論文 97-22-3-3

다중 엔트로피를 기반으로 하는 새로운 결정 트리 생성기 MEC

正會員 전 병 환*, 김 재 희**

MEC; A New Decision Tree Generator Based on Multi-base Entropy

Byung Hwan Jun*, Jaihie Kim** *Regular Members*

※본 논문은 정보통신부에서 시행한 대학기초연구지원사업비에 의하여 연구되었음

요 약

이 논문에서는 속성의 이산화 및 선택을 위해 다중 엔트로피(multi-base entropy)의 차를 일관된 평가 기준으로 사용하고 인접한 두 구간을 융합해가면서 최선의 이산화를 결정하는 새로운 결정 트리 생성기 MEC를 제안한다. 제안한 생성기의 성능을 알아보기 위해, 엔트로피를 기반으로 하는 평가 기준을 사용하고 이산화 방식에서 차이를 보이는 기존의 생성기들과 비교하였다. 실험 결과, 제안한 생성기는 학습 집합을 구성하는 속성값이 이산적인지 연속적인지에 관계없이 동일한 에러율에서 최소의 단말 노드수를 갖는 가장 효율적인 분류기를 생성하는 것으로 나타났다.

ABSTRACT

A new decision tree generator MEC is proposed in this paper, which uses the difference of multi-base entropy as a consistent criterion for discretization and selection of attributes. To evaluate the performance of the proposed generator, it is compared to other generators which use criteria based on entropy and adopt different discretization styles. As an experimental result, it is shown that the proposed generator produces the most efficient classifiers, which have the least number of leaves at the same error rate, regardless of whether attribute values constituting the training set are discrete or continuous.

*국립공주대학교 전자계산학과
**연세대학교 전자공학과
論文番號:96252-0819
接受日字:1996年 8月 19日

Ⅰ. 서 론

결정 트리의 생성은 데이터로부터 지식을 취득하는 효과적인 학습 방법인 동시에, 복잡하고 전역적인 분류 규칙을 간단하고 지역적인 판단의 모임으로 대신하는 효율적인 인식 방법을 제공한다. 초기의 연구에서는 주로 이산적이거나 이산화된 속성들 중에서 하나를 선택하는 문제에 주목해서 결정 트리 생성기를 개발해왔다. 그런데 속성의 분류 능력은 어떻게 이산화하느냐에 따라 크게 달라지기 때문에, 최근에는 연속적인 속성을 이산화하는 방법에 대해서도 많은 연구가 진행되고 있다. 그러나 기존의 연구에서는 속성의 선택과 이산화를 별개의 문제로 간주하고 있기 때문에, 속성의 이산화에서는 속성의 선택에 적용하는 평가 기준과 다른 별도의 평가 기준을 사용하고 있으며, 연속적인 속성만이 이산화의 대상이 되고 이산적인 속성은 그대로 사용하고 있다.

이 연구에서는 보다 일반적인 관점에서 속성의 이산화와 선택을 이해하고자 한다. 첫째, 속성의 이산화는 값구간이 다르게 분할되는 속성들 중에서 하나를 선택하는 문제이기 때문에, 넓은 의미에서 속성을 선택하는 문제라 할 수 있다. 따라서 동일한 평가 기준에 의해 각 속성의 이산화가 결정되고 그 중 하나의 속성이 선택되는 것이 타당하다고 할 수 있다. 둘째, 주어진 샘플 집합을 분류하기 위해 이산적인 속성을 적용하는 경우에도, 동일한 부류로 구성되는 일부 인접한 구간들을 융합하여 구간수를 줄인다면, 생성되는 결정 트리의 크기를 감소시킬 수 있을 뿐만 아니라 학습에 사용하지 않은 데이터에 대한 에러율도 감소시킬 수 있다. 따라서 속성의 값이 연속적인지 혹은 이산적인지에 관계없이 동일한 방법으로 이산화하는 것이 합리적이라고 할 수 있다. 이제, 결정 트리의 생성에 속성을 사용하기 전에 유념해야 할 점은, 그 속성이 갖는 값이 이산적인지 연속적인지가 아니라, 값들이 순서를 갖고 있는지 아니면 각 값들이 상호 독립적인지를 알아보는 것이다. 예로써, '기온'이라는 속성이 갖는 값이 저온, 상온, 고온이라고 하면, 각 값들이 순서가 있어서 저온과 상온은 인접하지만 저온과 고온은 인접하지 않음을 쉽게 알 수 있다. 반면, '종'이라는 속성이 호랑이, 사자, 치타라는 값을 갖는다고 하면, 각 값들은 독립적이다. 이럴 경우, 어떠한 값도 인접할 수 없다고 가정할 수도 있고 반대로 임의의 두 값은 항상 인접한다고 가정할 수도 있다.

이 논문에서는 속성의 값이 갖는 특성에 관계없고 일관된 평가 기준으로 속성의 이산화 및 선택을 수행하는 새로운 결정 트리 생성기 MEC(Multi-base Entropy based Classification)를 제안한다. 먼저, 기존의 엔트로피 개념을 확장하여 주어진 샘플 집합을 분류하는데 소요되는 속성의 분지수에 따라 정보량을 상대적으로 측정하는 다중 엔트로피(multi-base entropy)[1]를 정의하였으며, 분할 전과 후에 대한 다중 엔트로피의 차를 속성의 이산화 및 선택을 위한 평가 기준으로 사용한다. 실제로 다중 엔트로피의 차를 사용하면, Quinlan이 제안한 엔트로피의 차(Gain)나 이를 값의 정보량으로 나눈 비율(Gain ratio)에 비해 분류 성능이 향상되는 것으로 나타났다[1]. 한편, 속성을 이산화하는 방식으로는 가능한 모든 분할을 한 후, 가장 분할할 필요가 없는 인접한 두 구간을 융합해가면서 주어진 평가 기준에 의해 최선의 이산화를 선택하는 방식을 채택하고 있다.

끝으로, 제안한 결정 트리 생성기의 성능을 평가하기 위해, 엔트로피를 기반으로 하는 평가 기준을 사용하고 이산화 방식에서 큰 차이를 보이는 여러 생성기들과 비교하였다. 먼저, 샘플의 속성값을 그대로 분지하는 방식으로는 평가 기준으로 Gain을 사용하는 ID3[2]와 Gain ratio를 사용하는 ID3-IV[2]가 있다. GID3[3]와 GID3*[4]도 기본적으로는 속성값을 그대로 분지하되 각 값의 분지 여부를 평가하고 적절하지 못한 값들은 모아서 하나의 디폴트 가지(default branch)로 분지함으로써 분지되는 가지의 수를 다소 줄이고 있다. 이때 GID3*는 인위적으로 결정되는 한계치에 의존하지 않도록 하기 위해 GID3를 개선한 생성기이다. 이진 분할하는 방식으로는 평가 기준으로 직교성(orthogonality)을 사용하는 O-BTREE[5]나 대조(contrast)를 사용하는 Ct-2[6] 등도 있으나, 이 논문에서는 ID3와 동일한 평가 기준을 사용하는 생성기를 ID3-BIN[5]이라 부르고, 비교 실험에 참가시켰다. 한편, 주어진 분할의 종료 조건이 만족될 때까지 이진 분할을 반복함으로써 다진 분할하는 방식으로는 D-2[7], C4[8], MDLPC[9] 등이 있는데, 여기서는 최소묘사길이원칙(minimum description length principle)으로부터 유도되는 조건을 사용하는 MDLPC를 선정하였다.

Ⅱ. 결정 트리 생성기 MEC

이 장에서는 다중 엔트로피의 차를 속성의 이산화 및 선택의 일관된 평가 기준으로 사용하고, 인접한 두 구간을 융합해가면서 속성을 이산화하는 새로운 결정 트리 생성기 MEC를 제안한다.

다중 엔트로피(multi-base entropy)는 분류에 적용하는 속성의 용량(capacity)에 따라 정보량을 상대적으로 측정한다. 여기서 속성의 용량이란 최대로 분류해낼 수 있는 부류의 수를 의미하는데, 이는 곧 주어진 샘플 집합에 적용하는 속성의 분지수이다. 샘플 집합 S가 k개의 부류로 구성되고 각 부류의 확률이 $p_1, p_2, ..., p_k$라고 하자. 여기서, 모든 p_i는 0보다 크고, $\sum_{i=1}^{k} p_i = 1$이다. 만일, 속성 A에 의해 집합 S가 부분 집합들 $S_1, S_2, ..., S_n$로 분할된다고 하면, 분할 전의 집합 S에 대한 다중 엔트로피 Mul(A, S)와 분할 후의 부분 집합들에 대한 다중 엔트로피 M(A, S)는 각각 식 (1), 식 (2)와 같이 정의된다.

$$Mul(A, S) = -\sum_{i=1}^{k} p_i \log_n p_i \quad (1)$$

$$M(A, S) = -\sum_{i=1}^{n} P_i \sum_{j=1}^{k_i} p_{ij} \log_n p_{ij},\ P_i = \frac{|S_i|}{|S|} \quad (2)$$

따라서, 집합 S의 분할에 속성 A를 적용함으로써 얻는 이득인 다중 엔트로피의 차는 식 (3)과 같다.

$$\triangle M(A, S) = Mul(A, S) - M(A, S) \quad (3)$$

결국, 다중 엔트로피는 정보량을 계산하는 수식에서 로그의 밑수를 2로 고정시키는 것이 아니라, 주어진 샘플 집합을 분할하는 속성의 분지수를 로그의 밑수로 사용하고 있다. 이와 같이 정의된 다중 엔트로피의 차는 보다 세분화되는 분할일수록 선호하는 엔트로피 차의 특성을 극복하는데 효과적인 것으로 나타났다[1]. 이때, Mul(A, S), M(A, S), △M(A, S)는 필요한 경우에 로그의 밑수 b와 분할 후와 전의 분지수 n, m을 포함하여 각각 $Mul_b(A, S)$, $M_b(A, S, n)$, $\triangle M_b(A, S, n:m)$와 같이 상세하게 표기하도록 한다.

결정 트리의 성장은 다음과 같이 재귀적인 과정으로 수행된다. 먼저, 성장의 종료 조건이 만족되면 최선의 부류(class)를 할당한다. 그렇지 않으면, 최선의 이산화된 속성을 선택하여 샘플 집합을 분할하고, 각 부분 집합에 대해 성장을 계속한다. 이때, 성장의 종료 조건은 주어진 집합의 모든 샘플들이 동일한 부류에 속하거나, 혹은 어떠한 속성으로도 분할이 되지 않는 경우를 의미한다. 그림 1은 MEC에 의한 속성의 이산화 및 선택에 대한 세부 흐름도이다.

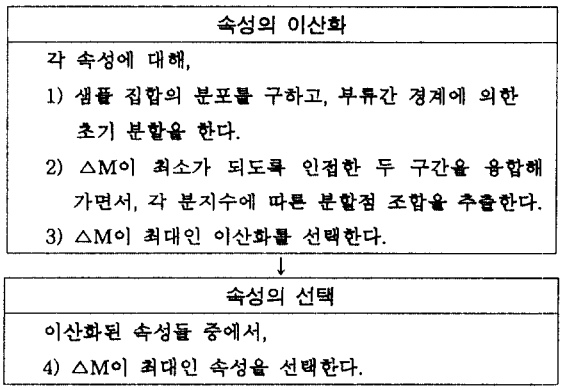


그림 1. MEC에 의한 속성의 이산화 및 선택 과정
Fig. 1. Discretization and selection process of attributes

1. 부류간 경계에 의한 초기 분할

각 속성에 대해 샘플 집합을 분포시키고 모든 부류간의 경계를 분할하면, 분할 후의 다중 엔트로피 M(A, S)는 최소치가 되고, 보다 세분화시키는 분할을 하더라도 그대로 최소치를 유지한다. 그러나 분할 전의 다중 엔트로피 Mul(A, S)는 분할이 세분화될수록 계속 감소하기 때문에, 초기 분할을 위해 모든 부류간의 경계를 분할하는 것 이상의 분할은 불필요하다. 이와 같은 현상은 학습 집합에 몇 개의 부류들이 있고 어떻게 분포하는지에 관계없으며, 분할점을 찾는 알고리듬의 효율성을 향상시키는데 도움이 된다. 여기서, 경계점(boundary point)의 의미는 정의 1과 같다[7].

[정의 1]

속성 A의 값에 따라 정렬된 샘플열에서 두 샘플 e_1, $e_2 \in S$이 서로 다른 부류에 속하고, A의 범위(range)에 있는 값 T에 대해 $A(e_1) < T < A(e_2)$의 관계가 성립하

며, $A(e_1) < A(e') < A(e_2)$을 만족하는 다른 샘플 e'가 존재하지 않으면, T는 경계점(boundary point)이라 한다.

예를 들어, 부류 O에 속하는 샘플 $o_1, o_2, ..., o_9$와 부류 X에 속하는 샘플 $x_1, x_2, ..., x_{11}$로 구성되는 학습 집합이 임의의 속성 A에 의해 그림 2와 같은 분포를 나타낸다고 하자

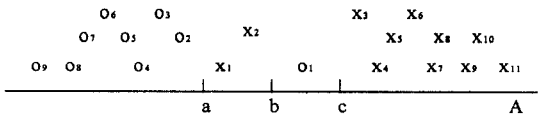


그림 2. 학습 집합의 분포
Fig. 2. Distribution of a training set

이와 같은 분포에서 모든 부류간의 경계를 초기 분할점으로 채택하면 그림 3과 같이 4개의 구간으로 분할된다.

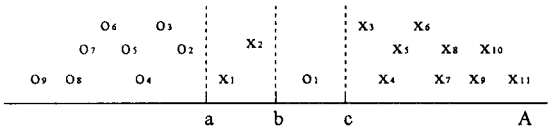


그림 3. 학습 집합의 초기 분할 (분지 수=4)
Fig 3. Initial partitioning of a training set (branch number =4)

2. 인접 구간의 반복적인 융합

이산화를 위해서는 초기 분할점 집합의 일부로 구성되는 가능한 모든 분할점 조합에 대해 분할 전후의 다중 엔트로피의 차 △M을 구하고 이중 최대의 값을 갖는 분할점 조합을 선택하면 된다. 그러나, 가능한 모든 분할점 조합을 재구성하여 평가하기 위해서는 매우 많은 처리 시간이 소요된다.

이 연구에서는 초기 분할부터 시작하여 가장 분할할 필요가 없는 인접한 두 구간을 융합해가면서 최선의 분지수와 분할점 조합을 선택하는 효율적인 방법을 제안한다. 여기서, 가장 분할할 필요가 없는 인접한 두 구간이란, 임의의 인접한 두 구간을 융합하여 생기는 가능한 모든 이산화에 대해 다중 엔트로피 차를 측정했을 때 그 값이 최대가 되게 하는 구간쌍을 의미한다.

다음 그림 4는 임의의 속성 A에 의해 샘플 집합 S가 n개의 부분 집합으로 분할되어 있을 때, 이 중 임의의 인접 구간을 융합하여 구간수를 줄이는 것을 보여주고 있다. 이때, $\triangle M_n(A, S, n:1)$은 융합 전의 이산화에 대한 다중 엔트로피의 차이고, $\triangle M_{n-1}(A, S, n-1:1)$은 융합 후의 이산화에 대한 다중 엔트로피의 차이고, $\triangle M_n(A, S, n:n-1)$는 융합 전후에 대한 다중 엔트로피의 차이며, 정리 1은 이들간의 관계를 나타내고 있다.

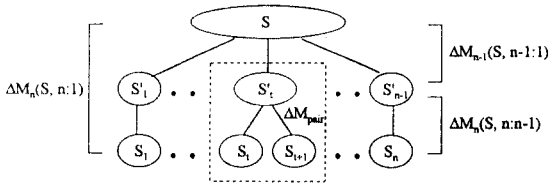


그림 4. 인접한 두 구간을 융합하는 경우의 다중 엔트로피의 차
Fig. 4. Difference of multi-base entropy when merging adjacent two intervals

[정리 1]

임의의 속성 A에 의해 샘플 집합 S가 정렬된 구간열을 따라 n개의 부분 집합 $S_1, S_2, ..., S_n$로 분할된다고 하자. 이 중 임의의 인접한 두 구간의 부분 집합 S_t와 S_{t+1}를 S'_t로 융합하여 n-1개의 부분 집합 $S'_1, S'_2, ..., S'_{n-1}$로 되었을 때, 융합 후의 이산화에 대한 다중 엔트로피의 차 $\triangle M_{n-1}(S, n-1:1)$, 융합 전의 이산화에 대한 다중 엔트로피의 차 $\triangle M_n(S, n:1)$, 그리고 융합 전후에 대한 다중 엔트로피의 차 $\triangle M_n(S, n:n-1)$은 다음 식 (4)와 같은 관계가 성립한다.

$$\triangle M_{n-1}(A, S, n-1:1) = \frac{\triangle M_n(A, S, n:1) - \triangle M_n(A, S, n:n-1)}{\log_n(n-1)} \quad (4)$$

증명:

$\triangle M_n(A, S, n:1) = Mul_n(A, S) - M_n(A, S, n)$

$\triangle M_n(A, S, n:n-1) = M_n(A, S, n-1) - M_n(A, S, n)$

그러므로,

$$\begin{aligned}&\triangle M_n(A, S, n:1)-\triangle M_n(A, S, n:n-1)\\&=[Mul_n(A, S)-M_n(A, S, n)]-[M_n(A, S, n-1)\\&\quad -M_n(A, S, n)]\\&=Mul_n(A, S)-M_n(A, S, n-1)\\&=\log_n(n-1)\cdot[Mul_{n-1}(A, S)-M_{n-1}(A, S, n-1)]\\&=\log_n(n-1)\cdot\triangle M_{n-1}(A, S, n-1:1)\end{aligned}$$

□

식 (4)에서 융합 전의 이산화에 대한 다중 엔트로피의 차 $\triangle M_n(A, S, n:1)$와 분모 $\log n(n-1)$은 어떠한 인접한 두 구간을 선택하더라도 항상 일정한 값을 갖는다. 따라서, 융합 후의 이산화에 대한 다중 엔트로피의 차 $\triangle M_{n-1}(A, S, n-1:1)$은 단지 융합 전후에 대한 다중 엔트로피의 차 $\triangle M_n(A, S, n:n-1)$에 의해 결정됨을 알 수 있다. 그러므로, 구간의 수를 하나 줄이면서 다중 엔트로피의 차를 최대로 하는 분할을 선택하기 위해서는, $\triangle M_n(A, S, n:n-1)$를 최소로 하는 구간쌍을 선택하여 융합하면 된다.

다음 정리 2는 인접한 두 구간만에 대해 융합 전후에 대한 지역적인 다중 엔트로피의 차 $\triangle M_{pair}$와 구간쌍의 확률 P_{pair}만을 이용하여 $\triangle M_n(A, S, n:n-1)$을 효율적으로 계산할 수 있음을 보여주고 있다. 여기서, $\triangle M_{pair}$는 앞의 그림 4에 나타나 있다.

[정리 2]

임의의 속성 A에 의해 샘플 집합 S가 정렬된 구간열을 따라 부분 집합들 $S_1, S_2, ..., S_n$로 분할되어 있다고 하자. 이 중 임의의 인접한 두 구간의 부분 집합 S_t와 S_{t+1}의 융합 전후에 대한 다중 엔트로피의 차 $\triangle M_n(A, S, n:n-1)$은 다음 식 (5)와 같이 계산된다. 여기서, $\triangle M_{pair}$는 두 구간만이 존재한다고 간주하고 계산하는 지역적인 다중 엔트로피의 차를 의미한다.

$$\triangle M_n(A, S, n:n-1)=(\log_n 2)P_{pair}\cdot\triangle M_{pair} \qquad (5)$$

$$P_{pair}=P_t+P_{t+1}=\frac{|S_t|+|S_{t+1}|}{|S|}$$

$$\triangle M_n=\triangle M_2(A, S_t\cup S_{t+1}, 2:1)$$

증명:

$$\triangle M_n(A, S, n:n-1)=M_n(A, S, n-1:1)-M_n(A, S, n:1)$$

$$M_n(A, S, n-1:1)=-\sum_{i=1}^{n-1}P_i'\sum_{j=1}^{k_i'}p_{ij}'\log_n p_{ij}'$$

$$M_n(A, S, n:1)=-\sum_{i=1}^{n}P_i\sum_{j=1}^{k_i}p_{ij}\log_n p_{ij}$$

여기서, $1\le i\le k-1$에 대해,

$$\text{if } i<t: P_i'=P_i, \quad p_{ij}'=p_{ij}$$

$$\text{if } i=t: P_i'=P_i+P_{i+1}, \quad p_{ij}'=p_{ij}+p_{(i+1)j}$$

$$\text{if } i>t: P_i'=P_{i+1}, \quad p_{ij}'=p_{(i+1)j}$$

$$\begin{aligned}&=-\Big(\sum_{i=1}^{n-1}P_i'\sum_{j=1}^{k_i'}p_{ij}'\log_n p_{ij}'-\sum_{i=1}^{n}P_i\sum_{j=1}^{k_i}p_{ij}\log_n p_{ij}\Big)\\&=-\Big(P_t'\sum_{j=1}^{k_t'}p_{tj}'\log_n p_{tj}'-P_t\sum_{j=1}^{k_t}p_{tj}\log_n p_{tj}\\&\quad -P_{t+1}\sum_{j=1}^{k_{t+1}}p_{(t+1)j}\log_n p_{(t+1)j}\Big)\end{aligned}$$

이때, 두 구간만에 대한 지역적인 값은 구별이 쉽도록 위 첨자 *를 사용한다.

$$P_{pair}=P_t'=P_t+P_{t+1}$$

$$P_t'=\frac{|S_t'|}{|S|}$$

$$P_t=\frac{|S_t|}{|S|}=\frac{|S_t'|}{|S|}\frac{|S_t|}{|S|}=P_t'P_1^*=P_{pair}P_1^*$$

$$P_{t+1}=\frac{|S_{t+1}|}{|S|}=\frac{|S_t'|}{|S|}\frac{|S_{t+1}|}{|S|}=P_t'P_2^*=P_{pair}P_2^*$$

$$p_j^*=p_{tj}',\ p_{ij}^*=p_{(t+i-1)j}$$

$$\begin{aligned}&=-P_{pair}\Big(\sum_{j=1}^{k^*}p_j^*\log_n p_j^*-\sum_{i=1}^{2}P_i^*\sum_{j=1}^{k_i'}p_{ij}^*\log_n p_{ij}^*\Big)\\&=-(\log_n 2)P_{pair}\Big(\sum_{j=1}^{k^*}p_j^*\log_2 p_j^*-\sum_{i=1}^{2}P_i^*\sum_{j=1}^{k_i'}p_{ij}^*\log_2 p_{ij}^*\Big)\\&=-(\log_n 2)P_{pair}\cdot\triangle M_{pair}\end{aligned}$$

□

예로써, 앞의 그림 3에서 각 분할점 a, b, c에 대해 융합 전후의 다중 엔트로피의 차를 각 $\triangle M(a)$, $\triangle M(b)$, $\triangle M(c)$라 할 때, 정리 2의 식 (5)에 의해 계산된 결과는 다음과 같다.

$$\triangle M(a)=(\log_4 2)P_{pair}\cdot\triangle M_{pair}$$

$$= -\log_4 2 \cdot \frac{10}{20} \cdot [(\frac{8}{10} \log_2 + \frac{8}{10}$$

$$+ \frac{2}{10} \log_2 \frac{2}{10}) - 0] = 0.180$$

$\triangle M(b) = 0.069$

$\triangle M(c) = 0.117$

따라서, $\triangle M$이 최소인 분할점 b를 융합하여 그림 5와 같이 분지수가 3인 분할점 조합 {a, c}를 선택한다.

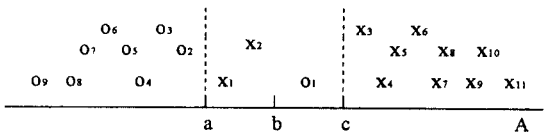


그림 5. 분할점 b에 인접한 두 구간의 융합 (분지수 = 3)
Fig. 5. Merging two intervals adjacent to a point b (branch number = 3)

마찬가지로, 분할점 a와 c에 대한 다중 엔트로피의 차 $\triangle M(a)$와 $\triangle M(c)$를 구하면 다음과 같다.

$\triangle M(a) = 0.164$

$\triangle M(c) = 0.070$

그림 6은 동일한 방식으로 분할점 c가 융합되어 2개의 구간으로 분할된 이산화를 나타낸다.

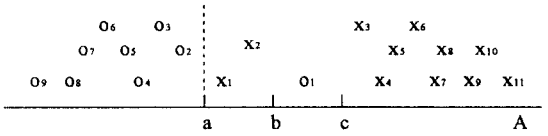


그림 6. 분할점 c에 인접한 두 구간의 융합 (분지수 = 2)
Fig. 6. Merging two intervals adjacent to a point c (branch number = 2)

3. 최선의 이산화 선택 및 속성 선택

분지수에 따른 각 분할점 조합에 대해 $\triangle M$을 구하고, 이 중 최대가 되는 이산화를 선택한다. 앞의 예에서 각 분지수에 따라 결정된 분할에 대한 $\triangle M$을 각각 $\triangle M(\{a, b, c\})$, $\triangle M(\{a, c\})$, $\triangle M(\{a\})$이라 하면 계산된 결과는 다음과 같다.

분지수 = 4일 경우 : $\triangle M(\{a, b, c\})$

$$= -[(\frac{9}{20} \log_4 \frac{9}{20} + \frac{11}{20} \log_4 \frac{11}{20}) - 0] = 0.496$$

분지수 = 3일 경우 : $\triangle M(\{a, c\}) = 0.539$

분지수 = 2일 경우 : $\triangle M(\{a\}) = 0.744$

따라서, 속성 A에 대해서는 $\triangle M$이 최대가 되는 분할점 a만을 분할하는 이산화가 채택된다. 그리고, 이와 같은 과정으로 이산화된 모든 후보 속성들 중에서 $\triangle M$이 최대인 속성을 선택한다.

Ⅲ. 실험 결과

결정 트리를 생성하기 위해, 표 1과 같이 UCI Repository에서 제공하는 총 10 종류의 데이터베이스를 사용하였다. 이 중 5개의 데이터베이스는 속성값이 이산적인 값으로 구성되어 있고, 나머지 5개는 연속적인 값으로 구성되어 있다.

U. M. Fayyad 등은 학습 집합으로부터 생성된 결정 트리의 단말 노드의 수가 적을수록 확률적인 의미에서 전체 노드의 수, 속성 테스트의 수, 단말 노드당 평균 샘플 수는 물론 에러율 측면에서 향상된 성능을 갖는다는 것을 입증하였다[10]. 따라서, 이 실험에서는

표 1. 데이터베이스
Table 1. Databases

DB \ 특성		부류수	속성수	샘플수
이산적	Tic-tac-toe Endgame Database	2	9	958
	Fitting Contact Lenses Database	3	4	24
	Small Soybean Database	5	35	47
	Zoo Database	7	16	101
	Flag to Religion Database	8	23	194
연속적	Ionosphere Database	2	34	351
	Iris Plants Database	3	4	150
	Wine Recognition Database	3	13	178
	Glass Identification Database	6	9	214
	Image Segmentation Database	7	19	210

동일한 학습 집합에 대해 각 생성기에 의해 결정 트리를 생성하고 단말 노드의 수를 조사하였다.

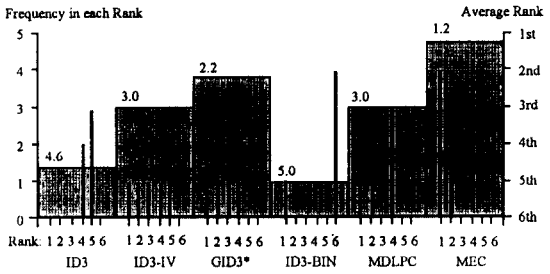
제안한 생성기 MEC를 비롯하여 실험에 참가한 ID3, ID3-IV, GID3*, ID3-BIN 등은 모든 객체들이 동일한 부류로 구성되거나 어떠한 속성으로도 구분이 되지 않는 경우에 분류를 중지하는 기본적인 성장의 종료 조건을 갖고 있다. 그런데 MDLPC만은 반복적인 이진 분할을 중단시킬 수 있는 별도의 종료 조건을 추가로 두고 있다[9]. 따라서 부류의 구분이 가능한 속성이 존재하는 경우에도 더 이상의 분할을 하지 않고 그 샘플 집합을 단말 노드로 인정하는 경우가 발생할 수 있다. 이럴 경우 단말 노드수의 차이가 곧 생성기의 성능 차이라 할 수 없다. 따라서, 나머지 생성기들은 상호간에 비교를 위해서는 단말 노드수를 그대로 비교하면 되지만, MDLPC와 비교하기 위해서는 MDLPC에 의해 생성된 분류기와 동일한 혹은 유사한 에러율로 절단(pruning)을 한 후에 단말 노드수를 비교해야 한다. 이를 위해, 단말이 아닌 각 노드에 대해 절단 전후의 에러차와 그 노드를 루트로 하는 부분 트리의 단말 노드수를 구하여 단말 노드당 평균 에러차를 계산하고, 이 값이 최소인 노드를 절단한다. 이와 같은 과정을 원하는 에러율이나 근접한 에러율이 될 때까지 반복한다.

표 2. 결정 트리 생성기의 성능 비교
Table 2. Performance comparison of decision tree generators

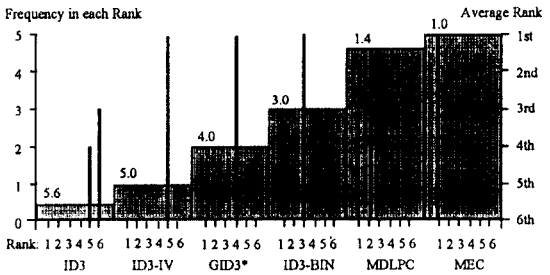
DB	성능	결정 트리 생성기					
		ID3	ID3-IV	GID3*	ID3-BIN	MDLPC	MEC
Tic-tac-toe	에러율(%)	0	0	0 (24.2)	0	23.5	0 (23.5)
	단말 노드수	218	200	169 (10)	409	8	81 (7)
	등위	5	4	3	6	2	1
Lenses	에러율(%)	0	0	0	0 (12.5)	12.5	0 (12.5)
	단말 노드수	9	9	8	7 (3)	3	7 (3)
	등위	5	5	4	1	1	1
Soybean	에러율(%)	0	0	0	0	0	0
	단말 노드수	5	4	4	6	5	4
	등위	4	1	1	6	4	1
Zoo	에러율(%)	0	0	0	0	0	0
	단말 노드수	14	11	11	34	13	10
	등위	5	2	2	6	4	1
Flag	에러율(%)	2.6 (60.3)	2.6 (64.9)	2.6	2.6	68.0	2.6
	단말 노드수	139 (8)	123 (2)	109	148	2	116
	등위	4	3	1	6	4	2
Ionosphere	에러율(%)	0	0	0	0 (3.4)	3.4	0 (4.3)
	단말 노드수	296	276	217	91 (65)	13	33 (9)
	등위	6	5	4	3	1	1
Iris	에러율(%)	0	0	0	0	2.7	0 (2.7)
	단말 노드수	64	43	26	14	4	9 (4)
	등위	6	5	4	3	1	1
Wine	에러율(%)	0	0	0	0 (2.8)	2.8	0 (2.2)
	단말 노드수	147	147	117	72 (59)	6	9 (5)
	등위	5	5	4	3	2	1
Glass	에러율(%)	0	0	0	0 (22.0)	22.0	0 (41.1)
	단말 노드수	194	192	164	125 (43)	16	76 (6)
	등위	6	5	4	3	1	1
Image	에러율(%)	0	0	0	0 (3.8)	3.8	0 (3.3)
	단말 노드수	202	202	198	101 (85)	18	20 (12)
	등위	5	5	4	3	2	1

표 2는 이러한 방법으로 생성기들의 성능을 비교하고 등위를 구한 결과이다. MDLPC와의 비교를 위해, 절단된 트리에 의한 실험치는 괄호 안에 나타내었으며, 우열을 가릴 수 없는 경우에는 공동 등위를 부여하였다.

그림 7은 총 10 개의 데이터베이스를 데이터의 유형에 따라 두 그룹으로 나누고, 각 결정 트리 생성기에 대한 등위별 분포와 평균 등위(average rank)를 나타낸 것이다. 그래프의 수평축은 각 결정 트리 생성기마다 1등부터 6등까지 등위가 나열되어 있다. 왼쪽의 수직축은 각 등위별 빈도수를 의미하며 가는 막대로 표시한다. 오른쪽의 수직축은 생성기의 평균 등위를 의미하며 투명한 굵은 막대로 나타낸다. 여기서 총 r개의 결정 트리 생성기를 비교할 때, 임의의 생성기 G에 대해 각 등위 i에서의 빈도수가 $f_G(i)$라고 하면 평균 등위는 다음 식 (6)과 같이 계산된다.



(a) 이산적인 속성값으로 구성되는 5개의 DB를 사용한 경우



(b) 연속적인 속성값으로 구성되는 5개의 DB를 사용한 경우

그림 7. 결정 트리 생성기의 등위별 빈도수 및 평균 등위
Fig. 7. Frequency in each rank and average rank of decision tree generators

$$\text{Average rank}(G) = \frac{\sum_{i=1}^{r} (1 \times f_G(i))}{\sum_{i=1}^{r} f_G(i)} \quad (6)$$

결과적으로, 제안한 방법 MEC가 데이터의 유형에 관계없이 일관되게 최상위 등위를 유지하고 있음을 알 수 있다. 이는 연속적인 속성과 이산적인 속성이 혼합되어 있는 데이터베이스에 대해서도 좋은 성능을 보일 수 있음을 의미한다. 반면, 기존의 생성기들은 데이터베이스의 유형에 따라 성능이 크게 달라진다. 예로써, GID3*의 경우 이산적인 특성을 갖는 DB에 대해서는 평균 등위가 2.2이지만, 연속적인 특성을 갖는 DB에 대해서는 4.0 등위에 그치고 있다. 다른 예로, MDLPC는 이산적인 속성값으로 구성되는 DB에 대해서는 3.0 등위이지만, 연속적인 속성값으로 구성되는 DB에 대해서는 1.4 등위를 유지하고 있다.

이상과 같이, 결정 트리의 생성에서 MEC가 향상된 성능을 보이는 이유는 다음과 같이 요약될 수 있다. 첫째, 속성의 이산화 및 선택을 위해 일관된 평가 기준을 사용한다는 점이다. 기존의 결정 트리 생성기에서는 속성의 이산화와 선택을 별개의 문제로 취급하였으며, 대체로 속성의 이산화와 선택에 다른 기준이 적용되어 왔다. 그러나 속성의 이산화도 다르게 분할된 속성들 중에서 하나를 선택하는 문제이므로 동일한 기준에 의해 속성의 이산화와 선택이 결정되는 것이 보다 일관성 있는 방법이라 할 수 있다. 둘째, 다중 엔트로피의 차는 기존의 엔트로피 차의 한계를 효과적으로 극복하고 있다는 점이다. 즉, 보다 세분화된 분할일수록 무조건 선호하는 것이 아니라, 적정한 분지수로 분할을 한 후 동일한 속성으로 계속 분할할지 아니면 다른 속성으로 분할할지를 공정히 경쟁시킬 수 있기 때문에 보다 최적에 근접한 결정 트리를 생성시킬 수 있다. 셋째, 속성의 이산화를 위해, 인접한 두 구간을 융합해가면서 최선의 이산화를 결정하는 점진적인 방식을 사용한다는 점이다. 이러한 방식의 경우, 적정의 구간수를 미리 정할 필요가 없으며, 이산화를 선택하는 과정을 종료하기 위한 별도의 어떠한 휴리스틱한 기준도 필요로 하지 않는다.

Ⅳ. 결 론

이 논문에서는 다중 엔트로피(multi-base entropy)의 차를 속성의 이산화 및 선택을 위한 일관된 평가 기준으로 사용하는 새로운 결정 트리 생성기 MEC를 제안하였다.

제안한 결정 트리 생성기 MEC는 인접한 두 구간을 융합해가면서 다중 엔트로피의 차에 의해 최선의 분지수와 분할점 조합을 선택하는 효율적인 이산화 방식을 채택하고 있다. 또한, 각 노드에서 적정의 분지수로 속성을 이산화한 후에, 하위 노드에서 그 속성으로 계속 분할할지 아니면 다른 속성으로 분할할지를 공정하게 경쟁시킬 수 있다. 만일 더 좋은 다른 속성이 존재하지 않으면 상위 노드에서 사용한 속성을 다시 사용하게 된다. 이러한 방법은 각 노드에서 속성의 이산화를 한번에 결정하는 방법에 비해, 지역적인 최적에서 벗어나 보다 전역적인 최적에 근접하는 결정 트리를 생성시킬 수 있다.

실험 결과, 동일한 학습 집합으로 결정 트리를 생성할 때, 제안한 결정 트리 생성기 MEC는 데이터베이스의 유형에 관계없이 기존의 생성기들에 비해 보다 적은 단말 노드수로 구성되는 분류기를 생성시키는 것으로 나타났다.

참 고 문 헌

1. 전병환, 결정 트리 생성에서 속성의 이산화 및 선택을 위한 다중 엔트로피, 연세대학교 본 대학원 전자공학과 박사학위논문, 1996년 6월.
2. J. R. Quinlan, "Induction of Decision Trees," *Machine Learning*, Vol. 1, pp.81-106, 1986.
3. J. Cheng, U. M. Fayyad, K. B. Irani and Z. Quian, "Improved Decision Trees: A Generalized Version of ID3," *Proc. of 5th Int. Conf. on Machine Learning*, San Mateo, CA: Morgan Kaufmann, pp. 100-108, 1988.
4. U. M. Fayyad, "Branching on Attribute Values in Decision Tree Generation," *Proc. of 12th National Conference on Artificial Intelligence*, Seattle Washington, pp.601-606, Jul. 31-Aug. 4, 1994.
5. U. M. Fayyad and K. B. Irani, "The Attribute

Selection Problem in Decision Tree Generation," *Proc. of 10th National Conference on AI*, pp. 104-110, Jul. 1992.

6. T. V. Merckt, "Decision Trees in Numerical Attribute Spaces," *Proc. of 13th International Joint Conference on Artificial Intelligence*, pp.1016-1021, Aug. 1993.
7. J. Catlett, "On Changing Continuous Attributes into Ordered Discrete Attributes," *European Working Session on Learning*, 1991.
8. J. R. Quinlan, P. J. Compton, K. A. Horn and L. Lazarus, "Inductive Knowledge Acquisition: A Case Study," In Quinlan J. R., *Applications of Expert Systems*, Addison-Wesley, Sydney, pp. 157-173, 1987.
9. U. M. Fayyad and K. B. Irani, "Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning," *Proc. of 13th International Joint Conference on Artificial Intelligence*, pp. 1022-1027, Aug. 1993.
10. U. M. Fayyad and K. B. Irani, "What Should Be Minimized in a Decision Tree?," *Proc. of the 8th National Conference on Artificial Intelligence*, pp. 749-754, 1990.

전 병 환(Byung Hwan Jun) 정회원

1989年 2月 : 연세대학교 전자공학과 공학사.
1991年 8月 : 연세대학교 대학원 전자공학과 공학석사.
1996年 8月 : 연세대학교 대학원 전자공학과 공학박사.
1997年~현재 : 국립공주대학교 전자계산학과 전임강사.
※주관심분야 : 패턴인식, 문자인식, 분산처리 등.

김 재 희(Jaihie Kim) 정회원

1979年 2月 : 연세대학교 전자공학과 공학사.
1982年 7月 : Case Western Reserve University(USA) 공학석사.
1984年 5月 : Case Western Reserve University(USA) 공학박사.
1984年~현재 : 연세대학교 전자공학과 교수.
※주관심분야 : 인공지능, 패턴인식, 문자인식, 얼굴인식, 정보융합 등.