유전 알고리즘을 이용한 선형 결정 함수의 결정 및 이진 결정 트리 구성에의 적용

정 순원, 박 귀태
서울 성북구 안암동 5가 고려대학교 전기공학과

A determination of linear decision function using GA and its application to the construction of binary decision tree

Soon-Won Jung, Gwi-Tae Park
Dept. of Electrical Eng. Korea Univ. 5-1 Anam-dong, Sungbuk-gu, Seoul

Abstract

In this paper a new determination scheme of linear decision function is proposed. In this scheme, the weights in linear decision function is obtained by genetic algorithm.

The result considering balance between clusters as well as classification error can be obtained by properly selecting the fitness function of genetic algorithm in determination of linear decision function and this has the merit in applying this scheme to the construction of binary decision tree.

The proposed scheme is applied to the artificial two dimensional data and real multi dimensional data. Experimental results show the usefulness of the proposed scheme.

Ⅰ. 서론

패턴 인식 시스템에 있어서 가장 중요한 기능은 분류 에러 없이 주어진 패턴이 어느 부류(class)에 속하는가를 결정하는 것이다. 이러한 기능을 행하기 위한 일반적인 접근 방법 중의 하나는 결정 함수(decision function)의 개념을 사용하는 것이다. 실제로 패턴 인식에 쓰이는 많은 방법들이 결정 함수 혹은 결정 경계(decision boundary)를 결정하는 것과 관련되어 있다.[1]

그들중 통계적 패턴 인식 방법은 기존의 확률 개념을 도입한 방법으로서 개념 자체가 통계치에 바탕을 두고 있으므로, 인식 절차의 복잡성, 베이시안 통계에서의 계산의 복잡성을 수반하며 특히 인식에 결정적인 영향을 미치는 특징의 실제적인 통계 특성을 구하기 힘들다는 문제 점이 있다. 한편 반복적인 학습 알고리즘에 의해 훈련 패턴으로부터 결정 경계가 자동적으로 생성되는 신경회로망 분류기의 경우 그 근본 원리가 기울기 강하(gradient decent) 방법에 기초하고 있어 국부 극점에 빠질 위험성이 존재한다.[2]

본 논문에서는 이러한 문제를 해결하기 위하여 최적화 기법의 하나로써 최근 많은 관심을 모으고 있는 유전 알고리즘을 이용하여 선형 결정 함수를 구하는 방법을 제안한다. 먼저 선형 결정 함수의 가중치들을 유전 알고리즘의 이진 스트링에 직접 대응시켜 구하는 방법의 문제점들을 살펴 보고 이의 해결을 위해 새로운 방법을 제안하였다. 또한 유전 알고리즘 내의 적합도 함수를 이진 결정 트리 구성에 적합하도록 적절히 설정하여 분류 에러뿐 아니라 분할된 군집의 균형까지도 고려한 결과를 얻을 수 있음을 보여준다. 마지막으로, 제안되는 방법을 임의로 만들어 낸 2차원 데이터뿐만 아니라 표준 데이터로 많이 사용되는 Iris 데이터에 적용하여 그 유용성을 보였다.

Ⅱ. 유전 알고리즘을 이용한 선형 결정 함수의 결정

2.1 유전 알고리즘

유전 알고리즘은 1970년대 미국의 John Holland 교수에 의해 정립된 이론으로 자연의 유전학과 자연 선택의 원리에 근거한 최적해 탐색 방법이다.[3] 기존의 최적해 탐색이 국부 탐색을 하는데 반해 유전 알고리즘은 여러 해를 동시에 탐색하는 전역 탐색을 함으로서 전역적인 최적 해를 찾을 확률이 기존의 최적화 탐색에 비해 큰 것이 특징이다. 일반적인 유전 알고리즘은 이진 부호화 기법에 의해 생물과 같이 복제(reproduction), 교배(crossover), 돌연변이(mutation)를 거쳐 다음 세대의 자손을 만들어 낸다.

2.2 선형 결정 함수

가상의 두 패턴 부류를 나타내는 그림 1을 고려해 보자. 이 그림을 보면 두 개의 패턴 집단을 하나의 직선으로 분리할 수 있음을 알 수 있다.

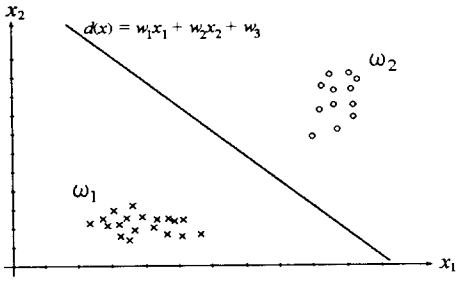

그림 1 선형 결정 함수

그림 1에서 $d(\mathbf{x}) = w_1x_1 + w_2x_2 + w_3 = 0$을 분리선의 방정식이라고 하자. 만일, 패턴 부류 ω_1에 속하는 임의의 패턴 $\mathbf{x}$를 $d(\mathbf{x})$에 대입해 보면 양의 값이 나옴을 알 수 있으며, ω_2에 속하는 임의의 패턴을 $d(\mathbf{x})$에 대입해 보면 음의 값이 되므로 $d(\mathbf{x})$를 결정 함수로 사용할 수 있다. 이러한 2차 선형 결정 함수를 n차의 경우로 쉽게 확장할 수 있으며 그 형태는 다음과 같다.

$$d(\mathbf{x}) = w_1x_1 + w_2x_2 + \cdots + w_nx_n + w_{n+1} = \mathbf{w}_0\ \mathbf{x} \tag{1}$$

여기서 $\mathbf{x} = (x_1, x_2, \cdots, x_n, 1)'$과 $\mathbf{w} = (w_1, w_2, \cdots, w_n, w_{n+1})$을 각각 확장된 패턴 벡터(augmented pattern vector)와 확장된 가중치 벡터(augmented weight vector)라고 한다. 혹은 간략하게 패턴 벡터와 가중치 벡터라고 한다.

2.3 유전 알고리즘을 이용한 선형결정함수의 결정

유전 알고리즘을 이용하여 선형 결정 함수를 결정하는데 있어서 먼저 생각해 볼 수 있는 방법은 선형 결정 함수의 가중치 $w_1, w_2, \ldots, w_{n+1}$을 유전 알고리즘의 이진 스트링에 직접 대응시켜 구하는 방법이라 할 수 있다. n차원의 경우 가중치 각각에 대응되는 $(n+1)$개의 세그먼트를 가지는 이진 스트링을 생각해 볼 수 있다. 기지의 데이터(labelled data)가 주어졌을 때 결정 함수는 분류 에러를 최소로 하도록 결정되어야 하므로 유전 알고리즘 내의 적합도 함수를 다음과 같은 형태로 설정할 수 있다.

$$fitness = \frac{1}{1 + error} \tag{2}$$

즉, 각 개체에 대한 이진 스트링을 복호화(decoding)하여 $(n+1)$개의 가중치를 구하고 이로부터 결정 함수를 구성한 후, 샘플 데이터를 구해진 결정 함수에 대입하여 분류 에러를 구한다. 식(2)에 의해 각 개체에 대한 적합도를 계산하면 분류 에러가 작을수록 적합도 값은 커지며 이는 또한 이에 대응되는 이진 스트링이 다음 세대로 복제될 가능성이 크다는 것을 의미한다.

그러나 이와 같은 방법으로 결정 함수의 가중치를 이진 스트링에 직접 대응시킬 경우 크게 두 가지 문제점들이 발생된다. 첫째, 실제 데이터를 나눌 수 없는 결정 경계가 많이 존재하게 되어 전체 집단중 필요 없는 개체의 수가 많이 존재하게 된다. 둘째, 각 축에 대한 기울기와 절편 값이 골고루 분포되지 못하고 한 값에 집중되는 현상이 발생한다.

이러한 문제점들을 해결하기 위하여 다음과 같이 새로운 방법을 제안한다. 먼저, 주어진 데이터 집합 $P = \{p_1, p_2\ \ldots\ p_n\}$을 생각해 보자. 단, $p_k \in R^q$이며 q는 특징의 개수이다. 식(7)과 같이 각 특징의 값들을 모두 포함하는 구간의 하한, 상한인 l과 r을 정의하자.

$$l(j) = \min_i\ p_{ij}$$
$$r(j) = \max_i\ p_{ij} \tag{3}$$

이차원의 경우 j는 $j=1, 2$의 값을 가지며 l과 r을 기초로 전체 데이터를 포함하는 최소의 사각형을 구할 수 있다. 이 사각형 범위 내에서 임의로 두점을 선택하면 이 점들을 동시에 지나는 한 직선을 구할 수 있으며 이 직선의 기울기와 절편으로부터 w_1, w_2, w_3를 구할 수 있게 된다. 이를 n차원 공간으로 확장하면 임의로 n개의 점을 선택하여 이들을 지나는 초 평면(hyperplane)을 구할 수 있으며 이로부터 $w_1, w_2, \ldots, w_{n+1}$을 구할 수 있다. 이 개념을 유전 알고리즘의 이진 스트링에 대응시키면 이진 스트링내의 각 세그먼트는 n차원 공간상의 각 축의 한 위치를 가리키며 n개의 세그먼트는 한 점을 가리키게 된다. 즉, n차원의 경우 n개의 점을 선택해야 하므로, 필요한 세그먼트의 개수는 n^2이 되며, 각 세그먼트의 길이를 m-bit라 하면 이진 스트링의 총길이는 n^2m-bit이다.

유전 알고리즘을 이용하여 결정 함수의 가중치를 구하는 전체 알고리즘은 다음과 같다.

i) 스트링 집단을 초기화시킨다.
ii) 각 스트링을 임의의 좌표 점으로 변환한다.
iii) 위에서 기술한 방법으로 좌표 점들이 이루는 초평면에 대응되는 가중치를 구한다.
iv) 결정된 결정 함수로부터 분류 에러를 구하고 식(2)를 이용하여 적합도를 계산한다.
v) 원하는 적합도에 도달한 개체가 존재하면 알고리즘 수행을 끝낸다.
vi) 각 개체에 대한 적합도를 기반으로 하여 유전 알고리즘의 세 가지 연산을 수행한다.
vii) 최대 세대수에 도달하였으면 알고리즘을 끝내고 전체 세대중 가장 좋은 적합도를 가지는 스트링을 최종 결과로 취한다. 그렇지 않으면 ii)로 간다.

2.4 이진 결정 트리 구성에의 응용

다 부류에 대해 이진 결정 트리를 구성할 경우 한 노드에서 나누어진 두 군집(cluster)들은 어느 한쪽으로 치우침 없이 비슷한 수의 부류를 가지고 있는 고른 분포를 유지하는 것이 좋다.[4,5] 이러한 개념을 바탕으로 식(4)

과 같은 적합도 함수를 고려해 보자.

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

식(4)에서 *error*는 분류 에러, *balance*는 군집간의 균형 계수를 의미한다. 또한 w_e, w_b는 각각의 파라미터에 가중을 주기 위한 가중치(weights)이다. 가중치 w_e, w_b를 조절함에 의해 트리 구성의 결과가 달라질 수 있다. 예로서 w_b에 큰 값을 준다면 분류 에러는 발생하더라도 군집간의 균형이 좋은 트리 구조를 얻을 수 있는 확률이 크게 된다.

이때 균형 계수를 다음과 같이 정의하며 균형 계수가 작을 수록 부류 군집들 간의 최적의 균형 관계를 유지할 수 있고 전체적으로 패턴 인식을 위한 매칭 횟수가 줄어들게 된다.[5]

$$balance = \sqrt{\frac{\sum_{j=1}^{2}(n_j - \frac{n}{2})^2}{(\frac{n}{2})^2}} \quad (12)$$

여기서 n은 입력 패턴의 수, n_j는 j번째 노드에 속하는 패턴의 수이다.

각 노드에 대한 전체 알고리즘은 앞서 살펴본 결정함수를 구하는 과정 중 iv만 다음과 같이 바꾸면 된다.

iv) 결정된 결정 함수로부터 분류 에러, 밸런스를 구하고 식(4)를 이용하여 적합도를 계산한다.

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

III. 실험 결과

3.1 제안되는 알고리즘의 2차원 데이터에의 적용

임의로 만든 세 가지 2차원 데이터중 첫 번째 데이터와 식(4)의 적합도 함수를 유전 알고리즘에 적용하여 구한 결정 경계, $d(\mathbf{x})=0$를 그림 4에 나타내었다.

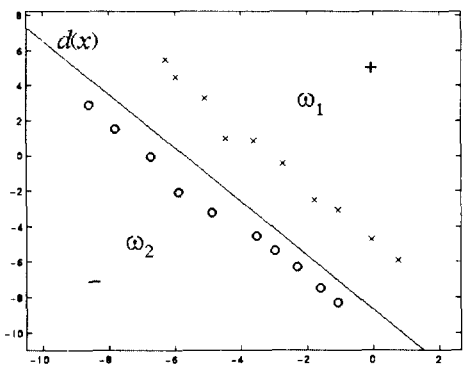

그림 2 샘플 데이터 1에 대한 결정 경계

그림에서 알 수 있듯이 ω_1에 속하는 임의의 패턴들을 $d(\mathbf{x})$에 대입하였을 때 양의 값을 가지며, ω_2에 속하는 패턴에 대해서는 음의 값을 가진다. 식(5)에 유전 알고리즘을 통해 구한 결정 함수를 나타내었다.

$$d(\mathbf{x}) = 0.1752x_1 + 0.1152x_2 + 1 \quad (5)$$

그림 3에 나타내어진 데이터는 두 부류가 아닌 세 부류의 데이터이며 각각 '10'개, '6'개, '4'개의 패턴을 포함하고 있다. $d_i(\mathbf{x})$를 ω_i와 다른 패턴 부류들을 분리시키는 결정 함수라 하자. 만일 적합도 함수를 식(4)와 같이 설정하면 $d_1(\mathbf{x})$, $d_2(\mathbf{x})$, $d_3(\mathbf{x})$가 모두 적절한 결정 함수로서 적합도는 최대 값인 '1'이 된다. 또한 직관적으로 알 수 있듯이 $d_1(\mathbf{x})$보다는 $d_2(\mathbf{x})$ 혹은 $d_3(\mathbf{x})$가 결정 함수로 선택될 확률이 크다는 것을 알 수 있다. 그러나 식(4)를 적합도 함수로 설정하면 분리된 두 군집의 균형까지도 고려하게 되므로 $d_1(\mathbf{x})$가 결정 함수로 선택될 수 있도록 유전 알고리즘은 수렴한다.

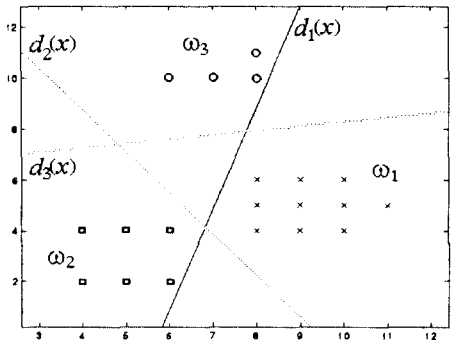

그림 3 샘플 데이터 2에 대한 결정 경계

식(6)에 식(4)를 적합도 함수로 설정하여 구한 결정 함수, $d_1(\mathbf{x})$를 나타내었다. w_e와 w_b는 모두 '1'로 설정하고 실험을 하였다.

$$d_1(\mathbf{x}) = -0.1731x_1 + 0.0437x_2 + 1 \quad (6)$$

마지막으로 간단한 이진 결정 트리를 구성하기 위한 샘플 데이터의 예를 그림 4에 나타내었다.

루트 노드에서의 결정 함수, $d_0(\mathbf{x})$는 그림에서 알 수 있듯이 두 종류가 존재할 수 있으며 유전 알고리즘의 수행 결과 그림과 같이 구해졌다. 일단 전체 데이터가 두 부분으로 분리된 후 계속하여 분리해 나가면 ω_1과 ω_2를 나누는 결정 함수는 $d_{11}(\mathbf{x})$, 마지막으로 ω_3과 ω_4에 대해서는 $d_{12}(\mathbf{x})$가 구해졌다. 식(7)에 이들 세 결정 함수를 나타내었다. 이 경우에도 w_e와 w_b는 모두 '1'로 설정하고 실험을 하였다.

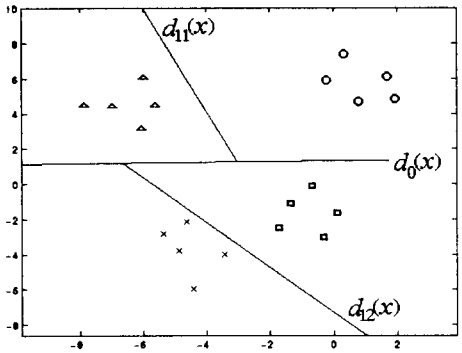

그림 4 샘플 데이터 3에 대한 결정 경계

$$d_0(\mathbf{x}) = 0.0146x_1 - 0.7577x_2 + 1$$
$$d_{11}(\mathbf{x}) = 0.1747x_1 + 0.1372x_2 + 1$$
$$d_{12}(\mathbf{x}) = 0.3838x_1 + 0.1311x_2 + 1 \qquad (7)$$

3.2 제안되는 알고리즘의 Iris 데이터에의 적용

실제 데이터는 표준 데이터로 많이 사용하는 세 종류의 붓꽃(Iris)으로서 각각 'setosa', 'versicolor', 그리고 'virginica'이다. 각 집단마다 50개의 샘플이 있으며 4가지의 특징을 가진다. 특징들은 꽃받침(sepal)과 꽃잎(petal)의 길이와 폭이다. 그림 7에 꽃잎의 길이와 폭을 두 축으로 하여 Iris 데이터를 2차원 평면상에 표시하였다. 여기에서 'x', 'o', '*'는 각각 setosa, versicolor, virginica를 나타낸다.

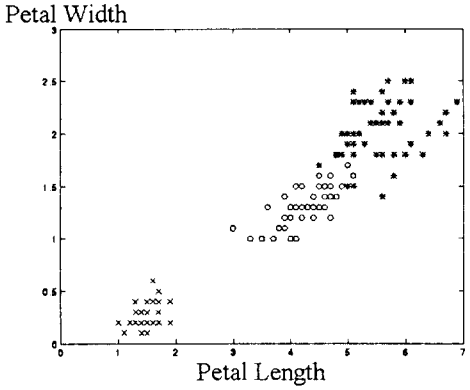

그림 5 Petal length와 width에 대한 Iris 데이터의 도시

그림 6에 Iris데이터들에 대해 구성한 이진 결정 트리를 나타내었다. 분류 에러는 C_0에서 '0', 그리고 C_{12}에서 '2'가 얻어졌다. Iris 데이터의 경우 I_2, I_3는 선형 분리가 되지 않는 것으로 알려져있다. 식(8)에 각 노드에서의 결정 함수를 나타내었다. 여기서 d_{ij}는 군집 C_{ij}에서의 결정 함수를 나타낸다.

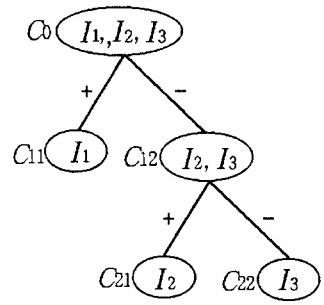

그림 6 Iris 데이터에 대해 구성한 이진 결정 트리

$$d_0(\mathbf{x}) = -0.25x_1 + 0.21x_2 - 0.09x_3 - 0.04x_4 + 1.0$$
$$d_{12}(\mathbf{x}) = -0.11x_1 - 0.10x_2 + 0.11x_3 - 0.17x_4 + 1.0 \qquad (8)$$

IV. 결론

본 논문에서는 최적화 기법의 하나인 유전 알고리즘을 이용하여 선형 결정 함수를 결정해 보았다. 제안된 방법을 이진 결정 트리에 응용하는 방법을 살펴보았으며 이를 여러 패턴에 적용하여 좋은 결과를 얻을 수 있었다. 제안되는 방법의 장점은 유전 알고리즘내의 적합도 함수를 적절히 설정함에 따라 분류 에러뿐 아니라 분할된 군집의 균형까지도 고려한 결과를 얻을 수 있다는 것이다.

제안되는 방법의 문제점으로서 특징의 차수가 n일 경우 한 개체의 세그먼트 개수가 n^2이 되어 특징의 개수가 늘어나면 유전 알고리즘의 계산 시간이 많이 늘어나는 점을 들 수 있다.

앞으로의 연구 과제로는 위의 문제점을 효과적으로 해결하기 위한 방안의 연구와 특징의 선정 문제까지도 함께 고려한 이진 결정 트리 구성에 관한 연구 등을 들 수 있다.

참고 문헌

[1] J.T.Tou and R.C.Gonzalez, *Pattern Recognition Principles*, Addison-Wesley, 1974.

[2] Y.H.Pao, *Adaptive Pattern Recognition and Neural Networks*, Addison-Wesley, 1989

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

[4] 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.

[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.