論文96-33B-6-11

유전 알고리즘을 이용한 이진 결정 트리의 설계와 응용

(A design of binary decision tree using genetic algorithms and its applications)

鄭 淳 元 *, 朴 貴 泰 **

(Soon-Won Jung and Gwi-Tae Park)

요 약

본 논문에서는 새로운 이진 결정 트리(binary decision tree)의 설계 방법을 제안하였다. 이 방법에서는 유전 알고리즘과 FCM(fuzzy c-means) 알고리즘을 사용하여 이진 결정 트리를 구성하게 된다. 각 노드에서는 유전 알고리즘에 의해 분류 에러, 군집간의 균형, 사용되는 특징의 개수에 반비례하는 적합도 함수를 최적화 시키는 최적의(optimal) 혹은 최적 근처의(near optimal) 특징 부집합(subset)을 선정하게 된다. 유전 알고리즘에 있어서의 이진 스트링은 특징 부집합을 결정하며, FCM 알고리즘의 결과인 퍼지 분할 행렬(fuzzy partition matrix)로부터 구한 분류 결과 - 분류 에러, 밸런스 - 는 다음 세대의 재생산(reproduction)에 영향을 미치게 된다. 제안되는 방법을 필기체 영문자와 타이어 접지면 패턴 인식에 적용하여 보았으며, 실험 결과는 본 방법이 유용함을 보여준다.

Abstract

A new design scheme of a binary decision tree is proposed. In this scheme a binary decision tree is constructed by using genetic algorithm and FCM algorithm. At each node optimal or near-optimal feature subset is selected which optimizes fitness function in genetic algorithm. The fitness function is inversely proportional to classification error, balance between cluster, number of feature used. The binary strings in genetic algorithm determine the feature subset and classification results - error, balance - from fuzzy partition matrix affect reproduction of next generation. The proposed design scheme is applied to the tire tread patterns and handwritten alphabetic characters. Experimental results show the usefulness of the proposed scheme.

I. 서 론

많은 부류의 패턴들을 분류하는데 있어서 결정 트리 분류기(decision tree classifier)는 광범위하게 쓰이는 기법이다. 트리 분류기와 같은 다단계(multi-stage) 분류기의 기본 개념은 복잡하고 전역적인 결정을 몇 개의 단순하고 국부적인 결정으로 나누어 빠르고, 정확한 의사 결정을 가능하게 한다는 것이다.[1] 결정 트리 분류기를 설계하기 위해서는 먼저 결정 트리를 구성해야 한다.

결정 트리를 구성하는데 있어 고려해야 할 사항은 크게 두 가지로, 하나는 각 노드에 대한 하부노드의 개수이고 다른 하나는 각 노드에서의 특징 선정(feature selection) 문제이다. 본 논문에서는 하부 노드의 개수가 2인 이진 트리를 구성하였다. 각 노드간의 중복의 문제와 균형의 문제가 있지만, 이진 트리를 이용하면 기존의 어떠한 단일 단계 분류기도 트리 분류기로 변

* 正會員, 高麗大學校 電氣工學科
(Dept. of Electrical Eng. Korea Univ.)
** 正會員, 高麗大學校 電氣工學科, 서울대 ERC-ACI 硏究 委員
(Dept. of Electrical Eng. Korea Univ., ERC-ACI Researcher, Seoul Univ.)

接受日字:1995年9月6日, 수정완료일:1996年5月3日

환할 수 있으며, 구성상 구조가 간단하다는 장점을 가지므로 이진 트리를 많이 사용한다[2].

한편 특징의 선정 문제는 패턴 인식에 있어서 매우 중요한 주제 중의 하나이다. 분류에 꼭 필요한 특징만을 선정하여 분류기에 사용하면 분류 정밀도(classification accuracy)가 높아지고 특징 추출 및 분류에 소요되는 계산 시간이 줄어들게 된다. 또한 단일 단계(one stage)인 경우에 비해 트리 구조인 경우 특징 선택이 더 용이해 진다. 이진 트리 분류기에 대한 많은 연구가 있었으며[1,2,3,4,5], 특히 80년대 이후 퍼지 개념이 도입된 FCM 군집화 알고리즘을 이용한 이진 트리 분류기에 대한 연구들이 있었다.[6,7,8,9] [6]에서는 트리의 각 노드에서의 특징 선정 문제를 고려하지 않았다. [7]에서는 이진 트리의 구성과 특징 선정이 동시에 이루어 졌다. 각 노드에서 이진 군집화를 위하여 처음에 선정된 모든 특징들이 사용되며 일단 군집화가 된 후 분류 척도(separation index)를 이용하여 단 한 개의 특징이 선정된다. 그러나 각 노드에서 이러한 방법으로 구한 한 개의 특징이 데이터 전체적인 관점에서 볼 때 전체 특징을 대표한다고 보기 힘들고 단 하나의 특징만을 사용함으로 인해 인식률이 저하되어 실제 적용상의 어려움이 있다. 또한 이 방법은 노이즈(outlier pattern)에 너무 민감하다는 단점이 있다. 그러나 q개의 특징 중에서 최적의 특징을 추출하는 경우 성능 시험을 위한 특징의 부분 집합의 개수는 (2^q-1)개가 되며 특징의 개수가 조금만 커져도 실제 적용하기에는 어려움이 많다.

이러한 특징 선정 문제를 최적화 기법의 하나로써 최근 많은 관심을 모으고 있는 유전 알고리즘을 이용하여 해결하려는 시도가 있었으며 만족할 만한 결과를 보여주고 있다[10]. 그러나 [10]은 근본적으로 본 논문에서 제시하는 이진 결정 트리의 설계에 적용하기에는 부적당하다고 할 수 있다. 3장에서 자세히 다루겠지만 [10]에서 제시한 방법은 전체 특징 개수에 대해 선정될 특징의 개수를 미리 정하고 최적의 특징 부집합을 구하며, 분류 에러를 최소화시키는 특징들을 구하였다 하더라도 일반적으로 트리 구조에서 요구하는 분할된 군집간의 균형 문제를 고려하지 않은 것이므로 이진 결정 트리의 설계에 적용하기에는 부적당하다고 할 수 있다. 또한 위 방법들은 모두 효율적인 이진 트리를 구성하기 위해 필요한 각 노드에서의 균형 문제를 고려하지 않았다.

본 논문에서는 유전 알고리즘을 이진 결정 트리의 설계에 효과적으로 적용시키기 위하여 유전 알고리즘 내의 이진 스트링을 특징 부집합에 적절히 대응시키는 방법을 제시하고 분류 에러, 군집간의 균형, 선정된 특징의 개수 등을 포함하는 적합도 함수를 정의하여 각 노드에서의 특징 선정, 균형 유지 등의 문제를 해결하고자 한다. 한편 각 노드에서의 분류 에러, 균형 계수(balance coefficient)는 FCM 군집화 알고리즘을 이용하여 구한 퍼지 분할 행렬로부터 얻어진다. 또한 제안되는 방법을 필기체 영문자와 타이어 접지면 패턴 인식에 적용하여 만족할 만한 결과를 얻을 수 있었다.

II. 유전 알고리즘

유전 알고리즘은 1970년대 미국의 John Holland 교수에 의해 정립된 이론으로 자연의 유전학과 자연 선택의 원리에 근거한 최적해 탐색 방법이다[11]. 기존의 최적해 탐색이 국부 탐색을 하는데 반해 유전 알고리즘은 여러 해를 동시에 탐색하는 전역 탐색을 함으로서 전역적인 최적해를 찾을 확률이 기존의 최적화 탐색에 비해 큰 것이 특징이다. 유전 알고리즘의 성능은 실제 파라미터의 부호화 기법(coding technique), 복제(reproduction), 교배(crossover), 돌연변이(mutation)등의 유전 연산자(genetic operator)와 평가 함수(evaluation function)의 설정 등에 크게 의존한다[12].

일반적인 이진 부호화 기법에 의해 생물과 같은 재생산, 교배, 돌연변이를 거쳐 다음 세대의 자손을 만들어 내는 과정은 다음과 같다.

i) 부호화 및 초기화(coding and initialization)
생물의 유전 정보를 담고 있는 염색체와 같이 유전 알고리즘에서는 염색체에 해당하는 파라메터를 부호화 한다. 일반적인 부호화 방법은 파라메터를 유한 길이의 이진 스트링으로 부호화하며 랜덤하게 N개의 문자열을 생성하여 초기 해집단(initial population)을 구한다.

ii) 적합도 평가(fitness evaluation)
각 문자열을 디코딩하여 목적 함수에 대한 적합도를 계산한다.

iii) 복제
자연 선택의 개념을 기반으로 높은 적합도를 가진 문자열에 대하여 다음 세대로 복제될 확률을 높게 한다.

iv) 교배

두 문자열을 임의로 선정하여 문자열 안에 있는 유전자 정보를 서로 교환하여 새로운 정보를 갖는 문자열을 만든다.

v) 돌연변이

문자열 안에 있는 유전자의 일부를 임의로 바꾸어 새로운 정보를 갖는 문자열을 만든다.

위와 같은 과정을 반복하여 유전 알고리즘은 최적의 해를 탐색해 나간다. 유전 알고리즘에 관한 자세한 내용은 [11]에 잘 기술되어 있다.

Ⅲ. FCM 알고리즘 및 유전 알고리즘을 이용한 이진 결정 트리의 설계

1. FCM 알고리즘

FCM 알고리즘은 주어진 데이터 집합, $X = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n\}$($\vec{x}_k \in R^q$, q는 특징의 개수)에 대하여 특정 내적 노옴자를 거리 척도로 사용하여, 정의된 어떤 목적 함수가 근사적 최소 값에 도달되도록 퍼지 분할 행렬 U와 군집의 중심값 $V=\{\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_c\}$를 반복 계산법에 의해 구하는 최적화 퍼지 군집화 알고리즘이다[13]. 일반적으로 FCM에서 사용되는 목적 함수는 다음과 같다.

$$J_m(U, V) = \sum_{k=1}^{n}\sum_{i=1}^{c} (u_{ik})^m (d_{ik})^2 \qquad (1)$$

$$d_{ik}^2 = \| \vec{x}_k - \vec{v}_i \|^2$$

여기서 m은 지수 가중치이며 c는 군집의 개수이므로 이진 트리를 구성할 경우 c=2가 된다.

입력 데이터가 기지의(labelled) 패턴일 경우 FCM 알고리즘의 결과로서 나오는 퍼지 분할 행렬 U로부터 분류 에러, 군집의 균형 정도를 구할 수 있으며 이러한 값들은 유전 알고리즘에서의 적합도 함수 값을 결정하는 중요한 요소가 된다. 또한 FCM알고리즘을 통하여 구한 군집의 중심값 V는 일반 K-*means* 알고리즘을 이용하여 구한 중심값보다 노이즈 패턴에 의한 영향을 덜 받게 된다.

2. 유전 알고리즘을 이용한 이진 결정 트리의 설계

본 논문에서는 이진 결정 트리를 설계하는데 있어서 각 노드에서 필요한 특징을 구하기 위하여 최적화 기법의 하나인 유전 알고리즘을 이용하였다.

전체 특징 개수를 q라하고 각 노드에서 사용할 수 있는 최대 특징 개수를 q_s라 하자. 유전 알고리즘에서 사용되는 이진 스트링을 k-bit라 하면 이 스트링이 나타낼 수 있는 가짓수는 모두 2^k개가 된다. 만일 [10]에서처럼 $k=q$라하고 각 bit의 '0', '1'을 각 특징의 '유', '무'로 대응시킬 경우 두 가지 문제점이 발생하게 된다.

첫째 문제는 $q_s<q$인 경우 이진 스트링중 '1'의 개수가 q_s보다 큰 경우 이를 늘 보정시켜야 한다는 문제가 있다. 이는 집단 초기화에서뿐만 아니라 교배, 돌연 변이 시에도 늘 발생되는 문제이다.

두 번째는 위와 같은 문제를 bit보정에 의하여 해결한다고 하더라도 유전 알고리즘을 통해 선정되는 특징의 개수간에 불평등이 생기게 된다. 전체 q개의 특징중에서 i개의 원소를 가지는 특징 부집합이 선정되는 경우의 수를 C_i라 정의하자.

$$C_i = {}_qC_i \qquad (2)$$

확률적으로 l개의 원소를 가지는 특징 부집합이 선정될 확률, P_l은 다음과 같이 주어진다.

$$P_l = \frac{{}_qC_l}{\sum_{i=1}^{q_s} {}_qC_i} \qquad (3)$$

예로서 q=20, q_s=10일 경우 1개, 5개, 10개의 특징을 가지는 특징 부집합이 선정될 확률은 각각 P_1=3.24 $\times 10^{-5}$, P_5=2.51 $\times 10^{-2}$, P_{10}=0.3이 되어 심한 불평형 상태를 보인다. 따라서 위 두 문제를 해결하기 위해 다음과 같이하여 특징 개수 선정의 평형을 유지하도록 한다.

먼저 k-bit가 나타낼 수 있는 총 가짓수 2^k를 q_s등분하여 각 구간을 각 특징 개수에 대응하도록 하였고, 각 구간을 그 구간에 속하는 특징 부집합의 종류로 다시 나누어 사용되는 특징에 대응하게 하였다. 표 1에 k=16, q=8, q_s=4일 경우 그 과정을 나타내었다.

예로서 집단중 한 스트링을 10진수로 디코딩하여 17,000의 값이 나왔다면 표 1 에서 알 수 있듯이 특징 부집합 $\{f_1, f_3\}$에 해당하는 의미임을 알 수 있다.

유전 알고리즘에서의 이진 스트링을 특징 부집합에 대응시키는 과정을 설명했으므로 이제 유전 알고리즘과 FCM 알고리즘을 결합하는 방법에 대해 알아보자.

표 1. 이진 스트링과 특징량 부집합의 대응 과정

Table 1. Correspondence between binary string and feature subset.

Number of features	Range	Sub-range	Feature subset	Number of subset
1	0 ~ 16383 (65536/4)	0 ~ 2047	$\{f_1\}$	${}_8C_1=8$
		2048 ~ 4095	$\{f_2\}$	
		⋮	⋮	
		14336 ~ 16383	$\{f_8\}$	
2	16384 ~32767	16384 ~ 16969	$\{f_1, f_2\}$	${}_8C_2=28$
		16970 ~ 17555	$\{f_1, f_3\}$	
		⋮	⋮	
		32182 ~ 32767	$\{f_7, f_8\}$	
3	32768 ~ 49151	32768 ~ 33061	$\{f_1, f_2, f_3\}$	${}_8C_3=56$
		33062 ~ 33353	$\{f_1, f_2, f_4\}$	
		⋮	⋮	
		48858 ~ 49151	$\{f_6, f_7, f_8\}$	
4	49152 ~ 65535	49152 ~ 49386	$\{f_1, f_2, f_3, f_4\}$	${}_8C_4=70$
		49387 ~ 49621	$\{f_1, f_2, f_3, f_5\}$	
		⋮	⋮	
		65301 ~ 65535	$\{f_5, f_6, f_7, f_8\}$	

유전 알고리즘을 수행하는 첫 단계는 앞서 설명한 바와 같이 미리 정한 비트 수를 가진 N개의 스트링을 랜덤하게 초기화시키는 것으로부터 시작된다. 먼저, 초기화된 각 스트링은 표 1과 같은 과정을 거쳐 특징 부집합으로 대응되게 된다. 즉 각 스트링에 대응되는 특징 부집합이 구해지게 된다. 유전 알고리즘의 일반적 응용중의 하나인 함수값 극점 탐색에 있어서는 스트링 자체가 어떤 값의 의미를 가지나 여기서는 그 변형된 형태로서 스트링이 어떤 값을 의미하는 것이 아니라 특징 부집합을 의미하게 된다. 이러한 변형된 대응 방법은 유전 알고리즘의 여러 응용에서 볼 수 있다. 일단 N개의 특징 부집합이 선정되면 이들 특징 부집합을 가지고 각각에 대하여 FCM 군집화 알고리즘을 수행하게 되며, 이 결과로서 각 군집의 중심과 퍼지 분할 행렬을 구할 수 있다. 퍼지 분할 행렬은 군집화된 데이터의 공간적인 구조를 반영하며 이 행렬로부터 군집시 발생된 에러의 개수, 균형 계수 등을 구할 수 있다.

유전 알고리즘의 두 번째 단계는 각 이진 스트링을 의미있는 실제 값으로 변환시키고 이를 미리 설정한 적합도 함수에 대입시켜 적합도를 계산하는 것이다. 이와 유사하게, 사용된 특징 부분집합의 원소의 개수, 즉 특징의 개수와 퍼지 분할 행렬로부터 구한 에러, 균형 계수 등을 반영하는 적합도 함수를 설정하고 각각의 군집화 결과에 대해 적합도를 계산한다.

이러한 방법으로 각 스트링에 대한 특징 부집합 대응, 적합도 계산이 끝나며 구해진 적합도를 기반으로 하여 유전 알고리즘의 세가지 연산인 복제, 교배, 돌연변이를 수행한다. 적합도가 크다는 것은 이진 스트링에 대응되는 특징 부집합이 패턴을 잘 군집화 시킨다는 것을 의미하며, 또한 유전 알고리즘에 있어서 다음세대로 복제될 가능성이 크다는 것을 의미한다.

유전 알고리즘 수행중 적합도를 구하기 위한 적합도 함수는 다음과 같이 정의하였다.

$$fitness = \frac{1}{1 + w_e \cdot error + w_b \cdot balance + w_f \cdot (feature - 1)} \quad (4)$$

식(4)에서 *error*는 분류 에러, *balance*는 군집간의 균형 계수, *feature*는 그 노드에서 사용된 특징의 개수를 의미한다. 또한 w_e, w_b, w_f는 각각의 파라미터에 가중을 주기 위한 가중치(weighting)이다. 만일 한 노드에서 분류 에러와 균형 계수가 '0'이고 사용된 특징의 개수가 최소 값인 '1'이면 적합도는 최대값 '1'을 가지게 된다. 한편 가중치 w_e, w_b, w_f를 조절함에 의해 트리 구성의 결과가 달라질 수 있다. 예로서 w_b에 큰 값을 준다면 분류 에러와 특징 개수를 좀 희생하더라도 군집간의 균형이 좋은 트리 구조를 얻을 수 있는 확률이 크게 된다. 가중치에 대한 특별한 범위를 정량적으로 정하기는 힘들며 본 논문에서는 Ⅳ장에서 보이는 바와 같이 '1'에서 '10'사이의 값으로 실험을 행하였다.

트리 분류기의 군집화 과정중 각 부류들은 어느 한 쪽으로 치우침 없이 같은 수의 부류를 가지고 있는 고른 분포로 유지되는 것이 좋다[14]. 이때 균형 계수를 부류들의 평균수와 생성된 새로운 군집에서의 부류들의 수의 편차로서 다음과 같이 정의하며 균형 계수가 작을수록 부류 군집들 간의 최적의 균형 관계를 유지할 수 있고 전체적으로 패턴 인식을 위한 매칭 횟수가 줄어들게 된다.

$$balance = \sqrt{\frac{\sum_{j=1}^{h}(n_j - \frac{n}{h})^2}{(\frac{n}{h})^2}} \quad (5)$$

여기서 h는 노드의 수, n은 입력 패턴의 수, n_j는 j번째 노드에 속하는 패턴의 수이다. 본 논문에서는 이진 트리를 구성하므로 h는 2가된다.

각 노드에 대한 전체 알고리즘은 다음과 같다.

i) 스트링 집단을 초기화시킨다.

ii) 위에서 기술한 방법으로 각 스트링을 특징의 개수와 종류로 변환한다.

iii) 전체 특징 중에서 선정된 특징을 기초로 FCM 알고리즘을 수행한다.

iv) 클러스터링의 결과로부터 분류 에러, 밸런스를 구하고 식(4)를 이용하여 적합도를 계산한다.

v) 원하는 적합도에 도달한 개체가 존재하면 알고리즘 수행을 끝낸다.

vi) 각 개체에 대한 적합도를 기반으로 하여 유전 알고리즘의 세 가지 연산인 복제, 교배, 돌연변이를 수행한다.

vii) 미리 정한 최대 세대수에 도달하였으면 알고리즘 수행을 끝내고 전체 세대 중 가장 좋은 적합도를 가지는 스트링을 최종 결과로 취한다. 그렇지 않으면 ii)로 간다.

위와 같은 과정을 각 노드에 대해 실행하여 전체 트리 구조가 완성될 때까지 반복한다. 즉, 각 최종 노드에 모든 패턴 클래스가 독립적으로 나타나면 트리 구성 과정은 끝나게 된다.

Ⅳ. 실험 결과

4장에서는 3장에서 제안한 방법으로 이진 결정 트리를 구성하여 두 종류의 패턴에 대하여 실험을 행하고 그 타당성을 보이고자 한다. 두 종류의 패턴은 각각 필기체 영문자 패턴과 타이어 접지면 패턴이다. 유전 알고리즘은 그 특성상 일반적으로 세대수가 증가할수록 적합도가 증가하는 양상을 보인다. 그러나 유전 알고리즘의 특성상 알고리즘을 실행하는 도중에 가장 좋은 적합도를 가지는 개체가 나타날 수 있다. 따라서 본 실험에서는 실행할 세대 수와 개체 수를 각 경우에 대해 일정하게 정해 놓고 유전 알고리즘의 실행중 가장 높은 적합도를 가지는 개체, 즉 이진 문자열을 선정하였다.

1. 제안되는 알고리즘의 영문자 분류에의 적용

1) 실험에 사용된 영문자와 특징 추출

필기체 문자 분류에 사용된 문자의 종류는 영문자중 대문자 A~Z, 26종이며 각 문자당 다섯 개씩 총 130개의 문자를 스캐너로 취득하였다. 나열된 여러 문자들을 x, y축 상에 투영하여 개별 문자들로 분리해 이치 영상으로 만들었다. 그림 1에 필기체 영문자의 예를 보인다.

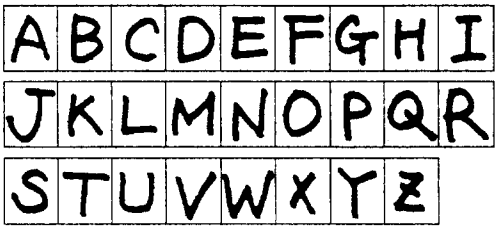

그림 1. 필기체 영문자
Fig. 1. Handwritten alphabetic characters.

문자의 특성을 나타내는 특징은 문자의 모멘트, 방향성분, 화소값 변화 등 여러 가지 방법으로 추출할 수 있으나 본 논문에서는 인식을 위한 특징을 다음과 같이 간단하게 추출하였다. 문자를 둘러싸는 최소의 사각형에 대하여 각 면을 4등분하는 3점으로부터 문자까지의 거리를 구한다. 상, 하, 좌, 우 4면에 대하며 각각 3개씩의 특징을 추출하면 그림 2와 같이 각각의 문자에 대해 12개의 특징이 얻어진다. 그림 2를 보면 이 과정을 쉽게 알 수 있다. 표 2에 각 영문자에 대한 특징의 예를 보인다.

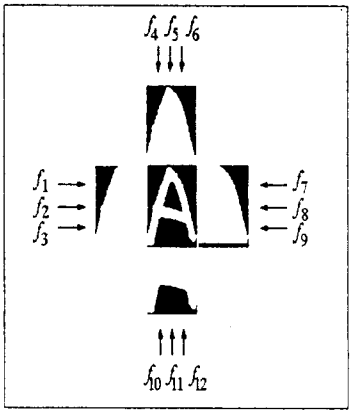

그림 2. 필기체 영문자의 특징 추출
Fig. 2. Feature extraction of handwritten alphabetic character

2) 실험 조건

실험에 사용된 영문자 패턴은 모두 26종류이며 각 종류 마다 5개의 샘플을 가지고 있어 총 130개의 패턴을 사용하였다. 각 노드에서 사용할 수 있는 최대 특징 개수, q_s는 전체 특징의 1/3인 '4'로 하였다. 유전 알고

리즘을 수행하는데 있어서 세대수는 100으로 하였으며, 집단 수와 P_c , P_m은 각각 40, 1.0, 0.1로 하였다. 또한 가중치 w_e, w_b, w_f중 w_e를 '10'으로하고 나머지는 '1'로 가중을 준 경우와, 가중치 모두를 '1'로 준 경우 두 가지로 나누어 실험을 하였다.

표 2. 샘플 영문자 패턴의 특징 벡터
Table 2. Feature vectors for the sample alphabetic character patterns.

Patterns \ Features	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	f_{11}	f_{12}
A	7	3	0	12	1	11	7	2	3	4	7	5
B	0	2	4	1	2	6	6	1	2	9	0	2
C	3	0	3	4	0	1	1	11	5	3	0	2
D	1	2	3	0	0	2	3	0	8	6	1	6
E	1	0	0	1	0	9	8	9	8	0	0	0
F	2	0	0	3	0	0	8	12	16	2	7	19
:							:					
Z	3	6	4	1	0	0	2	8	7	8	0	2

3) 실험 결과

그림 3, 5에 위 두 경우에 대하여 구성한 이진 결정 트리를 나타내었고 표 3, 4에 각 노드에 속한 패턴과 분류 에러, 균형 계수, 선정된 특징을 나타내었다. 예상할 수 있듯이 표 3을 살펴보면 C_{36}, C_{37}, C_{51}, C_{63}등에서와 같이 비슷한 모양을 가지는 문자는 최종 분류 단계에서 분류가 되었으며 분류 에러도 이러한 상황에서 많이 발생됨을 알 수 있다. 특히 첫 번째의 경우처럼 w_e에 가중치를 많이 준 결과 약간의 에러 감소가 있었다.

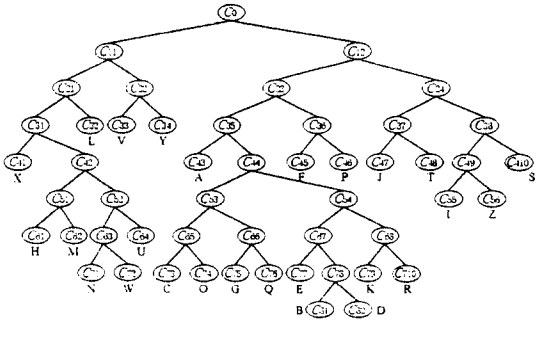

그림 3. w_e=10, w_b=1, w_f=1에 대하여 구성된 이진 결정 트리
Fig. 3. Constructed binary decision in the case of w_e=10, w_b=1, w_f=1.

그림 4에 첫 번째의 경우에 대한 C_0노드에서의 각 세대에 따른 최대 적합도와 평균 적합도를 도시하였다. 최대 적합도는 몇 세대 지나지 않아 최댓값 '0.2262'에 도달하였으며, 그 이후 두 경우 모두 최대 적합도는 변동이 없고 평균 적합도만이 진동을 하며 점차 증가하는 양상을 보여 준다.

그림 6에서는 본 논문에서 제시한 방법을 사용하지 않고 각 노드에서 선정된 특징 부집합이 아닌 전체 특징을 모두 사용하여 구성된 트리 구조를 보여 준다. 분류 에러는 '39'개로 에러율 '0.3'을 나타내고 있으며 그림 3, 5에 비해 매우 좋지 않은 결과를 보여준다.

표 3. 그림 3에 대한 각 노드에서의 패턴, 에러, 균형 계수, 선정된 특징
Table 3. Patterns, error, balance coefficient, extracted feature at each node for Fig. 3.

node	pattern	error	balance	feature
C_0	A Z	0	0.4216	$f_4f_5f_7f_{10}$
C_{11}	H,L N,U-Y	0	0.7857	f_1f_{12}
C_{12}	A-G,I-K,O-T,Z	1(S_5)	0.6156	$f_2f_5f_7$
C_{21}	H,L-N,U,W,X	0	0.1.0102	f_5f_7
C_{22}	V,Y	0	0	f_{11}
C_{23}	A-G,K,O-R	0	0.9428	f_8f_9
C_{24}	I,J,S,T,Z	0	0.2357	f_{12}
C_{31}	H,M,N,U,W,X	0	0.9428	f_2
C_{35}	A-E,G,K,O,Q,R	0	1.1314	f_4
C_{36}	F,P	0	0	f_7
C_{37}	J,T	0	0	f_{10}
C_{38}	I,S,Z	0	0.6010	f_4f_9
C_{42}	H,M,N,U,W	0	0.2828	$f_2f_3f_6f_{11}$
C_{44}	B F,G,K,O,Q,R	0	0.1571	f_1f_5
C_{49}	I,Z	0	0	f_{10}
C_{51}	H,M	0	0	f_6f_{12}
C_{52}	N,U,W	0	0.4714	f_{12}
C_{53}	C,G,O,Q	1(G_1)	0.1414	f_{11}
C_{54}	B,D,E,K,R	0	0.2828	$f_{10}f_{11}$
C_{63}	N,W	2($W_{2,3}$)	0.5657	f_3
C_{65}	C,O	0	0	f_8
C_{66}	G,Q	0	0.1571	$f_1f_6f_7f_{11}$
C_{67}	B,D,E	0	0.4714	$f_6f_{11}f_{12}$
C_{68}	K,R	0	0	f_5
C_{78}	B,D	2($D_{3,4}$)	0.5657	f_9

2. 제안되는 알고리즘의 타이어 접지면 패턴 분류에의 적용

1) 실험에 사용된 타이어 접지면 패턴과 특징 추출

타이어 접지 패턴 분류를 위한 데이타 취득 부분은

타이어 접지면의 영상 취득 과정 및 특징 추출을 위한 전처리 과정으로 구성된다.

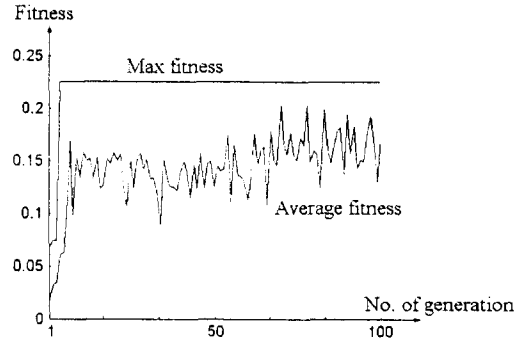

그림 4. 노드 C_0에서 유전 알고리즘 수행중의 최대 적합도와 평균 적합도

Fig. 4. Maximum fitness and average fitness in genetic algorithm at node C_0.

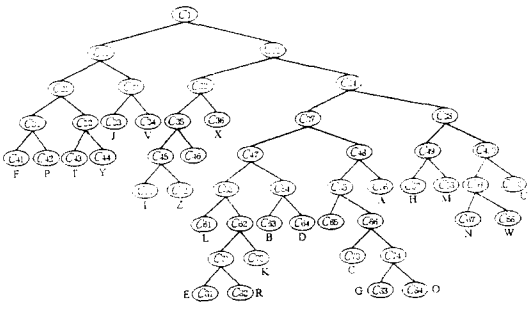

그림 5. w_e=1, w_b=1, w_f=1에 대하여 구성된 이진 결정 트리

Fig. 5. Constructed binary decision in the case of w_e=1, w_b=1, w_f=1.

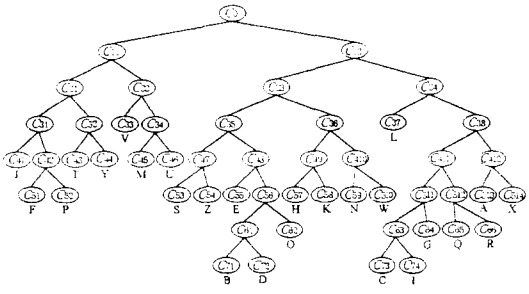

그림 6. 각 노드에서 전체 특징을 모두 사용한 경우 구성된 이진 결정 트리(에러율=0.3)

Fig. 6. Constructed binary decision tree by using all the available features(error rate=0.3).

전처리 과정은 서브 샘플링과 외곽선 검출로 이루어지며, 전자는 영상 취득후 처리해야 할 데이타량을 줄이기 위하여 3×3 마스크로 마스킹을 취한 후 4화소당 한 화소를 취함으로써 행해진다. 한편, 화상의 이치화 과정은 윤곽선 검출을 동시에 행할 수 있는 알고리즘을 사용하였다. 그림 7(a),(b), 그림 8(a),(b)은 각각 샘플 패턴중 두종류의 원 영상과 전처리 후의 영상을 나타낸다.

표 4. 그림 5에 대한 각 노드에서의 패턴, 에러, 균형 계수, 선정된 특징

Table 4. Patterns, error, balance coefficient, extracted feature at each node for Fig. 5.

node	pattern	error	balance	feature
C_0	A~Z	1(G_1)	0.7397	$f_9f_{10}f_{12}$
C_{11}	F,J,P,T,V,Y	0	0.4714	f_2f_{11}
C_{12}	A~E,G~I,K~O,Q~S,U,W,X,Z	1(S_3)	0.8714	f_2f_8
C_{21}	F,P,T,Y	0	0	f_{10}
C_{22}	J,V	0	0	f_2f_{11}
C_{23}	I,S,X,Z	0	0.6699	f_{11}
C_{24}	A~E,G, H,K~O,Q,R,U,W	0	0.5191	$f_2f_7f_8f_{12}$
C_{31}	F,P	0	0	f_2f_{10}
C_{32}	T,Y	0	0	$f_2f_9f_{11}$
C_{33}	I,S,Z	0	0.601	f_4f_9
C_{37}	A~E,G,K,L,O,Q,R	0	0.1571	f_1f_5
C_{38}	H,M,N,U,W	0	0.2828	$f_2f_4f_9f_{11}$
C_{45}	I,Z	0	0	f_{10}
C_{47}	B,D,E,K,L,R	1(K_5)	0.3771	$f_1f_8f_{11}f_{12}$
C_{48}	A,C,G,O,Q	0	0.8250	f_2
C_{49}	H,M	0	0	f_9f_{11}
C_{410}	N,U,W	0	0.4714	f_{12}
C_{53}	E,K,L,R	0	0.6699	f_8
C_{54}	B,D	2($D_{3,4}$)	0.5657	f_9
C_{55}	C,G,O,Q	0	0.6699	f_6
C_{56}	N,W	2($W_{9,3}$)	0.5657	f_3
C_{62}	E,K,R	0	0.6061	f_7
C_{66}	C,G,O	0	0.4041	f_8
C_{71}	E,R	0	0	f_6
C_{73}	G,O	0	0.1571	f_7

위와 같이 전처리된 영상으로부터 알 수 있듯이 전처리된 패턴의 각도 성분은 패턴 분류에 유용한 특징으로 이용될 수 있음을 알 수 있다. 윤곽선의 방향에 대한 정보는 이치화 영상에 대한 계조치 동시 발생 행렬(gray-level cooccurrence matrix)을 통해 표현될 수 있으며 이 행렬은 그 특성상 행과 열의 개수가 계조치의 단계와 같아지므로 이치화 영상에 대해 2×2 행렬이 된다[15]. 특히 행렬 요소중 (2, 2) 요소는 윤

곽선의 양에 대한 정보를 가지고 있으므로 이를 특징으로 사용할 수 있다. 본 실험에서는 타이어에 주로 존재하는 8방향의 각도 성분에 대해 계조치 동시 발생 행렬을 구하고 이로부터 (2, 2) 요소를 추출하여 특징으로 사용하였다. 표 5에 각 패턴 클래스에 속한 샘플 패턴의 특징을 나타내었다.

그림 7. 타이어 접지면 패턴의 원영상
(a) 타이어 접지면 패턴 p_1 (b) 타이어 접지면 패턴 p_6
Fig. 7. Original images of tire tread patterns.
(a) Tire tread pattern p_1 (b) Tire tread pattern p_6

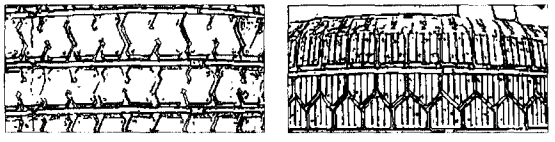

그림 8. 타이어 접지면 패턴의 전처리된 영상
(a) 전처리 후의 패턴 p_1 (b) 전처리 후의 패턴 p_6
Fig. 8. Preprocessed images of the tire tread patterns.
(a) p_1 pattern after preprocessing (b) p_6 pattern after preprocessing

표 5. 샘플 타이어 접지면 패턴의 특징량 벡터
Table 5. Feature vectors for the sample tire tread patterns.

Patterns \ Features	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
p_1	1714	931	1750	1743	111	926	2410	2327
p_2	822	1163	2195	2134	345	1551	2342	2433
p_3	525	1801	4271	3885	983	2237	3927	4616
p_4	1427	836	1599	1697	205	882	2055	2137
p_5	803	1285	2378	2252	288	1282	2344	2262
p_6	1216	895	1907	2046	2680	1218	2267	2415

2) 실험 조건

실험에 사용된 타이어 접지면 패턴의 부류는 모두 6종류이며 각 종류당 10개의 샘플을 가지고 있다. 각 패턴 클래스를 { P_1, P_2, . . . P_6 }로 나타내고, p_k는 P_k에 속하는 패턴이라 하자. q_s는 앞서 문자 인식 실험에서와 같이 '4'로 하였다. 유전 알고리즘을 수행하는데 있어서 세대수는 50, 집단수는 20, P_c , P_m은 각각 1.0 , 0.1 로 하였다. 또한 가중치 w_e, w_b, w_f중 w_b를 '10'으로하고 나머지는 '1'의 가중을 준 경우와, 가중치 모두를 '1'로 준 경우 두 가지로 나누어 실험을 하였다.

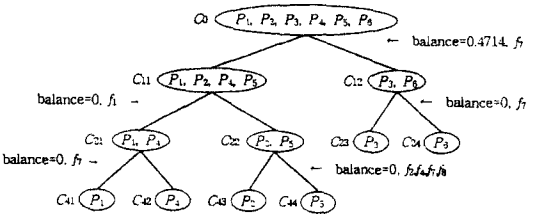

그림 9. w_e=1, w_b=10, w_f=1에 대하여 구성된 이진 결정 트리
Fig. 9. Constructed binary decision in the case of w_e=1, w_b=10, w_f=1.

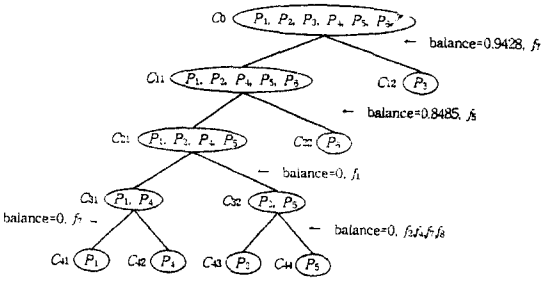

그림 10. w_e=1, w_b=1, w_f=1에 대하여 구성된 이진 결정 트리
Fig. 10. Constructed binary decision in the case of w_e=1, w_b=1, w_f=1.

3) 실험 결과

그림 9, 10에 위 두 경우에 대하여 구성한 이진 결정 트리와 각 노드에 속한 패턴과 분류 에러, 균형 계수, 선정된 특징을 나타내었다. 최대 세대수는 50으로 하였으나 앞에서의 영문자의 경우보다 특징의 개수 자체가 적어서 불과 수세대 이후에 최대 적합도 값을 갖는 특징 부집합을 선정할 수 있었다. 특히 첫 번째의 경우처럼 w_b에 가중치를 많이 준 결과 두 번째의 경우보다 균형이 좋은 트리 구조를 얻을 수 있었다.

V. 결 론

본 논문에서는 유전 알고리즘과 FCM알고리즘을 이용하여 이진 결정 트리를 구성하여 보았다. 트리의 각 노드에서 이진 트리 구조에 적합한 특징 부집합을 구하는 방법에 대해 살펴보았으며 이를 타이어 접지면

패턴과 필기체 영문자 인식에 적용하여 만족할 만한 결과를 얻을 수 있었다. 제안되는 특징 선정 방법의 장점은 유전 알고리즘에 의해 각 노드에서의 분류 에러, 군집간의 균형, 선정된 특징의 개수를 고려한 적합도 함수를 설정하고 이로부터 적절한 분류를 위해 필요한 특징 부집합의 선정을 행할 수 있다는 것이다. 단, 유전 알고리즘 자체의 문제로서 우세한 개체의 영향력에 의해 작은 확률이지만 국소 최소값을 얻게되는 문제점이 있다.

앞으로의 연구 과제는 이진 트리뿐 아니라 데이터의 구조를 잘 반영하는 n진 트리를 위의 알고리즘을 개선하여 구현해 보는 것과, 더 많은 특징을 가지는 패턴에 적용해 보아 그 특성을 확인하여 보는 것이라 하겠다.

참 고 문 헌

[1] S.R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology", *IEEE Trans. Sys., Man, & Cybern.*, vol. 21, pp. 660-674, May/Jun. 1991.

[2] J.K.Mui and K.S. Fu, "Automated classification of nucleated blood cells using a binary tree classifier", *IEEE Trans. Patt. Anal. Mach. Intl.*, vol. 5, pp. 429-443, Sep. 1980.

[3] Y.K.Lin and K.S.Fu, "Automatic classification of cervical cells using a binary tree classifier", *Pattern Recognition*, vol. 16, pp. 69-80, 1983.

[4] H.J.Payne and W.S.Meisel, "An algorithm for constructing optimal binary decision trees", *IEEE Trans. Computers*, vol. 26, pp 905-916, Sep. 1977.

[5] G.H. Landeweerd, T.Timmers, and E.S. Gelsema, "Binary tree versus single level tree classification of white blood cells", *Pattern Recognition*, vol. 16, pp. 571-577, 1983.

[6] D.Dimitrescu, "Hierarchical pattern classification", *Fuzzy Sets and Systems*, vol. 28, pp. 145-162, 1988.

[7] B.B.Devi, "Binary tree design using fuzzy isodata", *Pattern Recognition Letter*, vol. 4 pp. 13-18, Feb. 1986.

[8] 정 순원, 박 귀태, "FCM 알고리즘을 이용한 이진 결정 트리의 구성및 타이어 접지면 패턴 인식에의 적용", 1994년도 한국 퍼지 시스템 학회 추계학술대회, pp. 146-151, 1984

[9] S.W.Jung, S.W.Bae, and G.T.Park, "A design scheme for a hierarchical fuzzy pattern matching classifier and its application to the tire tread pattern recognition", *Fuzzy Sets and Systems*, vol. 65, pp 311-322, 1994.

[10] W.Siedlecki and J.Sklansky, "A note on genetic algorithms for large-scale feature selection", *Pattern Recognition Letters*, vol. 10, pp 335-347, Nov. 1989.

[11] D.E.Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*, Reading, MA:Addison-Wesley, 1989.

[12] M.Srinivas, and L.M.Patnaik, "Adaptive probabilities of crossover and mutation in genetic algorithms", *IEEE Trans. Trans. Sys., Man, & Cybern.*, Vol. 24, No. 4, pp 656-667, Apr 1994.

[13] J.C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithm*. New York, Plenum Press, 1981.

[14] C.Y.Suen and W.R.Wang, "ISOETRP : An interactive clustering algorithm with new object", *Pattern Recognition*, Vol. 7, No. 4, pp 211-219, 1984.

[15] 정 순원, "이치화 영상에 대한 계조치 동시 발생 행렬을 이용한 타이어 접지 패턴의 분류", 석사 학위 논문, 고려 대학교 전기공학과 1992

저 자 소 개

鄭 淳 元(正會員) 第 32卷, B編 第 11號 參照
현재 고려 대학교 대학원 전기공학과 박사과정 재학중

朴 貴 泰(正會員) 第 31卷, B編 第 6號 參照
현재 고려 대학교 대학원 전기공학과 교수, 서울대 ERC-ACI 연구 위원