단측 순수성에 의한 나무모형의 성장에 대하여

김 용 대*, 최 대 우**

On the Tree Model grown by one-sided purity

Yongdai Kim, Daewoo Choi

요 약

의사결정 나무라고 불리우기도 하는 나무모형은 결과 해석의 용이성으로 데이터마이닝의 분류예측 모형으로써 큰 각광을 받고 있다. 현재 나무모형으로 가장 많이 사용되는 Breiman *et al.*의 CART나 Quinlan의 C4.5 모두 생성된 노드들의 자료 구성이 목표변수를 기준으로 수준 구성비 측면에서 순수해지도록 진행된다. 그러나 CRM에 있어 가장 흔한 주제인 해지예측을 위한 모델링을 실시하는 경우 관심의 대상인 해지자가 전체 자료에 극히 일부를 차지하여, 기존의 분할 방법에서와 같이 모든 노드의 순수성을 고려하기란 불가능하다.

Buja와 Lee는 이와 같이 소수의 관심에 대상이 되는 부류를 찾아내기 위한 나무모형 생성방법을 소개하였다. 즉, 해지자 관리가 중요한 경우 해지자와 비해지자 구분을 진행하는 기존의 방법과는 달리 전체 자료 중 해지자를 집중적으로 찾아가는 탐색적 분할 기준인 단측 순수성(one-sided purity)을 제안하였다.

본 연구에서는 단측 순수성에 의한 나무모형을 모 PC통신 회사의 해지자 자료에 적용하여 기존의 방법과 비교하였고 몇 가지 시뮬레이션 자료를 통해 단측 순수성의 문제점과 앞으로 해결하여야 할 과제에 대하여 살펴보았다.

Key words: tree model, split-rule, one-sided purity

1. 서론

의사결정 나무(decision tree)라고 불리우는 나무모형은 결과 해석의 간결함으로 통계분석, 기계학습 분야뿐 아니라 마케팅, 그리고 CRM(Customer Relationship Management) 내의 데이터마이닝 분석기법으로 애용되고 있다. 그러나 분류(classification)를 위한 나무모형에서 흔히 사용되는 Gini나 entropy 지수와 같은 분할기준은 생성되는 두 노드의 균형을 고려한 것이므로 목표변수(target variable)의 수준구성이 극단적인 경우 사용하기 곤란하다.

예를 들어, 이동통신 회사의 월별 해지자를 예측하여 고객을 관리하는 경우를 생각해 보자. 일반적으로 월별 해지율은 2% 이내로 분류모형이 모든 고객은 서비스를 해지하지 않는다고 예측한다면 예측 정확도가 98%에 달하게 된다. 즉, 관심의 대상이 극소수인 상황에서 기존의 나무모형을 이용하여 우수한 분류모형을 도출하기는 무척 힘든 것이다. 이와 같이 데이터마이닝을 실제로 활용하는데 있어 소수의 관심 대상을 분류하는 모형을 도출하여야 하는 경우가 많은 것이다.

본 연구에서는 분류모형에 있어 목표변수의 관심 수준 구성비가 극소수인 경우 적용할 수 있는 단측 순수성(one-sided purity)에 대하여 분류 정확도면과 생성하는 나무모형의 형태를 관찰한 후 기존 분할법칙과의 차이점과 향후 개선방안에 대하여 살펴 보았다.

2. 기존 분할법칙

나무모형은 각 노드에서 분할이 일어나면서

* 한국외국어대학교 정보통계학과 전임강사 (kimy@stat.hufs.ac.kr)

** 한국외국어대학교 정보통계학과 조교수 (dachoi@dreamwiz.com)

자라나게 된다. 자료의 분할과정은 분할 후 각 노드에 속하는 자료의 순수도(purity)가 가장 크게 증가하도록 진행되는데, 이를 위해 분할 기준이 되는 변수와 분할의 위치를 결정하여야 한다. 순수도의 증가란, 분할 후 각 노드에 속하는 자료의 구성이 어느 한 수준에 속하는 자료의 비율이 높다는 의미이다.

나무모형에서 순수도의 최대 증가란 분할 후 불순도(impurity)의 감소 폭이 가장 크다는 사실과 동일하다. 그러한 맥락에서 노드에서의 분할원칙을 다음과 같이 정리할 수 있다.

2.1 불순도 함수

나무모형 T의 전체 불순도 $D(T)$는 다음과 같이 구한다.

$$D(T) = \sum_{g \in G} \phi(g) p(g)$$

여기서 G는 나무모형 T의 종료 노드의 집합이고 $p(g)$는 종료 노드 g에 속할 확률이다. 예를 들어, 전체 자료 100개중 종료 노드 g에 20개의 자료가 포함된다면 $p(g) = 20/100 = 0.2$이다. 그리고 $\phi(g) = \phi(p_1(g),\ldots,p_J(g))$는 불순도 함수이다. 여기서 $p_j(g),(j=1,\ldots,J)$들은, 예를 들어, 노드 g에 속하는 자료들에 대한 해당 부류(class)의 비율 (혹은 그 추정치 $\hat{p}_j(g)$)이라고 생각할 수 있다. 즉, 남자와 여자의 비율이 4:1이라면 $(\hat{p}_1(g),\hat{p}_2(g)) = (4/5,1/5)$라고 표현할 있다. 좀더 자세한 불순도 함수 $\phi(g)$에 대한 사항은 Breiman *et al.*(1984)을 참조하길 바란다.

종료 노드 g의 불순도를 D_g는 다음과 같이 표현된다.

$$D_g = \phi(g) p(g). \qquad (1)$$

노드 g에서 두 개의 노드로 분할하기 위해서는 아래의 값이 최대가 되도록 두 개의 노드 g_L와 g_R로 분할된다.

$$D_g - D_{g_L} - D_{g_R}$$

즉, 노드 g에서 두 개의 노드로 분할되면서 각 노드의 불순도 D_{g_L}과 D_{g_R}이 감소될 것이다. 최적의 분할은 D_{g_L}과 D_{g_R} 값이 최대로 감소하여 $D_g - D_{g_L} - D_{g_R}$이 커지도록 이루어져야 할 것이다. 식 (1)의 관계를 이용하여 정리하면 다음과 같다.

$$D_g - D_{g_L} - D_{g_R} = \phi(g)p(g) - \phi(g_L)p(g_L) - \phi(g_R)p(g_R) \qquad (2)$$

노드 g에서 노드 g_L과 g_R로 분할된 확률을 P_L, P_R이라고 하면 다음과 같이 표현된다.

$$P_L = p(g_L)/p(g), \quad P_R = p(g_R)/p(g).$$

결국 식 (2)는 아래의 식 (3)과 같이 재표현 될 수 있다.

$$D_g - D_{g_L} - D_{g_R} = [\phi(g) - \phi(g_L)P_L - \phi(g_R)P_R] \cdot p(g) \qquad (3)$$

식 (3)을 살펴보면 대괄호 안의 계산된 수가 클수록 분할 후 전체 불순도가 크게 감소할 것이라는 것을 알 수 있다. 다시 말해 노드 g에서의 분할은 불순도 함수 값의 감소 폭이 최대가 되도록 실행되어야 한다.

2.2 불순도 함수의 종류

함수 ϕ로 표현될 수 있는 불순도 함수로는 다음과 같은 것들이 있다.

① Gini 지수

$$\phi(g) = \sum_j \hat{p}_j(g)(1 - \hat{p}_j(g)).$$

② Entropy 지수

$$\phi(g) = -\sum_j \hat{p}_j(g) \log \hat{p}_j(g).$$

③ Deviance

$$\phi(g) = -2\sum_j n_j \log \hat{p}_j(g).$$

Gini 지수는 CART에서, entropy 지수는 C4.5에서 사용되는 분할 기준이다. Deviance는 AT&T research에 의해 개발된 것으로 S-PLUS라는 통계 소프트웨어에 구현되어 있다.

3. 단측 순수성

2절에서 살펴본 바와 같이 기존 나무모형의 분할은 임의 노드 g로부터 두 개의 노드 g_L와 g_R로 분할될 때, 두 노드의 불순도가 최대한 감소되도록 진행된다. 그러나 목표변수의 구성비가 불균형이 심한 경우 생성되는 두 노드의 순수성을 동시에 고려한 분할기준 하에서는 정확도 향상을 기대할 수 없다. 이와 같은 문제에 활용될 수 있는 나무모형의 분할기준으로써 Buja와 Lee(1999)의 단측 순수성(one-sided purity)와 단측 극단성(one-sided extreme)을 제안하였다.

Buja 와 Lee는 데이터마이닝 분석의 분류과제에 있어 분류 예측하려고 하는 목표변수 중 특정 수준에 대한 설명에 관심이 있는 경우 생성되는 두 노드의 순수성을 동시에 고려하기보다는 관심의 대상이 되는 특정 수준의 발굴에 중점을 둔 분할 기

준에 대하여 연구한 것이다. 즉, 이동통신사의 경우 해지자, 자동차 보험사에 있어 장기보험까지 가입하는 고객의 발굴 등이 관심의 대상이 명확한 데이터 마이닝 과제에 해당된다.

임의의 노드 g 에서 분류 수준이 두 개(0과 1)인 경우의 분할 기준 단측 순수성 다음과 같다.

$$\min(P_L^0, P_L^1, P_R^0, P_R^1) \qquad (4)$$

여기서 P_L^0 은 생성되는 왼쪽 노드의 수준 0의 확률(여기서는 구성비율)이며 나머지 기호도 같은 요령으로 이해하면 된다. 위의 기준 (4)는 다음과 동일하다.

$$\min(P_L^0 P_L^1, P_R^0 P_R^1)$$

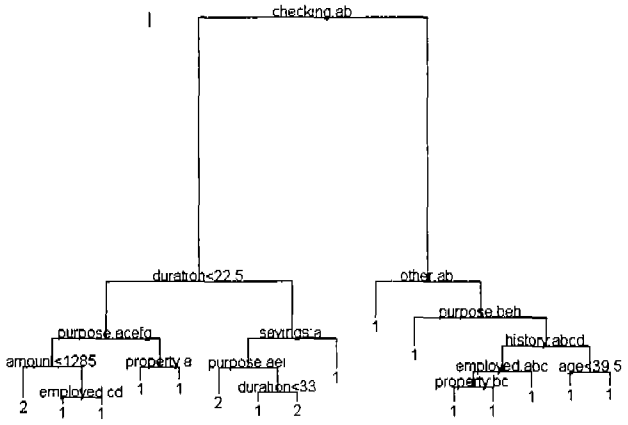

그림-1 Deviance에 의해 생성된 나무모형

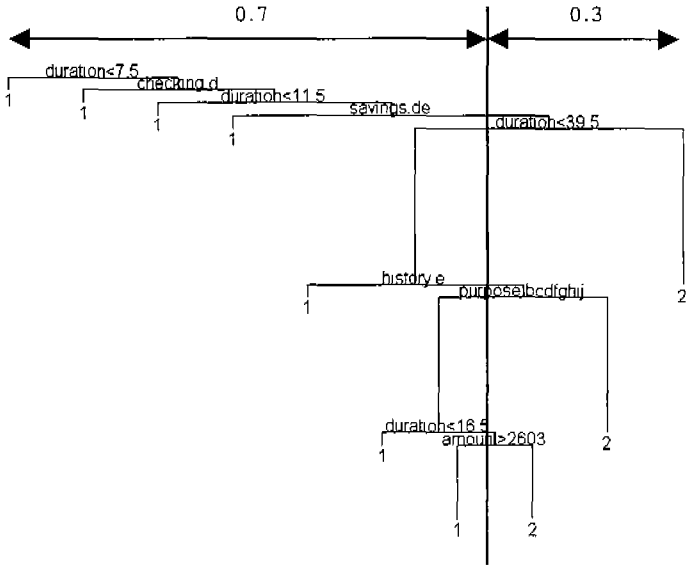

그림-2 단측 순수성에 의해 생성된 나무무형

4. 실제자료 및 실험자료 분석

4.1 German 신용평가 자료분석

German 신용평가 자료는 총 1,000개의 관측치로 이루어져 있고 20개의 설명변수 중 7개는 이산형이고 나머지는 연속형 변수이다. 목표변수는 관측치의 신용도로 신용우량, 불량 두 개의 수준으로 이루어져 있다.

그림-1은 deviance에 의해 생성된 나무 모형이고 그림-2는 단측 순수성에 의해 도출된 모형이다. 그림-2의 경우 신용우량과 불량을 예측하는 종료 노드들이 나무의 중심 축으로부터 좌, 우로 배열되어 있는데 S-PLUS의 나무모형의 그래프에 의하면 나무 중심축으로부터 가장 가장자리의 좌, 우측의 종료노드와의 거리 비율은 7:3이다.

4.2 국내 PC통신 해지자 분석

전체 자료 3만 건 중 514건만이 해지자 이다. 그림-3은 가지치기 이전의 deviance에 의해 도출된 나무모형이다. 그림에서 분할에 의해 생성된 가지의 길이는 향상된 순수성의 정도를 나타낸 것으로 deviance의 경우 분할 초기에 순수성이 상대적으로 급격히 증가한 후 분할이 진행됨에 따라 불순도의 감소가 없어 더 이상의 분류 정확도 향상을 기대할 수 없을 것으로 보인다. 이와 같은 경우 가지치기의 과정을 거치게 되면 대부분의 가지들이 제거되어 간결한 나무모형이 제공된다.

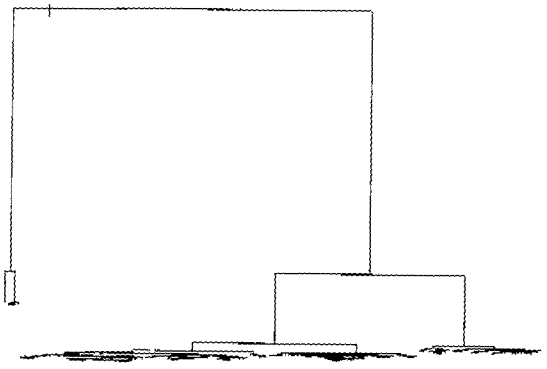

그림-3 Deviance에 의해 생성된 PC 통신 해지자 예측 모형-가지치기 전

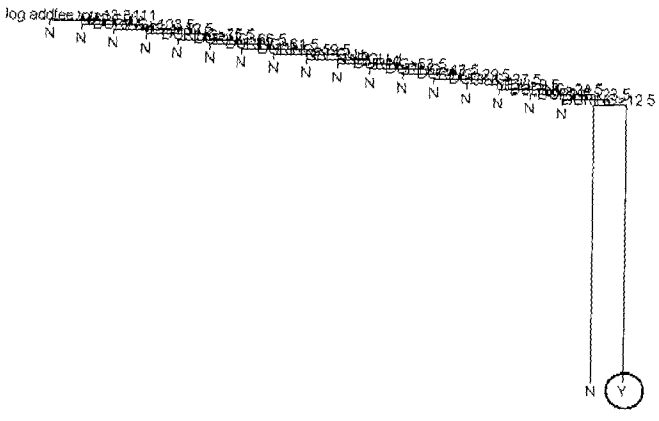

그림-4 단측 순수성에 의해 생성된 나무무형

그림-4는 단측 순수성에 의해 생성된 나무모형이다. 자료의 초기 분할에서는 종료 노드가 "N"이라고 표시된 비해지자를 분리하고 분할단계의 마지막 부분에서 해지자 "Y"로 예측하는 종료노드가 생성된다. 분할 초기의 짧은 가지들이 생성하는 조건들은 PC통신의 총 사용기간과 3개월간 부가사용료의 합의 단 두 연속형 변수에 의해 생성된 것이다. 상당히 많은 조건들이 연결되어 있는 것으로 보이나 해지자의 특성을 분할하는 구간이 한쪽 방향으로 범위가 축소되는 형태로 조건이 형성되므로 최종 해석이 간결하다. 또 다른 특징으로는 deviance와는 달리 순수성의 급격한 상대적 증가가 분할이 진행되어 관심의 대상이 되는 수준의 발굴이 가능한 상황에서 일어난다는 것이다. 그러한 이유로 분할의 마지막 부분에서의 가지가 길게 나타나는 것이다.

Deviance와 단측 순수성에 의해 생성된 나무모형의 정확도를 비교하기 위하여 3만개의 자료 중 1/3를 검증자료로 분리하여 모형적합에는 사용하지 않았다. 그리고 다음의 세가지 기준에 의해 분류 정확도를 비교하였다.

① 적중률(%Response)
(옳게 예측된 해지자 수)/(해지자로 예측된 수)

② 탐색률(%Captured)
(옳게 예측된 해지자 수)/(자료 중 실제 해지자 수)

③ 오분류 비율(MER; Misclassification Error Rate)
1-(옳게 예측된 수)/(검증자료 수)

표-1은 3만개의 자료에 대하여 훈련 및 검증자료를 나누는 표본추출을 5회 반복 실시하여 각 자료에 대하여 정확도를 측정한 결과이다. Deviance에 의한 나무모형의 경우 10-fold cross validation에 의해 최적의 나무 크기를 결정하였다. 적중률(%Response)의 괄호 안의 숫자는 추천된 해지자 수를 나타낸다.

	Deviance에 의한 나무모형				단측 순수성에 의한 나무모형		
	나무 크기	%Response	%Captured	1-MER	%Response	%Captured	1-MER
1	12	0.952 (186)	0.375	0.9695	0.895 (172)	0.326	0.9663
2	9	0.950 (201)	0.387	0.9686	0.865 (185)	0.324	0.9641
3	8	0.946 (202)	0.393	0.9693	0.911 (179)	0.335	0.9661
4	15	0.944 (215)	0.401	0.9684	0.901 (182)	0.324	0.9639
5	9	0.947 (195)	0.405	0.9701	0.884 (189)	0.346	0.9662
평균		0.9478	0.3922	0.9692	0.8912	0.331	0.9653

표-1 Deviance와 단측 순수성에 의한 나무모형의 정확도 비교

표-1에 의하면 단측 순수성을 이용한 해지자 예측이 deviance에 기초한 나무모형에 의한 예측보다 정확성이 떨어진다는 것을 알 수 있다.

4.3 실험자료 분석

단측 순수성에 의한 나무모형의 성질을 이해하기 위하여 다음의 실험자료(simulated data)를 통해 분석하여 보았다.

2차원 구간 $[-1,1]\times[-1,1]$ 사이에서 1,000개의 난수를 발생한 후 다음의 조건에 의해 목표변수의 분류 수준을 정의한다.

① 실험자료-1
만약

$$(x-0.5)^2+(y-0.5)^2<0.25^2 \text{ 이거나}$$
$$-0.9<x<-0.5\ ;\ -0.5<y<-0.1 \text{ 이면}$$

Class 2이고 나머지는 Class 1으로 한다.

② 실험자료-2

만약 $y<2x-1$ 이면

Class 2이고 나머지는 Class 1으로 한다.

③ 실험자료-3

만약 $y>x+2-\sqrt{2}$ 이거나

$y<x-2+\sqrt{2}$ 이면

Class 2이고 나머지는 Class 1으로 한다.

그림-5, 6, 7은 단측 순수성에 의해 도출된 각 실험자료의 나무모형이다.

실험자료-1의 경우 Class 2로 이루어진 2개의 분리된 군집(cluster)이 그림-5에서 확연히 발굴될 수 있다. 실험자료-3의 경우도 2개의 군집이 존재하나 각 군의 경계가 각 축에 대하여 겹쳐있고 나무모형의 분할이 축에 수직으로 진행되는 한계에 의해 단측 순수성에 의한 나무모형이 실험자료-1에 비해 확연하지 못한 군의 존재를 보여주고 있다.

PC통신 해지자료의 경우와 마찬가지로 다수에 해당되는 수준을 분리하는 조건에 의해 분할이 진행된 후 소수의 수준을 발굴하는 분할조건이 도출된다.

각 실험자료에 대한 MER을 비교하면 다음과 같다.

	단측 순수성	Deviance
실험자료-1	1.5%	0.7%
실험자료-2	3.7%	1.9%
실험자료-3	11.7%	5.8%

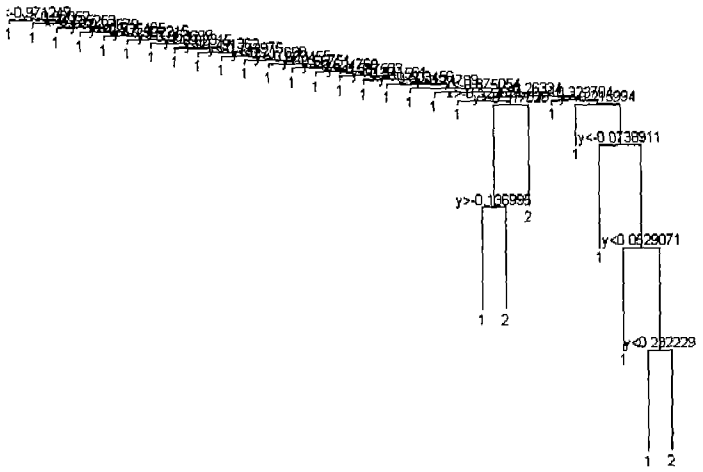

그림-5 단측 순수성에 의한 실험자료-1의 분석결과

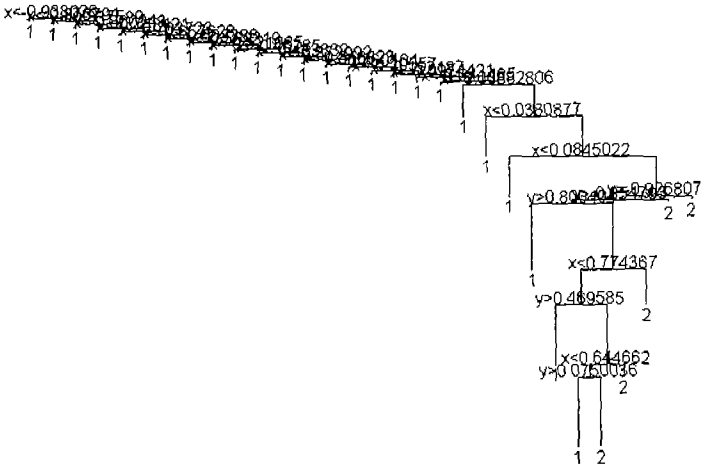

그림-6 단측 순수성에 의한 실험자료-2의 분석결과

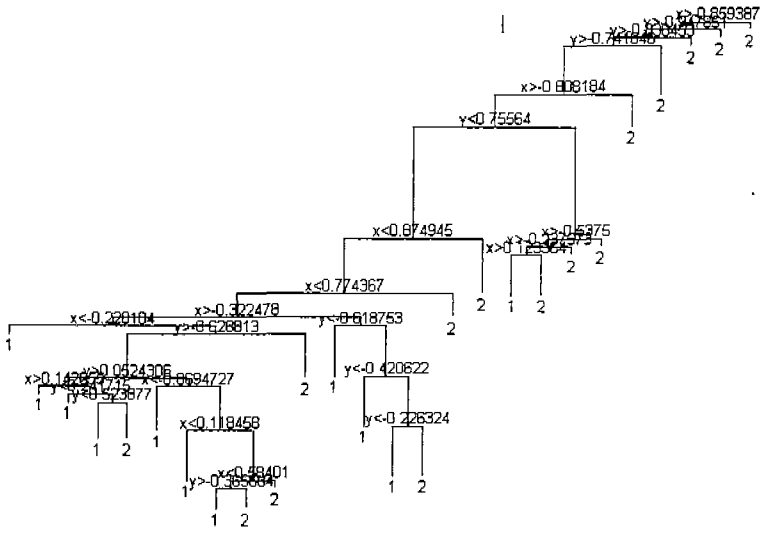

그림-7 단축 순수성에 의한 실험자료-3의 분석결과

5. 결언 및 향후 연구과제

지금까지의 결과에 의하면 단측 순수성이 분류 정확도의 향상에 기여한다고 할 수 없다. 이는 단측 순수성을 제안하였던 Buja와 Lee가 언급한 것과 일치하는 것으로 단측 순수성이나 단축 극단성의 목적은 관심 대상의 수준을 쉽게 설명하는데 있다고 하였다. 그러나 다음과 같은 면에서 단측 순수성이 유용하다고 할 수 있으며 지속적 연구를 통해 분류 정확도 향상에도 기여할 것으로 생각된다.

① 소수 관심의 대상을 설명하는 조건을 쉽게 해석할 수 있다.
: 단측 순수성의 경우 분할단계 초기에 목표변수 수준 중 다수를 분리하는 분할이 연속적으로 일어난다. 이 분할 조건은 마치 다차원의 자료로부터 순수성이 높은 축, 단일 수준으로 이루어져 있는 초다면체(hyper cube)를 분리하는 것과 동일한 것이다. 이와 같은 과정은 전체 자료 분포에서 발굴된 조건에 의해 껍질을 벗기는 것에 비유할 수 있어 Friedman과 Fisher(1997)은 껍질벗기 과정(peeling process)이라 하였다. Friedman과 Fisher의 PRIM (Patient Rule Induction Method)은 단측 순수성과는 달리 사용자가 입력한 비율만큼 각 변수의 극단부분을 제거해 가면 목표변수의 수준을 탐색해 간다. 결국 단축 순수성의 껍질벗기 과정과 같은 탐색적인 과정을 통해 관심 대상의 수준을 설명하는 분할조건(segmentation rule)을 발굴할 수 있는 것이다.

② 자료 내의 목표변수의 각 수준에 대한 군집상황을 파악할 수 있다.
: 제 4절에서의 실험 자료들로부터 단측 순수성에 의한 나무모형이 군집(cluster)의 존재 및 공간상의 분포형태를 보여줌을 알 수 있었다. 이와 같은 결과는 단측 순수성이 deviance나 Gini 처럼 불순도 함수와 같은 정의된 기준의 변화에 의해 분할되는 것과는 달리 목표변수 수준을 탐색적 방법, 즉 껍질벗기 과정을 통해 접근하기 때문이다.

③ 단측 순수성과 deviance 혹은 Gini를 혼합하여 사용하는 혼성모형(hybrid model)을 고려할 수 있다.
: 단측 순수성의 경우 분할단계 초기에 신중한 껍질벗기 과정을 통해 목표변수 구성 중 다수를 차지하는 수준을 분리하여 나간다. 특히 수준의 불균형이 심한 경우 다수의 수준을 분리하는 과정을 통해 수준의 구성비가 균형을 이루어가는 것이다. 그러나 수준의 균형이 이루어진 후 단측 순수성으로는 분류예측의 정확도를 기대할 수 없는 것이다. 제 4절의 실험자료에 대한 분석결과에서도 볼 수 있듯이 수준의 구성비 차이가 크지 않을수록, 분류문제에 있어 단축 순수성이 deviance에 비하여 정확하지 않음을 알 수 있다. 그러므로 자료 분할초기에는 단측 순수성을 이용하고 수준의 구성비가 균형을 이루면 기존의 Gini나 deviance 처럼 생성되는 두 노드의 순수성을 동시에 고려하는 분할기준을 사용하는 것이 좋을 것이다.

그 외의 연구과제로는 단측 순수성에 적용할

수 있는 가지치기 방법 혹은 나무모형의 성장을 중지하는 최적의 나무크기를 판단하는 알고리즘등의 개발이 중요한 연구 과제라고 할 수 있다.

참고문헌

[1] Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1983) Classification and regression trees, Wadsworth International, Belment, CA.

[2] Buja, A., and Lee,Y.-S. (1999), Data mining criteria for tree-based regression and classification, http://www.research.att.com/~andreas/papers/trees.ps.gz

[3] Friedman, J. H., and Fisher, N. I. (1997), Bump hunting in high-dimensional data. Tech. report, Dept. of Statsitcs, Stanford University.