The Korean Communications
in Statistics Vol. 10 No. 2, 2003
pp. 471-478

A Recursive Partitioning Rule for Binary Decision Trees[1)]

Sang Guin Kim[2)]

Abstract

In this paper, we reconsider the Kolmogorov-Smirnoff distance as a split criterion for binary decision trees and suggest an algorithm to obtain the Kolmogorov-Smirnoff distance more efficiently when the input variable have more than three categories. The Kolmogorov-Smirnoff distance is shown to have the property of exclusive preference. Empirical results, comparing the Kolmogorov-Smirnoff distance to the Gini index, show that the Kolmogorov-Smirnoff distance grows more accurate trees in terms of misclassification rate.

Keywords : Komogorov-Smirnoff distance, Binary decision tree, Split criterion.

1. 서론

입력변수(input variable)들에 의해 목표변수(target variable)의 범주(class)를 가능한 정확히 분류(Classification)하고 예측(prediction)하기 위한 의사결정나무(decision tree)는 적절한 분할기준(split criterion)에 의하여 주어진 자료를 반복적으로 분할해 나가면서 나무를 성장시키고, 성장된 나무는 분류 정확도와 단순화를 위하여 가지치기(pruning)를 수행하여 최종적인 나무를 결정하는 두 단계를 거치게 된다. 따라서 나무를 성장시키기 위하여 주어진 마디(node)에서 고려할 수 있는 분할들 중에서 가장 적절한 분할을 선택하기 위한 분할기준은 생성된 의사결정나무의 분류 예측력을 결정하는 가장 중요한 요소 중의 하나이다. 그러므로 주어진 자료로부터 고려할 수 있는 여러 분할 중에서 목표변수의 범주를 보다 잘 분류 혹은 예측하게 하는 분할을 찾기 위하여 많은 연구자들에 의해 여러 분할기준들이 제안되고 있으며, 이들은 일반적으로 좋은 분할, 즉 분할된 자료에 목표변수의 특정 범주가 다른 범주에 비해 많이 포함되어 있는 순수한 분할(purer partition)에서 큰 값을 갖는 측도들이라는 공통점을 가지고 있다.

여러 제안된 분할기준들로는 Kass(1980)와 Shih(1999)등과 같이 가능한 분할들의 분류 적합도(goodness of fit)를 측정하기 위한 측도들과 Clark와 Prgibon(1992), Quinlan(1993)등과 같이 엔트로피에 기반을 둔 분할 기준 등이 있다. 또한 Breiman, Freidman, Olshen과 Stone(1984)은 분리된 마디의 불순도(impurity) 감소량을 측정하기 위하여 Gini 기준을 제안하였고 Mola와 Sicilliano(1997)는 이와 동일한 분할을 결정하지만 분할의 선택을 보다 빠르게 수행할 수 있는 계

1) This paper was supported by a fund of a foreign dispatch program, 2001, Kyonggi university.

2) Professor, Division of Economics, Kyonggi University, Suwon, 442-760, Korea,
E-mail : sgkim@kyonggi.ac.kr

산의 효율성을 위하여 Gini 기준의 함수인 Goodman-Kruskal의 예측력지수(predictablility index)를 분할기준으로 제안하였다. 그리고 분할기준이 갖추어야 할 여러 조건들 중에서 배타적 선호특성(exclusive preference property)을 만족하는 Taylor와 Silverman(1993)이 제안한 MPI(mean posterior improvement) 기준 등이 있다.

이와는 달리 Friedman(1977), Gordon과 Olshen(1978) 그리고 Rounds(1980)는 목표변수의 특정 범주가 주어진 누적분포함수(conditional cumulative distribution)의 차이가 가장 큰 입력변수 값에 의해 분할을 결정하기 위하여 Komogorov-Smirnoff 거리를 분할기준으로 제안하였다. 특히 Utgoff와 Clouse(1996)은 Kolmogorove-Smirnoff 거리를 목표변수가 다범주인 경우로 확장한 보다 일반적인 분할을 선택하기 위한 알고리즘을 제안하였다. 그러나 그들의 연구는 입력변수가 연속형 변수인 경우에 촛점이 맞추어져 있고, 입력변수가 범주형 변수인 경우에도 연속형 변수에서 제안한 방법이 직접 적용될 수 있다고 언급하고 있다. 또한 제안된 방법에 의한 Kolmogorove-Smirnoff 거리의 특성을 판단하기 위하여 자신들이 제안한 ITI 알고리즘에 의하여 나무를 최대한 성장시킨 후에 역시 Gain ratio에 의하여 최대한 성장된 나무와의 비교 결과를 제시하였다.

본 논문에서는 입력변수가 셋 이상의 다범주를 갖는 범주형 변수인 경우에 보다 효율적으로 Kolmogorove-Smirnoff 거리를 얻기 위한 알고리즘을 제안할 것이다. 그리고 제안된 방법과 분할기준으로서 Kolmogorove-Smirnoff 거리가 갖는 특성을 파악하기 위하여 여러 자료를 이용한 분석 결과 또한 제시할 것이다. 먼저, 2절에서는 분할기준으로서 Kolmogorove-Smirnoff 거리를 Friedman(1977)의 정의에 따라 재고찰하고 입력변수가 셋 이상의 다범주를 갖는 범주형 변수인 경우에 보다 효율적으로 Kolmogorove-Smirnoff 거리를 얻기 위한 방법을 제안하고자 한다. 또한 Kolmogorove-Smirnoff 거리가 분할기준이 갖추어야할 바람직한 조건인 배타적 선호특성을 만족한다는 것을 Taylor와 Silverman(1993)이 제시한 가상자료를 통해 밝힐 것이다. 3절에서는 Merz와 Murphy(1996)에서 인용한 여러 자료들을 이용하여Kolmogorove-Smirnoff 거리와 Gini 기준에 의해 나무를 최대한 성장시킨 후에 두 기준에 의해 생성된 나무의 특징을 비교하였고, Breiman, Freidman, Olshen과 Stone(1984)이 제안한 cost-complexity 방법을 이용하여 가지치기를 수행한 후 두 기준이 갖는 특징을 토론하였다. 마지막으로 4절에서는 본 논문의 결과를 정리하였다.

2. 이진의사결정나무의 성장을 위한 Kolmogorov-Smirnoff 거리

주어진 마디 t에서 나무의 성장은 두 하위마디(subnode) t_L과 t_R만을 갖는 이진 의사결정나무와 목표변수가 A와 B의 주 범주만을 갖는다고 하자. 또한 두 범주 A와 B에 대하여 이들 범주가 주어진 입력변수 X의 조건부 분포함수 $f_A(x)$, $f_B(x)$와 이들의 누적분포함수 $F_A(x)$와 $F_B(x)$를 고려하기로 한다.

$\vec{x}$를 연속형 입력변수 X가 가질 수 있는 모든 값을 가진 집합이라고 하면 Friedman(1977)에 의하여 다음을 만족하는 α를 입력변수 X의 최적 분할점(optimal cutpoint)으로 생각할 수 있다.

$$\max_{\alpha \in \overrightarrow{x}} \{|F_A(\alpha) - F_B(\alpha)|\} \quad (1)$$

식 (1)은 Kolmogorov-Smirnoff 거리가 이다. 따라서 주어진 입력변수에 의한 분할은 두 누적분포함수값의 차이 중에서 가장 큰 값인 Kolmogorov-Smirnoff 거리를 갖게 하는 값에 의해 결정할 수 있다. 즉, $x < \alpha$인 자료들은 마디 t_L에 속하며 그렇지 않은 경우에는 마디 t_R에 속하게 하여 나무를 성장시킬 수 있다. 이때 식 (1)의 값은 경험적 누적분포함수 $\widehat{F_A}(\alpha)$와 $\widehat{F_B}(\alpha)$에 의해 얻을 수 있다.

목표변수가 A와 B두 범주만을 가지고 있으므로 마디 t에 주어지 자료로부터 분할점 α에 의해 2×2분할표를 구성할 수 있다. 즉, 분할점 α에 의해 왼쪽마디에 속하는 비율을 p_{AL}과 p_{BL} 그리고 오른쪽 마디에 속하는 비율을 p_{AR}과 p_{BR}과 같이 표현하기로 한다. 이로부터 연속형 입력변수의 분할점 α에 대하여 $p_{AL} = \widehat{F_A}(\alpha)$ 그리고 $p_{BL} = \widehat{F_B}(\alpha)$의 관계가 성립한다. 그러므로 α를 최적 분할점이라고 하면 식 (1)의 Kolmogorov-Smirnoff 거리는 단순히 왼쪽마디에 속하는 두 목표변수의 범주 비율의 차이의 절대값인 $|p_{AL} - p_{BL}|$과 같아진다.

만일 입력변수가 범주에 순서가 부여된 순위범주를 갖는 범주형 변수, 다시 말해 순위변수(ordinal variable)라면 연속형 입력변수에 대한 Kolmogorov-Smirnoff 거리는 쉽게 확장될 수 있다. 분할점 α는 각 순위범주가 될 것이고 분할점 α에서의 $\widehat{F_A}(\alpha)$ 그리고 $\widehat{F_B}(\alpha)$는 각각 두 목표변수의 범주 A와 B에 대하여 분할점 보다 순서가 낮은 비율들의 합과 분할점 보다 순서가 높은 범주 비율들의 합으로 나타날 것이다. 결국 순위변수에 대한Kolmogorov-Smirnoff 거리는 연속형 변수에서와 같이 $|p_{AL} - p_{BL}|$에 의해 얻을 수 있다.

입력변수가 두 가지, '가'와 '나', 범주를 갖는 명목변수(nominal variable)인 경우를 고려하기로 한다. 이 경우에 분할점은 x='가' 혹은 x='나'가 될 것이고 두 범주 '가'와 '나'에는 순서가 부여되지 않았으므로 식 (1)의 Kolmogorov-Smirnoff 거리는 다음과 같이 표현될 수 있다.

$$\max\{|\widehat{F_A}(\text{가}) - \widehat{F_B}(\text{가})|, |\widehat{F_A}(\text{나}) - \widehat{F_B}(\text{나})|\} = \max\{|p_{A\text{가}} - p_{B\text{가}}|, |p_{A\text{나}} - p_{B\text{나}}|\} \quad (2)$$

그러나 입력변수가 셋 이상의 범주를 갖는 명목변수라면 식 (2)는 직접 적용될 수 없다. 이러한 경우에 Utgoff와 Clouse(1996)는 주어진 입력변수의 범주들로부터 모든 가능한 두 부분집합을 열거(enumeration)한 이후에 이들 두 부분집합의 비율로부터 식 (2)를 통하여 Kolmogorov-Smirnoff 거리를 얻을 수 있다고 언급하고 있다. 그러나 만일 입력변수가 J개 범주 $C=\{1,2,...,J\}$를 가지고 있다면 서로 다른 원소를 갖는 두 범주집합은 $2^{J-1}-1$개가 존재하고, 이는 각 $2^{J-1}-1$개 부분집합에 대하여 식 (2)를 계산하여야 하는 것을 의미한다.

이러한 계산문제를 해결하기 위하여 다음과 같은 경험적 기준을 고려하기로 한다.

$$C_L = \{j : p_{Aj} \geq p_{Bj}\}, \quad (3)$$

여기서 p_{ij}는 목표변수가 i 범주이고 입력변수는 j인 경우의 비율을 나타낸다. 식 (3)으로부터 두 번째 범주는 $C_R = C - C_L$과 같이 얻을 수 있다. 결국 $C_L \cap C_R = \phi$이고 $C = C_L \cup C_R$가 될 것이다.

마디 t에 주어진 자료에서 목표변수의 주변비율 $p_A = \sum_{j=1}^{J} p_{Aj}$와 $p_B = \sum_{j=1}^{J} p_{Bj}$는 주어진 값이고 또한 입력변수의 비율 $p_j = p_{Aj} + p_{Bj}$ 역시 주어진 값이므로 식 (3)에 의한 범주 집합 C_L에 대하여 $p_{AC_L} \geq p_{BC_L}$인 관계를 만족한다. 더욱이 $p_{AC_L} + p_{AC_R} = p_{BC_L} + p_{BC_R} = 1$ 이므로 $p_{AC_R} \leq p_{BC_R}$이다. 따라서 식 (2)는 식 (3)에 의한 범주 C_L에 의해

$$\max\{|p_{AC_L} - p_{BC_L}|, |p_{AC_R} - p_{BC_R}|\} \tag{4}$$

와 같이 표현할 수 있다. 예를 들어 <표 1>과 같이 입력변수가 세 범주를 가진 가상 자료를 고려하기로 한다.

	범주		
	가	나	다
Class A	0.60	0.10	0.30
Class B	0.35	0.45	0.20

<표 1> 입력변수가 세 범주를 갖는 가상자료

<표 1>에서 각 칸의 값은 해당 범주조합의 비율을 나타낸다. 모든 가능한 부분집합을 고려할 경우에 식 (2)의 Kolmogorov-Smirnoff 거리는 범주 {가, 다} 그리고 범주 {나}인 경우에 '0.35'인 것을 알 수 있다. 이제 식 (3)의 경험적 기준을 고려하면 $C_L = \{가, 다\}$ 따라서 $C_R = \{나\}$이므로 식 (4)에 의해 <표 1>의Kolmogorov-Smirnoff 거리는 0.35로 쉽게 얻을 수 있다.

목표변수가 다범주인 경우에 Utgoff와 Clouse(1996)는 Brieman, Friedman, Olshen과 Stone(1984)이 Twoing 기준을 위해 제안한 초범주(super class)를 적용하여 Kolmogorov-Smirnoff 거리를 얻는 방법을 제안하고 있다. 제안된 식 (4) 역시 목표변수가 셋 이상인 경우 직접 적용할 수 없으므로 역시 먼저 초범주를 결정한 뒤에 경험적 기준 (3)을 적용하는 방법을 고려할 수 있다.

여기서 Taylor와 Silverman(1984) 그리고 Shih(1999)에 의해 분할기준이 갖추어야 할 배타적 선호특성을 설명하기 위해 제시한 <표 2>와 같은 가상자료를 고려해보기로 한다.

	왼쪽 마디	오른쪽 마디
Class A	40	0
Class B	20	0
Class C	0	10
Class D	0	10

(가) 분할점 α_1

	왼쪽 마디	오른쪽 마디
Class A	40	0
Class B	3	17
Class C	10	0
Class D	7	13

(나) 분할점 α_2

<표 2> 분할점 α_1과 α_2에 따른 가상자료

	오분류률		잎의 수	
	Gini	K-S	Gini	K-S
Post-operative	0.5402 (0.0534)	0.3793 (0.0520)	21.4 (3.4383)	20.1 (2.4244)
Heart Disease	0.2555 (0.0265)	0.2518 (0.0264)	30.8 (2.4404)	31.2 (2.9364)
Iris	0.0600 (0.0194)	0.0733 (0.0213)	4.9 (0.0739)	6.0 (0.9428)
Wine	0.0843 (0.0208)	0.0786 (0.0202)	8.1 (2.3781)	6.1 (2.8451)
Balance scale	0.2064 (0.0162)	0.2016 (0.0161)	61.9 (3.9847)	61.8 (2.3944)

<표 3> 6개 자료에 대한 추정 오분류률과 잎의 수

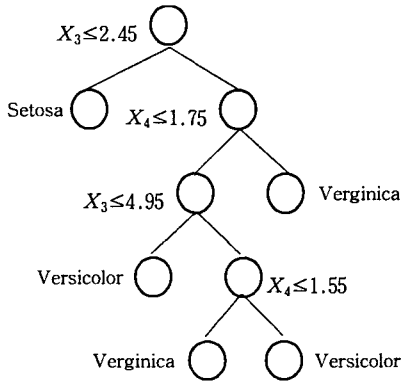
<표 2>에서 분할점 α_1에 의한 분할이 분할점 α_2에 의한 분할에 비해 보다 바람직한 분할이라는 것은 너무도 명확하다. 그러나 이들 두 분할에 대하여 Gini 기준은 각각 0.1979와 0.2081로 얻어져 분할점 α_2에 의한 분할을 선택하게 한다. 그러나 초범주는 <표 2>의 (가)에 대해서는 각각 $\{A, B\}$와 C, D 그리고 (나)에 대해서는 A, C와 B, D로 얻어지므로 이들에 의한 Kolmogorov-Smirnoff 거리는 각각 1과 0.75로 구할 수 있고, 결국 분할점 α_1에 의한 분할을 선택하게 하므로 분할기준으로서의 Kolmogorov-Smirnoff 거리는 배타적 선호특성을 만족하는 것을 알 수 있다.

3. Gini 기준과 Kolmogrov-Smirnoff 거리와의 비교

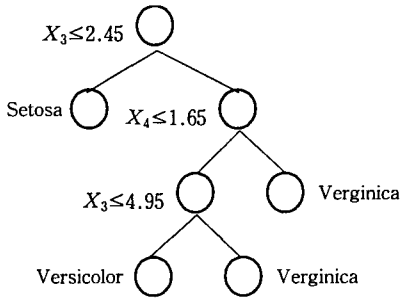
이진 의사결정나무의 생성을 위한 분할기준으로서의 Kolmogorov-Smirnoff 거리의 특성을 파악하기 위하여 Merz와 Murphy(1996)에서 발췌한 Post-operative 등 5개 자료를 이용하여 Gini 기준과 비교하고자 한다. 두 기준에 의한 나무는 잎(leaf)에 세 개 관찰값이 속할 때 까지 성장시켰으며, 각 기준에 의한 오분류률(misclassification rate)은 10-fold 교차타당성(cross validation)방

법에 의하여 추정하였다. 이들 자료에 대한 추정 오분류률과 잎의 수에 관한 결과가 <표 3>에 정리되어 있다. <표 3>에서 잎의 수는 각 fold에서 성장된 나무의 잎의 수들의 평균이며, 표의 각 칸에 오분류률과 잎의 수 하단 괄호안의 값은 표준편차를 나타낸다.

<표 3>으로부터 5개 자료 중엣 Iris 자료를 제외한 나머지 자료에서는 Kolmogorov-Smirnoff 거리에 의해 생성된 나무의 오분류률이 적게 추정된 것을 알 수 있다. 마찬가지로 Iris 자료를 제외한 나머지 자료의 완전히 성장된 나무의 잎의 수는 Gini 기준에 의해 성장된 나무의 잎에 수에 비하여 큰 차이가 나지 않는 것을 알 수 있다. 이러한 사실로부터 분석에서 사용된 자료에서는 Kolmogorov-Smirnoff 거리가 Gini 기준에 비해 대체적으로 분류와 예측을 위해 우수한 나무를 생성하는 것으로 판단할 수 있다. 그러나 <표 4>를 살펴보도록 한다. <표 4>는 <표 3>의 각 나무에 대하여 Brieman, Friedman, Olshen과 Stone(1984)이 제안한 cost-complexity 방법에 의한 가지치기를 수행한 결과이다.



(가) Gini



(나) Kolmogorov-Smirnoff Distance

<그림 1> Iris 자료에 대한 1-SE 방법에 의해 선택된 나무

	오분류률		잎의 수	
	Gini	K-S	Gini	K-S
Post-operative	0.2874	0.3103	2	2
Heart Disease	0.2074	0.2000	16	11
Iris	0.0600	0.0533	5	4
Wine	0.0843	0.0955	10	6
Balance scale	0.1986	0.2064	11	8

<표 4> cost-complexity 방법에 의한 가지치기 결과

<표 4>로부터 가지치기를 수행한 후에 Kolmogorov-Smirnoff 거리에 의해 생성된 나무의 잎의 수가 적게 나타난 것을 알 수 있다. 그러나 잎의 수가 적은 간결한 나무를 생성한 것에 비하여 <표 3>에서의 결과와는 상이하게 Gini 기준에 의해 생성된 나무에 비해 오분류률은 조금 높게 추정된 것을 발견할 수 있다. 한 가지 주목할 점은 Iris 자료에 대하여 완전히 성장된 나무로부터 가지치기를 수행한 후에 Kolmogorov-Smirnoff 거리에 의한 나무가 오분류률도 적고 잎의 수도 적은 간결한 나무를 생성했다는 것이다.

<그림 1>은 Iris 자료의 두 기준에 의해 생성된 가지치기를 수행한 후의 나무를 나타낸다. <표 4>의 결과에서와 같이 Iris 자료인 경우에 Kolmogorov-Smirnoff 거리에 의해 결정된 나무는 잎의 수가 적은 것을 알 수 있다. 특히 분할에 사용된 입력변수는 모두 같고 단지 오른쪽 첫 가지에서 사용된 분할점만이 다르게 선택된 것만으로 Kolmogorov-Smirnoff 거리에 의한 결과가 잎의 수는 적고 오분류률은 적은 좋은 나무를 생성해준 것을 알 수 있다.

5. 결론

본 연구에서는 이진 의사결정나무의 성장을 위하여 주어진 마디에서 가장 적절한 분할을 선택하기 위한 분할기준으로 Kolmogorov-Smirnoff 거리를 재조명하고, Utgoff와 Clouse(1996)의 결과를 확장하여 범주형 입력변수인 경우에 Kolmogorov-Smirnoff 거리의 계산을 효율적으로 수행할 수 있는 방법을 제안하였다. 또한 분할기준으로서의 Kolmogorov-Smirnoff 거리는 Taylor와 Silverman(1993)이 제시한 배타적 선호특성을 만족한다는 것을 그들이 제시한 가상자료를 통해 경험적으로 밝혔다.

Kolmogorov-Smirnoff 거리와 제안된 계산방법에 의해 생성된 나무의 특징을 살펴보기 위하여 Merz와 Murphy(1996)에서 발췌한 5개 자료를 분석한 결과 완전히 성장된 나무에서의 Gini 기준과의 비교에서는 대체적으로 오분류률이 적게 추정된 결과를 얻을 수 있었다. 그러나 cost-complexity 가지치기를 수행한 후의 나무에서는 Gini 기준에 비해 잎의 수가 적은 나무를 생성해주는 것을 알 수 있었다.

참고문헌

[1] 전병환, 김창수, 송홍엽, 김재희(1997). A new criterion in selection and discretization of attributes for generation of decision trees, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19, 1371-1375.

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

[3] Clark, L. A. and Pregibon, D.(1992). Tree-based models, In: J. M. Chambers and T. J. Hastie (eds) *Statistical Models in S*, Wadsworth & Brooks/Cole.

[4] Friedman, J. H.(1977). A recursive partitioning decision rule for nonparametric classification. *IEEE Transactions of Computers*, C-26, 404-408.

[5] Goodman, L.A. and Kruskal, W.H.(1954). Measures of association for cross-classifications, *Journal of the American Statistical Association*, 49, 732-764.

[6] Gordon, L. and Olshen, R. A.(1978). Asymptotically efficient, computationally feasible solutions to the classification problem. *Annals of Statistics*, 6, 515-533.

[7] Kass, G. V.(1980). An exploratory technique for investigation large quantities of categorical data, *Applied Statistics*, 29, 119-127.

[8] Merz, C. J. and Murphy, P. M.(1996). *UCI Repository of Machine Learning Databases*, Department of Information and Computer Science, University of California, Irvine, CA.

[9] Mola, F. and Siciliano, R.(1997). A fast splitting procedure for classification trees, *Statistics and Computing*, 7, 209-216.

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

[11] Rounds, E. M. (1980). A combined nonparametric approach to feature selection and binary decision tree design. *Pattern Recognition*, 12, 313-317.

[12] Shih, Y. S.(2001). Selecting the best splits for classification trees with categorical variables, *Statistics & Probability Letters*, 54, 341-345.

[13] Taylor, P. C. and Silverman, B. W.(1993). Block diagrams and splitting criteria for classification trees, *Statistics and Computing*, 3, 147-161.

[2003년 7월 접수, 2003년 8월 채택]