응용통계연구 제 17권 2호, 2004년, pp. 347-357

의사결정나무에서 분리 변수 선택에 관한 연구*

정성석[1] 김순영[2] 임한필[3]

요 약

의사결정나무에서 분리 변수를 선택하는 것은 매우 중요한 일이다. C4.5는 변수 선택에 있어 연속형 변수로의 변수 선택 편의가 심각하고, QUEST는 연속형 변수와 관련해서 정규성 가정이 위반될 경우 변수 선택력이 떨어진다. 본 논문에서는 통계적 로버스트 검정 알고리즘을 제안하고, 모의 실험을 통하여 C4.5, QUEST 그리고 제안된 알고리즘의 효율성을 비교하였다. 실험 결과 제안된 알고리즘이 변수 선택 편의와 변수 선택력 측면에서 로버스트함을 알 수 있었다.

주요용어: 분류나무, 그룹화, Peizer & Pratt 변환, 변수 선택 편의, 변수 선택력.

1. 서론

컴퓨터의 발달과 자료의 축적으로 인하여 많은 양의 자료를 분석하여 데이터 내에 존재하는 관계, 패턴, 규칙 등의 우리가 알지 못했던 사실들을 탐색하고 찾아내어 모형화 함으로써 유용한 지식을 추출하는 데이터 마이닝(data mining)에 관한 연구가 활발히 진행되고 있다.

데이터 마이닝의 여러 기법 중에서 자주 사용되는 의사결정나무는 목표 변수가 범주형인 경우는 분류 나무(classification tree), 연속형인 경우는 회귀 나무(regression tree)라고 한다. 분류 나무는 분석용 자료(training data)를 나무 구조로 반복적 분할을 수행함으로써 형성된다. 분류 나무 형성 시 중요한 것은 목표 변수를 잘 분리해주는 변수를 선택하는 것이다. 이는 분리의 측도인 분리 기준(split criterion)에 영향을 크게 받는다.

분류 나무 알고리즘 중 C4.5는 모든 예측 변수들을 이득 비율(gain ratio)이라는 분리 기준을 사용하여 전체 탐색법으로 분리 변수를 선택한다. 일반적으로 같은 연관성을 가진 예측 변수라면 분리 변수로 선택될 비율이 동일해야 한다. C4.5는 분리 변수 선택과 분리점 선택이 동시에 이루어지기 때문에, 예측 변수의 형태에 따라 분리 변수로 선택될 비율이 달라지는 변수 선택 편의가 발생한다(Loh 와 Shih, 1997). QUEST에서는 먼저 유의한 변수를 선택하고, 선택된 변수에 대해서 이차판별분석(quadratic discriminant analysis)을 실시

* 이 논문은 2001년도 전북대학교의 지원 연구비에 의하여 연구되었음.

1) (561-756) 전라북도 전주시 덕진구 덕진동 1가 644-14, 전북대학교 통계정보과학과 교수
E-mail: sschung@chonbuk.ac.kr

2) (561-756) 전라북도 전주시 덕진구 덕진동 1가 644-14, 전북대학교 통계정보과학과 박사과정
E-mail: rabbit@chonbuk.ac.kr

3) (561-756) 전라북도 전주시 덕진구 덕진동 1가 644-14, 전북대학교 통계정보과학과 석사과정 졸업
E-mail: hanpil@hotmail.com

하여 분리점을 선택하는 단계적 수행으로 변수 선택 편의를 현저하게 감소시켰다. 그러나 QUEST는 분리 변수 선택시 연속형 변수에 대해서는 정규성의 분포를 가정하는 ANOVA F-검정, 범주형 변수에 대해서는 카이제곱검정을 사용하기 때문에, 연속형 변수가 정규성 가정에 심각하게 위배되는 경우에 변수 선택력이 떨어지는 문제가 발생한다.

본 연구에서는 QUEST의 변수 선택 알고리즘 중 연속형 변수를 그룹화하여 범주형 변수로 변환하여 변수의 형태를 동일하게 한 후, 카이제곱 검정을 이용하여 분리 변수를 선택하는 로버스트 검정 알고리즘을 제시하였다. 이 방법은 변수 선택 편의가 심하지 않고 변수 선택력이 향상되고, 예측변수의 분포에 무관하게 로버스트(robust)한 경향이 있었다. 로버스트 검정 알고리즘의 변수 선택 성능을 기존의 알고리즘과 비교하고, 각 특성을 알아보기 위해서 Loh와 Shih(1997)가 제시한 방법에 정규성 가정이 위배된 경우를 추가하여 모의실험을 수행하였다.

2절에서는 분류 나무와 분리 기준에 대해서 알아보고, 3절에서는 C4.5와 QUEST의 분리 변수 선택과 분리점 선택의 특성을 알아본 후 통계적 로버스트 검정 알고리즘을 제시하였다. 4절에서는 모의 실험을 통한 변수 선택 알고리즘의 성능을 비교하고, 마지막으로 5절에서는 결론과 향후 연구 방향에 대해 논의하였다.

2. 분리 방법

분류 나무는 분석용 자료를 분리 규칙에 의해서 나무 구조의 형태로 반복적으로 분할하여 관심 대상을 몇 개의 하위 집단으로 분류하거나 예측하는 기법으로, 마디(node)라고 불리는 구성 요소들로 이루어져 있다. 분류 나무의 주요 생성 요소는 분리 기준의 선택, 정지 규칙, 가지치기 그리고 클래스 할당이다. 대표적인 분류 나무 알고리즘으로는 CHAID(Kass, 1980), CART(Breiman etc., 1984), FACT(Loh와 Vanichsetakul, 1988), C4.5(Quinlan, 1993), QUEST(Loh와 Shih, 1997), CRUISE(Kim, 1998) 등이 있다. 분류 나무는 분리 기준에 따라 알고리즘의 특성이 결정된다.

분리기준은 하나의 부모마디로부터 자식마디들이 형성될 때, 분리 변수의 선택과 범주의 병합이 이루어지는 분리점 선택의 기준을 의미한다. 즉, 어떤 변수를 이용하여 어떻게 분리하는 것이 목표변수의 클래스를 가장 잘 구별해 주는지를 파악하여 자식마디를 형성해야 되므로, 분리(split)의 원리는 부모 마디보다 자식 마디를 순수하게 하는 분리 변수와 분리점을 선택하는 것이다. 이러한 분리 기준은 마디의 순수도를 계산하는 불순도 함수(impurity function)을 사용하는 방법과 카이제곱검정이나 ANOVA F-검정을 이용하여 변수를 선택 후 분리점을 선택하는 방법이 있다.

보통 불순도 함수로는 지니지수(Gini index)와 엔트로피(entropy)를 많이 사용하며, 이를 이용하여 예측 변수에 대해 분리가 가능한 모든 분리점에서 불순도 함수를 계산한 후 불순도 함수의 감소량이 최대가 되는 분리를 찾는다. 이와 같은 전체 탐색법에 의해 분리를 수행하는 알고리즘은 CART와 C4.5가 있다.

Loh와 Shih(1997)에 의하면 전체 탐색법은 계산량이 많고, 변수 선택시 편의(bias)가 발생하는 문제가 있다. 제한 없이 모든 분리점에 대한 검색을 수행하기 때문에 변수가 가지

고 있는 분리의 수가 많을수록 선택 확률도 커지게 된다. 예를 들어 목표 변수와 관련 정도가 같은 예측변수들 가운데 개별값이 많은 연속형 변수가 범주형 변수보다 더 잘 선택되고, 범주가 많은 범주형 변수가 이진 변수보다 더 잘 선택되는, 즉 변수의 형태에 따라 선택될 확률이 편중되는 현상이 나타난다. 이러한 현상으로 인해서 분류 나무의 구조로부터 믿을 만한 결론을 이끌어 내는데 어려움이 따른다.

이러한 전체 탐색법의 단점을 보완한 것이 변수 선택과 분리점 선택을 나누어서 시행하는 방법이다. 변수 선택 단계에서는 각 예측 변수들의 카이제곱 통계량이나 F-통계량의 유의확률을 계산하여 가장 작은 유의확률을 가지는 변수를 선택한다. 다음으로 분리점 선택 단계에서는 선택된 변수에 대해서 선형판별분석이나 이차판별분석을 통해서 분리점을 찾는다. 이와 같은 통계적 검정을 이용하는 알고리즘은 FACT, QUEST, CRUISE가 있다.

3. 분리 변수 선택 알고리즘

분류 나무는 분리 변수 선택과 분리점 선택의 반복적 수행과 가지치기를 통해서 생성된다. 이러한 과정 중에서 분리 변수 선택은 중요한 변수를 선택하여 의미 있는 분리를 수행하기 위한 사전 단계로 분류 나무 생성에 가장 중요한 부분이다. Loh와 Shih(1997)와 Kim과 Loh(2001)는 분류 나무의 변수 선택 알고리즘에 대한 성능을 비교하기 위해 변수 선택 편의와 변수 선택력을 정의하였다.

분리 변수 선택 편의는 예측 변수 모두가 목표 변수에 대한 정보를 전혀 가지고 있지 않은 경우에 각 변수가 분리 변수로 선택될 확률이 같아야 하는데 적어도 하나 이상이 같지 않은 경우를 의미한다.

변수 선택력은 목표 변수에 대한 정보를 가지고 있는 변수들과 의미 없는 변수들이 혼합되어 있는 경우에 정보를 가지고 있는 예측 변수가 분리 변수로 선택될 확률을 말한다. 만일 의미 없는 변수가 분리 변수로 선택된다면 정보를 가지고 있는 변수가 분리 변수로 선택된 경우보다 더욱 많은 분리를 수행하게 될 것이다. 이러한 경우에는 가지치기의 과정을 수행한다고 할지라도 잘못된 구조를 가지고 있는 나무가 생성되거나 전혀 나무가 생성되지 않을 것이다(Loh와 Shih, 1997). 따라서, 좋은 알고리즘은 변수 선택 편의가 거의 발생하지 않고, 정보를 가지고 있는 예측 변수의 변수 선택력이 커야 할 것이다.

본 절에서는 전체 탐색법을 사용하는 C4.5와 통계적 검정법을 사용하는 QUEST의 분리 변수 선택 알고리즘에 대해서 살펴보고, 통계적 로버스트검정(SRT)알고리즘을 제안하였다.

3.1. C4.5의 변수 선택

C4.5는 분리기준으로 엔트로피를 변형시킨 이득비율을 이용하여 변수 선택과 분리점 선택이 동시에 이루어지는 전체 탐색법을 사용하고 연속형인 경우에 이지분리를, 범주형인 경우에 다지분리를 한다. 즉, 모든 가능한 분리에 대해 이득비율을 계산하고 그 이득 비율이 최대가 되도록 분리한다.

어떠한 변수 선택 알고리즘도 변수 선택 편의는 전혀 발생하지 않을 수 없지만 C4.5는 연속형 변수로의 편의의 정도가 매우 심함이 알려져 있다(Quinlan, 1996). 즉, C4.5는 전체 탐색을 통해서 변수 선택을 하기 때문에 목표 변수와의 연관성보다는 변수의 형태에 더욱 의존하는 단점을 가지고 있다.

3.2. QUEST의 변수 선택

QUEST는 통계적 검정 방법을 이용하는 알고리즘으로, 분리 변수의 선택과 선택된 분리 변수에서 분리점 선택으로 나누어 이지분리를 수행한다. 분리기준으로는 연속형 변수는 ANOVA F-통계량의 유의확률을 계산하고 범주형 변수는 카이제곱 검정 통계량의 유의확률을 계산하여 가장 작은 유의확률을 갖는 변수를 선택한다. 이렇게 선택된 분리 변수에 대해 이차판별분석을 수행하여 분리점을 선택한다.

Loh와 Shih(1997)는 QUEST에서 목표 변수와 예측 변수가 상호 독립일 경우, 각 예측 변수가 분리 변수로 선택될 비율이 무시할 정도의 작은 편의를 가지고 있고, QUEST가 CART나 FACT보다 변수 선택력이 뛰어나다고 말하고 있다. 그러나 QUEST는 연속형 변수에 대해 ANOVA F-검정을 사용하기 때문에 정규성이나 등분산의 가정에 어긋날 경우에는 문제가 생긴다. 실제로 클래스의 구분이 확실한데도 평균과 분산이 비슷한 경우 의미 없는 변수로 취급하기가 쉽고, 한쪽의 클래스가 왜도(skewness)가 심하거나 첨도(kurtosis)가 큰 경우 ANOVA F-검정은 큰 효과를 발휘하지 못한다.

즉, QUEST의 변수선택 방법은 변수선택 편의가 거의 없으나, 연속형 변수에 대해서 정규성의 분포 가정이 크게 위배될 때는 변수 선택력이 현저히 감소한다.

3.3. 통계적 로버스트 검정 알고리즘

변수선택 편의가 심하지 않으며 연속형 변수에 대해서 분포와 무관하게 변수 선택력을 향상시키기 위해, QUEST의 변수선택 알고리즘 형태를 그대로 유지하면서 연속형 변수를 그룹화하여 카이제곱 검정을 이용하는 통계적 로버스트 검정 알고리즘(SRT)을 고려해 보자.

알고리즘 수행 단계는 먼저 범주형 변수와 연속형 변수의 자료 형태를 동일하게 만들기 위해 적절한 빈(bin)의 개수 M을 선택하고, 예측 변수 X의 분위수(quantile value) Q_i, $i = 1, \ldots, M$를 구한다. 연속형 변수의 경우 $X \in (Q_{i-1}, Q_i]$이면 $X = i$로 범주화시킨다. 예측변수가 순서형이고 X의 개별값(distinct value)의 개수를 M_d라 할 때, $M_d < M$인 경우 $X = X_{(j)}$, $j = 1, \ldots, M_d$이면 $X = j$로 범주화시킨다. $M_d \geq M$인 경우는 $X \in (Q_{i-1}, Q_i]$, $i = 1, \ldots, M$이면 $X = i$로 범주화시켜 그룹화한후, 목표 변수와 예측 변수의 분할표에 대한 카이제곱 통계량을 구한다. 목표변수 클래스의 수를 J라할 때, 예측변수X_k, $k = 1, \ldots, K$에 대한 카이제곱 검정 통계량은 다음과 같다.

$$\chi^2(X_k) = \sum_{j=1}^{J} \sum_{i=1}^{M} \frac{(O_{ijk} - E_{ijk})^2}{E_{ijk}} \sim \chi^2((M-1)(J-1)).$$

카이제곱 통계량이 자유도에 의존하기 때문에 Kim과 Loh(2001)가 사용했던 방법과 같이 Peizer & Pratt 변환을 하면 $N(0,1)$에 근사하게 된다.

χ^2통계량의 Peizer-Pratt변환은 다음과 같다. $\nu=(M-1)\times(J-1)$이고 $W=\chi^2+\nu+1$일 때,

$$z=\frac{W-1/3}{|W|}\sqrt{(\nu-1)\log\left(\frac{\nu-1}{\chi^2}\right)+W}\ .$$

각 예측변수에 대해 Peizer-Pratt변환된 z_k, $k=1,\ldots,K$의 값을 구하고 가장 큰 변환된 값을 갖는 예측변수를 분리변수로 사용한다.

4. 모의 실험

모의 데이터를 이용하여 C4.5, QUEST 그리고 SRT에 대한 변수 선택 성능을 비교하였다.

모의 실험에 사용될 자료의 분포는 표4.1과 같이 하였고 실제 모의 실험 설계를 위해 분류 나무는 Java 언어로 구현된 Weka의 J48패키지를 이용하여 C4.5에 QUEST와 SRT를 추가하여 구현하였다(Witten과 Frank, 1999).

표 4.1: 모의 실험에 사용된 분포들

표시	설명
Cat. U_k	정수 $1,\ldots,k$를 취하는 범주형 균일분포, $\Pr(X=i)=\frac{1}{k}$, $\forall i=1,\ldots,k$.
Ord. U_k	정수 $1,\ldots,k$를 취하는 순서형 균일분포, $\Pr(X=i)=\frac{1}{k}$, $\forall i=1,\ldots,k$.
$Exp(a,\theta)$	지수분포, $f(x)=\frac{1}{\theta}\exp\{-\frac{(x-a)}{\theta}\}I(x>a)$.
$DE(\theta)$	이중지수분포, $f(x)=\frac{1}{2\theta}\exp\{-\frac{\lvert x\rvert}{\theta}\}$.
$Gam(a,\,\alpha,\theta)$	감마분포, $f(x)=\frac{(x-a)^{\alpha-1}}{\Gamma(\alpha)\theta^{\alpha}}\exp\{-\frac{(x-a)}{\theta}\}I(x>a)$.
$LN(a,\,\mu,\sigma^2)$	로그정규분포, $f(x)=\frac{1}{\sqrt{2\pi}\sigma(x-a)}\exp\{-\frac{(ln(x-a)-\mu)^2}{2\sigma^2}\}I(x>a)$.
$T(\nu)$	Student T분포, $f(x)=\frac{\Gamma[(\nu+1)/2]\{1+(x^2/\nu)\}^{-(\nu+1)/2}}{\sqrt{\pi\nu}\,\Gamma(\nu/2)}$.
$N(\mu,\sigma^2)$	정규분포, $f(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp\{-\frac{(x-\mu)^2}{2\sigma^2}\}$.
Ord. A_4	정수 $1,\ldots,4$를 취하는 순서형 분포 A, $\Pr(X=1)=\Pr(X=2)=\Pr(X=3)=\frac{2}{9}$, $\Pr(X=4)=\frac{1}{3}$.
Ord. B_4	정수 $1,\ldots,4$를 취하는 순서형 분포 B, $\Pr(X=1)=\Pr(X=2)=\Pr(X=3)=\frac{1}{5}$, $\Pr(X=4)=\frac{2}{5}$.

모의 실험은 크게 두 가지로 나누어서 수행했다. 첫째, 변수 선택 편의에 관한 부분으로 Loh와 Shih(1997)가 제시한 모의 실험과 동일한 설정을 하였다. 둘째, 변수 선택력에 관한 부분으로 Loh와 Shih(1997)가 제시한 모의 실험과 함께 왜도와 첨도와 관련된 부분을 덧붙여 설정했다. 구체적인 모의 실험 설정은 표4.2와 같다.

표 4.2: 모의 실험 설정

변수선택	실험 내용	설정		
		X_1	X_2	
		$N(0,1)$	$T(2)$	
		$N(0,1)$	$Exp(0,1)$	
		$T(2)$	$Exp(0,1)$	
		$N(0.1)$	*Ord.* U_4	
변수		$T(2)$	*Ord.* U_4	
	Loh와 Shih	$Exp(0,1)$	*Ord.* U_4	
선택	(1997)의 실험	*Cat.* U_4	*Cat.* U_{15}	
		Cat. U_4	$N(0,1)$	
편의		*Cat.* U_{15}	$N(0,1)$	
		Cat. U_4	$Exp(0,1)$	
		Cat. U_{15}	$Exp(0,1)$	
		Cat. U_4	*Ord.* U_4	
		X_1		$X_2, \cdots, X_{20}$
		class1	class2	
		Ord. U_4	*Ord.* A_4	$N(0,1)$
		Ord. U_4	*Ord.* A_4	$T(2)$
	Loh와 Shih	*Ord.* U_4	*Ord.* A_4	$Exp(0,1)$
	(1997)의 실험	*Cat.* U_4	*Cat.* B_4	*Cat.* U_{15}
변수		$N(0,1)$	$N(0.25,1)$	*Cat.* U_{15}
		$Exp(0,1)$	$Exp(0,1.3)$	*Cat.* U_{15}
선택력		*Ord.* U_4	*Ord.* B_4	*Cat.* U_{15}
		$N(0,1)$	$Exp(-1,1)$	*Cat.* U_{15}
	왜도 변경	$N(0,1)$	$Gam(-1.4,2,0.7)$	*Cat.* U_{15}
		$N(0,1)$	$LN(-2.9,1,0.3^2)$	*Cat.* U_{15}
		$N(0,1)$	$DE(1)$	*Cat.* U_{15}
	첨도 변경	$N(0,1)$	$T(2)$	*Cat.* U_{15}
		$N(0,1)$	$T(1)$	*Cat.* U_{15}

표4.2의 왜도 변경 모의 실험은 목표 변수가 클래스 1인 경우에는 좌우 대칭인 표준정규분포의 정보를 X_1에 주고, 클래스 2인 경우에는 치우침이 있는 분포로부터 생성된 정보를 X_1에 주었다. 또한 치우침이 있는 분포와 표준정규분포의 평균과 분산이 같아지도록 설정했다. 의미 없는 변수들은 범주수가 15인 범주형 분포로부터 생성했다.

표4.2의 첨도 변경 모의 실험은 목표 변수가 클래스 1인 경우에는 표준정규분포의 정보

를 X_1에 주고, 클래스 2인 경우에는 꼬리가 표준정규분포보다 두터운 분포로부터 생성된 정보를 X_1에 주었다.

변수 선택 알고리즘에 있어서 QUEST는 유의 수준을 0.05로 설정하여 수행했다. SRT를 수행하기 위한 M의 값은 이윤모(2002)에서 논의된 10으로 선정하여 변수 선택을 수행했다.

모의 실험의 시행 횟수는 10,000번으로 설정하고 모의 실험의 결과는 정해진 변수(X_1)에 대한 선택 비율을 나타내도록 했다. 따라서 선택 비율에 대한 모의 실험의 표준오차는 0.005보다 작다.

4.1. 변수 선택 편의

각 알고리즘의 변수 선택 편의를 알아보기 위한 모의 실험 설정 요인은 변수의 형태와 변수에 대한 개별값의 개수이다. 표4.2의 변수 선택 편의에 대한 모의 실험은 두 개의 변수, 즉 X_1과 X_2에 대해서 (i) 연속형과 연속형, (ii) 범주형과 범주형 그리고 (iii) 범주형과 연속형으로 설정하고 각 알고리즘에 의한 X_1의 선택 비율이 계산되도록 하였다.

표4.3은 변수 선택 편의에 대한 모의 실험 결과이다.

표 4.3: Loh와 Shih(1997)의 변수 선택 편의에 대한 모의 실험 결과

X_1	X_2	$\Pr(X_1)$		
		C4.5	QUEST	SRT
$N(0,1)$	$T(2)$	0.495	0.493	0.502
$N(0,1)$	$Exp(0,1)$	0.505	0.501	0.501
$T(2)$	$Exp(0,1)$	0.498	0.516	0.500
$N(0,1)$	*Ord.* U_4	0.879	0.502	0.497
$T(2)$	*Ord.* U_4	0.882	0.517	0.497
$Exp(0,1)$	*Ord.* U_4	0.882	0.499	0.500
Cat. U_4	*Cat.* U_{15}	0.028	0.500	0.505
Cat. U_4	$N(0,1)$	0.174	0.495	0.493
Cat. U_{15}	$N(0,1)$	0.942	0.491	0.496
Cat. U_4	$Exp(0,1)$	0.182	0.489	0.502
Cat. U_{15}	$Exp(0,1)$	0.941	0.496	0.501
Cat. U_4	*Ord.* U_4	0.492	0.493	0.498
Cat. U_{15}	*Ord.* U_4	0.991	0.499	0.497

목표 변수의 클래스에 대한 정보를 두 예측 변수가 가지고 있지 않기 때문에 알고리즘이 변수 선택 편의가 없다면 변수 선택 과정에서 X_1은 0.5의 확률로 선택될 것이나 C4.5는 변수 선택 편의가 매우 심함을 알 수 있다. 연속형 변수와 개별값이 작은 연속형 변수 사이

에서 편의가 심하게 발생하고 범주형 변수 사이에서도 범주가 많은 쪽의 변수로의 편의가 심하다. 또한 연속형 변수보다 범주가 많은 범주형 변수로의 편의가 심하게 발생하고 있다. 반면 QUEST나 SRT는 작은 변수 선택 편의를 가지고 있다. SRT는 QUEST에 비해 변수 선택 편의가 조금이나마 작게 발생하고 있음을 볼 수 있다.

4.2. 변수 선택력

표4.2의 변수 선택력 실험에서 3가지 모두, 한 개의 예측 변수에는 목표 변수의 클래스에 대한 정보를 주고 나머지 19개의 예측 변수들은 목표 변수와 무관하게 설정을 했다. 정보를 가지고 있는 예측 변수가 (i) 개별값이 적은 연속형 변수인 경우, (ii) 범주 수가 적은 범주형 변수인 경우 그리고 (iii) 개별값이 많은 연속형 변수인 경우로 나누어 설정했다.

표 4.4: Loh와 Shih(1997)의 변수 선택력에 대한 모의 실험 결과

X_1		$X_2, \cdots, X_{20}$	$\Pr(X_1)$		
class1	class2		C4.5	QUEST	SRT
Ord. U_4	*Ord.* A_4	$N(0,1)$	0.004	0.161	0.162
Ord. U_4	*Ord.* A_4	$T(2)$	0.004	0.172	0.156
Ord. U_4	*Ord.* A_4	$Exp(0,1)$	0.004	0.170	0.156
Cat. U_4	*Cat.* B_4	*Cat.* U_{15}	0.115	0.406	0.410
$N(0,1)$	$N(0.25,1)$	*Cat.* U_{15}	0.096	0.385	0.159
$Exp(0,1)$	$Exp(0,1.3)$	*Cat.* U_{15}	0.098	0.427	0.159
Ord. U_4	*Ord.* B_4	*Cat.* U_{15}	0.072	0.417	0.410

표4.4는 Loh와 Shih(1997)가 계획한 모의실험에 의한 변수 선택력의 결과이다. C4.5는 모든 경우에 QUEST나 SRT보다 변수 선택력이 매우 떨어진다. SRT는 QUEST와 비교하여 연속형 변수에서만 변수 선택력이 떨어지고 나머지는 비슷한 수준이다. Loh와 Shih(1997)의 모의 실험에서는 QUEST가 변수 선택력이 제일 좋았다.

표 4.5: 왜도 변경을 통한 변수 선택력에 대한 모의 실험 결과

X_1		$X_2, \cdots, X_{20}$	$\Pr(X_1)$		
class1	class2		C4.5	QUEST	SRT
$N(0,1)$	$Exp(-1,1)$	*Cat.* U_{15}	0.096	0.385	0.159
$N(0,1)$	$Gam(-1.4,2,0.7)$	*Cat.* U_{15}	0.403	0.065	0.395
$N(0,1)$	$LN(-2.9,1,0.3^2)$	*Cat.* U_{15}	0.064	0.059	0.165

표4.5에서 QUEST가 C4.5나 SRT에 비해 변수 선택력이 떨어짐을 알 수 있다. 표4.5의

첫 번째 모의 실험에서 확연히 구분이 가는 정규분포와 지수분포임에도 불구하고 QUEST는 ANOVA F-검정을 이용한 변수 선택을 하기 때문에 변수 선택력이 떨어짐을 알 수 있다.

표 4.6: 첨도 변경을 통한 변수 선택력에 대한 모의 실험 결과

X_1		$X_2, \cdots, X_{20}$	$\Pr(X_1)$		
class1	class2		C4.5	QUEST	SRT
$N(0,1)$	$DE(1)$	*Cat.* U_{15}	0.369	0.755	0.750
$N(0,1)$	$T(2)$	*Cat.* U_{15}	0.305	0.387	0.317
$N(0,1)$	$T(1)$	*Cat.* U_{15}	0.928	0.415	0.846

표4.6에서 C4.5는 꼬리가 두텁지 않은 이중 지수분포와 관련이 있는 경우에만 변수 선택력이 약간 떨어졌다. QUEST는 꼬리가 매우 두터운 코쉬분포의 경우를 제외하고는 C4.5와 SRT보다 변수 선택력이 뛰어났다. SRT는 꼬리의 두터움에 상관없이 상대적으로 만족할 만한 변수 선택력을 보여주고 있다.

결론적으로 C4.5는 변수의 형태에 따라 변수 선택 편의가 심각하게 발생하고 반면 클래스의 구분이 확실한 경우에는 변수 선택력이 높음을 알 수 있다. 또한 QUEST는 변수 선택 편의가 작고 변수 선택력에 있어서 평균과 분산이 같고 정규성 가정이 위배된 경우 현저하게 떨어짐을 알 수 있고 SRT의 경우는 변수 선택 편의가 작지만 연속형 변수의 정보 손실로 인하여 변수 선택력이 떨어지는 경우가 있음을 알 수 있다.

5. 결론 및 향후 연구 방향

본 연구에서는 분류 나무에서 분리 변수를 선택하는데 있어서 편의가 작고 분포 가정과 무관한 변수 선택 알고리즘을 제안하고 기존의 변수 선택 알고리즘과 비교하여 특성과 장단점을 알아보는 모의실험을 실시하였다.

실험 결과 SRT이 변수 선택 편의가 작고 변수 선택력이 C4.5나 QUEST에 비해서 상대적으로 로버스트함을 알 수 있었다. 이는 분포의 가정이 없다는 분류 나무의 본래 특성을 유지하면서 기존 알고리즘에서 지적된 변수 선택의 문제를 보완했다는데 의의가 있다.

그렇지만 SRT도 역시 연속형 변수의 범주화로 인해 정보 손실을 초래하고 그로 인해서 변수 선택력이 떨어지는 경우도 있다.

이 실험 분리 변수 선택의 문제만을 고려하였지만 변수 선택 편의가 작을지라도 분리점 선택이 잘못된 경우에는 의미 없는 분류 나무가 형성 될 수도 있기 때문에 향후 분리점 선택의 문제도 다루어져야 할 것이다. QUEST의 경우는 분리점 선택에 이차판별분석을 사용하나, 이것 또한 정규성 가정에 근거한 아이디어이므로 분포에 무관한 방법을 적용시켜서 분류 나무를 완성하는 것도 의미가 있을 것이다.

마지막으로, 데이터 마이닝과 관련된 변수들은 결측값(missing value)이 많은 편인데 결측값을 가지는 중요한 변수가 있을 때 결측값에 의존하지 않고 변수 선택력이 좋아지는 방

법에 대한 연구도 필요할 것이다.

참고문헌

이윤모(2002). A study on bias problems in constructing classification trees, 박사학위논문. 서울대학교.

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

Kass, G. V. (1980). An Exploratory technique for investigating large quantities of categorical data, *Applied Statistics*, **29**, 119-127.

Kim, H. (1998). Multiway Split Classification Trees, Ph.D. Thesis, University of Wisconsin - Madison.

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

Loh, W. Y. and Shih, Y. S. (1997). Split selection method for classification trees, *Statistica Sinica*, **7**, 815-840.

Loh, W. Y. and Vanichsetakul, N. (1988). Tree-structured classification via generalized discriminant analysis (with discussion), *Journal of the American Statistical Association*, **83**, 715-728.

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

Quinlan, J. R. (1996). Improved use of continuous attribute in C4.5, *Journal of Artificial Intelligence Research*, **4**, 77-90.

Witten, I. H. and Frank, E. (1999). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann.

[2003년 7월 접수, 2004년 1월 채택]

A Study on Selection of Split Variable in Constructing Classification Tree*

S.S.Chung [1] S.Y.Kim [2] H.P.Lim [3]

ABSTRACT

It is very important to select a split variable in constructing the classification tree. The efficiency of a classification tree algorithm can be evaluated by the variable selection bias and the variable selection power. The C4.5 has largely biased variable selection due to the influence of many distinct values in variable selection and the QUEST has low variable selection power when a continuous predictor variable doesn't deviate from normal distribution. In this thesis, we propose the SRT algorithm which overcomes the drawback of the C4.5 and the QUEST. Simulations were performed to compare the SRT with the C4.5 and the QUEST. As a result, the SRT is characterized with low biased variable selection and robust variable selection power.

Keywords: Classification tree, Grouping, Peizer & Pratt transformation, Variable selection bias, Variable selection power.

* This paper was supported by research funds of Chonbuk National University.

1) Professor, Department of Statistical Informatics, Chonbuk National University, 664-14 1 ga Duckjin-Dong Duckjin-Gu Chonju Chonbuk, 561-756
E-mail: sschung@chonbuk.ac.kr

2) Dotoctoral Student, Department of Statistical Informatics, Chonbuk National University, 664-14 1 ga Duckjin-Dong Duckjin-Gu Chonju Chonbuk, 561-756
E-mail: rabbit@chonbuk.ac.kr

3) Graduated with Master's degree in Department of Statistical Informatics, Chonbuk National University, 664-14 1 ga Duckjin-Dong Duckjin-Gu Chonju Chonbuk, 561-756
E-mail: hanpil@hotmail.com