유전 알고리즘을 이용한 이진 결정 트리의 설계와 영문자 인식에의 응용

A design of binary decision tree using genetic algorithms and its application to the alphabetic character

정 순원, 김 경민, 박 귀태
Soon-Won Jung, Kyoung-Min Kim, Gwi-Tae Park

서울 성북구 안암동 5가 고려대학교 전기공학과
Dept. of Electrical Eng. Korea Univ. 5-1 Anam-dong, Sungbuk-gu, Seoul

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 or feature subset among all the available features is selected based on fitness function in genetic algorithm which is inversely proportional to classification error, balance between cluster, number of feature used. The proposed design scheme is applied to the handwritten alphabetic characters. Experimental results show the usefulness of the proposed scheme.

Ⅰ. 서론

많은 부류(multi-class)의 패턴들을 분류하는데 있어서 결정 트리 분류기(decision tree classifier)는 광범위하게 쓰이는 기법이다.[1] 결정 트리를 구성하는데 있어 고려해야 할 사항은 크게 두 가지로, 하나는 각 노드에 대한 하부(descendant) 노드의 갯수이고 다른 하나는 각 노드에서의 특징량 선정(feature selection) 문제이다. 본 논문에서는 하부 노드의 갯수가 2인 이진 트리를 구성하였다. 한편 특징량의 선정 문제는 패턴 인식에 있어서 매우 중요한 주제 중의 하나이다. 분류에 꼭 필요한 특징량만을 선정하여 분류기에 사용하면 분류 정밀도(classification accuracy)가 높아지고 특징량 추출및 분류에 소요되는 계산 시간이 줄어들게 된다. 그러나 q개의 특징량 중에서 최적의(optimal) 특징량을 추출하는 경우 성능 시험을 위한 특징량의 부분 집합의 갯수는 (2^q-1)개가 되며 특징량의 갯수가 조금만 커져도 실제 적용하기에는 어려움이 많다. 이러한 특징량 선정 문제를 최적화 기법의 하나로써 최근 많은 관심을 모으고 있는 유전 알고리즘을 이용하여 해결하려는 시도가 있었으며 만족할 만한 결과를 보여주고 있다.[2] 그러나 [2]는 근본적으로 본 논문에서 제시하는 이진 결정 트리의 설계에 적용하기에는 부적당하다고 할 수 있다. 3장에서 자세히 다루겠지만 [2]에서 제시한 방법은 전체 특징량 갯수에 대해 선정될 특징량의 갯수를 미리 정하고 최적의 특징량 부집합을 구하며, 분류 에러를 최소화시키는 특징량들을 구하였다 하더라도 일반적으로 트리 구조에서 요구하는 분할된 군집간의 균형 문제를 고려하지 않은 것이므로 이진 결정 트리의 설계에 적용하기에는 부적당하다고 할 수 있다. 본 논문에서는 유전 알고리즘을 이진 결정 트리의 설계에 효과적으로 적용시키기 위하여 유전 알고리즘내의 이진 스트링을 특징량 부집합에 적절히 대응시키는 방법을 제시하고 분류 에러, 군집간의 균형, 선정된 특징량의 갯수등을 포함하는 적합

도 함수를 정의하여 각 노드에서의 특징량 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 영문자 인식에 적용하여 만족할 만한 결과를 얻을 수 있었다.

II. 유전 알고리즘

유전 알고리즘은 1970년대 미국의 John Holland 교수에 의해 정립된 이론으로 자연의 유전학(natural genetics)과 자연 선택(natural selection)의 원리에 근거한 최적해 탐색 방법이다.[3] 일반적인 이진 부호화 기법(binary coding technique)에 의해 생물과 같은 재생산, 교배, 돌연변이를 거쳐 다음 세대의 자손(offspring)을 만들어 내는 과정은 다음과 같다.

i) 부호화및 초기화(coding and initialization)
ii) 적합도 평가(fitness evaluation)
iii) 복제
iv) 교배
v) 돌연변이

III. FCM 알고리즘및 유전 알고리즘을 이용한 이진 결정 트리의 설계

3.1 FCM 알고리즘

FCM 알고리즘은 주어진 데이타 집합, $X=\{ \vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n \}$($\vec{x}_k \in R^q$, q is the number of features)에 대하여 정의된 어떤 목적 함수가 근사적 최소 값에 도달되도록 퍼지 분할 행렬(fuzzy partition matrix) U와 군집의 중심값 $V=\{ \vec{v}_1, \vec{v}_2, \ldots, \vec{v}_c \}$를 반복 계산법에 의해 구하는 최적화 퍼지 군집화 알고리즘이다.[4] FCM에서 사용되는 목적 함수는 다음과 같다.

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

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

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

3.2 유전 알고리즘을 이용한 이진 결정 트리의 설계

본 논문에서는 이진 결정 트리를 설계하는데 있어서 각 노드에서 필요한 특징량을 구하기 위하여 최적화 기법의 하나인 유전 알고리즘을 이용하였다. 전체 특징량 개수를 q라하고 각 노드에서 사용할 수 있는 최대 특징량 갯수를 q_s라 하자. 유전 알고리즘에서 사용되는 이진 스트링을 k-bit라 하면 이 스트링이 나타낼 수 있는 가짓수는 모두 2^k개가 된다. 만일 [2]에서처럼 $k=q$라하고 각 bit를 각 특징량에 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. 두 번째는 위와 같은 문제를 bit보정에 의하여 해결한다고 하더라도 유전 알고리즘을 통해 선정되는 특징량의 갯수간에 불평등이 생기게 된다. 확률적으로 l개의 원소를 가지는 특징량 부집합이 선정될 확률, P_l은 다음과 같이 주어진다.

$$P_l = \frac{{}_qC_l}{\sum_{i=1}^{q_s} {}_qC_i} \qquad (2)$$

따라서 위 두 문제를 해결하기 위해 다음과 같이하여 특징량 갯수 선정의 평형을 유지하도록 한다. 먼저 k-bit가 나타낼 수 있는 총 가짓수 2^k를 q_s등분하여 각 구간을 각 특징량 갯수에 대응하도록 하였고, 각 구간을 그 구간에 속하는 특징량 부집합의 종류로 다시 나누어 사용되는 특징량에 대응하게 하였다. 표 1에 k=15, q=8, q_s=2일 경우 그 과정을 나타내었다.

표 1 이진 스트링과 특징량 부집합의 대응 과정

Table 1 Correspondence between binary string and feature subset

특징량 갯수	구 간	부분 구간	특징량 부집합	부집합 종류
1	0 ~ 16383	~ 2047	f_1	$_8C_1$=8
		~ 4095	f_2	
		⋮	⋮	
		~ 16383	f_8	
2	~ 32767	~ 16969	f_1, f_2	$_8C_2$=28
		~ 17555	f_1, f_3	
		⋮	⋮	
		~ 32767	f_7, f_8	

위와 같은 과정을 거쳐 이진 스트링에 대응되는 특징량 부집합을 선택한 후 이들 특징량을 이용하여 FCM 군집화 알고리즘을 수행하게 된다. FCM 군집화의 결과로서 나오는 나오는 퍼지 분할로부터 분류 에러의 갯수, 균형 계수 등을 구할 수 있다. 유전 알고리즘 수행중 적합도(fitness)를 구하기 위한 적합도 함수(fitness function)는 다음과 같이 정의하였다.

$$fitness = \frac{1}{1 + w_e \cdot e + w_b \cdot b + w_f \cdot (f-1)} \tag{3}$$

식(3)에서 e는 분류 에러, b는 군집간의 균형 계수, f는 그 노드에서 사용된 특징량의 갯수를 의미한다. 또한 w_e, w_b, w_f는 각각의 파라미터에 가중을 주기 위한 가중치(weighting)이다. 한편 균형 계수는 부류들의 평균수와 생성된 새로운 군집에서의 부류들의 수의 편차로서 다음과 같이 정의한다.[5]

$$Balance = \sqrt{\frac{\sum_{j=1}^{h}(n_j - \frac{n}{h})^2}{(\frac{n}{h})^2}} \tag{4}$$

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

각 노드에 대한 전체 알고리즘은 다음과 같다.

i) 스트링 집단(population)을 초기화

ii) 스트링을 특징량의 갯수와 종류로 변환

iii) 선정된 특징량으로 FCM 알고리즘을 수행

iv) FM의 결과인 퍼지 분할 행렬로부터 에러, 밸런스를 구하고 적합도를 계산

v) 원하는 적합도에 도달한 개체(individual)가 존재하면 알고리즘 수행을 끝낸다.

vi) 적합도를 기반으로 하여 유전 알고리즘의 복제, 교배, 돌연변이를 수행한다.

vii) 최대 세대수에 도달 하였으면 전체 세대 중 가장 좋은 적합도를 가지는 스트링을 최종 결과로 취하고 알고리즘 수행 종료, 그렇지 않으면 ii)로 간다.

위와 같은 과정을 각 노드에 대해 실행하여 전체 트리 구조가 완성될 때까지 반복한다.

IV. 실험 결과

4장에서는 3장에서 제안한 방법으로 이진 결정 트리를 구성하여 필기체 영문자 패턴에 대하여 실험을 행하고 그 타당성을 보이고자 한다.

4.1. 제안 알고리즘의 영문자 분류에의 적용

4.1.1 실험에 사용된 영문자와 특징량 추출

필기체 문자 분류에 사용된 문자의 종류는 영문자중 대문자 A~Z, 26종이며 각 문자당 다섯개씩 총 130개의 문자를 취득하였다.

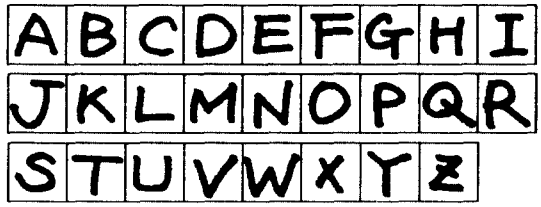

그림 1 필기체 영문자

Fig.1 Handwritten alphabetic characters

한편 문자의 특성을 나타내는 특징량 12개를 그림 2와 같이 각 면의 세 점으로부터 문자까지의 거리로 추출하였다.

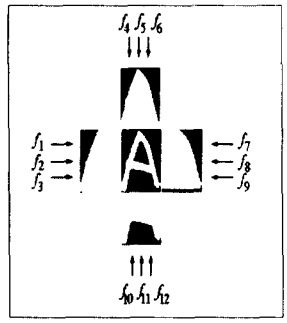

그림 2 필기체 영문자의 특징량 추출

Fig. 2 Feature extraction of handwritten alphabetic character

4.1.2 실험 조건

각 노드에서 사용할 수 있는 최대 특징량 갯수, q_s는 전체 특징량의 1/3인 '4'로 하였다. 세대수(number of generation)는 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 주었다.

4.1.3 실험 결과

그림 3에 구성된 이진 결정 트리를 나타내었고 표 2에 각 노드에 속한 패턴과 분류 에러, 균형계수, 선정된 특징량을 나타내었다. 예상할 수 있듯이 표 2를 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. 그림 4에 C_0노드에서의 각 세대에 따른 최대 적합도(maximum fitness)와 평균 적합도(average fitness)를 도시하였다.

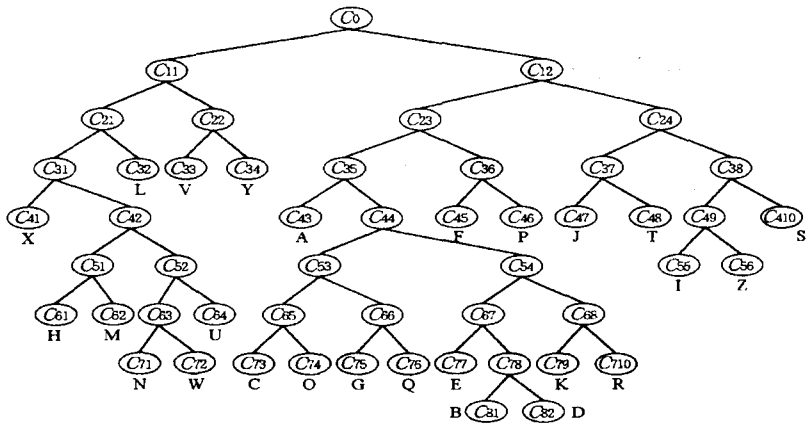

그림 3 구성된 이진 결정 트리
Fig. 3 Constructed binary decision

표 2 그림 3에 대한 각 노드에서의 패턴, 에러, 균형 계수, 선정된 특징량
Table 2 Patterns, error, balance coefficient, extracted feature at each node for Fig. 3

node	pattern	error	balance	feature	node	pattern	error	balance	feature
C_0	A-Z	0	0.4216	f_4,f_5,f_7,f_{10}	C_{44}	B-E,G,K,O,Q,R	0	0.1571	f_1,f_3
C_{11}	H,L-N,U-Y	0	0.7857	f_3,f_{12}	C_{49}	I,Z	0	0	f_{10}
C_{12}	A-G,I-K,O-T,Z	1(S_5)	0.6156	f_2,f_5,f_7	C_{51}	H,M	0	0	f_6,f_{12}
C_{21}	H,L-N,U,W,X	0	0.1.0102	f_5,f_7	C_{52}	N,U,W	0	0.4714	f_{12}
C_{22}	V,Y	0	0	f_{11}	C_{53}	C,G,O,Q	1(G_1)	0.1414	f_{11}
C_{23}	A-G,K,O-R	0	0.9428	f_8,f_9	C_{54}	B,D,E,K,R	0	0.2828	f_{10},f_{11}
C_{24}	I,J,S,T,Z	0	0.2357	f_{12}	C_{63}	N,W	2($W_{2,3}$)	0.5657	f_3
C_{31}	H,M,N,U,W,X	0	0.9428	f_2	C_{65}	C,O	0	0	f_8
C_{35}	A-E,G,K,O,Q,R	0	1.1314	f_4	C_{66}	G,Q	0	0.1571	f_4,f_6,f_7,f_{11}
C_{36}	F,P	0	0	f_7	C_{67}	B,D,E	0	0.4714	f_6,f_{11},f_{12}
C_{37}	J,T	0	0	f_{10}	C_{68}	K,R	0	0	f_5
C_{38}	I,S,Z	0	0.6010	f_4,f_9	C_{78}	B,D	2($D_{3,4}$)	0.5657	f_9
C_{42}	H,M,N,U,W	0	0.2828	f_2,f_4,f_6,f_{11}					

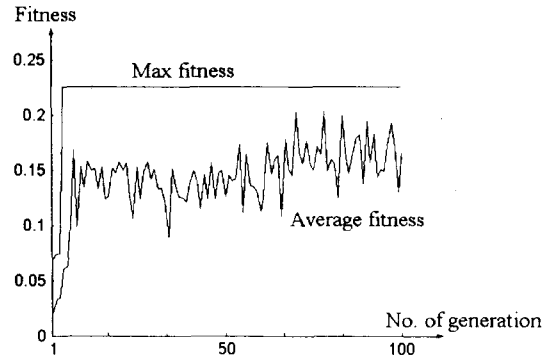

그림 4 노드 C_0에서의 최대 적합도와 평균 적합도
Fig. 4 Maximum and average fitness at node C_0

V. 결론

본 논문에서는 유전 알고리즘과 FCM알고리즘을 이용하여 이진 결정 트리를 구성하여 보았다. 트리의 각 노드에서 이진 트리 구조에 적합한 최적 혹은 최적 근처의 특징량 부집합을 구하는 방법에 대해 살펴보았으며 이를 필기체 영문자 인식에 적용하여 만족할 만한 결과를 얻을 수 있었다. 제안되는 특징량 선정 방법의 장점은 유전 알고리즘에 의해 각 노드에서의 분류 에러, 군집간의 균형, 선정된 특징량의 갯수를 고려한 적합도 함수를 설정하고 이로부터 적절한 분류를 위해 필요한 특징량 부집합의 선정을 행할 수 있다는 것이다.

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

참고 문헌

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

[2] W.Siedlecki and J.Sklansky, "A note on genetic algorithms for large-scale feature selection", *Pattern Recognition Letters*, vol. 10, pp335-347, 1989.

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

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

[5] C.Y.Suen and W.R.Wang, "ISOETRP : An interactive clustering algorithm with new object", Pattern Recognition, Vol.7, No.4, pp211-219, 1984.