The Korean Communications
in Statistics Vol. 10 No. 3, 2003
pp. 627-635

# Bias Reduction in Split Variable Selection in C4.5[1)]

Sung-Chul Shin[2)], Yeon-Joo Jeong[3)], Moon Sup Song[4)]

## Abstract

In this short communication we discuss the bias problem of C4.5 in split variable selection and suggest a method to reduce the variable selection bias among categorical predictor variables. A penalty proportional to the number of categories is applied to the splitting criterion *gain* of C4.5. The results of empirical comparisons show that the proposed modification of C4.5 reduces the size of classification trees.

*Keywords* : classification tree, entropy, gain, gain ratio

## 1. 서론

의사결정나무는 훈련용 데이터를 이용하여 나무구조 형식의 예측모형을 만들고, 구축된 모형을 이용하여 새로운 개체의 예측변수로부터 목표변수의 값을 예측하게 된다. 의사결정나무에서 특히 목표변수가 범주형인 경우에 분류나무(classification tree)라고 한다. 분류나무 기법은 비모수적 방법으로서 예측변수로는 연속형과 범주형 모두 사용할 수 있으며, 결과의 해석이 용이한 장점이 있다.

분류나무는 주어진 분리기준(splitting criterion)에 따라 훈련용 데이터를 반복적으로 분리함으로써 얻어진다. 따라서 분리기준 또는 분리기준에 사용되는 최적화 측도에 의해 알고리즘의 특성이 결정된다. 예를 들어 CART(Breiman, Friedman, Olshen, and Stone, 1984)에서는 Gini 지수를 이용한 전체탐색법(exhaustive search method)을 주로 사용하며, C4.5(Quinlan, 1993) 또는 C5.0에서는 이득비율(gain ratio)을 최적화 측도로 사용하여 전체탐색법을 시행한다.

전체탐색법에서는 각 변수의 모든 가능한 분리방법에 대하여 주어진 측도의 값을 계산하고, 이 측도를 최적화시키는 분리방법을 선택한다. 그러나 전체탐색법에서는 변수선택 편의(bias)가 심각한 것으로 알려져 있다. 즉, 목표변수와의 연관성이 같다는 조건 하에서, 범주형 변수에서는 범주의 개수가 많은 변수가 그렇지 않은 변수보다 더 자주 선택되며, 연속형 변수에서는 별개값(distinct value)을 많이 갖는 변수가 그렇지 않은 변수보다 더 자주 선택되는 문제가 있다.

1) This research was supported in part by the Brain Korea 21 Project.
2) Consulting Division/Consultant, Systembusiness Co., Ltd., Seoul 150-872, Korea
E-mail : shin508@systembusiness.co.kr
3) CB Business Department/Associate, Korea Information Service, Seoul 150-010, Korea
E-mail : yeon1743@empal.com
4) Professor, Department of Statistics, Seoul National University, Seoul 151-742, Korea
Email : songms@plaza.snu.ac.kr (Corresponding Author)

이와 같은 편의의 문제는 분류나무 전체의 오분류율에는 큰 영향을 주지 않지만, 분류나무를 필요 이상으로 크게 만들고 따라서 분류나무의 해석을 어렵게 하는 단점이 있다. 본 논문에서는 C4.5에서 변수선택 편의 문제를 살펴보고, 편의를 줄이는 한 방법을 제시하고자 한다.

C4.5와 C5.0은 특히 상업용 패키지인 Clementine에 설치되어 있어 널리 사용되고 있다. 송문섭·윤영주(2001)는 Clementine의 C5.0을 포함한 여러 알고리즘에서 변수선택 편의를 비교하였으며, Huh and Lee(2003)는 Clementine에서 입력변수 중요도를 개선하는 방안을 제시하기도 하였다.

분류나무의 변수선택 편의에 관한 연구결과는 많은 저자들에 의해 발표되었다. 예를 들어 Quinlan(1996), Loh and Shih(1997), Kim and Loh(2001), 송문섭·윤영주(2001), Dobra and Gehrke(2001), Lee and Song(2002) 등에서 분류나무의 편의 문제를 다루었다. Loh and Shih(1997)와 Kim and Loh(2001), Lee and Song(2002)에서는 불편성의 성질을 갖게 하기 위하여 분리변수 선택과 분리점 선택을 두 단계로 나누는 방법을 제안하였다. Dobra and Gehrke(2001)는 분리변수 선택에서 편의를 수량화하여 정의하고, 분리기준의 $p$-값을 이용한 분리변수 선택법을 제안하였다. 특히 Gini 지수를 이용한 분리기준에 제안된 방법을 적용하여 불편성의 성질을 가질 수 있음을 보였다. 그러나 이 방법은 분리기준의 분포를 이용해야 하므로 실제 사용에서는 근사분포를 구해야 하는 문제점이 있다.

2절에서는 Quinlan(1993)의 C4.5 알고리즘에서 변수선택 방법과 편의의 문제점을 살펴보고, 연속형 변수에 대한 편의를 줄이기 위하여 Quinlan(1996)이 제안한 MDL(Minimum Description Length)을 이용한 분리기준을 살펴본다. 3절에서는 범주형 변수에 대한 수정된 분리기준을 제안한다. 4절에서는 제안된 분리기준을 모의실험 자료와 실제 자료에서 기존의 방법과 비교한다. 비교연구의 결과에 의하면 본 논문에서 제안한 방법은 변수선택 편의를 개선할 수 있으며, 실제 자료에서도 분류나무의 크기를 상당한 수준으로 줄이고 있음을 알 수 있었다.

## 2. C4.5와 C4.5 Rel 8

Quinlan(1993)이 제안한 C4.5는 기계학습(machine learning) 분야에서 기본 분류자(classifier)로 널리 사용된다. C4.5의 특징은 연속형 변수에 대하여는 이진분리(binary split)를 시행하고 범주형 변수에 대하여는 다원분리(multiway split)를 시행하며, 분리측도로는 이득비율을 사용하는 것이다. 이 알고리즘을 좀더 자세히 알아보기 위하여 예측변수를 $X$, 목표변수를 $y$, $y$가 취하는 계급(class)의 개수를 $C$라고 하자. 즉, $y \in \{1,2,\cdots,C\}$이다. 분리를 원하는 한 노드(node)에서 개체(case)들의 집합을 $D$라 하고, $D$에 속하는 개체들의 개수는 $|D|$ 또는 $N$으로 나타내기로 한다. 또한 $D$에서 $j$번째 계급에 속하는 개체들의 비율을 $P(D,j)$로 나타내기로 한다.

집합 $D$에 속하는 개체들에 대하여 각 개체가 속하는 계급을 인식하기 위한 평균 정보량을 $Info(D)$로 나타내고 다음과 같이 정의한다.

$$Info(D) = -\sum_{j=1}^{C}[P(D,j)\times \log_2(P(D,j))]$$

이 정보량은 흔히 집합 $D$의 엔트로피(entropy)라 부르기도 한다. 또한 $X$에 기초한 분리에서 집

합 $D$가 $k$개의 부분집합 $D_1, \cdots, D_k$로 분할될 때, 각 부분집합의 엔트로피의 가중평균을 $Info_X(D)$로 나타내면, 분리에 의해 얻어지는 정보량의 이득(gain)은 다음과 같다.

$$Gain(D, X) = Info(D) - Info_X(D)$$
$$= Info(D) - \sum_{i=1}^{k}\left[\frac{|D_i|}{|D|} \times Info(D_i)\right]$$

이와 같은 정보 이득은 $X$에 의해 분할되는 부분집합의 개수인 $k$에 큰 영향을 받으므로 이에 대한 조정이 필요하다. 집합 $D$가 $X$에 의해 $k$개의 부분집합으로 분할될 때 잠재적으로 발생하는 정보량을 분리정보(splitting information)라 하며, 다음과 같이 정의된다.

$$Split(D, X) = -\sum_{i=1}^{k}\left[\frac{|D_i|}{|D|} \times \log_2\left(\frac{|D_i|}{|D|}\right)\right]$$

이 분리정보는 $k$가 증가할수록 커지는 정보량이다. 따라서, $Gain(D, X)$를 $Split(D, X)$로 나누면 표준화의 효과를 기대할 수 있으며, 이를 이득비율(gain ratio)이라 한다. 즉,

$$GainRatio(D, X) = \frac{Gain(D, X)}{Split(D, X)}$$

C4.5에서는 이득비율 $GainRatio(D, X)$를 최대로 하는 분리방법을 찾게 되며, C5.0에서는 범주형 변수에 대한 범주의 병합 및 재분리를 디폴트로 시행한다. C4.5에서 분리 기준은 다음과 같다.

1. 각 연속형 예측변수에 대하여 '$X \le c$' (분리점 $c$는 연속된 두 데이터의 중앙값) 형태의 모든 분리 방법에 대하여 $GainRatio$를 계산한다.
2. 각 범주형 예측변수에 대하여는 $X$의 범주로 분할하는 분리 방법에 대하여 $GainRatio$를 계산한다.
3. 1과 2의 모든 분리 방법에서 $GainRatio$를 최대로 하는 방법으로 집합 $D$를 분할한다.

연속형 변수 $X$가 $d$개의 별개값을 갖는다면 분리점 $c$는 $d-1$개가 있다. 즉 $d-1$개의 각 경우에 대하여 집합 $D$는 $D_1$과 $D_2$로 분할되고 $GainRatio$가 계산된다. 따라서 목표변수 $y$와 연관성이 같은 두 연속형 변수에 대하여 $d$가 작은 경우보다는 큰 경우에 더 잘 뽑히는 편의 문제가 발생한다.

편의의 정도를 알아보기 위하여 간단한 모의실험을 시행하였다. 크기 $N=200$인 자료집합에서 예측변수 $X_1$과 $X_2$는 각각 10개와 200개의 별개값을 갖는 연속형 변수이며, 목표변수 $y$는 $X$ 값에 무관하게 계급값 1과 2를 50%씩 임의로 배정하였다. 이와 같은 자료집합에 C4.5를 적용하는 실험을 다시 200회 반복시행한 결과 $X_1$과 $X_2$가 분리변수로 선택된 비율은 <표 1>과 같다.

<표 1> 연속형 변수의 편의

| 변수 | $X_1$ | $X_2$ |
|---|---|---|
| 선택비율 | 0.185 | 0.815 |

$X_1$과 $X_2$는 모두 목표변수와 무관하므로 각각이 대략 0.5의 비율로 선택되어야 하지만 모의실험 결과로는 별개값이 많은 $X_2$가 0.815의 비율로 선택되어 편의가 심각함을 보여준다.

이와 같은 문제점을 보완하기 위하여 Quinlan(1996)은 MDL을 이용한 개선 방안을 제시하였다. 즉, 별개값이 많은 변수에 대하여는 MDL에 기초한 벌점(penalty)을 부여함으로써 편의를 수정하는 것이다. 연속형 변수가 $d$개의 별개값을 가질 경우에 '$X \le c$' 형식의 분리 방법이 $d-1$개 존재하며, 이로 인하여 $\log_2(d-1)$ 비트의 정보량이 추가로 발생한다. 따라서 $Gain(D,X)$에서 개체당 추가 정보량인 $\log_2(d-1)/N$을 벌점으로 부여하는

$$Gain'(D,X) = Gain(D,X) - \frac{\log_2(d-1)}{N}$$

을 새로운 이득의 측도로 제안하고, $Gain'$을 $Split$으로 나눈 $GainRatio$를 분리기준으로 사용하였다. 이와 같은 수정으로 편의가 제거되는 것은 아니지만 많이 개선되었으며, 이를 C4.5 Rel 8이라 부르기로 하였다. Quinlan(1996)은 Rel 8과 Rel 7을 UCI Repository(Blake and Merz, 1998)에 있는 20개의 자료집합에서 비교하였으며, 비교 결과에 의하면 분류나무의 크기는 평균 88%로 감소되고 오류율도 96%로 감소됨을 보였다.

## 3. 수정된 분리기준의 제안

C4.5에서 변수선택 편의의 문제는 연속형에서뿐만 아니라 범주형에서도 존재한다. 즉, 범주의 개수가 많은 경우에 분리정보 $Split(D,X)$에 의해 $Gain(D,X)$를 표준화시키는 것만으로는 편의 수정이 부족하다. 이를 확인하기 위하여 간단한 모의실험을 실시하였다. 예측변수 $X_1$은 범주의 개수가 5개, $X_2$는 범주의 개수가 $k=10, 20$인 경우에 목표변수 $y$는 $X$의 값과 무관하게 계급값 1과 2를 50%씩 임의로 배정하였다. 이와 같이 얻어진 크기 200인 자료집합에 C4.5를 적용하는 실험을 200회 반복시행한 결과 $X_1$과 $X_2$가 선택된 비율은 <표 2>와 같다.

<표 2> 범주형 변수의 편의

| $k$ \ 변수 | $X_1$ | $X_2$ |
|---|---|---|
| 10 | 0.220 | 0.780 |
| 20 | 0.095 | 0.905 |

<표 2>에서 보는 바와 같이 $X_2$의 범주의 개수가 증가할수록 $X_2$로의 변수선택 편의가 심각하게 증가함을 알 수 있다. 이와 같은 편의를 줄이기 위하여 범주형에서도 이득 $Gain$을 수정하는 방안을 제시하고자 한다.

2절에서 소개한 Quinlan(1996)의 제안은 $Gain$에 별개값의 개수에 비례하는 벌점을 주어 변수선택 편의를 완화한 것이다. 마찬가지 개념으로 범주형 예측변수에서도 $Gain$에 범주의 개수에

비례하는 벌점을 부여한 후에 *GainRatio*를 구하는 방안을 고려할 수 있다.

Akaike 정보기준 AIC는 로그-가능도 손실함수의 개념으로 다음과 같이 정의된다.

$$\mathrm{AIC}(R) = -2\ln(\text{ maximum likelihood}) + 2R$$

여기서 $R$은 모수의 개수로서 $R$이 커질수록 벌점이 증가하는 개념이며, AIC를 최소화하는 축소모형을 최적의 모형으로 선택한다. AIC에서 모수의 개수 $R$을 예측변수 $X$의 범주의 개수 $k$로 간주하면 이득함수 *Gain*에서도 같은 개념을 도입할 수 있다.

먼저 *Gain*을 최대가능도(maximum likelihood) 함수로 나타내 보자. 자료집합 $D$에서 한 개체가 $j$번째 계급에 속할 확률을 $p_j = P(D,j)\,(j=1,\cdots,C)$라 하고 $j$번째 계급에 속하는 개체의 개수를 $n_j$라 하면, $(n_1,\cdots,n_C)$는 다항분포에 따른다. 따라서 밑이 2인 로그를 사용할 경우에 $\theta=(p_1,\cdots,p_C)$의 로그 최대가능도 함수는 다음과 같이 나타낼 수 있다.

$$\ell(\hat{\theta}) \propto \sum_{j=1}^{C} n_j \times \log_2 \hat{p}_j$$

마찬가지 방법으로 예측변수 $X$에 의해 $D$가 $k$개의 부분집합 $D_1,\cdots,D_k$로 분할될 때, 각 $D_i$에서 $j$번째 계급의 확률을 $p_{ij} = P(D_i,j)$라 하고, $\theta_i = (p_{i1},\cdots,p_{iC})$의 로그 최대가능도 함수를 $\ell(\hat{\theta}_i)$라 하면

$$Gain(D,X) \propto -\frac{1}{N}[\,\ell(\hat{\theta}) - \sum_{i=1}^{k} \ell(\hat{\theta}_i)]$$

의 관계가 있다.

이와 같은 관계에 착안하여 범주형 예측변수 $X$에 대한 이득을 다음과 같이 수정한다.

$$Gain'(D,X) = Gain(D,X) - r \times \frac{k-2}{N}$$

여기서 $r$은 조절상수이다. 예측변수가 이진변수일 때는 수정항이 0이 되어 *Gain'*은 *Gain*과 같으며, $k$가 증가할수록 *Gain'*은 *Gain*보다 작게된다.

이론적으로 최적인 $r$값은 찾기가 어려우므로 모의실험을 이용하였다. 즉, $r$값을 결정하기 위하여 $r=1.0$부터 $0.1$씩 감소시키며 모의실험을 시행한 결과 $r=0.7$에서 편의가 가장 적은 것으로 나타났으며, 따라서 본 논문에서는 $r=0.7$을 사용하여 *Gain*의 값을 수정하기로 하였다. 즉, 범주형 변수에 대한 *Gain*값은

$$Gain'(D,X) = Gain(D,X) - 0.7 \times \frac{k-2}{N}$$

로 수정하고, *Gain'*을 분리정보인 *Split*으로 나눈 *GainRatio*를 사용할 것을 제안한다. 이와 같이 수정된 C4.5를 'Modified version of C4.5'라는 의미에서 C4.5_M이라 부르기로 하고, 다음절에서는 제안된 방법을 C4.5 Rel 8과 비교하고자 한다.

# 4. 비교연구

## 4.1 모의실험에 의한 비교

3절에서 제안된 C4.5_M을 C4.5 Rel 8과 비교하기 위하여 다음과 같은 모의실험을 하였다. 여기서 모든 프로그램은 Unix C로 구현하였다.

범주형 예측변수에 대한 편의를 알아보기 위하여 $X_1$은 2개의 범주, $X_2$는 15개의 범주를 갖는 크기 200 및 500인 자료집합에 목표변수 $y$는 1과 2를 50%씩 임의로 배정하였다. 이 자료집합에 C4.5 Rel 8과 C4.5_M을 적용하는 실험을 200번 반복시행한 결과 $X_1$과 $X_2$가 선택된 비율은 <표 3>과 같다. 표에 나타난 결과로는 제안된 방법인 C4.5_M이 C4.5 Rel 8보다 편의를 줄이고 있음을 알 수 있다.

<표 3> 범주형 변수의 편의 비교

| | $N=200$ | | $N=500$ | |
|---|---|---|---|---|
| | $X_1$ | $X_2$ | $X_1$ | $X_2$ |
| C4.5 Rel 8 | 0.110 | 0.890 | 0.150 | 0.850 |
| C4.5_M | 0.530 | 0.470 | 0.525 | 0.475 |

그러나 이와 같은 수정으로 범주형과 연속형 사이의 변수 선택 편의를 완전히 해소시키지는 못하고 있다. 이를 확인하기 위하여 다음과 같은 모의실험을 시행하였다. $X_1$은 15개의 범주를 갖는 범주형, $X_2$는 200개 및 500개의 별개값을 갖는 연속형인 자료집합에서 목표변수 $y$는 1과 2를 50%씩 임의로 배정하였다. 이와 같은 자료집합에 C4.5 Rel 8과 C4.5_M을 적용하는 실험을 200번 반복시행한 결과 $X_1$과 $X_2$가 선택된 비율은 <표 4>와 같다. 표에 나타난 결과에 의하면, C4.5 Rel 8과 C4.5_M 모두가 연속형으로 편의를 갖고 있으며, 제안된 방법보다는 C4.5 Rel 8이 더 심각함을 알 수 있다. 이와 같은 편의를 해결하는 문제는 앞으로의 연구과제가 될 수 있다.

<표 4> 변수의 형태에 따른 편의 비교

| | $N=200$ | | $N=500$ | |
|---|---|---|---|---|
| | $X_1$ | $X_2$ | $X_1$ | $X_2$ |
| C4.5 Rel 8 | 0.085 | 0.915 | 0.130 | 0.870 |
| C4.5_M | 0.275 | 0.720 | 0.345 | 0.655 |

## 4.2 실제 자료에 의한 비교

실제 데이터에서 C4.5 Rel 8과 C4.5_M을 비교하기 위하여 UCI Repository에서 범주형 변수가

많은 6개의 자료집합을 선택하였다. 각 자료집합의 특징은 <표 5>와 같다.

<표 5> 자료집합의 특징

| 자료집합 | 크기 | 목표변수 계급 개수 | 변수 개수 | |
|---|---|---|---|---|
| | | | 범주형 | 연속형 |
| anneal | 798 | 6 | 32 | 6 |
| breast-cancer | 699 | 2 | 9 | 0 |
| hayes-roth | 150 | 3 | 4 | 0 |
| house-votes | 435 | 2 | 16 | 0 |
| monk | 1296 | 2 | 6 | 0 |
| tic-tac-toe | 958 | 2 | 9 | 0 |

선택된 자료집합에 10겹(10-fold) 교차타당성 방법으로 C4.5 Rel 8과 C4.5_M을 적용시킨 후에 평균 나무크기(tree size)와 오류율을 구하였다. 또한 두 방법을 비교하기 위하여 나무크기와 오류율 각각에서 *Ratio*를 다음과 같이 구하였다.

$$Ratio = \frac{\text{C4.5\_M}}{\text{C4.5 Rel 8}}$$

C4.5에서는 주어진 정지규칙을 만족할 때까지 나무를 키운 다음에, 최적인 나무를 선택하기 위하여 가지치기(pruning)를 시행할 수 있다. 따라서 나무크기와 오류율을 가지치기 전과 후에 각각 구하였으며, 결과는 <표 6>에 수록되어 있다.

<표 6> 실제 자료에서 나무크기와 오류율의 비교

| 자료집합 | 가지치기 | 나무크기 | | | 오류율(%) | | |
|---|---|---|---|---|---|---|---|
| | | C4.5_M | C4.5 Rel 8 | *Ratio* | C4.5_M | C4.5 Rel 8 | *Ratio* |
| anneal | 전 | 36.1 | 44.4 | 0.813 | 4.1 | 4.1 | 1.000 |
| | 후 | 27.4 | 31.4 | 0.873 | 4.3 | 4.3 | 1.000 |
| breast-cancer | 전 | 80.0 | 129.0 | 0.620 | 5.9 | 6.7 | 0.881 |
| | 후 | 24.0 | 32.0 | 0.750 | 5.3 | 5.3 | 1.000 |
| hayes-roth | 전 | 26.5 | 30.1 | 0.880 | 31.1 | 31.1 | 1.000 |
| | 후 | 24.9 | 24.9 | 1.000 | 31.8 | 31.8 | 1.000 |
| house-votes | 전 | 26.2 | 26.2 | 1.000 | 3.7 | 3.7 | 1.000 |
| | 후 | 10.6 | 10.6 | 1.000 | 3.2 | 3.2 | 1.000 |
| monk | 전 | 153.6 | 153.0 | 1.004 | 41.5 | 41.5 | 1.000 |
| | 후 | 29.4 | 29.4 | 1.000 | 37.3 | 37.3 | 1.000 |
| tic-tac-toe | 전 | 179.2 | 179.2 | 1.000 | 12.8 | 12.8 | 1.000 |
| | 후 | 130.0 | 130.0 | 1.000 | 14.7 | 14.7 | 1.000 |
| 평균 | 전 | | | **0.89** | | | **0.98** |
| | 후 | | | **0.94** | | | **1.00** |

실제 데이터에서 비교한 결과에 의하면 제안된 방법은 C4.5 Rel 8에 비하여 오류율을 감소시키지는 못하지만 나무크기가 가지치기 전에는 89%로 가지치기 후에는 94%로 감소되었음을 볼 수 있다.

## 5. 결 론

분류나무의 구성에 사용되는 C4.5 알고리즘에서 변수선택 편의를 개선하는 방안을 제안하였다. Quinlan(1996)은 연속형 변수에 대하여 변수선택 편의를 감소시키는 방안을 제안하였다. 본 논문에서는 범주형 변수에서도 범주의 개수가 많은 변수로의 편의가 존재함을 확인하고, 이를 개선하는 방안을 제시하였다. 즉, 분리에서 얻어지는 이득의 측도인 *Gain*에서 범주형 예측변수의 범주수에 비례하는 벌점(penalty)를 부여하여 *Gain*값을 조정하였다. 제안된 방법은 모의실험 자료에서 편의를 현격하게 감소시켰으며, 범주형 변수가 많은 실제 자료에서 나무크기를 감소시키는 효과가 있었다.

일반적으로 편의는 오류율과 직접 관계되지는 않을 수 있다. 예를 들어 주어진 노드에서 $X_1$ 대신에 $X_2$가 선택되었다면, 하위 노드에 가서 다시 $X_1$이 선택되어 오류율 자체에는 큰 변화를 주지 않을 수 있다. 그러나 변수선택 편의로 인하여 나무크기가 증가하고 나무구조의 해석을 어렵게 하는 문제점이 있으므로 편의를 감소시키기 위한 노력이 필요하다. 본 논문에서 제안한 방법은 편의를 제거하는 것은 아니며, 다만 범주형 변수에서 편의를 줄이는 효과가 있다.

C4.5와 C5.0에서는 연속형과 범주형 사이에서도 편의가 존재한다. C4.5 Rel 8에서는 별개값이 많은 연속형으로의 편의가 존재하지만, C5.0에서는 범주의 개수가 많은 범주형으로의 편의가 존재한다. 따라서 이들에 대한 연구는 앞으로의 과제가 될 수 있다.

## 참 고 문 헌

[1] 송문섭, 윤영주 (2001), 데이터마이닝 패키지에서 변수선택 편의에 관한 연구, 「응용통계연구」, 제14권, 475-486.

[2] Blake, C.L. and Merz, C.J. (1998), UCI repository of machine learning databases (http://www.ics.uci.edu/~mlearn/MLRepository.html), Irvine, CA: University of California, Department of Information and Computer Science.

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

[4] Dobra, A. and Gehrke, J. (2001), Bias correction in classification tree construction, *Proceedings of the Seventeenth International Conference on Machine Learning*, 90-97.

[5] Huh, M.H. and Lee, Y.G. (2003), Input variable importance in supervised learning models, *The Korean Communications in Statistics*, Vol. 10, 239-246.

[6] Kim, H. and Loh, W.Y. (2001), Classification trees with unbiased multiway splits, *Journal*

*of the American Statistical Association,* Vol. 96, 589-604.

[7] Lee, Y.M. and Song, M.S. (2002), A study on unbiased methods in constructing classification trees, *The Korean Communications in Statistics*, Vol. 9, 809-824.

[8] Loh, W.Y. and Shih, Y.S. (1997), Split selection methods for classification trees, *Statistica Sinica,* Vol. 7, 815-840.

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

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

[ 2003년 7월 접수, 2003년 9월 채택 ]