Journal of Korean
Data & Information Science Society
2005, Vol. 16, No. 2, pp. 195~205

A Study on Split Variable Selection Using Transformation of Variables in Decision Trees[1)]

Sung S. Chung[2)] · Ki H. Lee[3)] · Seung S. Lee[4)]

Abstract

In decision tree analysis, C4.5 and CART algorithm have some problems of computational complexity and bias on variable selection. But QUEST algorithm solves these problems by dividing the step of variable selection and split point selection.

When input variables are continuous, QUEST algorithm uses ANOVA F-test under the assumption of normality and homogeneity of variances. In this paper, we investigate the influence of violation of normality assumption and effect of the transformation of variables in the QUEST algorithm. In the simulation study, we obtained the empirical powers of variable selection and the empirical bias of variable selection after transformation of variables having various type of underlying distributions.

Keywords : Decision Tree, Normality, QUEST, Transformation

1. 서론

데이터마이닝은 통계학, 수학, 전산학 등의 다양한 방법을 이용하여 대용량의 자료에서 변수 또는 자료간의 관계, 패턴, 성향, 규칙 등을 찾아내 모형화하고 이로부터 의미 있는 정보를 추출해내는 일련의 과정이다. 데이터마이닝 기법 중 대표적인 방법의 하나인 의사결정나무 기법은 나무 구조를 이용하여 자료의 분류와 예측을 하는 방법으로, CHAID(Kass, 1980), CART(Breiman et al., 1984), C4.5(Quinlan, 1993), QUEST(Loh and Shih, 1997) 알고리즘 등이 주로 사용되고 있다.

---

1) 본 연구는 과학기술부 주관 특정연구개발과제의 연구비 지원에 의해 수행되었음(과제번호 : M10427020001-04L2702-00110)
2) 제1저자 : 전라북도 전주시 덕진구 덕진동 1가 644-14, 전북대학교 통계정보과학과 교수
E-mail: sschung@chonbuk.ac.kr
3) 전라북도 전주시 완산구 효자동 3가 1200, 전주대학교 경영학부 교수
E-mail: khlee@jj.ac.kr
4) 전라북도 전주시 덕진구 덕진동 1가 644-14, 전북대학교 통계정보과학과 박사과정
E-mail: pinokio129@chonbuk.ac.kr

이 중 C4.5와 CART는 모든 가능한 입력변수에서 순수도(purity measure)가 최대가 되는 변수와 분리점(split point)을 동시에 선택해 나무 구조를 생성해 나가는 방법으로, 입력변수의 형태에 따라 변수의 선택비율이 달라지고 분리 시간도 많이 걸린다고 알려져 있다(Loh and Shih,1997). 반면 QUEST에서는 먼저 유의한 변수를 선택하고, 선택된 변수에 대해서 이차판별분석(quadratic discriminant analysis)을 실시하여 분리점을 선택하는 단계적 수행으로 변수 선택 편의(bias)와 나무 생성시간을 현저하게 감소시켰다.

QUEST는 분리변수 선택 시 연속형 변수에 대해서는 분포의 정규성을 가정하는 ANOVA F-검정, 범주형 변수에 대해서는 카이제곱검정을 사용하기 때문에, 연속형 변수가 정규성 가정에 심각하게 위배되는 경우에 변수 선택력이 떨어지는 문제가 발생한다. 정성석 등(2004)은 연속형 변수가 정규성 가정에 심각하게 위배되는 경우에, 연속형 변수를 범주형 변수로 변환하여 QUEST 알고리즘을 적용함으로써 변수선택 편의를 줄이고 변수 선택력을 향상시켰다.

본 연구에서는 QUEST 알고리즘을 이용하여 의사결정 나무를 생성할 때 연속형 변수의 정규성 가정이 분리 변수 선택 시 미치는 영향과 변수변환 방법을 이용하여 자료를 변환한 후 변수선택 편의와 변수 선택력의 변화를 모의실험 통하여 살펴보았다.

2. QUEST 알고리즘과 변수변환 방법

2.1 QUEST 알고리즘

QUEST(Loh와 Shih, 1997) 알고리즘은 의사결정나무 중 목표변수가 범주형인 분류나무(classification tree)에 적용하며, 입력변수의 형태에 따라 연속형 입력변수는 ANOVA F-통계량을, 범주형 입력변수는 카이제곱 검정통계량을 이용하여 유의확률(p-값)을 계산하여 분리변수를 결정해 나간다.

다음은 QUEST 알고리즘의 분리 변수 선택 알고리즘이다.

단계 1. 연속형 입력변수는 ANOVA F-검정을 이용하여 유의확률을 계산한다.

단계 2. 범주형 입력변수는 분할표를 이용하여 카이제곱 독립성 검정의 유의확률을 계산한다.

단계 3. 단계 1, 단계 2에서 가장 작은 유의확률이 다중비교의 본페로니(Bonferroni) 임계값보다 작으면 그에 대응되는 변수를 분리변수로 선택한다.

단계 4. 그렇지 않으면 연속형 입력변수에 대해 절대편차에 기초한 ANOVA F-검정을 실시하여 가장 작은 유의확률이 수정된 본페로니 임계값보다 작으면 대응하는 새로운 변수를 분리변수로 선택한다. 그러한 변수가 없으면 단계 1, 2에서 가장 작은 유의확률을 갖는 변수를 분리변수로 선택한다.

분리변수가 선택된 후 분리점은 범주형 입력변수에 대해 크림코오드(CRIMCOORD:

Gnanadesikan, 1977)방법으로, 연속형 입력변수일 경우에는 이차판별분석(quadratic discriminant analysis)을 이용하여 결정한다.

2.2 변수변환 방법

모집단의 분포를 가정하고 유도된 통계적 분석기법은 분포 가정이 만족하지 않았을 때 비모수적인 방법을 이용하거나 자료에 일정한 변환을 실시하여 가정을 만족시킨 후 분석을 실시하는 것이 보통이다. 변수변환 방법은 기존의 자료에 적당한 멱수(power)를 적용하여 자료를 가정에 맞는 분포로 변환시켜 주는 방법으로, 멱수는 자료의 형태를 고려하여 제곱근(square root), 로그(log), 역수(inverse) 등이 주로 쓰인다. 여기서 변환 방법으로 박스-칵스((Box-Cox, 1964), 모듈러스(Modulus), 여-존슨(Yeo-Johnson, 2000) 방법들을 살펴보도록 한다.

① 박스-칵스 변환

박스-칵스에 의해 제안되어진 이 방법은 분포에 대한 모형이 비대칭적일 때 효율성이 뛰어난 변환방법으로 변환을 위한 식은

$$X^{(\lambda)}=\begin{cases}\dfrac{(X^{\lambda}-1)}{\lambda} & \text{if } \lambda\neq 0,\\ \log(X) & \text{if } \lambda=0.\end{cases}$$

이다. 그러나 이 방법은 변환 범위가 양수($X>0$)로 한정되어 실수 전 구간에서 변환이 불가능하다는 변수 범위의 제약을 지니고 있다.

② 모듈러스 변환

모듈러스 변환 방법은 John과 Draper(1980)에 의해 제안되었으며, 박스-칵스 변환 방법이 가지는 변수 범위 제약을 가지지 않는 변환방법으로 변환을 위한 식은 다음과 같다.

$$X^{(\lambda)}=\begin{cases}\text{sign}(X)\dfrac{((|X|+1)^{\lambda}-1)}{\lambda}, & \text{if } \lambda\neq 0,\\ \text{sign}(X)\log(|X|+1), & \text{if } \lambda=0.\end{cases}$$

단, $\text{sign}(X)=\begin{cases}+1, & \text{if } X\geq 0,\\ -1, & \text{if } X<0\end{cases}$ 이다.

모듈러스 변환 방법은 대칭적인 구조를 가지는 자료에서는 매우 효율이 뛰어난 방법으로 알려져 있지만, 다른 구조를 갖는 자료에 대해서는 효율이 떨어진다고 알려져 있다.

③ 여-존슨 변환

여-존슨 변환은 모듈러스 변환을 기반으로 자료의 범위에 따라 멱 모수(power parameters)에 변화를 주는 방법으로 자료의 범위에 따른 변환 식은 다음과 같다.

$$X^{(\lambda)}=\begin{cases}\dfrac{(X+1)^{\lambda}-1}{\lambda}, & \text{if } X\geq 0,\ \lambda\neq 0,\\ \log(X+1), & \text{if } X\geq 0,\ \lambda=0,\\ -\dfrac{(-X+1)^{2-\lambda}-1}{(2-\lambda)}, & \text{if } X<0,\ \lambda\neq 2,\\ -\log(-X+1), & \text{if } X<0,\ \lambda=2.\end{cases}$$

여-존슨 변환 방법은 박스-칵스 변환이 갖는 변수 범위에 대한 제약이 없고, 모듈러스 변환이 갖는 대칭성이 만족되지 않는 자료에 대해서 효율면에서 향상된 것으로 알려져 있다.

2.3 모수 추정 및 QUEST의 적용

변수변환에 이용되는 멱수(power)는 일정한 분포 하에서 생성된 독립인 n_i개의 자료 $\boldsymbol{X}_i=(X_{i1},\ldots,X_{in_i})'$, $(i=1,\ldots,l)$를 초기 값 λ을 통해 변환시켜 새로 생성된 $\boldsymbol{X}_i^{(\lambda)}=(X_{i1}^{(\lambda)},\ldots,X_{in_i}^{(\lambda)})'$, $(i=1,\ldots,l)$가 근사적으로 정규분포를 따른다는 가정 하에 추정되어진다. 즉, 변수 $\boldsymbol{x}=(x_{11},\ldots,x_{1n_1},\ldots,x_{l1},\ldots,x_{ln_l})'$가 아래와 같은 정규분포하의 가능도함수로 표시될 수 있다.

$$L_N(\boldsymbol{\theta}\mid \boldsymbol{x})=(2\pi\sigma^2)^{-N/2}\exp\Big[-\frac{1}{2\sigma^2}\sum_{i=1}^{l}\sum_{j=1}^{n_i}(x_{ij}^{(\lambda)}-\mu_i)^2\Big]J(\lambda).$$

여기서 $\boldsymbol{\theta}=(\boldsymbol{\mu}', \sigma^2, \lambda)'$, $\boldsymbol{\mu}=(\mu_1,\ldots,\mu_l)'$, $N=\sum_{i=1}^{l}n_i$ 이다.

$\boldsymbol{x}^{(\lambda)}=(x_{11}^{(\lambda)},\ldots,x_{1n_1}^{(\lambda)},\ldots,x_{l1}^{(\lambda)},\ldots,x_{ln_l}^{(\lambda)})'$일 때, 위 식에서 포함된 자코비안 값은 각 변환 방법에 따라 다음과 같이 주어진다.

$$J(\lambda)=\left|\frac{\partial \boldsymbol{x}^{(\lambda)}}{\partial \boldsymbol{x}'}\right|=\begin{cases}\prod_{i=1}^{l}\prod_{j=1}^{n_i}(x_{ij})^{(\lambda-1)I(x_{ij}>0)}, & \text{Box}-\textit{Cox Transformation}\\ \prod_{i=1}^{l}\prod_{j=1}^{n_i}(|x_{ij}|+1)^{(\lambda-1)}, & \text{Modulus } \textit{Transformation}\\ \prod_{i=1}^{l}\prod_{j=1}^{n_i}(|x_{ij}|+1)^{sign(x_{ij})(\lambda-1)}, & \text{Yeo}-\textit{Johnson Transformation}\end{cases}$$

주어진 λ에 의해 생성된 변수 $X_i^{(\lambda)}$의 평균과 분산은 다음과 같이 추정된다.

$$\widehat{\mu}_i(\lambda) = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}^{(\lambda)}. \tag{2.1}$$

$$\widehat{\sigma}^2(\lambda) = \frac{1}{N}\sum_{i=1}^{l}\sum_{j=1}^{n_i}[x_{ij}^{(\lambda)} - \widehat{\mu}_i(\lambda)]^2. \tag{2.2}$$

(2.1)과 (2.2)의 추정된 값을 이용하여 아래의 로그가능도함수가 부분 최적화시켜주는 $\widehat{\lambda}$을 찾아 그 값을 최적의 멱수로 결정한다.

$$l_{\max}(\lambda) = -\frac{N}{2}(\log(2\pi)+1) - \frac{N}{2}\log(\widehat{\sigma}(\lambda)^2) + \log J(\lambda).$$

그리고 QUEST 알고리즘에서 연속형 입력변수의 유의확률을 계산 시 이용되는 ANOVA F-검정 통계량은 다음과 같다.

$$F_N(\lambda) = (l-1)^{-1} RSS_H(\lambda) / S^2{}_N(\lambda). \tag{2.3}$$

단, $RSS_H(\lambda) = \sum_{i=1}^{l}\{\overline{x}_i^{(\lambda)} - \overline{x}^{(\lambda)}\}^2$는 F-검정 통계량에서 처리 제곱합을 나타내며, $S_N^2(\lambda) = \sum_{i=1}^{l}\sum_{j=1}^{n_i}[x_{ij}^{(\lambda)} - \overline{x}_i^{(\lambda)}]^2/(N-l)$는 잔차제곱합, $\overline{x}_i^{(\lambda)} = \sum_{j=1}^{n_i} x_{ij}^{(\lambda)}/n_i$는 범주별 평균, $\overline{x}^{(\lambda)} = \sum_{i=1}^{l} n_i\,\overline{x}_i^{(\lambda)}/N$는 전체 평균을 의미한다.

3. 모의실험

의사결정나무는 분리변수 및 분리점 선택을 반복적 수행하여 나무를 생성한 후 가지치기(pruning)를 통해 최적화된 나무를 생성한다. 여기서 분리변수의 선택은 최적의 나무를 만들기 위한 초기단계로써 분류 알고리즘의 효율성을 결정하는데 중요한 역할을 차지한다.

그러나 QUEST 알고리즘은 분류변수의 선택 시 연속형 입력변수가 정규성을 만족시키지 못하는 경우 분류변수 선택에 문제점을 지니고 있다. 본 장에서는 QUEST 알고리즘이 연속형 변수의 정규성 가정에 대한 민감도를, Loh and Shih(1997)와 Kim and Loh(2001)에 의해 제안된 변수선택 편의와 선택력을 통하여 살펴보았다.

3.1 실험 과정 및 설계

실제 모의실험 설계를 위한 분류나무 및 변수변환은 Java 언어로 구현된 Weka의 J48을 이용한 알고리즘에 QUEST와 변수변환을 추가하여 구현하였다.

실험 과정은 몬테칼로 모의실험으로 생성된 자료의 다변량 정규성을 살펴보기 위해 마르디아(Mardia, 1974) 계수를 이용하여 변수별 왜도와 첨도를 유의수준 5%에서 조사하였으며, 첨도 가정이 기각되면 모듈러스 변환을, 왜도의 가정이 기각되면 여-존슨 변환을 적용하여 변수 변환을 실시하였다. 마르디아의 왜도와 첨도 검정은 SAS에서 매크로(http://ftp.sas.com/techsup/download/stat/multnorm.sas)로 제공되고 있는 아래의 식을 이용하였다.

$$\frac{\frac{k}{6n}\left\{\sum_{i=1}^{n}(x_i-\bar{x})^3\right\}^2}{S^3} \sim \chi^2\left(\frac{l(l+1)(l+2)}{6}\right).$$

$$\frac{\frac{\sum(x_i-\bar{x})^4}{nS^2}-l(l+2)}{\sqrt{8l(l+2)/n}} \sim N(0,1).$$

여기서 마르디아 계수를 계산하는 i번째 변수 $x_i(i=1,\ldots,n)$의 평균과 표준편차는 각각 $\bar{x}=\sum_{i=1}^{n}x_i/n$, $S=\sum_{i=1}^{n}(x_i-\bar{x})^2/n$이고, $k=\frac{(l+1)(n+1)(n+3)}{\{n(n+1)(l+1)-6\}}$(단, $l=1$)로 계산된다.

모의실험에 이용된 변수 선택 편의와 변수 선택력 정의와 설계는 아래와 같다.

① 변수 선택 편의(variable selection bias)

변수 선택에서 편의란, 같은 조건을 갖는 변수들은 동일한 확률 하에 분리변수로 선택되어야 함에도 불구하고, 변수의 다른 특성으로 인해 선택될 확률이 편중되는 것을 의미한다. 즉, Y가 목표변수, $X_1,\ldots,X_K$는 예측 변수, R은 변수선택 알고리즘이고, $X_i(i=1,\ldots,K)$와 Y가 독립일 때,

$\Pr(X_j \text{ is selected}|R) \neq \Pr(X_k \text{ is selected}|R)$

인 j와 k가 존재하면 R은 변수 선택 편의가 발생한다고 한다.

② 변수 선택력(variable selection power)

하나의 특정한 입력변수만이 목표변수에 대한 정보를 가지고 있을 경우, 변수선택에 사용된 알고리즘에 의해 정보를 갖고 있는 입력변수가 선택될 확률을 의미한다. 즉, X_j가 Y에 대한 정보를 갖고 있으며 $X_i(i=1,\ldots,K,\ i\neq j)$와 Y가 독립일 때, $\Pr(X_j \text{ is selected}|R)$을 R에 대한 변수 선택력이라 한다.

변수 선택 편의를 위한 실험은 각 입력변수가 목표변수의 범주에 동일한 확률을 갖도록 하여 목표변수에 대한 정보를 지니지 않도록 설계하였다. 입력변수는 정규 분포의 왜도, 첨도의 변화를 고려하여 정규분포, 지수분포, T-분포로 설정하였고, 목표변수는 3개의 범주를 가지도록 설정하였다. 실험은 각 관측치를 300, 1,000개로 나누어 10000회 반복하였다.

변수 선택력에 대한 실험은 실험의 단순화를 위해 3개의 입력변수에 대해 목표변수의 범주 수를 2개로 설정하였다. 또한 목표변수에 대한 정보를 지닌 입력변수를 X_1으로 설정하여 변수 선택력에 대한 정규성의 영향력 정도를 살펴보았다. 이에 더해 X_1의 위치의 변화에 따른 변수 선택력의 변화 정도도 조사해 보고자 설정된 변수를 '이동모형'(shifted model)화 시켰고, 관측치는 300개로 고정하였으며 10,000회 반복 실험하였다.

실험의 결과 선택 편의가 존재하지 않는다면, 변수선택 편의의 경우 변수변환을 실시한 전·후의 입력변수의 선택 확률이 $1/3$에 가까운 값을 가져야 하며, 변수 선택력의 경우는 목표변수에 대해 정보를 지닌 변수 X_1의 선택 확률이 다른 입력변수의 선택 확률보다 큰 값을 가져야 한다. 이를 위해 변수 선택 편의와 선택력의 경험적 확률은 총 모의실험 회수 중 각 입력변수 X_i가 분류 변수로 선택된 회수로 계산하였다.

3.2 모의실험 결과

<표 1>은 변수 선택 편의를 관측치 개수 300개와 1,000개로 나누어 10,000회 반복 실시한 변수 변환 전·후의 결과이다.

<표 1> 변수 선택 편의의 결과

데이터 수	변환 전			변환 후		
	$N(0,1)$	$Exp(1)$	$T(2)$	$N(0,1)$	$Exp(1)$	$T(2)$
$n=300$	0.325	0.334	0.341	0.332	0.333	0.335
$n=1000$	0.329	0.325	0.346	0.343	0.326	0.331

위의 결과를 통해 QUEST 알고리즘은 변환을 실시하기 전·후 세 분포의 선택 확률이 각 관측 수에서 $1/3$에 가까워 변수에 대한 선택 편의가 존재하지 않음을 알 수 있었다. 이는 QUEST 알고리즘이 변수선택 편의에 관해서는 입력변수의 분포에 영향을 거의 받고 있지 않음을 나타낸다.

<표 2> 변수 선택력 모형

모의 실험	X_1		X_2	X_3
	class 1	*class* 2		
Ⅰ	$N(0,1)$	$N(0.2,1)$	$LN(0,1)$	$T(2)$
Ⅱ	$LN(0,1)$	$LN(0,1)+0.4$	$N(0,1)$	$T(2)$
Ⅲ	$T(2)$	$T(2)+0.5$	$N(0,1)$	$LN(0,1)$

<표 2>는 변수 선택력을 위한 실험 설계표이다. 각 실험은 관측치를 300개로 고정하여 10,000회 반복하였으며 2범주인 목표 변수에 대한 정보를 지닌 입력변수 X_1은 분포 가정의 영향력을 살펴보고자 각각 정규분포, 대수 정규분포(log normal)와 T-분포로 설정하여 변수 변환 전의 선택확률이 약 70%가 되도록 설정하였다.

<표 3> 변수 선택력 결과

모의 실험	변환 전			변환 후		
	X_1	X_2	X_3	X_1	X_2	X_3
Ⅰ	0.700	0.148	0.152	0.708	0.146	0.146
Ⅱ	0.685	0.156	0.159	0.995	0.002	0.002
Ⅲ	0.688	0.154	0.158	0.915	0.042	0.043

<표 3>의 변수 선택력의 결과를 살펴보면 X_1에 정규분포를 이용하여 목표변수에 대한 정보를 가진 모의실험 Ⅰ은 변수 변환 전·후의 변수 선택력에 큰 차이를 보이지 않는 반면, X_1이 대수 정규분포(log normal)와 T-분포인 경우의 모의실험 Ⅱ, Ⅲ는 변수변환 전·후의 변수 선택력의 확률이 0.685, 0.688에서 0.995와 0.915로 크게 향상됨을 볼 수 있다.

다음 실험은 목표변수의 범주별로 입력변수에 위치를 변화한 '이동모형'(shifted model)에서 변수 선택력의 변화 정도를 살펴보고자 하였다. 이동정도에 따른 선택력 변화를 살펴보기 위해 '이동모형'(shifted model)의 첫 번째 모의실험은 X_1을 정규분포로 설정하여 위치 모수를 0.05~0.4까지 0.05씩 이동시켰으며, X_2와 X_3은 위치 이동 없이 <표 2>의 모의실험 Ⅰ과 동일하게 설정하였다. 다음 <표 4>는 정규분포에 대한 이동모형 결과이다.

<표 4> 정규분포의 변수 선택력

이동 변수	변환 전			변환 후		
	이동변수	$LN(0,1)$	$T(2)$	이동변수	$LN(0,1)$	$T(2)$
N(0.05,1)	0.353	0.316	0.331	0.361	0.320	0.319
N(0.1,1)	0.447	0.269	0.284	0.452	0.273	0.275
N(0.15,1)	0.574	0.208	0.217	0.581	0.209	0.209
N(0.2,1)	0.700	0.148	0.152	0.708	0.146	0.146
N(0.25,1)	0.817	0.089	0.094	0.814	0.095	0.091
N(0.3,1)	0.894	0.053	0.054	0.892	0.055	0.053
N(0.35,1)	0.945	0.027	0.028	0.944	0.028	0.028
N(0.4,1)	0.976	0.012	0.012	0.974	0.014	0.013

앞의 결과는 변수 변환 전·후의 X_1변수에 대한 선택확률이 크게 변화하지 않았음을 보여준다. 목표변수에 대한 정보를 지닌 입력변수가 정규분포인 경우에는 정보를 지니지 않은 입력변수들의 정규성이 만족되지 않더라도 변수변환 전·후에 변수 선택력에 큰 변화를 보이지 않았다.

두 번째 이동모형에 대한 모의실험은 X_1을 대수 정규분포로 설정하여 평균을 0.1~0.8까지 0.1씩 변화시켰으며, 정보를 지니지 않는 입력변수는 <표 2>의 모의실험 Ⅱ과 동일하게 설정하였다. <표 5>는 대수 정규분포에 대한 이동모형 결과이다.

<표 5> 대수 정규분포의 변수 선택력

이동 변수	변환 전			변환 후		
	이동변수	$N(0,1)$	$T(2)$	이동변수	$N(0,1)$	$T(2)$
LN(0,1)+0.1	0.358	0.314	0.327	0.532	0.235	0.232
LN(0,1)+0.2	0.453	0.267	0.279	0.833	0.085	0.082
LN(0,1)+0.3	0.567	0.210	0.223	0.966	0.017	0.017
LN(0,1)+0.4	0.685	0.156	0.159	0.995	0.002	0.002
LN(0,1)+0.5	0.788	0.107	0.105	0.999	0.000	0.000
LN(0,1)+0.6	0.867	0.070	0.064	1.000	0.000	0.000
LN(0,1)+0.7	0.919	0.043	0.038	1.000	0.000	0.000
LN(0,1)+0.8	0.954	0.025	0.021	1.000	0.000	0.000

이동모형 별 추이를 살펴보면 대수 정규분포의 평균을 0.2만큼 이동시켰을 때 X_1이

분리 변수로 선택될 확률이 변환 전·후로 0.453에서 0.833으로 크게 향상되었고, 평균의 이동이 0.3이상일 때 X_1의 선택확률이 변수변환 후에 0.95이상으로 크게 향상되어 나타남을 볼 수 있다.

<표 6> T-분포의 변수 선택력

이동 변수	변환 전			변환 후		
	이동변수	$N(0,1)$	$LN(0,1)$	이동변수	$N(0,1)$	$LN(0,1)$
T(2)+0.1	0.360	0.311	0.330	0.386	0.296	0.318
T(2)+0.2	0.420	0.280	0.300	0.530	0.229	0.241
T(2)+0.3	0.502	0.242	0.256	0.690	0.153	0.157
T(2)+0.4	0.597	0.199	0.204	0.826	0.086	0.088
T(2)+0.5	0.688	0.154	0.158	0.915	0.042	0.043
T(2)+0.6	0.764	0.116	0.120	0.967	0.016	0.016
T(2)+0.7	0.825	0.087	0.088	0.988	0.006	0.006
T(2)+0.8	0.871	0.064	0.065	0.996	0.002	0.002

<표 6>은 X_1을 T-분포로 설정하여 0.1~0.8사이에서 변수 선택력의 변화 정도를 살펴보았으며, 정보를 지니지 않는 입력변수는 <표 2>의 모의실험 Ⅲ과 같이 정규분포와 대수 정규분포를 위치이동 없이 사용하였다. T-분포를 이동모형으로 설정한 <표 6>의 결과를 통해 변수 변환이 정보를 지닌 입력변수의 변수 선택력에 영향을 주고 있음을 볼 수 있으며, 평균의 이동 변화량이 거의 없는 0.1인 경우를 제외하고 이동모수의 변화에 따라 변수 선택력이 전체적으로 고르게 향상되어 나타나고 있다.

이러한 실험을 통해 QUEST 알고리즘은 정보를 가진 입력 변수가 정규분포인 경우 다른 입력변수에 관계없이 편의를 갖지 않고 변수 선택력에도 우수함을 알 수 있었다. 그러나 정보를 가진 입력변수가 대수 정규분포(log normal)나 T-분포와 같이 정규성 가정을 위배할 때, 변수변환에 의해 정규분포로 바꾸지 않는다면 변수 선택력에서 낮은 효율을 보이고 있다.

4. 결론

본 논문은 데이터 마이닝의 모형화 방법 중 의사결정나무의 QUEST 알고리즘이 연속형 입력변수의 분포에 대해 분리 변수를 선택 시 갖게 되는 편의와 선택력을 살펴보고 변수 변환 방법을 통해 선택력을 높이고 편의를 줄일 수 있는가를 조사하였다.

모의실험을 통해 각 입력변수가 목표변수에 동일한 정보를 지니는 경우에 변수 선택 편의는 분포의 가정에 상관없이 존재하지 않아 변수선택 편의에 관해서는 QUEST 알고리즘이 우수하다는 것을 확인하였다. 그러나 목표변수에 대해 정보를 지닌 입력변수의 정규성 가정이 위배될 때, 그 변수가 분리변수로 선택되는 선택력은 상대적으

로 낮게 나타났다. 선택력을 높이기 위해 우선 정규성이 의심되는 변수를 정규분포로 변수변환하여 QUEST 알고리즘을 적용한 결과 선택력이 우수해지는 것을 알 수 있었다. 간단한 왜도, 첨도 검정에 의한 판단과 변수변환만으로도 분리변수의 선택력을 매우 높일 수 있음을 실증분석을 통하여 증명하였고 이러한 과정의 의사결정나무기법을 제안하는 바이다.

향후에는 이렇게 변수변환을 적용한 QUEST 알고리즘이 분리점 선택에 있어 어떠한 특성을 갖는지 검토해 볼 필요가 있을 것이라 생각되어진다.

참고문헌

1. 정성석, 김순영, 임한필 (2004). 의사결정나무에서 분리 변수 선택에 관한 연구, *응용통계연구*, 제17권 2호, 347-357.
2. Box, G. E. P. and Cox, C. R. (1964). An Analysis of Transformations, *Journal of Royal Statistical Society*, *series B*, 26, 211-252.
3. Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification and Regression Trees*, New York : Chapman and Hall.
4. Gnanadesikan, R. (1977). *Methods for statistical data analysis of multivariate observations*, New York, Wiley
5. In-Kwon Y, A & Johnson R. A. (2000). A New Family of Power to Improve Normality or Symmetry, *Biometrika*, 87, 954-959.
6. Johnson, J. A. and Draper, N. R. (1980). An Alternative Family of Transformations, *Applied Statistics*, 29, 190-197.
7. Kass, G. V. (1980). An Exploratory Technique for Investigating Large Quantities of Categorical Data, *Applied Statistics*, 29, 119-127.
8. Loh, W. Y. and Shih, Y. S. (1997). Split Selection Method for Classification Trees, *Statistica Sinica*, 7, 815-840.
9. Kim, H. and Loh, W. Y. (2001). Classification Trees with Unbiased Multiway Splits, *Journal of the American Statistical Association* Vol. 96, No. 454, 589-604.
10. Mardia (1974). Applications of Some Measures of Multivariate Skewness and Kurtosis in Testing Normality and Robustness Studies. *Sankhya B*, 36, 115-128.
11. Quinlan, J. R. (1993). *C4.5: Programs for machine learning,* Morgan Kaufmann.

[2004년 10월 접수, 2005년 2월 채택]