Journal of Korean
Data & Information Science Society
2005, Vol. 16, No. 4, pp. 769~780

Comparison of Binary Discretization Algorithms for Data Mining[1)]

Jong-Hwa Na[2)] · Jeong-Mi Kim[3)] · Wan-Sup Cho[4)]

Abstract

Recently, the discretization algorithms for continuous data have been actively studied. But there are few articles to compare the efficiency of these algorithms. In this paper we introduce the principles of some binary discretization algorithms including C4.5, CART and QUEST and investigate the efficiency of these algorithms through numerical study. For various underlying distribution, we compare these algorithms in view of misclassification rate and MSE. Real data examples are also included.

Keywords : Binary Splitting, Data Mining, Discretization Algorithm, Misclassification Rate, MSE, Threshold

1. 서론

최근까지 기계학습(machine learning)과 데이터마이닝(data mining)을 비롯한 영역에서 연속형 자료에 대한 범주화(discretization or categorization) 알고리즘에 대한 연구가 활발히 진행되어 왔다. 또한 다양한 통계분석 프로그램들은 각기 다른 범주화 알고리즘을 채택하여 사용하고 있으나, 이들 알고리즘간의 효율성에 대한 통계적 관점에서의 비교가 미흡하다. 본 논문에서는 중요한 이진분리 알고리즘들을 소개하고, 실제자료에의 적용과 함께 방대한 모의실험을 통한 알고리즘 간의 효율성에 대한 비교를 수행하였다. 또한 연속되는 후편의 논문 Kim, et al. (2005)에서는 2개 이상의 분리기준값을 허용하는 다원분리 알고리즘의 효율성 및 이진분리 알고리즘과의 비교를

---

1) This work was supported by the research grant of the Chungbuk National University in 2005.

2) First Author : Professor, Dept. of Information and Statistics & Institute for Basic Science Research, Chungbuk National University, Cheongju, Chungbuk, 361-763, Korea. E-mail: cherin@chungbuk.ac.kr

3) Doctoral Student, Dept. of Information and Statistics, Chungbuk National University, Cheongju, Chungbuk, 361-763, Korea.

4) Associate Professor, Dept. of MIS, Chungbuk National University, Cheongju, Chungbuk, 361-763, Korea.

동시에 수행하였다.

먼저 연속형 자료에 대한 범주화는 크게 이진분리(binary splitting)와 다원분리(multiway splitting) 알고리즘으로 나누어진다. 이진분리 알고리즘은 범주화 과정의 매 단계에서 이진의 형태로만 구분해 나가는 방법으로서 이와 관련된 연구로는 Breiman et al. (1984), Dougherty et al. (1995), Quinlan (1993, 1996), Kohavi et al. (1997), Loh and Shih (1997), Gestwicki (1997) 등이 있다. 다원분리 알고리즘과 관련된 연구로는 Berka (1993a, 1993b), Berka and Bruha (1995, 1998), Holte (1993)의 연구가 대표적이다. 이들 알고리즘 간의 효율성에 대한 비교 연구로는 Kralik and Bruha (1997), Kohavi and Sahami (1996), Dougherty et al. (1995), Wang and Goh (1997) 등이 있다. 그러나 이들 연구는 모두 목적(target) 변수의 값을 알고 있는 실제자료를 중심으로 연구가 진행되거나, 알고리즘의 수행속도 등의 관점에서의 비교 연구가 주를 이루고 있다. 본 논문은 이들 알고리즘에 대한 통계적 관점에서의 효율성 비교를 그 목적으로 한다. 2절에서는 대표적인 이진분리 알고리즘을 소개하고 실제자료에의 적용과정을 다루었다. 3절에서는 다양한 모의실험을 통한 알고리즘의 효율성 비교를 수행하였으며, 4절은 결론으로 구성되어 있다.

2. 이진분리 알고리즘과 실제자료에의 적용

2.1 연속형자료에 대한 이진분리 알고리즘

이 절에서는 연속형자료에 대한 대표적인 이진분리 범주화 알고리즘들을 간략히 소개한다. 본 논문의 전반에 걸쳐 사용되는 자료의 형태는 n개의 1차원 연속형자료로써 $x_1, x_2, \cdots, x_n$으로 표기한다. 해당 알고리즘에 대한 보다 자세한 절차는 관련논문을 참고하기 바란다.

[Algorithm I] C4.5 알고리즘

C4.5 알고리즘은 1993년 Quinlan에 의해 제안된 알고리즘이다. 먼저 C4.5 알고리즘의 전개에 필요한 기호는 다음과 같다.

· D : 데이터의 집합 · C_j : 목표변수의 j번째 범주 · $|D|$: D에 속한 총 개체의 수

· $P(D, j)$: D에서 목표변수의 j번째 범주에 속하는 개체의 비율

C4.5알고리즘을 통해 연속형 데이터를 범주화하는 과정은 다음과 같다.

【단계1】 데이터 D에서 목표변수의 j번째 범주 C_j에 속하는 개체를 구별하기 위한 평균 정보량을 나타내는 엔트로피계수(Entropy Index), $Info(D)$를 계산한다.

$$Info(D) = -\sum_{j=1}^{k} P(D,j) \times \log_2(P(D,j)).$$

단, $P(D,j)$는 D에서 목표변수의 j번째 범주에 속하는 개체의 비율을 나타낸다.

【단계2】 데이터 D에서 종속변수 X를 크기순으로 정렬하여 d개의 개별값 $\{x_1, \cdots, x_d\}$을 취하고, 이웃하는 두 값의 중간값인 $(x_i + x_{i+1})/2$을 분리기준값(threshold)으로 정한다. 이 때 가능한 분리 기준값은 $d-1$개가 된다.

【단계3】 분리기준값에 의해 데이터 D가 2개의 부분집합 D_1, D_2로 분할됨으로서 얻어지는 정보량은 다음과 같다. 여기서 $|D|$는 D에 속한 총 개체의 수를 나타낸다.

$$Info_X(D) = \sum_{i=1}^{2} \frac{|D_i|}{|D|} \times Info(D_i).$$

【단계4】 정보량의 감소를 나타내는 정보량의 이득(Gain)을 계산한다.

$$Gain(D, X) = Info(D) - Info_X(D).$$

【단계5】 데이터 D가 n개의 부분집합으로 분할될 때 추가적으로 발생하는 정보량인 분리정보(Split Information)를 계산한다.

$$Split(D,\ X) = -\sum_{i=1}^{2} \frac{|D_i|}{|D|} \times \log_2\left(\frac{|D_i|}{|D|}\right).$$

【단계6】 $Gain(D, X)$를 $Split(D, X)$로 나눈 값인 $Gain\ ratio$를 계산하고, 이 값이 가장 큰 분리기준값을 선택한다.

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

[Algorithm Ⅱ] CART 알고리즘

CART(Classification and Regression Trees)는 L. Breiman에 의해 1984년 개발된 의사결정나무(Decision Tree) 분석에서 많이 사용되는 알고리즘으로 변수의 선택과 분리가 동시에 일어나는 특징을 가지고 있다. 먼저 CART알고리즘의 전개에 필요한 기호는 다음과 같다.

- N : 표본의 개체수
- $N(t)$: 노드 t에서의 개체수
- N_j : 범주 j에 속한 개체수
- $N_j(t)$: 노드 t에서 범주 j에 속한 개체수
- $p(j, t) = N_j(t)/N$: 임의의 개체가 범주 j와 노드 t에 속할 확률
- $p(t) = \sum_{j=1}^{J} p(j, t) = N(t)/N$: 임의의 개체가 노드 t에 속할 확률
- $p(j|t)$: 임의의 개체가 노드 t에 속할 때 범주 j에 속할 조건부 확률

$$p(j|t) = p(j, t)/p(t) = N_j(t)/N(t),\quad \sum_{j=1}^{J} p(j|t) = 1.$$

CART 알고리즘을 통해 연속형 데이터를 범주화하는 과정은 다음과 같다.

【단계1】 부모마디 t의 지니지수(Gini index)를 계산한다.

$$i(t)=\sum_{i\neq j} p(i|t)p(j|t)=1-\sum_j p^2(j|t).$$

【단계2】 데이터 D에서 종속변수 X를 크기순으로 정렬하여 d개의 개별값 $\{x_1, \cdots, x_d\}$을 취하고, 이웃하는 두 값의 중간값인 $(x_i+x_{i+1})/2$을 분리기준값(threshold)으로 정한다. 이 때 가능한 분리 기준값은 $d-1$개가 된다.

【단계3】 각 분리기준값에서 불순도의 변화량을 계산한다.

$$\Delta i(s, t)=i(t)-\{p_L i(t_L)+p_R i(t_R)\}.$$

【단계4】 불순도 변화량이 최대가 될 때의 값 s를 분리기준값으로 선택한다.

$$\Delta i(s^*, t)=\max_s \Delta i(s, t).$$

[Algorithm Ⅲ] QUEST 알고리즘

QUEST(Quick, Unbiased, Efficient, Statistical Tree) 알고리즘은 1997년 Wei-Yin Loh와 Yu-Shan Shih에 의해 고안된 것으로 데이터를 범주화하는 과정은 다음과 같다.

【단계1】 두 집단 표본평균이 동일하면, 둘 중 개체수가 더 많은 집단을 주 집단 A로 정의하고, 다른 하나의 집단은 B로 정의한다.

【단계2】 $\bar{x}_A$와 s_A^2을 주 집단 A의 표본평균과 분산이라 하고, 마찬가지로 $\bar{x}_B$와 s_B^2을 B의 표본평균과 분산이라 하고, $p(A|t)=\sum_{j\in A} p(j|t)$와 $p(B|t)=1-p(A|t)$는 각 집단의 사전확률이라 하자.

【단계3】 【단계2】에 서 계산된 값들을 통해 다음과 같은 식을 세운다.

$$p(A|t)s_A^{-1}\phi\{(x-\bar{x}_A)/s_A\}=p(B|t)s_B^{-1}\phi\{(x-\bar{x}_B)/s_B\},$$

여기서 $\phi(\cdot)$은 표준정규분포의 밀도함수이다. 위의 식에서 양변에 log를 취한 후 x에 관한 이차식으로 표현하면 다음과 같다.

$$ax^2+bx+c=0. \tag{2.1}$$

단, $a=s_A^2-s_B^2$, $b=2(\bar{x}_A s_B^2-\bar{x}_B s_A^2)$,

$c=(\bar{x}_B s_A)^2-(\bar{x}_A s_B)^2+2\,s_A^2 s_B^2 \log[\{p(A|t)s_B\}/\{p(B|t)s_A\}]$

이다. 만약 $a=0$이고, $\bar{x}_A\neq\bar{x}_B$이면,

$$x=(\bar{x}_A+\bar{x}_B)/2-(\bar{x}_A-\bar{x}_B)^{-1}s_A^2 \log\{p(A|t)/p(B|t)\}$$

가 되고, $a=0$이고, $\bar{x}_A=\bar{x}_B$이면 식(2.1)의 이차식은 근을 갖지 않는다.

【단계4】 분리기준을 $X=d$ 라 하면, d는 다음과 같이 정의된다. (Ⅰ) 만약 $a=0$이면,

$$d=\begin{cases}(\bar{x}_A+\bar{x}_B)/2-(\bar{x}_A-\bar{x}_B)^{-1}s_A^2\log\{p(A|t)/p(B|t)\}\ , & \bar{x}_A\neq\bar{x}_B,\\ \bar{x}_A\ , & \bar{x}_A=\bar{x}_B.\end{cases}$$

(Ⅱ) 만약 $a\neq0$이면, (a) $b^2-4ac<0$이면, $d=(\bar{x}_A+\bar{x}_B)/2$가 된다. (b) $b^2-4ac\geq0$이면, (b1) 두 개의 근 $(2a)^{-1}\{-b\pm\sqrt{b^2-4ac}\}$중 $\bar{x}_A$에 더 가까운 값을 d로 정의한다. 단, 분리기준에 의해 데이터를 둘로 나누면 양쪽 모두 최소한 한 개 이상의 값을 갖고 있어야 한다. (b2) 그 밖의 경우 $d=(\bar{x}_A+\bar{x}_B)/2$으로 정의한다.

2.2 실제자료에의 적용

다음의 표2.1은 여러 기상 조건(조망, 기온, 습도, 바람)에 따라 골프를 칠 것인가의 여부를 조사한 Quinlan (1994)의 자료에서 분석에 필요한 기온 자료만을 발췌한 것이다.

<표 2.1> 골프데이터[Temp: 기온(단위: 화씨), Class: Play여부]

Temp	85	80	83	70	68	65	64	72	69	75	72	72	81	71
Class	D	D	P	P	P	D	P	D	P	P	P	P	P	D

(P: Play, D: Don't Play를 의미.)

이제 위의 자료에서 연속형 변수인 Temp에 대해 앞 절에서 소개한 범주화 알고리즘을 적용하기로 한다. 먼저 C4.5알고리즘을 통한 범주화 과정은 다음과 같다. [단계1]의 평균정보량을 계산하면

$$Info(D)=-\{(9/14)\times\log_2((9/14))+(5/14)\times\log_2((5/14))\}=0.9402$$

이다. [단계2]와 [단계3]의 과정으로, 예를 들어, 분리기준값(데이터 사이의 중간값) 70.5에서 정보량 $Info_X$을 계산해보면 다음과 같다.

$$Info_X(D)=-(1/5)\times\log_2(1/5)+(4/5)\times\log_2(4/5)$$
$$+(5/9)\log_2(5/9)\times(4/9)\log_2(4/9)=0.895$$

다음으로 분리기준값 70.5에서 [단계4]~[단계6]의 정보량의 이득(Gain), 분리정보, 이득비율을 계산하면 다음과 같다.

$$Gain(D,X)=0.045,\ \ Split(D,X)=0.940,\ \ Gain\ ratio(D,X)=0.048$$

같은 방법으로 모든 분리기준값에서의 계산결과는 다음의 표2.2와 같다.

<표 2.2> Temp의 각 분리기준값에서의 이득비율

분리기준값	64.5	66.5	68.5	69.5	70.5	71.5	73.5	77.5	80.5	82.0	84.0
이득비율	0.129	0.017	0.001	0.017	0.048	0.001	0.001	0.029	0.001	0.017	0.305

표2.2에서 Temp의 분리기준값이 84일 때 이득비율(Gain Ratio)이 최대가 되지만, 한쪽 마디의 최소의 개체수는 2이상이 되도록 분리하는 것을 원칙으로 하기 때문에 분리기준값 64.5, 84는 대상에서 제외된다. 이 두 값을 제외하면 분리기준값이 70.5일 때 이득비율(Gain Ratio)이 0.048로 최대가 되어 Temp의 분리기준값은 70.5가 된다.

다음으로, C4.5 알고리즘의 과정을 소개하면 다음과 같다. 먼저, [단계1]의 부모마디 t의 지니지수(Gini Index)를 계산하면 $i(t)=1-(5/14)^2-(9/14)^2=0.459$ 이 된다. [단계2]에서, 예를 들어, 분리기준값이 70.5 일 때의 불순도의 변화량은 다음과 같다.

$$\Delta i(70.5,\ t)=0.459-(5/14)\{1-(1/5)^2-(4/5)^2\} \\ -(9/14)\{1-(4/9)^2-(5/9)^2\}=0.0274$$

다음의 표2.3은 각 분류기준값에서 불순도의 변화량을 나타낸다.

<표 2.3> 각 분리기준값에서의 불순도 변화량

분리기준값	64.5	66.5	68.5	69.5	70.5	71.5	73.5	77.5	80.5	82.0	84.0
불순도변화량	0.020	0.007	0.000	0.009	0.027	0.001	0.001	0.016	0.000	0.007	0.064

위의 결과를 보면 분류기준값이 84.0일 때 불순도의 변화량이 가장 큰 값을 갖지만, C4.5에서와 마찬가지로 한쪽 마디의 최소 개체수가 2이상이 되어야 하므로 84.0은 분류기준값 후보가 되지 못한다. 84.0을 제외하면 70.5일 때 불순도의 변화량이 0.0274로 가장 크다. 따라서 70.5가 CART 알고리즘에서의 분류기준값으로 선택된다. 이는 C4.5 알고리즘과 동일한 결과를 얻게 된다.

마지막으로 QUEST 알고리즘을 적용하면 다음과 같다. Play 집단을 P, Don't Play 집단을 D라고 할 때, 각 단계에서의 계산결과는 다음과 같다.

$$p(P|t)=9/14,\quad p(D|t)=5/14,$$
$$\bar{x}_P=73,\quad \bar{x}_D=74.6,\quad s_P=6.164,\quad s_D=7.893.$$
$$a=s_P^2-s_D^2=-24.3,\quad b=2(\bar{x}_P s_D^2-\bar{x}_D s_P^2)=3426.9.$$
$$c=(\bar{x}_D s_P)^2-(\bar{x}_P s_D)^2+2\,s_P^2 s_D^2\log[\{p(P|t)s_D\}/\{p(D|t)s_P\}]=-116592.9.$$

위 식에서 $a\neq 0$이고, $b^2-4ac\geq 0$이므로 이차식의 근인 d는 다음과 같이 계산된다.

$$d=(2a)^{-1}\{-b\pm\sqrt{b^2-4ac}\}=57.35 \text{ 또는 } 83.64.$$

위의 결과에서 두 개의 근 중 $\bar{x}_P=73$에 더 가까운 83.64가 최종 분리기준값으로 선택된다.

3. 범주화 알고리즘들의 효율성 비교

앞 절에서 소개한 세 가지의 이진분리 범주화 알고리즘들의 효율성을 비교하기 위해 본 논문에서는 다양한 모집단의 분포 가정 하에서 모의실험을 실시하였다. 모의실험에 사용된 모집단의 분포로는 위치모형, 척도모형, 위치-척도모형 그리고 혼합된 형태의 모형을 고려하였다. 모의실험을 통한 효율성 비교에는 오분류율(misclassification rate)과 MSE(mean square error)의 두 가지 관점에서 각각 비교를 실시하였다. 각 모집단 모형으로부터 표본의 크기는 30으로 하고, 반복수는 10000으로 하였다. 본 논문에서의 모의실험은 Splus6 언어를 사용하였다.

3.1 오분류율을 통한 효율성 비교

이 절에서는 오분류율(misclassification rate)을 통한 각 알고리즘의 효율성을 비교한다. 먼저 오분류율의 계산과정은 다음과 같다. 예를 들어 표3.1의 (a)의 경우, 두 모집단과 $N(1,1)$으로부터 각각 30개의 난수를 생성하고, 이들의 목표변수를 각각 0과 1로 설정한다. 생성된 총 60개의 난수에 대해 각 범주화 알고리즘에 적용하여 분리기준값을 찾는다. 분리기준값으로부터 나누어지는 두 그룹의 자료에 대해 보다 비율이 높은 목표변수의 값을 할당한다. 각 알고리즘으로부터 구해진 분리기준값들에 의해 할당된 목표변수의 값(0 또는 1)과 실제 알고 있는 목표변수의 값을 비교하여 잘못 분류된 비율을 구하고, 이러한 과정을 10000번 반복하여 구해진 비율의 평균값을 오분류율로 정의한다.

표3.1에서 표3.4는 각각 모집단의 분포가 위치모형, 척도모형, 위치-척도모형, 혼합모형인 경우에 대한 이진분리 알고리즘의 오분류율을 나타내고 있다. 결과에서 (*)표시는 가장 효율이 뛰어남을 의미한다. 이 결과들을 살펴보면 위의 네 가지 모형 모두에서 CART 알고리즘의 효율이 오분류율의 관점에서 가장 뛰어남을 알 수 있다. C4.5와 QUEST 알고리즘의 효율은 큰 차이가 없는 것으로 나타났다.

<표 3.1> 위치모형에서의 오분류율

위치모형	두 모집단의 분포	오분류율		
		C4.5	CART	QUEST
정규분포	(a) $N(0,1)$ *vs* $N(1,1)$	0.339	0.275	0.306
	(b) $N(0,1)$ *vs* $N(3,1)$	0.051	0.049	0.064
분포	(c) $t(2)$ *vs* $t(2)+1$	0.378	0.302	0.386
	(d) $t(2)$ *vs* $t(2)+3$	0.121	0.117	0.158
균일분포	(e) $U(0,1)$ *vs* $U(0,1)+0.2$	0.392	0.355	0.391
	(f) $U(0,1)$ *vs* $U(0,1)+0.6$	0.182	0.167	0.196
지수분포	(g) $\mathrm{Exp}(1)$ *vs* $\mathrm{Exp}(1)+1$	0.189	0.185	0.285
	(h) $\mathrm{Exp}(1)$ *vs* $\mathrm{Exp}(1)+3$	0.025	0.025	0.042

<표 3.2> 척도모형에서의 오분류율

척도모형	두 모집단의 분포	오분류율		
		C4.5	CART	QUEST
정규분포	(a) $N(0,1)$ *vs* $N(0,2^2)$	0.400	0.368	0.391
	(b) $N(0,1)$ *vs* $N(0,4^2)$	0.318	0.307	0.324
로지스틱 분포	(c) $L(0,1)$ *vs* $L(0,2)$	0.404	0.369	0.394
	(d) $L(0,1)$ *vs* $L(0,4)$	0.328	0.313	0.335
지수분포	(e) $\mathrm{Exp}(1)$ *vs* $\mathrm{Exp}(1/2)$	0.408	0.345	0.361
	(f) $\mathrm{Exp}(1)$ *vs* $\mathrm{Exp}(1/4)$	0.285	0.243	0.336
이중 지수분포	(g) $DE(0,1)$ *vs* $DE(0,2)$	0.417	0.376	0.401
	(h) $DE(0,1)$ *vs* $DE(0,4)$	0.350	0.328	0.357

<표 3.3> 위치-척도모형에서의 오분류율

위치-척도 모형	두 모집단의 분포	오분류율		
		C4.5	CART	QUEST
정규분포	(a) $N(0,1)$ *vs* $N(1,2^2)$	0.342	0.304	0.327
	(b) $N(0,1)$ *vs* $N(3,4^2)$	0.204	0.193	0.249
로지스틱 분포	(c) $L(0,1)$ *vs* $L(1,2)$	0.385	0.341	0.366
	(d) $L(0,1)$ *vs* $L(3,4)$	0.270	0.251	0.275
지수분포	(e) $\mathrm{Exp}(1)$ *vs* $\mathrm{Exp}(1/2)+1$	0.193	0.192	0.217
	(f) $\mathrm{Exp}(1)$ *vs* $\mathrm{Exp}(1/4)+3$	0.026	0.026	0.035
이중 지수분포	(g) $DE(0,1)$ *vs* $DE(1,2)$	0.375	0.309	0.358
	(h) $DE(0,1)$ *vs* $DE(3,4)$	0.230	0.204	0.250

<표 3.4> 혼합모형에서의 오분류율

혼합모형	두 모집단의 분포	오분류율		
		C4.5	CART	QUEST
정규분포 관련	(a) $N(1,1)$ *vs* $t(2)$	0.358	0.288	0.338
	(b) $N(1,1)$ *vs* $\mathrm{Exp}(1)$	0.418	0.385	0.435
	(c) $N(1,1)$ *vs* $L(0,1)$	0.349	0.301	0.326
	(d) $N(1,1)$ *vs* $U(0,1)$	0.253	0.249	0.260
오염 정규분포 관련	(e) $0.05N(0,1)+0.95N(2,1)$ *vs* $N(1,1)$	0.398	0.319	0.348
	(f) $0.05N(0,1)+0.95N(3,1)$ *vs* $N(1,1)$	0.217	0.178	0.200
	(g) $0.05N(0,1)+0.95N(2,1)$ *vs* $t(2)+1$	0.430	0.337	0.389
	(h) $0.05N(0,1)+0.95N(2,1)$ *vs* $L(0,1)$	0.284	0.227	0.240

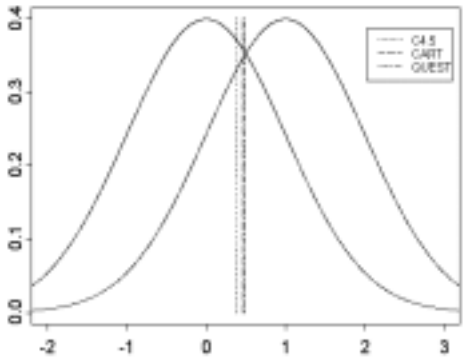
3.2 평균 분리기준값을 통한 비교

이 절에서는 각 알고리즘으로부터 구해지는 분리기준값들의 평균값을 통한 비교를 실시한다. 표3.5는 모의실험을 통해 구해진 분리기준값들의 평균을 나타내며, *표시는 참값에 가장 가까운 값을 의미한다. 표3.5에 제시된 결과들을 살펴보면 평균 분리기준값의 측면에서 CART 와 QUEST 알고리즘이 거의 대등한 정도(precision)를 보인다고 말 할 수 있다. 표3.5에 제시된 몇 가지 모형에 대한 평균 분리기준값을 그림3.1에 나타내었다. 이 그림에서 알 수 있듯이 모집단의 형태에 따라 각 알고리즘의 분리기준값에 큰 차이가 존재함을 알 수 있다.

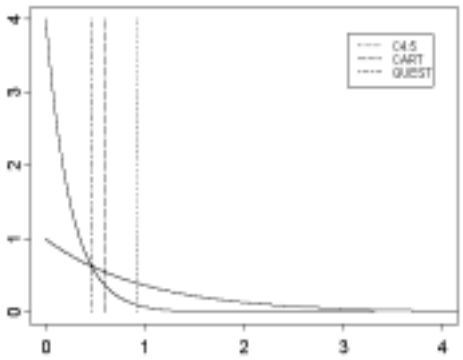
<표 3.5> 다양한 모집단에서의 평균 분리기준값

두 모집단의 분포	참값	평균 분리기준값		
		C4.5	CART	QUEST
(a) $N(0,1)$ *vs* $N(1,1)$	0.5	0.382	0.465	0.491*
(b) $N(0,1)$ *vs* $N(3,1)$	1.5	1.491	1.492*	1.510
(c) $t(2)$ *vs* $t(2)+1$	0.5	0.220	0.430	0.444*
(d) $t(2)$ *vs* $t(2)+3$	1.5	1.472	1.472	1.486*
(e) $\mathrm{Exp}(1)$ *vs* $\mathrm{Exp}(1/2)$	0.693	1.784	1.019*	1.028
(f) $\mathrm{Exp}(1)$ *vs* $\mathrm{Exp}(1/4)$	0.462	0.907	0.594	0.472*
(g) $\mathrm{Exp}(1)$ *vs* $\mathrm{Exp}(1/2)+1$	1	0.936	0.943	0.961*
(h) $\mathrm{Exp}(1)$ *vs* $\mathrm{Exp}(1/4)+3$	3	2.766*	2.766*	2.648
(i) $N(1,1)$ *vs* $t(2)$	0.351	-0.618	0.146*	-0.096
(j) $N(1,1)$ *vs* $U(0,1)$	1	1.041	1.022*	0.948

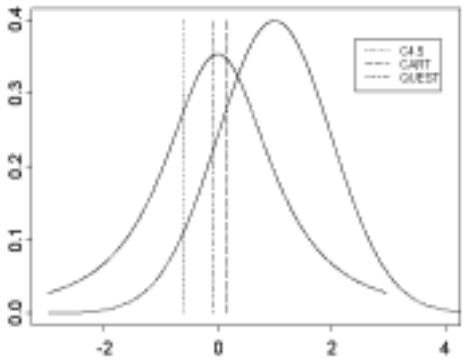
(a) N(0,1) vs N(1,1)



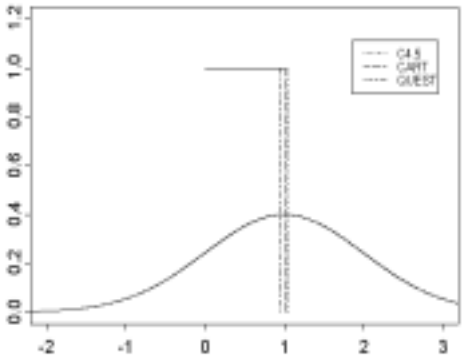
(f) E(1) vs E(1/4)



(i) N(1,1) vs t(2)



(j) N(1,1) vs U(0, 1)



<그림 3.1> 다양한 모집단에서의 평균 분리기준값

3.3 MSE를 통한 효율성 비교

이 절에서는 각 범주화 알고리즘들의 분리기준값의 정도를 알아보기 위한 또 다른 방법으로 아래의 식으로 정의되는 평균제곱오차(MSE)를 구하여 표3.6에 제시하였다. 여기서, MSE는 두 모집단을 가장 잘 분리하는 경계값(참값)과 각 알고리즘에 의해 계산된 분리기준값과의 차이의 제곱을 평균낸 것으로 다음과 같이 정의한다.

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (\theta - d_i)^2, \quad i = 1, 2, \cdots, n.$$

여기서, θ는 참값이고, d_i는 각 범주화 알고리즘으로부터 계산된 분리기준값이고, n은 모의실험에서의 반복횟수를 의미한다. 예를 들어 표3.5 (a)의 경우, 두 모집단 $N(0,1)$과 $N(1,1)$의 밀도함수의 교차점인 0.5가 오분류율을 최소로 하는 참값이 되며, 모의실험을 통해 계산된 분리기준값과의 차의 제곱들에 대한 평균값을 MSE로 정

의하였다.

표3.6에서 위치모형 정규모집단과 관련된 (a)와 (b)의 경우에는 QUEST알고리즘의 분리 효율이 뛰어남을 알 수 있고, 나머지 대부분의 경우에서는 CART 알고리즘의 분리효율이 다른 이진분리 알고리즘들 보다 뛰어난 것으로 나타났다.

<표 3.6> 범주화 알고리즘들의 MSE 비교

두 모집단의 분포	평균제곱오차(MSE)		
	C4.5	CART	QUEST
(a) $N(0,1)$ *vs* $N(1,1)$	1.504	0.344	0.041*
(b) $N(0,1)$ *vs* $N(3,1)$	0.149	0.126	0.030*
(c) $t(2)$ *vs* $t(2)+1$	6.076	0.517*	2.048
(d) $t(2)$ *vs* $t(2)+3$	0.306	0.199*	0.553
(e) $\mathrm{Exp}(1)$ *vs* $\mathrm{Exp}(1/2)$	1.686	0.430	0.149*
(f) $\mathrm{Exp}(1)$ *vs* $\mathrm{Exp}(1/4)$	0.305	0.072*	0.149
(g) $\mathrm{Exp}(1)$ *vs* $\mathrm{Exp}(1/2)+1$	0.006	0.005*	1.104
(h) $\mathrm{Exp}(1)$ *vs* $\mathrm{Exp}(1/4)+3$	0.078*	0.078*	0.134
(i) $N(1,1)$ *vs* $t(2)$	2.073*	0.454	0.492
(j) $N(1,1)$ *vs* $U(0,1)$	0.011	0.010*	0.014

4. 결론

본 논문에서는 최근 개발된 연속형 자료에 대한 이진분리 알고리즘들을 소개하고, 모의실험을 통해 그 효율성을 비교하였다. 모의실험의 결과를 종합해 보면, 오분류율과 MSE의 기준에서는 전반적으로 CART 알고리즘이 가장 우수한 것으로 나타났다. 평균 분리기준값의 측면에서는 CART와 QUEST가 대등한 효율을 보이고 있다. 다만 모집단이 정규분포와 관련된 경우에는 QUEST 알고리즘의 효율이 비교적 우수한 것으로 나타났다. 그러나 데이터마이닝의 영역에서처럼 실제 대량의 자료에 대해 적용될 경우 이들 알고리즘들의 선택문제에서는 본 논문에서 다루어진 통계적 관점의 효율성과 함께 데이터의 양에 따른 처리시간, 그리고 연속형 변수가 범주형에 비해 선택될 가능성이 상대적으로 높아지는 편의의 문제 등을 종합적으로 고려한 알고리즘의 선택이 바람직하다.

참고문헌

1. Berka, P. (1993a). Knowledge EXplorer : A tool for automated knowledge acquisition from data, Technical Report TR-93-03, Austrian Research Institute for AI, Vienna.
2. Berka, P. (1993b). Discretization of numerical attributes for Knowledge EXplorer, Technical Report LISP-93-03.

3. Berka, P. and Bruha, I. (1998). Discretization and grouping: preprocessing steps for data mining, Principles of Data Mining and Knowledge Discovery, 239-245.
4. Berka, P. and Bruha, I. (1995). Empirical comparisons of various discretization procedures, Technical Report LISP-95-04, Laboratory of Intelligent Systems.
5. Breiman. L, Freidman. J, Olshen. R, Stone. C. (1984). Classification and regression trees, Chapman and Hall, New York.
6. Dougherty, J., Kohavi, R. and Sahami, M. (1995). Supervised and unsupervised discretization of continuous features, Proceedings of the Twelfth International Conference, 194-202.
7. Gestwicki, P. (1997). ID3: History, implementation and applications, Manuscript.
8. Holte, R. C. (1993). Very simple classification rules perform well on most commonly used datatsets, Machine Learning, 11, 63-90.
9. Kim, J. S., Kim, J. M. and Na, J. H. (2005). Comparison of multiway discretization algorithms for data mining, submitted.
10. Kohavi, R. and Sahami, M. (1996). Error-based and entropy-based discretization of continuous features, Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 114-119.
11. Kohavi, R., Sommerfield, D. and Dougherty, J. (1997). Data mining using MLC++: A machine learning library in C++, Approved in International Journal on Artificial Intelligence Tools, 6, 537-566.
12. Kralik, P. and Bruha, I. (1997). Discretizing numerical attributes in a genetic attribute-based learning algorithm, Manuscript.
13. Loh, W. Y. and Shih, Y. S. (1997). Split selection methods for classification trees, 7, 815-840.
14. Quinlan. J. R. (1993). C4.5: Programs for machine learning, Morgan Kaufmann, San Mateo, CA.
15. Quinlan. J. R. (1996). Improved use of continuous attributes in C4.5, Journal of Artificial Intelligence Research, 4, 77-90.
16. Wang, K. and Goh, H. C. (1997). Minimum splits based discretization for continuous features, Manuscript.

[2005년 8월 접수, 2005년 9월 채택]