_{ij} = \frac{f_i \times f_j}{f_{..}} \ .$$

카이제곱 통계량이 자유도에 비해서 매우 작다는 것은 예측변수의 각 범주에 따른 목표변수의 분포가 서로 동일하다는 것을 의미한다. 따라서 예측변수가 목표변수의 분류에 영향을 주지 않는다고 결론지을 수 있다. 자유도에 대한 카이제곱 통계량 값의 크고 작음은 p값으로 표현될 수 있는데, 카이제곱 통계량 값이 자유도에 비해서 작으면 p값은 커지게 된다. 결국 분리기준을 카이제곱 통계량 값으로 한다는 것은 p값이 가장 작은 예측변수와 그 때의 최적분리에 의해서 자식마디를 형성시킨다는 것을 의미한다.

반면, 목표변수가 연속형인 경우 이용되는 F 통계량은 자유도 $(r-1, n-r)$인 반면 $F-$분포를 따르고, <표 2.2>과 같이 계산된다.

<표 2.2> 분산분석표

요인	자유도	평방합	평균평방	분산비
예측변수	$r-1$	$SST=\sum_{i=1}^{r} n_i(\overline{y_i}-\overline{y})^2$	$MST=SST/(r-1)$	$F=\frac{MST}{MSE}$
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 예측변수의 각 범주에 따른 목표변수의 평균치 차가 존재하지 않다는 것을 의미하며, 예측변수가 목표변수의 분류나 예측에 영향을 주지 않는다고 결론지을 수 있다. 자유도에 대한 F 통계량의 크고 작음은 p값으로 표현될 수 있는데 F 통계량이 자유도에 비해서 작으면 p값은 커지게 된다. 분리기준을 F 통계량 값으로 할 경우, 가장 작은 p값을 가지는 예측변수에 의해 자식마디를 형성하는 최적분리를 한다.

2.1.2 CART

CART(Classification and Regression Trees) 알고리즘은 의사결정나무를 형성하는데 있어서 가장 보편적인 알고리즘으로 1984년 Breiman 등에 의해 발표되었다. CART 알고리즘은 각 마디에서 자식마디로 분리 시 불순도(impurity)의 감소를 최대화 하는 분리기준을 선택한다. 지니 지수(Gini index)는 카이제곱 통계량과 마찬가지로 불순도(impurity) 또는 다양도(diversity)를 재는 측도 중의 하나이다. CART 알고리즘은 목표변수가 이산형 일 때 지니 지수를 이용하여 자료를 이진분리하며 목표변수가 연속형인 경우에는 분산(variance)의 감소량을 분리기준으로 자료를 이진분리하여 하위마디를 반복적으로 분할함으로써 값이 유사한 부분집합을 생성한다.

지니 지수는 다음의 식(2.1)으로 표현될 수 있다.

$$G = \sum_{j=1}^{c} P(j)(1-P(j)) = 1 - \sum_{j=1}^{c} P(j)^2 = 1 - \sum_{j=1}^{c} (\frac{n_j}{n})^2. \qquad (2.1)$$

여기서 n은 그 노드에 포함되어 있는 관찰치 수를 말하고 n_j는 목표변수의 j번째 범주에 속하는 관찰치 수를 말한다. $P(j)$는 각 마디에서 한 개체가 목표변수 j번째 범주에 속할 확률을 말한다. c는 범주 수이다. 식(2.1)에서와 같이 지니 지수는 임의의 한 개체가 목표변수의 j번째가 아닌 범주에서 추출되었고, 그 개체를 목표변수의 j번째 범주에 속한다고 오분류 할 확률을 모두 더하여 얻을 수 있다. 이는 n개의 원소 중에서 임의로 2개를 추출하였을 때, 추출된 2개가 서로 다른 그룹에 속해 있을 확률을 의미하기도 한다. 목표변수가 2개인 경우 지니 지수는

$$G = 2P(1)P(2) = 2(\frac{n_1}{n})(\frac{n_2}{n}), \tag{2.2}$$

로 표현될 수 있다. 식(2.2)은 카이제곱 통계량을 사용하는 것과 같은 결과를 갖는다. 또한 지니 지수의 감소량은 다음과 같이 계산된다.

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

여기서 n은 부모노드의 관찰치 수를 말하고, n_R과 n_L는 각각 자식노드의 관찰치 수를 의미한다. 즉 자식마디로 분리되었을 때의 불순도가 가장 작도록 자식마디를 형성하는 것이며 이는 다음의 식(2.3)과 같이 자식마디에서의 불순도의 가중합을 최소화하는 것과 동일하다.

$$P(L)G_L + P(R)G_R = \frac{n_L}{n} G_L + \frac{n_R}{n} G_R. \tag{2.3}$$

또한 목표변수가 연속형일 경우, 분산이나 평균을 활용하여 왼쪽과 오른쪽 자식마디의 관찰치 수를 가중한 분산평균을 사용한다. 먼저 기호를 표현하면 다음 식(2.4)와 같다.

$$\begin{aligned} \hat{\mu_L} &= \frac{1}{n_L} \sum_{n \in L} y_n & V_L &= \frac{1}{n_L} \sum_{n \in L} (y_n - \hat{\mu_L})^2 \\ \hat{\mu_R} &= \frac{1}{n_R} \sum_{n \in R} y_n & V_R &= \frac{1}{n_R} \sum_{n \in R} (y_n - \hat{\mu_R})^2. \end{aligned} \tag{2.4}$$

여기에서 L과 R은 각각 왼쪽과 오른쪽의 자식마디를 나타내며, 추정치들은

각각 왼쪽과 오른쪽 자식마디의 정규분포 하에서 구해진 최우추정치(MLE)들이다.

$$cart_{LR} = \frac{1}{n_L + n_R}[n_L V_L + n_R V_R] \qquad (2.5)$$
$$= \frac{1}{n_L + n_R}[\sum_{n \in L}(y_n - V_L)^2 + \sum_{n \in R}(y_n - V_R)^2].$$

식(2.5)에서처럼 전통적 분리기준은 왼쪽과 오른쪽 노드의 가중치 된 평균분산을 사용한다(이영섭,2003). 이는 자식마디에서 집단 내 분산(within variance)을 최소화 하는 것으로 분산의 감소량을 최대화 하는 것과 같다. 분산의 감소량을 최대화 하는 기준은 식(2.6)과 같이 표현된다.

$$\Delta V = V - \frac{n_L}{n} V_L - \frac{n_R}{n} V_R . \qquad (2.6)$$

2.1.3 C4.5

C4.5 알고리즘은 Quinlan(1993)에 의해 수정 발전된 의사결정 알고리즘이다. 이것의 초기버전인 ID3 (Interactive Dichotomizer 3, 1986) 알고리즘은 기계학습(machine learning) 분야에 많은 영향을 주었다. CART가 각 마디를 이진분리하며 나무구조를 만드는데 반하여 C4.5는 연속형 예측변수에 관해서는 이진분리를 하지만, 명목형 예측변수에 관해서는 각 범주가 하나의 마디를 가지는 다지분리 구조를 갖는 나무를 형성한다.

C4.5는 주어진 패턴을 올바르게 분류하면서 간단한 의사결정나무를 구성하기 위해 정보이론(information theory)에 따른 무질서도(entropy) 개념을 이용한다. 무질서도는 다른 종류의 패턴들이 섞여있는 정도를 정량적으

로 나타내는 것으로, 각 마디에 대응하는 부나무(subtree) 속에 서로 다른 집단(class)이 많이 섞여 있을수록 무질서도가 높고, 반대로 단일한 집단(class)으로 되어 있으면 무질서도가 낮게 된다. C4.5 알고리즘은 뿌리노드에서부터 무질서도를 가장 낮게 할 수 있는 속성을 선택하여 나무를 확장함으로써 의사결정나무를 구성한다.

C4.5는 정보(information)라는 개념을 사용한다. 메시지(message)의 확률이 p일 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.

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

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.

$$-\log_2(\frac{freq(C_j, S)}{|S|}),$$

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

$$Info(S) = \sum_{j}^{k} \left((\frac{freq(C_j, S)}{|S|}) \times \log_2 (\frac{freq(C_j, S)}{|S|}) \right),$$

위의 $Info(S)$와 비슷한 개념으로, T가 X에 의해 n개로 분할 된 후의 기대정보(expected information)를 구하려면 식(2.7)를 이용할 수 있다.

$$Info_X(T) = \sum_{i}^{n} \left(\frac{|T_i|}{T} \times Info(T_i) \right), \tag{2.7}$$

X에 의해 분할로 얻어진 정보(information)는 다음 식(2.8)에 의해 얻을 수 있다.

$$Gain(X) = Info(T) - Info_X(T), \tag{2.8}$$

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 분리변수가 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 못하다. 그래서 T에 있는 한 개체가 속하는 하부집합을 정의하는데 필요한 평균 정보의 양(split info)으로 정규화(normalize)시켜 줄 필요가 있다.

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

정보의 양(split info)은 T가 n개의 하부집합으로 분할됨에 따라 발생하는 정보(information)의 양이다. Gain을 정보의 양으로 나누어 주면, 분할에 의해 생성된 유용한 정보의 비율인 정보 비율(gain ratio)이 된다.

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

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

2.2 일반화 회귀나무

회귀나무를 형성하기 위한 다양한 방법론이 연구되었는데 그 중에 Chaudhuri 등(1995)이 제안한 일반화 회귀나무(generalized regression tree)가 있다. 본 장에서는 Chaudhuri 등(1995)이 제안한 일반화 회귀나무(generalized regression tree)의 분리변수 선택과 분리점에 대해 살펴보고자 한다. 먼저 일반화선형모형(generalized linear model)에 대해 알아보고 이를 적합한 일반화 회귀나무의 분리기준에 대해 알아본다.

2.2.1 일반화선형모형

일반화선형모형은 Nelder&Wedderburn(1972)에 의해 처음 체계화 된 것으로 지수족(exponential family)과 연결함수(link function)를 이용한 정규이론에 의한 선형모형이다. 다만, 오차의 분포를 정규분포를 포함하는

지수족의 여러 분포를 사용하며 목표변수의 기댓값과 예측변수의 선형결합(linear predictor)을 연결시키는 연결함수를 설정하여 일반화 하였다.

일반화선형모형은 임의의 성분, 체계적 성분, 연결 함수로 구성되어 있다. 임의의 성분은 $E(Y)=\mu$를 갖는 반응 변수 Y의 확률 분포를 말하고, 체계적 성분은 공변량 $x_1,x_2,\cdots,x_p$로 다음과 같은 선형결합 η로 표현된다.

$$\eta=\sum_{j=1}^{p}x_j\beta_j$$

연결함수 $g(\bullet)$는 식(2.9)와 같이 임의의 성분과 체계적 성분을 연결시킨다.

$$\eta_i=g(\mu) \qquad (2.9)$$

만약 임의의 성분이 지수족이라면 정준연결함수(canonical link function)라 불리는 연결함수가 존재한다. 정규분포는 항등함수, 포아송 분포는 로그함수, 이항분포는 성공확률에 대한의 로짓(logit)이 그 경우이다.

2.2.2 일반화 회귀나무의 분리기준

일반화 회귀나무 분리기준 알고리즘은 공변량 벡터들을 잔차의 부호에 따라 두 그룹으로 나누어 분리변수를 찾는다. 잔차의 부호로 나누어진 그룹의 공변량들을 전통적으로 구조화된 나무모형에서 발달되어진 t검정을 이용하여 순위화 한다. 이 때 순위가 가장 높은 공변량은 모형 적합의 결여가 가장 큰 것으로 해석하여 분리변수로 선택한다.

Chaudhuri 등(1995)은 예측변수가 연속형이고 목표변수가 포아송 분포

를 따르는 이산형 자료를 이용하여 회귀나무를 도출하고, 포아송 회귀나무라 명명하였다. 포아송 회귀나무는 $X_k \le a_k$ 또는 $X_k > a_k$ 형태를 통해 형성된다. 분리변수 k와 분리점 a_k의 선택과정은 다음과 같다.

Ⅰ (모형적합) 각 마디에 있는 포아송 자료를 회귀모형에 적합 시킨다. 여기서 $m = E(Y)$이고 $x_1, x_2, \cdots, x_K$는 K개의 공변량들로, 로그연결함수가 사용되었다.

$$\log(m) = \beta_0 + \sum_{k=1}^{K} \beta_k x_k .$$

Ⅱ (잔차계산) 각 마디에서 y_i에 대한 안스콤 잔차(anscombe residual)를 계산한다. $\hat{m}_i$는 m의 i번째 개체의 추정치이고 y_i는 Y_i의 관찰치라 하자. 이 때 안스콤 잔차는 다음과 같이 표현된다.

$$r_i = \frac{y_i^{2/3} - (\hat{m}_i^{2/3} - \frac{1}{9}\hat{m}_i^{-1/3})}{\frac{2}{3}\hat{m}_i^{1/6}}.$$

안스콤 잔차는 잔차의 중심을 평행이동 시킨다. 이는 잔차를 이용하여 두 그룹으로 분리할 때 개체들이 한 그룹으로 편중되는 것을 방지하기 위하여 사용되었다.

Ⅲ (잔차 분리) r_i에서 음의 잔차와 관련된 값은 그룹 1로 하고, 음이 아닌 잔차의 값은 그룹 2로 정한다. 이를 I_j라고 정의한다. $j = 1,2$

Ⅳ (평균과 관한 검정) 두 그룹에 대해서 이표본(two-sample) t검정을 이용하여 그룹 간 공변량들의 평균 차이가 있는지를 검정한다.

$$t_k^{(1)} = \frac{(\bar{x}_{k1} - \bar{x}_{k2})}{s_k\sqrt{I_1^{-1} + I_2^{-1}}}, k = 1, \cdots, K.$$

여기서 $\bar{x}_{kj}, s_{kj}^2$은 t마디에서 그룹 j에 속하는 자료값 X_k의 평균과 분산이며 s_k^2은 공통분산이다.

Ⅴ (변수선택) p값이 가장 작은 공변량을 선택하여 이를 분리변수로 하고 분리변수의 두 그룹의 평균값을 기준으로 두 개의 자식마디로 분리한다.

따라서 p값이 가장 작은 공변량 k가 분리변수로 선택되며, 분리점 a_k는 분리변수의 평균값으로 $a_k = \frac{(\bar{x}_{k1} + \bar{x}_{k2})}{2}$ 가 된다.

일반화 회귀나무의 분리기준 알고리즘은 CART의 분리기준 알고리즘과 몇 가지 차이점이 있다. CART의 경우 각 마디에 상수를 적용하여 분리하지만 일반화 회귀나무는 선형 또는 다항 회귀를 적합하여 분리를 한다. 또한 일반화 회귀나무 분리기준 알고리즘에서는 각 마디에 단 한 번 모형을 적합 하는 방법으로 분리를 한다. CART는 불순도가 가장 줄어 든 변수를 찾아 분리를 하게 된다. 불순도를 기준으로 한다면 로그선형모형을 적합 시킨 후 각 마디에서 부마디(subnode)로 분리 시 문제가 발생한다. 로그선형모형은 뉴턴-랩슨(Newton-Raphson)방법으로 모수를 추정하기 때문에 공

변량이 가장 줄어든 불순도를 찾기 위하여 매번 로그선형모형을 적합하여야 하므로 실용적이지 못한 방법이 된다. 또 다른 차이점은 CART는 잔차제곱합의 감소율을 분리기준으로 사용하지만 일반화 회귀나무는 잔차 분포를 이용하여 분리를 한다는 것이다. 이는 모형 적합이 만족스럽지 않을 경우 적합결여가 잔차에 반영되기 때문에 잔차 분포를 이용하여 분리를 하자는 의도에서 사용되었다. 예를 들어, 적합 시킨 모형에 적합결여가 생겼을 경우 잔차의 부호에 따라 두 집단으로 구분하면 두 그룹의 분포는 예측변수에 따른 평균차이가 발생한다. 두 그룹 간 평균차이가 가장 큰 예측변수를 분리변수로 선택하고, 두 그룹 간 분리변수의 평균값을 분리점으로 두 개의 마디로 분할을 반복하여 회귀나무를 형성한다.

제3장 영과잉 자료에 대한 의사결정나무

잔차 분포와 t검정을 이용한 일반화 회귀나무를 개선하여 영과잉(zero-inflated) 자료를 위한 영과잉 회귀나무 방법론을 제안한다. 본 장에서는 영과잉 포아송 회귀모형과 영과잉 음이항 회귀모형에 대하여 살펴보고 영과잉 회귀나무의 분리기준에 대해 알아본다.

3.1 영과잉 회귀모형

셀 수 있는 이산형 자료(count data)에 대한 분석모형으로 흔히 포아송 회귀모형을 적용한다. 그러나 실제로 시행횟수 혹은 발생건수들의 분포는 흔히 분산이 평균보다 큰 경우가 많다. 포아송 분포의 경우 평균과 분산이 동일하므로 과분산 된 자료에 포아송 분포를 사용하면 추정치가 편의(biased)되는 문제가 발생한다. 이러한 경우에는 종종 포아송 회귀모형 대신 음이항 회귀모형을 사용한다.

그러나 포아송 회귀모형이나 음이항 회귀모형과 같은 전통적인 추론방법을 이용한 모형들은 어떤 특정한 값에서 관측도수가 기대도수보다 높은 자료에 대하여 옳은 설명을 해 낼 수 없다. 즉, 영(zero)의 비율이 기존의 포아송 회귀모형 또는 음이항 회귀모형에 의해 기대되는 영의 비율보다 높게 관측된 경우에 두 모형들은 적절치 못한 적합이 된다. 영이 과도하게 관측된 자료를 영과잉(zero-inflated) 자료라 한다. 영과잉 자료를 포아송 회귀모형 혹은 음이항 회귀모형에 적합하여 통계적인 추정 및 검정을 한다면 이는 3종 오류를 범하는 결과를 초래 할 것이다. 이러한 문제를 해결하기 위해 분포의 혼합 형태인 영과잉 포아송 회귀모형(zero-inflated Poisson regression model:ZIP)과 영과잉 음이항 회귀모형(zero-inflated

Negative Binomial regression model:ZINB)이 제안되었다.

3.1.1 영과잉 포아송 회귀모형

영과잉 포아송 분포는 정상적인 포아송 확률분포보다 영의 값이 과도하게 관측된 분포를 말한다. 영과잉 포아송 회귀모형은 반응값이 영인 부분과 아닌 부분으로 나누어 베르누이 분포와 포아송 분포와의 혼합된 영과잉 포아송 분포를 이용한 모형이다. 영과잉 포아송 분포는 확률변수 y가 일정 단위당 나타나는 이산형 자료(count data)로 영만 나타나는 상태(perfect state)의 확률 값이 따로 정해진다. 이는 다음의 식(3.1)과 같이 표현할 수 있다.

$$\begin{array}{ll} y_i \sim 0 & p_i\text{의 확률}, \\ \sim Poisson(\lambda_i) & 1-p_i\text{의 확률}. \end{array} \tag{3.1}$$

여기서 $\boldsymbol{y} = (y_1, \cdots, y_n)'$는 반응벡터이고 독립이다. $0 \le p_i \le 1$ 값에서 주어지는 임의의 확률이고 λ_i는 포아송 분포의 평균으로 0보다 크다. 식(3.1)로부터 y_i의 확률질량함수(pmf)는 식(3.2)와 같이 표현된다.

$$\begin{aligned} P(y_i = 0) &= p_i + (1-p_i)e^{-\lambda_i} \\ P(y_i = k) &= (1-p_i)\frac{\lambda_i^k e^{-\lambda_i}}{k!}, \; k = 1,2,\cdots. \end{aligned} \tag{3.2}$$

식(3.2)의 영과잉 포아송 분포를 따를 경우 우도함수는 식(3.3)과 같이 나타낼 수 있다.

$$L(p_i,\lambda_i)=\prod_{i=1}^{n}\left[\left\{p_i+(1-p_i)e^{-\lambda_i}\right\}^{I_{(y_i=0)}}\times\left\{(1-p_i)\frac{\lambda_i^k e^{-\lambda_i}}{k!}\right\}^{I_{(y_i>0)}}\right]. \quad (3.3)$$

모수벡터 $\boldsymbol{\lambda}=(\lambda_1,\lambda_2,\cdots,\lambda_n)'$와 영의 확률벡터 $\boldsymbol{p}=(p_1,p_2,\cdots,p_n)'$는 다음의 식(3.4)와 같이 로그연결함수(log link function)와 로짓연결함수(logit link function)를 만족한다.

$$\begin{aligned}\log(\boldsymbol{\lambda})&=\boldsymbol{B\beta}\\ \log it(\boldsymbol{p})&=\log(\frac{\boldsymbol{p}}{1-\boldsymbol{p}})=\boldsymbol{G\gamma}\,.\end{aligned} \quad (3.4)$$

여기서 $\boldsymbol{B}$와 $\boldsymbol{G}$는 공변량들의 행렬이다. 위 모형에서 $\boldsymbol{\lambda}$와 $\boldsymbol{p}$가 함수적 관계가 없고 $\boldsymbol{B}=\boldsymbol{G}$ 일 때, 모형에 포함되는 모수의 수는 포아송 회귀모형보다 두 배 만큼 필요하다. $\boldsymbol{p}$가 공변량에 의존하지 않을 때 $\boldsymbol{G}$는 원소가 1인 벡터가 되어 영과잉 포아송 회귀모형은 포아송 회귀모형보다 단 한 개의 모수가 더 필요하다. 만약 $\boldsymbol{\lambda}$와 $\boldsymbol{p}$이 같은 공변량에 영향이 있다면(함수관계가 있다면) $\boldsymbol{\lambda}$의 함수만큼 $\boldsymbol{p}$를 생각할 수 있으므로 모수의 수가 줄어든다. $\boldsymbol{\lambda}$와 p가 어떻게 관계되었는지에 대한 두 모수의 사전정보가 $p_i=(1+\lambda_i^{\tau})^{-1}$로 알려져 있다면 식(3.4)의 영과잉 포아송 회귀모형은 아래의 식(3.5)과 같은 모형이 된다.

$$\begin{aligned}\log(\boldsymbol{\lambda})&=\boldsymbol{B\beta}\\ \log it(\boldsymbol{p})&=-\tau\boldsymbol{B\beta}.\end{aligned} \quad (3.5)$$

τ는 형태모수(shape parameter)이다. 위 식(3.5)의 영과잉 포아송 회귀모형은 일반화 선형모형을 만들기 위해 포아송 평균의 로그연결함수와 베르

누이 분포의 성공의 확률에 대해 로짓연결함수가 사용되었다. 따라서 식 (3.5)은 $ZIP(\tau)$로 정의된다.

영과잉 포아송 회귀모형에서 회귀계수의 추정은 $\boldsymbol{\lambda}$와 $\boldsymbol{p}$가 함수적 관계 여부에 따라 다르다. $\boldsymbol{\lambda}$와 $\boldsymbol{p}$가 함수적 관계가 없을 때 회귀계수 벡터 $\boldsymbol{\beta}$와 형태모수 γ에 대한 로그우도함수는 다음과 같다.

$$\begin{aligned} L(\boldsymbol{\gamma},\boldsymbol{\beta};\boldsymbol{y}) = \sum_{y_i=0} \log(e^{\boldsymbol{G_i\gamma}} + \exp(-e^{\boldsymbol{B_i\beta}})) \\ + \sum_{y_i>0} (y_i \boldsymbol{B_i\beta} - e^{\boldsymbol{B_i\beta}}) \\ - \sum_{i=1}^{n} \log(1+e^{\boldsymbol{G_i\gamma}}) \\ - \sum_{y_i>0} \log(y_i!). \end{aligned}$$

여기서 $\boldsymbol{G_i}$와 $\boldsymbol{B_i}$는 $\boldsymbol{G}$와 $\boldsymbol{B}$의 i번째 행이다.

$\boldsymbol{\lambda}$와 $\boldsymbol{p}$가 함수적 관계가 있을 때 회귀계수 벡터 $\boldsymbol{\beta}$와 형태모수 τ에 대한 우도함수는 다음과 같다.

$$\begin{aligned} L(\boldsymbol{\beta},\tau;\boldsymbol{y}) = \sum_{y_i=0} \log(e^{-\tau\boldsymbol{B_i\beta}} + \exp(-e^{\boldsymbol{B_i\beta}})) \\ + \sum_{y_i>0} (y_i \boldsymbol{B_i\beta} - e^{\boldsymbol{B_i\beta}}) \\ - \sum_{i=1}^{n} \log(1+e^{-\tau\boldsymbol{B_i\beta}}). \end{aligned}$$

ZIP와 $ZIP(\tau)$의 최우추정치는 점근적으로 정규분포를 따른다.

모형의 유의성 검정은 완전모형(full model)과 설정한 모형을 비교하여 유의한 차이가 있는지 여부로 평가한다.

$$2\log\lambda = 2[L(y;y) - L(\hat{\beta},\hat{\tau};y)] \,. \qquad (3.6)$$

식(3.6)는 귀무가설 하에서 두 배의 로그우도비로 점근적으로 카이제곱 분포을 따른다(Lambert,1992).

3.1.2 영과잉 음이항 회귀모형

영과잉 음이항 회귀모형은 로짓 모형과 음이항 모형의 혼합모형으로 반응값이 영인 부분은 로지스틱 분포(logistic distribution)를, 영이 아닌 부분은 음이항 분포(negative binomial distribution)를 따르는 것으로 간주한다. 이 모형은 영과잉 문제 뿐만 아니라 과대산포(overdispersion)문제를 동시에 고려할 수 있다.

영과잉 음이항분포는 영인 부분은 확률 ψ_i로 나타나며 영이 아닌 부분은 $(1-\psi_i)$의 확률로 음이항 분포를 따르게 된다. 따라서, 영과잉 음이항 회귀모형에서의 회귀분석은 먼저 영이 과잉된 부분의 확률 ψ_i과 $(1-\psi_i)$의 확률로 음이항 분포를 따르는 이산형 자료를 로짓모형(logit model)을 이용하여 구분하고 최우추정법(Maximum Likelihood Estimation)을 이용하여 모형을 추정한다(최창호 외, 2010). 과잉인 영의 확률 ψ_i과 $1-\psi_i$의 확률로 음이항 분포 $g(y_i|x_i)$ 따르는 y_i의 분포는 식(3.7)과 같이 나타낼 수 있다.

$$\begin{array}{ll} y_i \sim 0 & \psi_i\text{의 확률}, \\ \sim g(y_i|x_i) & 1-\psi_i\text{의 확률}. \end{array} \qquad (3.7)$$

식(3.7)로부터 y_i의 확률질량함수는 다음과 같다.

$$p(y_i|x_i,w_i)=\begin{cases}\psi_i+(1-\psi_i)g(y_i=0\,|\,x_i)\,, & y_i=0,\\ (1-\psi_i)g(y_i>0|x_i), & y_i=1,2,\cdots.\end{cases} \tag{3.8}$$

$$=\begin{cases}\psi_i+(1-\psi_i)\left(\dfrac{\theta}{\theta+\lambda_i}\right)^{\theta}, & y_i=0,\\ (1-\psi_i)\dfrac{\Gamma(\theta+y_i)}{\Gamma(\theta)\Gamma(y_i+1)}\left(\dfrac{\theta}{\theta+\lambda_i}\right)^{\theta}\left(\dfrac{\lambda_i}{\theta+\lambda_i}\right)^{y_i}, & y_i=1,2,\cdots.\end{cases}$$

여기서 $0\le\psi\le1$ 이고, $\theta(0\le\theta\le\infty)$는 과분산계수(over-dispersion coefficient)의 역수이다. 이러한 영과잉 음이항 분포를 따를 경우 우도함수는 다음 식(3.9)과 같이 나타낼 수 있다.

$$L(\beta,\psi,\theta)=\prod_{i=1}^{n}\left[\left\{\psi+(1-\psi)\left(\frac{\theta}{\theta+\exp(X_i^{'}\beta)}\right)^{\theta}\right\}^{I_{(y_i=0)}}\right.$$
$$\left.\times(1-\psi)\frac{\Gamma(\theta+y_i)}{\Gamma(\theta)\Gamma(y_i+1)}\left(\frac{\theta}{\theta+\exp(X_i^{'}\beta)}\right)^{\theta}\left(\frac{\exp(X_i^{'}\beta)}{\theta+\exp(X_i^{'}\beta)}\right)^{y_i}\right]^{I_{(y_i>0)}} \tag{3.9}$$

3.2 영과잉 회귀나무의 분리기준

영과잉 자료를 이용하여 회귀나무를 형성하므로 본 연구에서는 가중평균을 고려한 영과잉 회귀나무라 명명하였다. 가중평균을 고려한 영과잉 회귀나무는 Chaudhuri 등이 제안한 일반화 회귀나무 분리기준 알고리즘에서 분리점 선택 시 두 그룹의 평균값 대신 가중평균값을 사용한다. 가중평균을

사용하지 않고 Chaudhuri 등이 제안한 분리기준 알고리즘을 사용할 경우 영과잉 회귀나무라 명명하였다.

가중평균을 고려한 영과잉 회귀나무의 분리기준은 혼합모형인 영과잉 포아송 회귀모형 또는 영과잉 음이항 회귀모형을 적합하여 영과잉 자료를 분류한다. 이는 포아송 분포와 베르누이 분포 혹은 로지스틱 분포와 음이항 분포 사이의 공변량을 잔차 분포를 이용하여 두 그룹으로 분리하며 t검정으로 두 그룹 간의 평균차이가 가장 큰 공변량을 선택하고 가중평균을 분리점으로 하여 반복적으로 분할하므로 나무를 형성한다. 가중평균을 고려한 영과잉 회귀나무의 분리기준 알고리즘은 다음을 따른다.

Ⅰ (모형적합) 각 마디에 있는 자료를 영과잉 포아송 회귀모형 또는 영과잉 음이항 회귀모형에 적합한다.

Ⅱ (잔차계산) 각 마디에서 모형에 대한 잔차(residual)를 계산한다.

Ⅲ (잔차 분리) r_i에서 음의 잔차와 관련된 값은 그룹 1로 하고, 음이 아닌 잔차의 값은 그룹 2로 정한다. 이 때 I_j는 노드에서 그룹 j의 개체수이다.

Ⅳ (평균에 관한 검정) 두 그룹의 공변량에 대한 평균과 차이를 이표본(two-sample) t검정을 이용하여 실시한다.

$$t_k^{(1)} = \frac{(\bar{x}_{k1} - \bar{x}_{k2})}{s_k\sqrt{I_1^{-1} + I_2^{-1}}}, k = 1, \cdots, K.$$

여기서 $\bar{x}_{kj}, s_{kj}^2$은 각 마디에서 그룹 j에 속하는 자료값 X_k의 평균과

분산이며 s_k^2은 공통분산이다.

Ⅴ (변수선택) p값이 가장 작은 공변량을 선택하여 이를 분리변수로 하고 두 그룹의 가중 평균값을 기준으로 두 개의 자식마디로 분리한다.

분리점은 관찰치 수와 분산을 고려한 가중평균 분리점을 사용한다. 관찰치 수를 고려한 가중평균 $w_k^{(1)}$은 다음과 같다.

$$w_k^{(1)} = \frac{n_1 \bar{x}_{k1} + n_2 \bar{x}_{k2}}{n_1 + n_2},$$

분산을 고려한 가중평균 $w_k^{(2)}$은 다음과 같다.

$$w_k^{(2)} = \frac{(\frac{\bar{x}_{k1}}{s_{k1}^2} + \frac{\bar{x}_{k2}}{s_{k2}^2})}{\frac{1}{s_{k1}^2} + \frac{1}{s_{k2}^2}}.$$

제 4장 모의실험 및 적용

4.1 분리기준에 대한 모의실험

4.1.1 모의실험 설계

영과잉 자료에 대하여 제안한 분리방법이 옳은지 알아보기 위하여 R 프로그램을 이용하여 모의실험을 설계하여 수행한다. 모의실험은 두 가지 사항을 고려한다. 혼합된 분포 속에서 영과잉 포아송 분포를 분류해 내는 예측변수를 얼마나 잘 찾아내는지 확인하며, 가중평균을 고려한 분리점이 Chaudhuri 등이 제안한 단순 평균을 이용한 분리점 보다 향상된 방법인지를 확인한다. 모의실험의 수행과정은 다음과 같다.

Ⅰ (예측변수 생성) 정규분포 N(0,1)로부터 크기가 100인 임의의 설명변수 X_1과 X_2를 독립적으로 생성한다.

Ⅱ (목표변수 생성) 목표변수 Y는 예측변수 X_1에 의존하는 영과잉 포아송 분포로부터 크기가 100인 난수를 생성한다. 이 때 목표변수 Y는 Y_1과 Y_2로 구성된다. 즉, Y_1는 $X_1 < 0$ 고 $ZIP(\lambda_1, p_1)$에서 난수를 생성하며 Y_2는 $X_1 > 0$ 고 $ZIP(\lambda_2, p_2)$로부터 난수를 생성한다. 이때의 λ는 불완전한 상태(imperfect state)를 나타내는 포아송 분포의 평균이며, p는 영의 값만이 나타내는 완전한 상태(perfect state)의 확률 값이다. 여기서 λ_1과 p_1는 각각 3과 0.5로 고정하며 λ_2는 1부터 5까지 0.5씩 p_2는 0.1부터 0.9까지 0.1씩 증가시킨다. 이렇게 생성된 임의의 데

이터는 다음과 같은 구조를 가진다.

$Y \sim ZIP(\lambda, p)$	$X_1 \sim N(0,1)$	$X_2 \sim N(0,1)$
$Y_1 \sim ZIP(\lambda_1, p_1)$	if $X_1 < 0$	
$Y_2 \sim ZIP(\lambda_2, p_2)$	if $X_1 > 0$	

이 구조는 X_1에 따라 Y가 결정되었으므로 좋은 의사결정나무는 참 분리변수(true split variable)인 X_1의 선택확률이 높게 나타날 것이다. 또한 X_1변수가 영(0)보다 크거나 작음에 따라 Y를 결정했으므로 영 근처에서 분리가 일어나야 정확하게 분리를 해 내는 것이 된다.

Ⅳ (분리) 생성된 임의의 데이터에 가중평균을 고려한 영과잉 회귀나무의 분리기준을 적용하여 목표변수 Y에 대한 참 분리변수 X_1의 선택확률과 분리점을 계산한다.

Ⅴ (반복) 2,000번의 모의실험을 반복한다.

4.1.2 모의실험 결과

< 분리변수의 선택확률 >

영과잉 회귀나무의 분리기준과 CART의 분리기준에 대해 분리변수의 선택확률을 비교해 보자. 가중평균을 고려한 영과잉 회귀나무 분리기준과 영

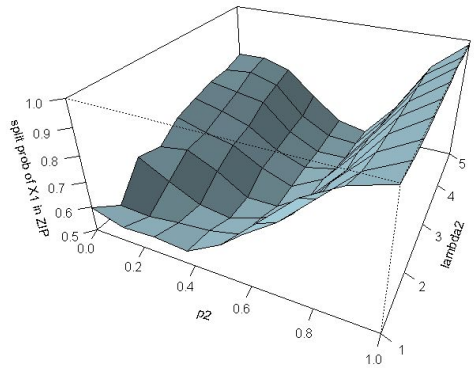
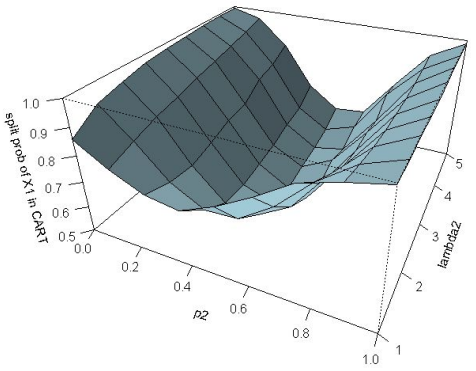
과잉 회귀나무 분리기준은 분리점만 다르므로 참 분리변수 X_1의 선택확률은 같다. <표 4.1>는 λ_2과 p_2의 변화에 대해 영과잉 회귀나무와 CART 분리기준의 X_1 선택확률을 나타낸 것이다.

<표 4.1> 영과잉 회귀나무와 CART 분리기준의 X_1 선택확률

λ_2	p_2	X_1 선택확률		λ_2	p_2	X_1 선택확률		λ_2	p_2	X_1 선택확률	
		영과잉 회귀나무	CART			영과잉 회귀나무	CART			영과잉 회귀나무	CART
1.0	0.1	0.5955	0.8590	2.5	0.1	0.6225	0.9705	4.0	0.1	0.7425	0.9900
1.0	0.2	0.5575	0.7830	2.5	0.2	0.6415	0.8620	4.0	0.2	0.7180	0.9235
1.0	0.3	0.5630	0.7290	2.5	0.3	0.5800	0.7060	4.0	0.3	0.6950	0.7985
1.0	0.4	0.5515	0.7095	2.5	0.4	0.5025	0.5585	4.0	0.4	0.5845	0.6210
1.0	0.5	0.6250	0.7650	2.5	0.5	0.5015	0.5140	4.0	0.5	0.5200	0.5390
1.0	0.6	0.7500	0.8390	2.5	0.6	0.5840	0.5825	4.0	0.6	0.5245	0.5645
1.0	0.7	0.8760	0.9230	2.5	0.7	0.7305	0.7450	4.0	0.7	0.6960	0.7240
1.0	0.8	0.9535	0.9630	2.5	0.8	0.8955	0.9095	4.0	0.8	0.8880	0.9045
1.0	0.9	0.9940	0.9945	2.5	0.9	0.9845	0.9845	4.0	0.9	0.9890	0.9865
1.5	0.1	0.5185	0.9235	3.0	0.1	0.6915	0.9785	4.5	0.1	0.7475	0.9950
1.5	0.2	0.5115	0.8170	3.0	0.2	0.6525	0.8840	4.5	0.2	0.7605	0.9420
1.5	0.3	0.5240	0.6930	3.0	0.3	0.6250	0.7330	4.5	0.3	0.7230	0.8180
1.5	0.4	0.5380	0.6215	3.0	0.4	0.5540	0.5730	4.5	0.4	0.6010	0.6390
1.5	0.5	0.5645	0.6300	**3.0**	**0.5**	**0.5080**	**0.4995**	4.5	0.5	0.5315	0.5635
1.5	0.6	0.6590	0.7180	3.0	0.6	0.5760	0.5570	4.5	0.6	0.5215	0.6030
1.5	0.7	0.7875	0.8435	3.0	0.7	0.7265	0.7335	4.5	0.7	0.6780	0.7800
1.5	0.8	0.9180	0.9385	3.0	0.8	0.8925	0.9015	4.5	0.8	0.8865	0.9115
1.5	0.9	0.9870	0.9895	3.0	0.9	0.9845	0.9810	4.5	0.9	0.9805	0.9845
2.0	0.1	0.6500	0.9485	3.5	0.1	0.7190	0.9850	5.0	0.1	0.7765	0.9960
2.0	0.2	0.5765	0.8455	3.5	0.2	0.6850	0.9030	5.0	0.2	0.7800	0.9630
2.0	0.3	0.5135	0.6835	3.5	0.3	0.6445	0.7445	5.0	0.3	0.7425	0.8600
2.0	0.4	0.4970	0.5630	3.5	0.4	0.5645	0.5880	5.0	0.4	0.6395	0.6915
2.0	0.5	0.5385	0.5525	3.5	0.5	0.5185	0.5210	5.0	0.5	0.5435	0.6015
2.0	0.6	0.6025	0.6285	3.5	0.6	0.5470	0.5870	5.0	0.6	0.5160	0.6155
2.0	0.7	0.7660	0.7920	3.5	0.7	0.7175	0.7395	5.0	0.7	0.7630	0.7725
2.0	0.8	0.9050	0.9220	3.5	0.8	0.8910	0.9020	5.0	0.8	0.9020	0.9180
2.0	0.9	0.9790	0.9885	3.5	0.9	0.9880	0.9860	5.0	0.9	0.9830	0.9855

$\lambda_2 = 3.0$ 이고 $p_2 = 0.5$일 때 λ_1과 p_1 역시 각각 3.0과 0.5로 동일하므로 Y_1과 Y_2는 같은 분포에서 난수가 발생된다. Y_1과 Y_2가 같은 분포에서 발생되었을 경우에 Y가 X_1과 X_2로 분리될 확률이 같아야 타당하다. <표 4.1>에서 $\lambda_2 = 3.0$ 이고 $p_2 = 0.5$일 때, 영과잉 회귀나무와 CART의 분리기준에서 X_1의 선택확률은 약 50%이다. 따라서 모의실험이 제대로 수행되고 있음을 알 수 있다.

모의실험 결과를 살펴보면 X_1이 선택되어 분리 될 확률은 영의 확률 p_2에 따라 다르게 나타난다. <그림 4.1>는 λ_2와 p_2의 변화에 대해 X_1의 선택확률을 3차원에 플롯 한 것이다. <그림 4.1>에서 p_2값이 클 때 두 분리기준이 X_1으로 분리 될 확률은 비슷하며, λ_2값이 커질수록 X_1의 선택확률은 커지지만 p_2에 비해 큰 영향은 없는 것으로 나타난다. p_2값이 작을 때 영과잉 회귀나무 분리기준은 X_1으로 분리될 확률이 CART 분리기준 보다 낮게 나타난다.

영과잉 회귀나무 분리기준	CART 분리기준
	

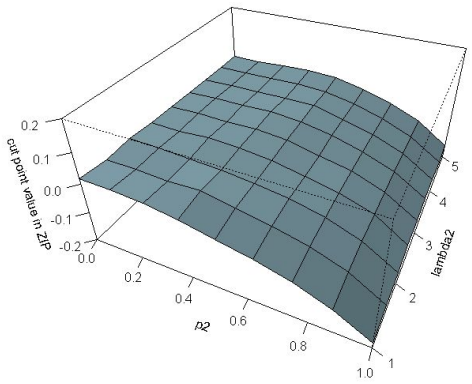
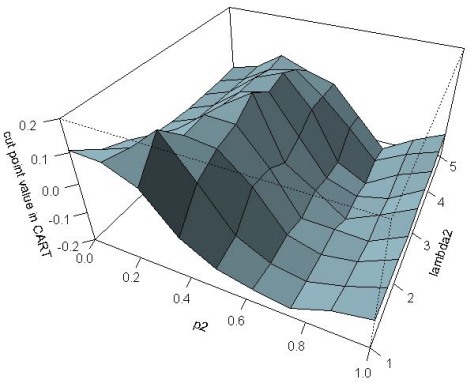
<그림 4.1> λ_2와 p_2의 변화에 대한 X_1 선택확률의 3차원 플롯

< 분리점 >

영과잉 회귀나무와 CART의 분리기준이 영에서 분리가 일어나는지 확인해 보자. 영과잉 회귀나무의 분리점은 Chaudhuri 등이 제안한 두 그룹의 평균값이다. <표 4.2>은 영과잉 회귀나무와 CART의 분리기준이 영에서 분리가 되는지를 확인하기 위하여 X_1이 분리변수로 선택 되었을 때의 분리점을 나타낸 것이다. <그림 4.2>는 <표 4.2>에 나타난 λ_2와 p_2의 변화에 대한 분리점을 3차원에 표현 한 것이다. <그림 4.2>에서 나타나는 영과잉 회귀나무 분리기준의 특징은 영 근처에서 분리가 일어나다가 p_2가 커지면서 영에서 조금 떨어진 곳에서 분리되는 양상을 보인다. 반면 CART 분리기준의 경우 p_2가 0.5일 때를 제외하고는 영에서 다소 떨어진 곳에서 분리가 일어난다. X_1의 선택확률과 분리점의 결과를 종합하여 볼 때 CART는 p_2가 0.5보다 작을 때와 높을 때 X_1의 선택확률은 높지만 영 근처에서 분리를 해내지 못한다. 반면 영과잉 회귀나무 분리기준에서는 p_2가 작을 때 CART 분리기준에 비해 X_1의 선택확률은 낮지만 이때의 분리점은 CART 분리기준을 적용했을 때 보다 영 근처에서 분리가 일어났음을 알 수 있다. 따라서 영과잉 자료에 대해 영과잉 회귀나무는 CART보다 분리변수 선택의 정확성은 떨어지지만, 정확성을 잃어버리지 않은 범위 내에서 효율적으로 분리를 한다.

<표 4.2> 영과잉 회귀나무와 CART 분리기준의 분리점

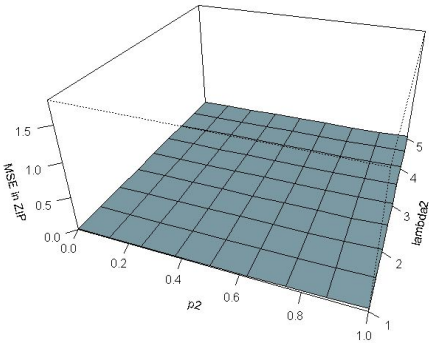
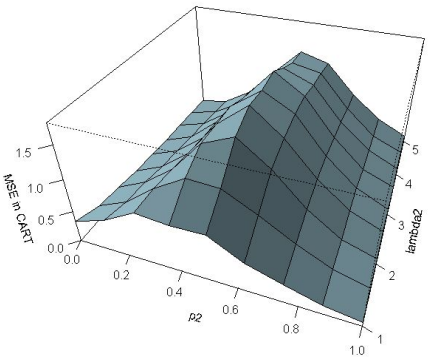
λ_2	p_2	분리점		λ_2	p_2	분리점		λ_2	p_2	분리점	
		영과잉 회귀나무	CART			영과잉 회귀나무	CART			영과잉 회귀나무	CART
1.0	0.1	0.0179	0.1086	3.0	0.6	0.0139	-0.0327	4.0	0.1	0.0123	-0.0452
1.0	0.2	0.0179	0.1042	3.0	0.7	0.0112	0.0193	4.0	0.2	0.0127	-0.0102
1.0	0.3	0.0113	0.0586	3.0	0.8	0.0160	0.0951	4.0	0.3	0.0138	0.0650
1.0	0.4	-0.0051	-0.0662	3.0	0.9	0.0098	0.0900	4.0	0.4	0.0126	0.1032
1.0	0.5	-0.0270	-0.1289	2.5	0.1	-0.0018	-0.0279	4.0	0.5	0.0109	-0.0031
1.0	0.6	-0.0480	-0.1531	2.5	0.2	-0.0228	-0.1756	4.0	0.6	-0.0147	-0.1823
1.0	0.7	-0.0760	-0.1699	2.5	0.3	-0.0442	-0.1778	4.0	0.7	-0.0409	-0.2367
1.0	0.8	-0.1176	-0.1211	2.5	0.4	-0.0832	-0.1827	4.0	0.8	-0.0826	-0.1664
1.0	0.9	-0.1808	-0.0970	2.5	0.5	-0.1537	-0.1207	4.0	0.9	-0.1510	-0.1203
1.5	0.1	0.0120	0.0444	2.5	0.6	0.0136	-0.0425	4.5	0.1	0.0128	-0.0420
1.5	0.2	0.0109	0.0803	2.5	0.7	0.0163	-0.0044	4.5	0.2	0.0123	-0.0125
1.5	0.3	0.0061	0.1644	2.5	0.8	0.0133	0.0733	4.5	0.3	0.0132	0.0458
1.5	0.4	0.0099	0.0357	2.5	0.9	0.0136	0.1369	4.5	0.4	0.0145	0.0696
1.5	0.5	-0.0108	-0.1048	3.0	0.1	0.0041	-0.0791	4.5	0.5	0.0083	-0.0080
1.5	0.6	-0.0306	-0.1617	3.0	0.2	-0.0217	-0.1492	4.5	0.6	-0.0101	-0.1056
1.5	0.7	-0.0586	-0.1813	3.0	0.3	-0.0428	-0.2250	4.5	0.7	-0.0409	-0.2122
1.5	0.8	-0.0984	-0.1489	3.0	0.4	-0.0838	-0.1707	4.5	0.8	-0.0829	-0.1547
1.5	0.9	-0.1645	-0.1029	3.0	0.5	-0.1528	-0.1070	4.5	0.9	-0.1540	-0.1118
2.0	0.1	0.0185	0.0140	3.5	0.1	0.0139	-0.0487	5.0	0.1	0.0115	-0.0326
2.0	0.2	0.0154	0.0604	3.5	0.2	0.0159	-0.0092	5.0	0.2	0.0119	-0.0065
2.0	0.3	0.0092	0.0879	3.5	0.3	0.0147	0.0653	5.0	0.3	0.0162	0.0576
2.0	0.4	0.0027	0.0733	3.5	0.4	0.0174	0.0890	5.0	0.4	0.0196	0.0204
2.0	0.5	-0.0068	-0.0185	3.5	0.5	0.0077	-0.0473	5.0	0.5	0.0125	-0.0120
2.0	0.6	-0.0209	-0.1504	3.5	0.6	-0.0132	-0.1747	5.0	0.6	-0.0099	-0.1177
2.0	0.7	-0.0493	-0.1814	3.5	0.7	-0.0424	-0.2127	5.0	0.7	-0.0388	-0.2267
2.0	0.8	-0.0899	-0.1622	3.5	0.8	-0.0839	-0.1853	5.0	0.8	-0.0817	-0.1587
2.0	0.9	-0.1556	-0.1090	3.5	0.9	-0.1490	-0.1165	5.0	0.9	-0.1535	-0.1096

영과잉 회귀나무 분리기준	CART 분리기준
	

<그림 4.2> λ_2와 p_2 의 변화에 대한 분리점

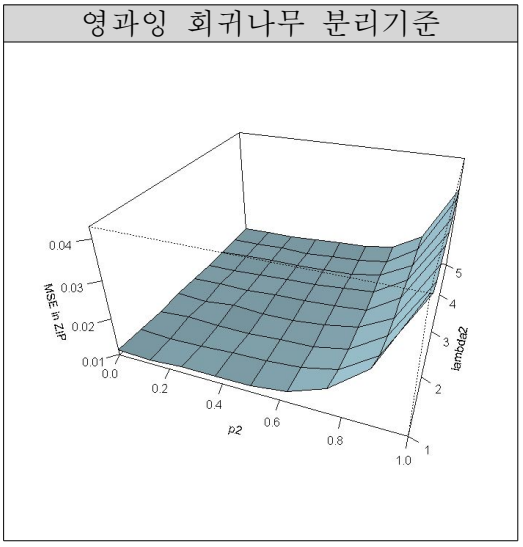
< 분리점의 MSE >

λ_2와 p_2의 변화에 대해 분리가 일어난 점과 실제 분리점을 알 수 있으므로 평균제곱오차(Mean Square Error)를 이용하여 두 분리기준의 오차를 비교할 수 있다. <그림 4.3>는 두 분리기준에 대하여 λ_2와 p_2의 변화에 따라 분리점에 대한 평균제곱오차를 3차원에 플롯 한 것이다. <그림 4.3>에서 영과잉 회귀나무 분리기준은 CART 분리기준에 비해 평균제곱오차가 확연히 낮음을 볼 수 있다.

영과잉 회귀나무 분리기준	CART 분리기준
	

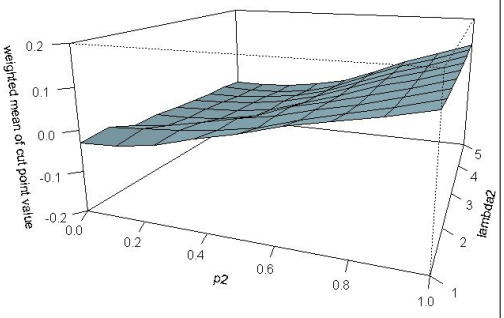
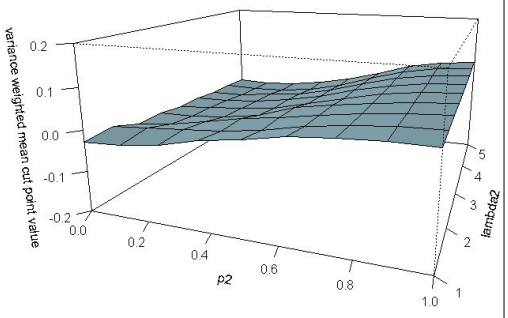
<그림 4.3> 분리점에 대한 MSE

<그림 4.3>의 영과잉 회귀나무 분리기준에서 분리점에 대한 MSE를 확대하여 보면 <그림 4.4>와 같이 p_2값이 커짐에 따라 분리점이 영에서 약간 멀어지는 것을 확인할 수 있다.

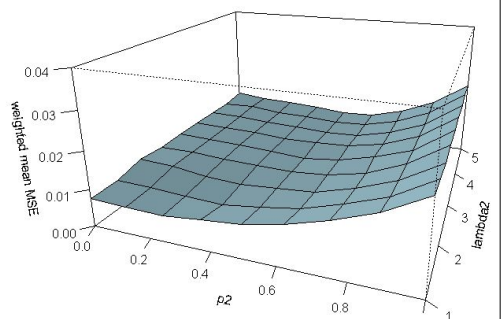
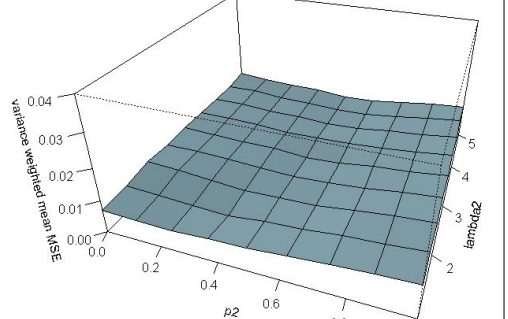


<그림 4.4> 확대한 영과잉 회귀나무의 분리점에 대한 MSE

<그림 4.2>의 영과잉 회귀나무 분리기준에서 p_2값이 커짐에 따라 분리점이 영에서 약간 멀어지는 것을 보완하기 위하여 가중평균 분리점을 고려해 볼 수 있다. <그림 4.5>는 관찰치 수와 분산을 고려한 가중평균 분리점을 사용하여 X_1이 영에서 분리가 일어나는지를 3차원 플롯으로 나타낸 것이다. <그림 4.6>은 가중평균 분리점에 대한 MSE를 나타낸 것이다. <그림 4.5>에서 가중평균의 분리점이 <그림 4.2>의 영과잉 회귀나무 분리점에 비하여 영 근처에서 보다 잘 분리되고 있으며, 특히 분산을 고려한 가중평균 분리점이 영에서 가장 잘 분리가 된다. 가중평균 분리점의 평균제곱오차 또한 그 값들이 굉장히 작으며, 관찰치 수를 고려한 가중평균보다는 분산을 고려한 가중평균의 평균제곱오차가 특히 작다. 따라서 Chaudhuri 등이 제안한 단순평균 분리점 보다는 가중평균을 이용한 분리점이 다소 안정적으로 분리됨을 알 수 있다.

관찰치 수를 고려한 가중평균	분산을 고려한 가중평균
	

<그림 4.5> 가중평균 분리점

관찰치 수를 고려한 가중평균	분산을 고려한 가중평균
	

<그림 4.6> 가중평균 분리점에 대한 MSE

4.2 실제자료의 적용

4.2.1 실제자료 소개

적용 자료는 통신사 고객의 사용행태 자료로 통신비 연체 건수를 목표변수로 고객의 사용행태에 따라 연체가 어떻게 발생하는지를 의사결정나무를 이용하여 분류, 예측하고자 한다. 자료에는 통신비 연체 건수, 청구금액, 가입기간, 정지 관련문의건수, 부가서비스 가입, 휴대전화 평균사용, 현재 휴대전화 사용, 인터넷 사용, 무통화가 포함되어있다. 자료의 수는 통신사를 이탈한 고객을 제외한 47,227개이다. 자세한 변수설명은 <표 4.3>와 같다.

<표 4.3> 자료의 변수 정의

변수이름		변수설명
목표변수	연체건수	통신비 연체건수
예측변수	청구금액	3개월 평균 청구금액, 단위 $
	가입기간	가입기간, 단위 일
	정지 관련문의	최근 3개월간 정지 관련문의건수
	부가서비스 가입	부가서비스 가입 건수
	휴대전화 평균 사용	휴대전화 평균사용일수, 단위 일
	현재 휴대전화 사용	현재 휴대전화 사용일수, 단위 일
	인터넷 사용	6개월 평균 인터넷 사용일수, 단위 일
	무통화	6개월간 무통화 월수, 단위 월

연체건수의 분포가 포아송 분포를 따르는 지에 대해 카이제곱 검정으로 적합도를 검정해 본 결과 자유도가 6이고 카이제곱 통계량이 205678로 값이 매우 크고, p값이 0.001로 유의수준 5%하에서 포아송 분포를 따른다는 귀무가설을 기각하므로 목표변수가 포아송 분포를 따른다고 할 수 없다. 따라서 목표변수의 관측값에 영이 많이 나타나는 경우에 적합한 영과잉 분포를 사용하였다. <표 4.4>는 연체건수에 대한 빈도표로 다른 연체 발생건수에 비하여 통신비 연체 경험이 없는 경우가 87.24%로 과잉되어 있음을 확인할 수 있다. 또한 연체건수가 6이상일 때 관측빈도가 다른 연체건수 관측빈도 보다 큰 것은 실제자료의 한계점이라 할 수 있다.

<표 4.4> 통신비 연체 건수 빈도표

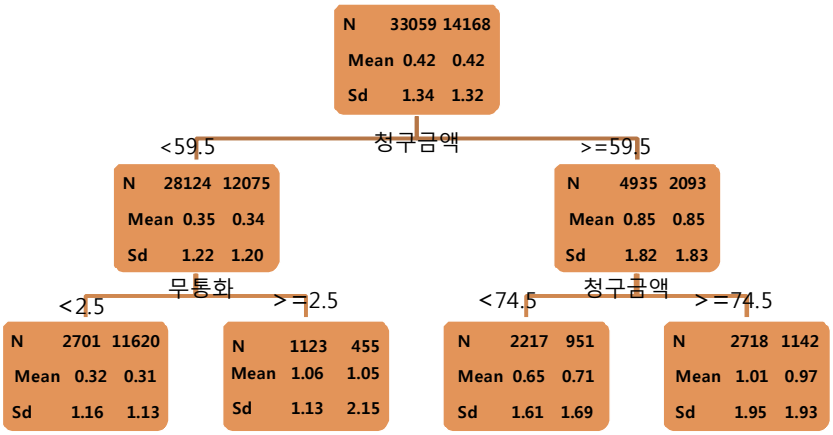
연체건수	관측빈도	백분율
0	41202	87.24
1	1973	4.18
2	883	1.87
3	569	1.20
4	442	0.89
5	338	0.70
6 이상	1840	3.90
합계	47227	100

4.2.2 전통적인 의사결정나무의 분석 결과

연체건수를 목표변수(target)로 8개의 예측변수를 입력변수(input)로 하여 CHAID, CART, C4.5를 이용하여 의사결정나무 분석을 수행하였다. 데이터는 훈련용 데이터 70%, 평가용 데이터 30%로 분할하여 나무모형을

개발하고 적용하였다. 그러나 CHAID와 C4.5의 경우 연체발생 빈도가 상당히 작아 분리가 일어나지 않는다. 반면 CART의 경우에만 의사결정나무모형이 도출된다. <그림 4.6>은 연체건수를 연속형 예측변수로 하여 구축한 의사결정나무 모형이다.

<그림 4.6>에서 통신비 연체는 3개월 평균 청구금액, 무통화 월수에 의해 최적분리가 이루어졌다. 첫 분리변수인 청구금액이 59.5$보다 큰 경우 평균값은 0.85이고 표준편차는 1.83으로 연체건수를 0.85로 예측하고 있으며, 청구금액이 높은 경우 청구금액이 낮은 경우에 비해 연체가 다소 높게 나타난다. 또한 청구금액이 59.5$보다 낮은 경우에 무통화가 2회 이상인 경우 연체건수의 평균이 1.05로 높은 연체를 보인다.



<그림 4.6> CART에 의한 의사결정나무

CART는 끝마디가 4개, 최종마디는 7개, 나무 깊이는 2단계로 나타난다. CART 나무모형의 설명력은 24.08% 다소 낮다.

4.2.2 영과잉 회귀나무의 분석 결과

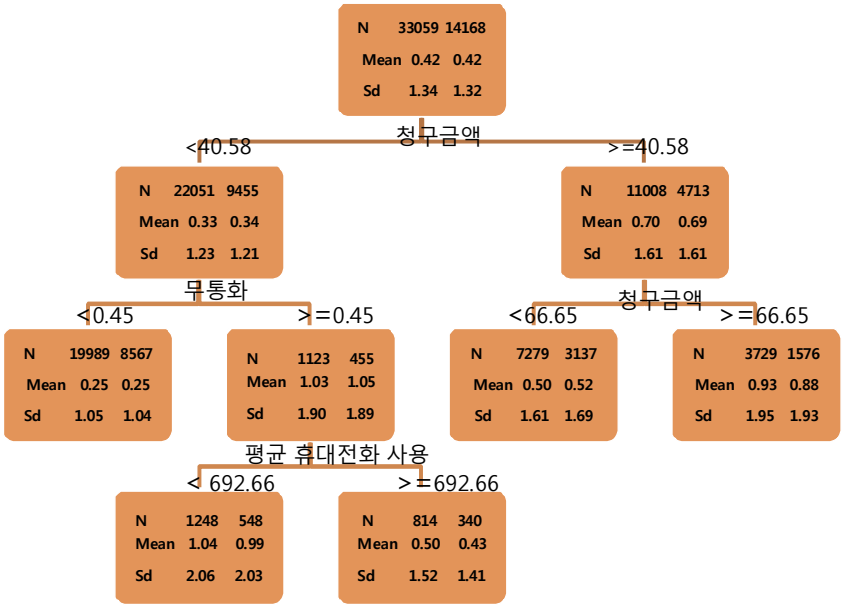
통신사의 연체건수를 자료를 영과잉 포아송 회귀모형에 적합 시킨 후 의사결정나무를 도출하였다. <표 4.4>는 영과잉 포아송 회귀 모형에 적합 시켜 얻은 추정계수와 표준오차 값이며 <그림 4.7>은 Chaudhuri가 제안한 방법론을 적용하여 도출한 영과잉 포아송 회귀나무이다. <그림 4.8>은 가중평균을 고려한 영과잉 회귀나무로 본 논문에서 제안한 방법론 중 분산을 고려한 가중평균 분리점을 사용하여 도출한 것이다.

<표 4.5> 영과잉 포아송 회귀 모수 추정값

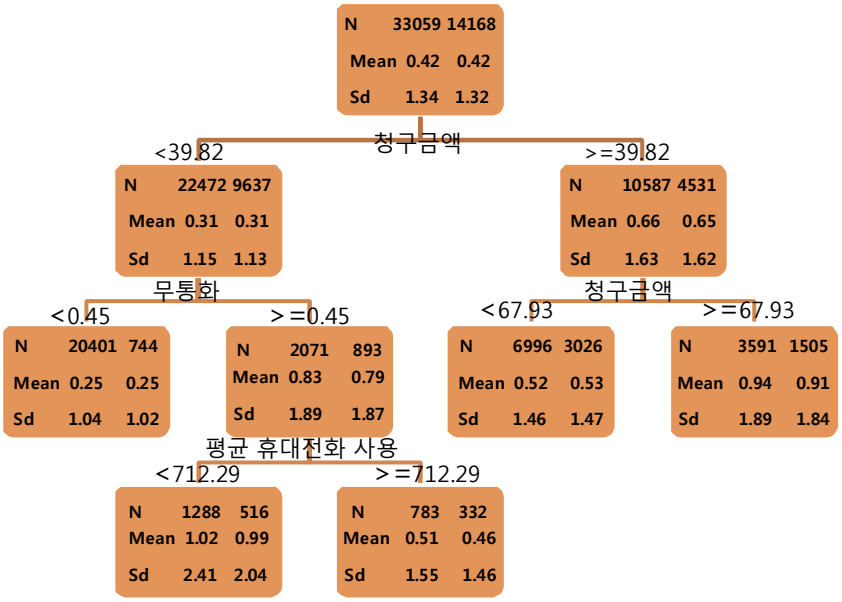
모수($\hat{\lambda}$)	추정계수	표준오차	$Pr>\|z\|$
상수	0.9867	0.01758	<0.0001
청구금액	0.0013	0.00022	<0.0001
가입기간	0.00002	0.00005	0.0007
정지 관련문의	-0.0151	0.02485	0.5415
부가서비스 가입	0.0405	0.00915	<0.0001
휴대전화 평균 사용	-0.0001	0.00002	0.0004
현재 휴대전화 사용	0.0001	0.00002	<0.0001
인터넷 사용	0.0014	0.00318	0.6569
무통화	0.0778	0.00559	<0.0001

모수($\hat{p}$)	추정계수	표준오차	$Pr>\|z\|$
상수	2.3410	0.03643	<0.0001
청구금액	-0.0001	0.00049	<0.0001
가입기간	0.0001	0.00002	<0.0001
정지 관련문의	-0.0792	0.04618	0.0865
부가서비스 가입	-0.1783	0.01753	<0.0001
휴대전화 평균 사용	0.0002	0.00005	0.0002
현재 휴대전화 사용	0.0001	0.00005	0.0239
인터넷 사용	-0.0021	0.00587	0.7193
무통화	-0.2568	0.01244	<0.0001

영과잉 포아송 회귀모형에서는 정지관련 문의건수와 인터넷 사용일수를 제외한 모든 변수가 유의한 값을 지니며, 통신비 연체 건수에 영향을 끼치는 변수라 할 수 있다. 의사결정나무 모형에서는 청구금액, 무통화 월수, 평균 휴대전화 사용일수로 최적분리가 일어난다. <그림 4.7>의 영과잉 회귀나무는 CART에 의한 의사결정나무모형과 비슷한 나무 구조를 가지고 있다. 청구금액이 40.58$보다 낮을 때, 현재 휴대전화 사용이 2년 미만이며 무통화 월수가 0.5이상인 경우의 통신비 연체가 높아지며, 청구금액이 66.65$ 이상일 때도 통신비 연체가 높아진다. 영과잉 회귀나무는 끝마디가 5개, 최종마디는 9개, 나무 깊이는 3단계로 CART에 의한 의사결정나무보다 약간 큰 나무를 형성하고 있다. 영과잉 회귀나무의 모형 설명력은 24.11%로 CART에 비하여 조금 높다. <그림 4.8>의 가중평균을 고려한 영과잉 회귀나무는 청구금액이 39.82$보다 낮을 때 현재 휴대전화 사용이 2년 미만이며 무통화 월수가 0.5이상인 경우의 통신비 연체가 높아지며, 청구금액이 67.93$ 이상일 때 또한 통신비 연체가 높아진다. 영과잉 회귀나무는 끝마디가 5개, 최종마디는 9개, 나무 깊이는 3단계로 CART에 의한 의사결정나무보다 약간 큰 나무를 형성하고 있다. 영과잉 회귀나무의 모형 설명력은 25.07%로 영과잉 회귀나무와 CART에 비하여 모형 설명력이 좋다.



<그림 4.7> 영과잉 회귀나무



<그림 4.8> 가중평균을 고려한 영과잉 회귀나무

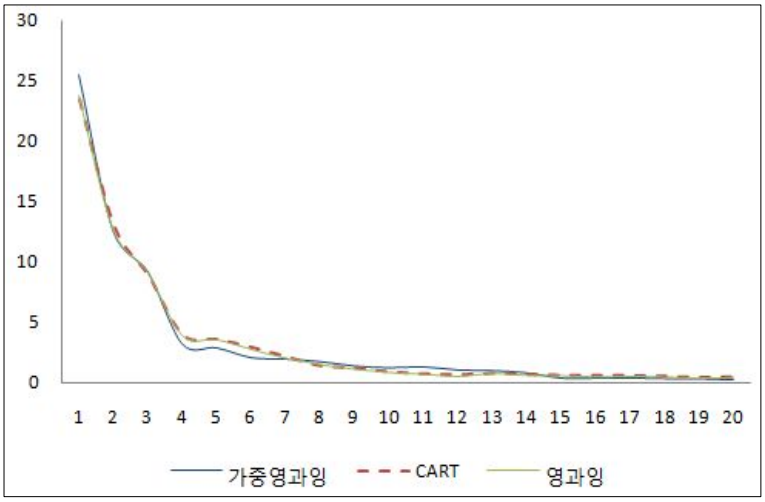
연체가 없는 경우는 $Y=0$, 연체가 발생한 경우를 $Y=1$로 하고, 각 개체에 대한 예측값으로 CART, 영과잉 회귀나무, 가중평균을 고려한 영과잉 회귀나무의 예측모형을 평가하였다. 그 결과는 <표 4.6>의 요약표로 나타난다.

<표 4.6> 누적 리프트 테이블

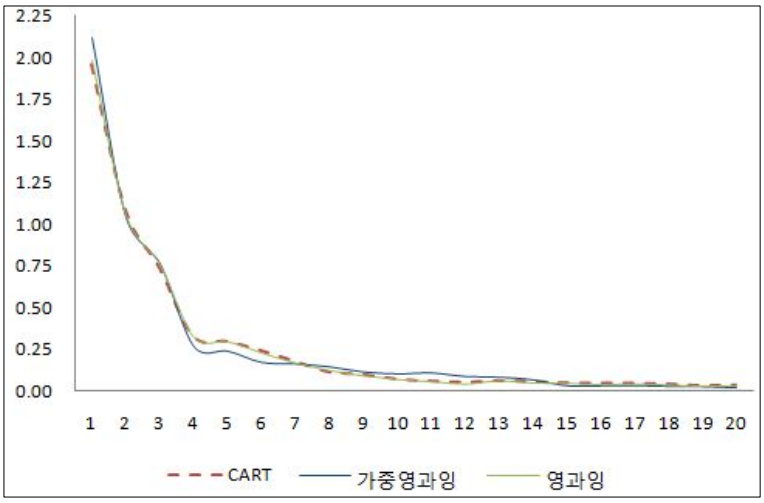
등급	빈도	CART의 반응률		영과잉 회귀나무의 반응률		가중평균을고려한 영과잉회귀나무의반응률	
	Y=1	%RESPONSE	Life	%RESPONSE	Life	%RESPONSE	Life
1	166	23.446	1.954	23.729	1.977	25.424	2.119
2	184	12.994	1.083	12.641	1.053	12.500	1.042
3	188	8.851	0.738	9.087	0.757	9.181	0.765
4	111	3.919	0.327	3.919	0.327	3.213	0.268
5	126	3.559	0.297	3.588	0.299	2.881	0.240
6	123	2.895	0.241	2.778	0.231	2.072	0.173
7	105	2.119	0.177	2.058	0.172	1.957	0.163
8	77	1.359	0.113	1.518	0.127	1.748	0.146
9	75	1.177	0.098	1.146	0.095	1.381	0.115
10	60	0.847	0.071	0.876	0.073	1.243	0.104
11	55	0.706	0.059	0.732	0.061	1.310	0.109
12	53	0.624	0.052	0.553	0.046	1.059	0.088
13	67	0.728	0.061	0.782	0.065	1.000	0.083
14	66	0.666	0.055	0.626	0.052	0.817	0.068
15	61	0.574	0.048	0.603	0.050	0.386	0.032
16	62	0.547	0.046	0.512	0.043	0.380	0.032
17	68	0.565	0.047	0.557	0.046	0.399	0.033
18	63	0.494	0.041	0.487	0.041	0.345	0.029
19	52	0.387	0.032	0.401	0.033	0.327	0.027
20	62	0.438	0.036	0.438	0.036	0.240	0.020

<그림 4.9>는 CART, 영과잉 회귀나무, 가중평균을 고려한 영과잉 회귀나무 모형들에 대한 연체건수의 반응률을 나타낸 것이다. 상위등급에서 높은 반응률을 보이고, 하위 등급에서는 굉장히 낮은 반응률을 보이므로 예측모형의 성능이 좋음을 알 수 있다. 상위 5%에서 CART는 약 23.45%, 영과잉 회귀나무는 23.73%, 가중평균을 고려한 영과잉 회귀나무에 약 25.42%의 연체자 포함되어 있다. 가중평균을 고려한 영과잉 회귀나무의 반응률이 영과잉 회귀나무와 CART 나무모형 보다 약간 좋은 반응률을 보인

다. <그림 4.10>는 기준 반응률에 비해 각 등급의 반응률이 얼마나 높은지 나타낸 리프트 그래프이다. 1등급에서는 기준 반응률 12%에 비해 약 2배 정도의 향상도를 보이며 가중평균을 고려한 영과잉 회귀나무의 향상도는 영과잉 회귀나무모형과 CART 나무모형 보다 더 좋다.



<그림 4.9> 예측모형의 반응률(%Response)



<그림 4.10> 리프트 그래프

제5장 결론 및 향후 연구과제

본 연구에서 영과잉 이산형 자료를 위한 의사결정나무분석을 수행하기 위하여 Chaudhuri 등(1995)이 제안한 일반화 회귀나무를 이용하여 가중평균을 고려한 영과잉 회귀나무를 제안하고 영과잉 이산형 자료를 분리해 보았다.

전통적으로 많이 쓰이는 CART의 분리기준과 가중평균을 고려한 영과잉 회귀나무의 분리기준을 이용하여 참 분리변수의 선택확률과 분리점에 대해 모의실험을 수행하였다. 모의실험 결과, 가중평균을 고려한 영과잉 회귀나무의 분리기준은 영의 값이 나타나는 확률(p_2)이 낮을 때 CART의 분리기준에 비해 참 분리변수 선택의 정확도가 떨어지지만 p_2가 커질수록 참 분리변수 선택의 정확도는 CART와 비슷하게 나타난다. 분리점이 영에서 분리되는지를 검토했을 때 CART와 Chaudhuri 등이 제안한 분리점보다는 가중평균을 고려한 분리점이 영 근처에서 더 정확하게 분리를 해냈다.

또한 실제 통신사의 통신비 연체건수 자료를 가지고 의사결정나무분석을 수행하였다. 그 결과 CHAID와 C4.5에서는 분리가 일어나지 않았으며 CART, 영과잉 회귀나무와 가중평균을 고려한 영과잉 회귀나무에서는 분리가 일어났다. 예측모형 평가 과정에서 가중평균을 고려한 영과잉 회귀나무 모형은 영과잉 회귀나무 모형과 CART 나무모형 보다 설명력(R^2)과 반응률(%Response)이 좋아서 가중평균을 고려한 영과잉 회귀나무 모형이 영과잉 회귀나무 모형과 CART 나무모형 보다 우수한 모형으로 나타났다.

본 연구에서 영과잉 이산형 자료에 대해 제안한 회귀나무는 전통적인 의사결정나무와 영과잉 회귀나무보다 효율적인 방법론인 것으로 판단된다. 나아가 본 연구에서는 독립변수가 연속형인 경우에만 고려하였으나 향후 범주형으로 확대하는 연구가 필요할 것으로 판단된다.

참 고 문 헌

[1] 이영섭(2003). Interesting Node Finding Criteria for Regression Trees, *Korean journal of applied statistics,* Vol.16, No.1, 45-53.

[2] 최종후, 한상태, 강현철, 김은석, 이성건 (2002). AnswerTree3.0을 이용한 데이터마이닝 예측 및 활용, SPSS 아카데미.

[3] 최창호, 안동환 (2010). 산업별 창업기업의 입지결정요인 분석, *대한국토·도시계획학회지*, 제 45권, 제 2호, 193-205.

[4] Breiman, L., Friedman J., Olshen R. A., & Stone, C. J. (1984). *Classification and Regression Tree*, Chapmanand Hall, New York, NY.

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

[6] Chaudhuri, P., Lo, W., Loh, W. Y. & Yang, C.C. (1995). Generalized Regression Trees, *Statistica Sinica*, Vol.5, 641-666.

[7] Chaudhuri, P., Huang, M-C., Loh, W. Y., Yao, R. (1994). Picewise-Polynomial Regression Trees, *Statistica Sinica,* Vol.4, 143-167.

[8] Lambert, D. (1992). Zero-Inflated Poisson regression, with an Application to Detects in Manufacturing, *Technometrics*, Vol.34, 1-14.

[9] Lee, S. B. & Park, S. Y. (2004). SUPPORT Application for Classification Trees, *Journal of Korean Data & Information Science Society*. Vol.15, No.3, 565-574.

[10] McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models*, 2nd edition, Chapman and Hall, London.

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

[12] Thearling, K. (1995). From Data Mining to Database Marketing, DIG White Paper.

ABSTRACT

A study on Decision Tree for Zero-Inflated Count Data

Bomi Choi
Department of Statistics
The Graduate School
Sungshin Women's University

Zero-inflated Poisson distribution is more applicable distribution in case of excess zeros than regular Poisson probability distribution as a discrete type probability distribution. The data of that has excess zeros come out in various fields, and there are many studies on the data. But it hard to find decision trees for zero-inflated count data. Occasionally conventional methods cannot be achieved for the data. So it is needed to build up a decision tree for zero-inflated count data.

In this paper, we establish a decision tree for zero-inflated count data. It will be constructed on basis proposed split criterion with Chaudhuri et al.(1995)'s generalized regression tree. We suggest a split point using weighted mean instead of mean. Simulation is performed to demonstrate the performance of the proposed method and to show the competitive decision tree. Finally we illustrate the proposed methodology with real data.

감사의 글

주옥같은 2년의 시간이 지나 졸업이라는 문 앞에 섰습니다. 힘든 시간들도 있었지만 따뜻한 관심과 애정 어린 질책 속에서 한층 성숙해질 수 있었습니다. 논문이라는 작은 결실을 맺을 수 있도록 도움을 주시고 응원해 주신 모든 분들께 감사의 마음을 전합니다.

먼저 논문의 시작에서부터 완성되기까지 많은 관심과 끊임없는 격려로 이끌어주시고 세심하게 지도해주신 이성건 지도교수님께 존경과 감사의 뜻을 전합니다. 항상 인자한 웃음으로 진리를 깨우쳐 주신 송일성 교수님, 따뜻한 격려의 말로 용기를 주신 이해용 교수님, 언제나 웃음을 잃지 않도록 해 주신 이우선 교수님, 따뜻한 사랑의 충고로 열심히 꾸짖어 주신 이종협 교수님, 배움의 열정과 노력을 깨닫게 해 주신 박만식 교수님께 진심으로 감사드립니다.

바쁜 와중에도 부족한 후배에게 많은 관심을 가져준 귀영언니, 향선언니, 아낌없이 충고해 주며 당당히 앞장서 나간 영은언니, 희라언니, 주현언니, 애란언니, 인경언니, 희원언니, 경혜언니에게 감사의 마음을 전합니다. 2년간의 대학원 생활을 동고동락한 정윤언니와 하얀, 명희언니, 슬지, 소영에게도 고마움을 전합니다.

기도로 후원해 주신 유영규 목사님, 이동순 담임목사님, 최동윤 목사님, 인옥경 전도사님께 감사드립니다. 힘들 때마다 웃게 해준 태양, 주원, 명준오빠, 선화언니, 지혜언니, 에이레네 식구들, 마하나임 중창단, 금요찬양단에게 미안한 마음과 감사의 마음을 전합니다.

마지막으로 한결같은 사랑과 정성으로 가장 가까운 곳에서 지켜봐주시고 기도해주신 사랑하는 부모님께 감사드립니다. 귀여운 아우님 재희에게도 고마움을 전합니다.

그 외에도 도움주신 많은 분들께 감사드립니다. 넓고 깊은 사람이 되도록 노력하겠습니다.