유전 알고리즘을 이용한 적응적 트리맵 설계

홍종선, 김대일, 장혜경, 김영호, 강대성
동아대학교 전자공학과

An Adaptive Tree Map Scheme using Genetic Algorithm

Jong-Sun Hong, Daeil Kim, Hye-Kyoung Jang, Young Ho Kim, Dae-Seong Kang
Department of Electronic Eng., Dong-A Univ.
E-mail: eva2012@empal.com

요약

본 논문에서는, 패턴 인식시 데이타의 최적의 특성을 구성할 수 있는 새로운 신경망 구조인 적응적 트리맵을 제안한다. 유전 알고리즘을 사용한 적응적 트리맵(adaptive tree map : ATM)은 데이터의 특징에 대한 중요도를 유전 알고리즘으로 구성하고, 특징의 우선순위에 따라 트리 구조를 도입하고 데이터의 유사성에 따라 신경망의 뉴런이 분리, 병합 될 수 있다. 패턴인식의 인식률에 영향을 미치는 인자 중에서 가장 중요한 특징은 연구자의 선택에 의하여 사용되거나 무시될 수 있으며, 반복적인 실험을 통하여 적절한 특징을 사용할 수 있으나 최적의 특징은 될 수 없다. 그러나 본 논문에서 제안한 ATM을 이용하면 블랙박스로 구성된 적응적인 시스템을 이용하여 원하는 출력을 얻을 수 있게 된다.

Ⅰ. 서론

일반적으로 음성, 문자, 숫자, 지문 등의 패턴 인식은 신경망을 사용하며, 입력된 데이터에 대한 고유한 특징을 분류하는 방식으로 인식한다. 기존의 신경망은 추출된 특징을 우선순의나 중요도에 관계없이 뉴런의 구성요소로 사용하거나 실험을 통하여 얻어진 가중치를 사용하며, 얻어진 특징은 모두가 동일하게 훈련과 인식에 사용한다. 따라서, 입력 데이터에서 n개의 특징을 추출한다면, n개에 대응하는 신경망의 뉴런 구조가 갖추어 진다. 본 논문에서는 이러한 특징을 효율적으로 이용하기 위해 이들의 중요도를 유전 알고리즘으로 구한다. 유전 알고리즘은 어떠한 문제에 대해 전역탐색 기법을 사용하여 최적의 해를 구하는 방식으로, 기존의 가중치를 사용하지 않거나 실험에 의해서 정해졌던 최적화 되지 않은 기존의 신경망 구조를 개선할 수 있다. 기존의 신경망에서는 n개의 특징 모두를 사용하여, 몇 개의 특징만으로 구분될 수 있는 데이타의 특성을 무시하게 된다. 예를 들어 1은 n개의 특징에서 첫 번째와 두 번째 특징만으로 분류가 가능하고, 8은 세 번째 특징까지 이용하여 분류가 가능하며, 나머지 숫자들은 그 이상 혹은 그 이하의 특징을 이용하여 분류가 가능할 경우, 기존의 신경망 구조에서는 이러한 특성이 무시된다. 또한 최소 M개의 뉴런들로 분류가 가능한 경우 M보다 큰 N×N의 신경망이 구성되면 공간 낭비가 심각할지라도 구조적인 특성으로 사용할 수밖에 없게 된다. 이는 N×N의 크기가 증가될수록 공간 낭비가 심각해질 것이다. N×N구조의 또 다른 심각한 문제는 구조가 고정되어 있기 때문에, 인식률이 낮은 경우 N의 수치를 증가시켜 재훈련이 요구되므로 인식률이 만족되기까지 시간 낭비가 많다는 점이다. ATM을 사용하면 오인식이 일어나는 단계에서 동적으로 뉴런의 개수를 증가시키며, 뉴런 사이의 벡터의 유사성으로 일정 문턱치(threshold)를 만족하면 병합할 수 있어 적응성이 우수하다. 따라서 ATM을 이용한다면 어떠한 데이터를 인식하는 경우 블랙박스로 구성된 적응적인 시스템을 이용하여 원하는 출력을 얻을 수 있게 된다.

Ⅱ. 본론

1. 유전 알고리즘

본 장에서 제안하는 유전 알고리즘[1][2]은 자연 생태계에서 우성 인자가 열성 인자보다 생존확률이 높다는 기본 규칙을 이용한 컴퓨팅 알고리즘으로, 공학 문제에서는 주어진 문제에 대해 해로 근접시키는 적합도 함수에 따라 우성과 열성으로 구분하여, 우성의 해가 발생 빈도가 높도록 만든 알고리즘이다. 유전 알고리즘을 사용하는 데 있어서 가장 중요한 문제는 염색체의 정의와 적합도 함수 정의이다.

신경망 알고리즘을 사용할 때, 가장 중요한 것은 입력 패턴 선정, 특징점 선정, 강도 벡터 선정, 매트릭스 개수 등의 여러 구성요소가 존재한다. 본 연구에서는 특징점의 우선순위로 강도 벡터의 값을 사용한다. 따라서, 신경망의 구성요소 중 강도 벡터를 유전 알고리즘을 통하여 최적화하였다. 초기화 과정은 연결강도에 대한 개체군을 무작위 값으로 구성한다.

개체의 구성요소는 연결 강도 벡터 $W_1, W_2, ..., W_n$에 대응되는 실수 값으로 구성되며, 개체는 M개의 개체군으로 초기에 무작위 값을 갖는다. 초기에 무작위 값을 갖도록 하는 것은 주어진 문제에 대하여 전구간에 대한 해를 찾기 위한 과정으로, 자연 생태계에서 무수한 유전 인자를 갖는 생물이 발생하는 것에 대응되고 적합도 계산 과정은 모든 개체를 주어진 문제에 대해서 적합한지를 판정하는 과정으로 적합도 함수 정의가 요구된다.

유전 알고리즘에서 무작위로 생성된 벡터 개체군들의 적합도는 인식률로서 판단할 수 있다. 생성된 강도 벡터 개체군 P_a의 인식률보다 P_b의 인식률이 더 좋다면, P_b가 주어진 입력에 대한 신경망의 강도 벡터로서 더욱 적합한 것으로 판단할 수 있다. 입력 데이타에 대한 출력을 알 수 있다고 가정하고, 입력 데이타의 각 분류별로 10%의 데이타들의 각 특징 평균을 뉴런으로 구성한다. 신경망의 인식 알고리즘으로 전체 입력 데이타의 특징과 연결강도를 이용하여 뉴런을 통해 인식된 인식률을 적합도 함수로 사용한다. 다음과 같은 연결 강도 적합도 함수가 정의된다.

$$f(x) = \frac{R(x)}{M} \qquad \text{(식 1)}$$

위의 수식에서, M은 입력한 전체 데이터이며, $R(x)$는 인식된 개수를 나타낸다. $R(x)$는 M개의 입력 데이터에서 분류별로 10%의 데이터를 무작위로 추출하여 이들에 대한 평균값으로 뉴런을 구성한 뒤, 한 개체의 연결강도를 적용하여 인식된 결과가 된다. 전체 개체군에 대해서 무작위로 구성되었던 뉴런은 동일하게 적용되어야 한다. 모든 개체군에 대해 적합도가 구해지면, 적합도 순위에 따라 정렬한다. 교배 과정에서는 실수형 데이타를 다루므로 유전 알고리즘의 최적 재교배 알고리즘을 선택하며, 생존 규칙은 룰렛 휠 방식을 사용한다. 돌연변이 과정은 최적 해가 아닌 범위에서 수렴되어지는 경우, 이를 극복하기 위한 유전 연산으로 유전 알고리즘의 전체 반복회수의 10% 주기마다 반복 적용한다. 재생산 연산에서는 교배되기 전의 개체군(N)과 교배·돌연변이 과정 이후 생성된 개체군(N)으로 구성된 전체($2\times N$)에서 적합도에 준하여 최상위 해 집단 N개를 선택한다. 유전 연산 교배, 돌연변이, 재생산 과정은 유전 알고리즘에서 해를 구하는데 적합한 알고리즘으로 본 논문에 선택한 방법 외에 다른 연산으로 대치될 수 있다.

2. ATM(adaptive tree map)

일반적으로 신경망으로 이용한 패턴 인식에서는 입력 데이터에 대한 모든 특징을 이용한다. 데이터의 특성에 관계없이 모든 특징이 인식에 사용되므로 비효율적이다. 이를 극복하기 위해 II 장에서 제안했던 유전 알고리즘을 이용하여 특징의 우선순위를 구하여, 데이터 분류를 효율적으로 개선할 수 있다. 예를 들어, 첫 번째 우선순위의 특징이 '최대 수직선 길이'인 경우, '1' '4' '7'이 갖는 숫자의 특징과 '0' '8'이 갖는 특징은 확실히 구분되며, 이후, 이들로 파생된 노드에서 다음 우선 순위의 특징을 사용하여 구분할 수 있으므로 인식의 정확도는 더욱 높아질 것이다. 그림 1은 이를 도식화한 것이다.

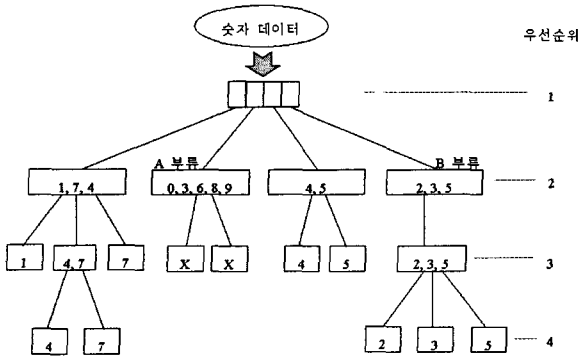

그림 1. 우선순위에 따른 트리 구조

3. ATM 훈련 알고리즘

ATM의 훈련 알고리즘은 인식을 위해 사용될 표본 데이타의 특성을 트리맵 구조로 형성하는 단계와 형성된 트리맵에서 출력 결과가 같은 데이타의 특징을 병합하는 과정을 갖고 있다. 병합에 사용되는 문턱치는 트리맵의 레벨에 따라 차등 적용되므로 인식에 유용성을 갖게 된다. 트리맵의 훈련 알고리즘은 반복 회수를 단 1회 요구하며, 병합 단계를 거치면 최적화된 트리맵이 형성된다.

훈련 알고리즘은 처음에 입력되는 데이터의 특징 개수에 따른 레벨별 맵을 구성하고, 레벨내의 첫 번째 노드를 설정한다. 입력되는 데이타들은 마지막 레벨의 특징을 사용하여 같은 출력을 내지 못하는 경우, 구분을 위한 새로운 특징이 요구되므로 프로세스를 중단한다. 해당 레벨의 노드 범위를 벗어나는 경우는 노드를 설정하고 새로운 맵을 구성한다. 입력되는 특징의 해당 레벨에서 범위의 교집합이 발생되면, 노드 범위를 업데이트하고, 현재 레벨 이하의 특징의 범위가 같은 경우는 서로의 노드를 병합시킨다. 그림 2에서 A입력이 1.0으로 현재 노드에 속하는 경우 업데이트를 수행하며, B와 같이 범위를 벗어나는 경우, 새로운 맵과 노드를 설정하게 된다. 또한 그림에서 X와 Y로 표현된 노드 범위의 교집합이 발생되면, 현재 레벨 이하의 맵 노드들이 서로 범위 내로 포함되며 동일한

출력이 발생되는 경우 하나의 노드로 병합한다.

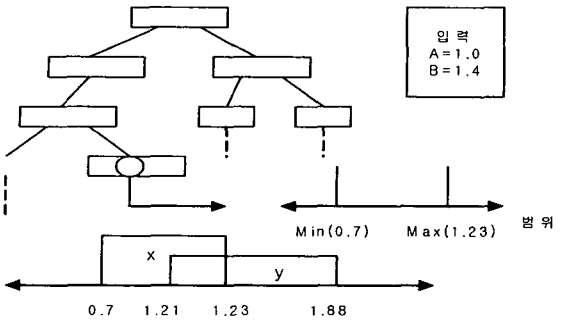

그림 2. 노드의 범위 조정

그림 3은 최적화 프로세스를 표현한 것으로, 상단의 그림은 현재 레벨이하에 하나의 노드만 갖는 맵을 줄여서 인식 과정에서 i개의 특징만 사용할 수 있도록 노드를 줄이는 과정이다. 또한, 하단의 그림은 현재 레벨이 하나의 노드만을 가지는 경우, 이하의 특징을 상단으로 편입시켜 다음 레벨이 인식에 사용될 수 있도록 하는 병합하는 과정이다.

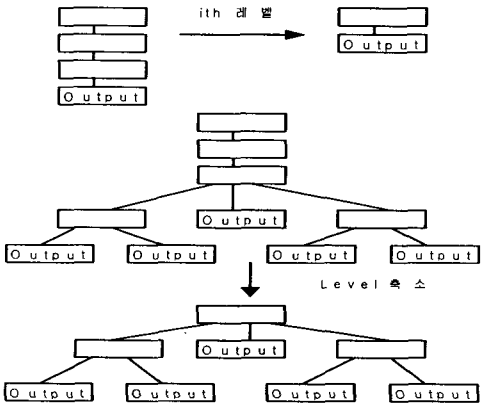

그림 3. ATM의 최적화 프로세스

III. 실 험

본 실험에서는 제안한 인식 알고리즘을 숫자 인식에 사용한다. 사용된 숫자 데이터는 44000개를 무작위로 추출하여, 훈련과 인식에 사용하였다.

그림 4에서 첫 번째 유형은 해당 숫자 번호마다 200개의 데이터가 저장된 텍스트 형식의 파일이고, 두 번째 유형은 0~9의 숫자가 100개씩 1000개로 팩된 데이터를 갖는 텍스트 파일이며, 세 번째 유형은 위의 두 가지 유형에 속하지 않는 나머지 데이터 1000개를 갖는 텍스트 파일이다. 실험에서 전처리 과정으로 입력된 데이터의 화질

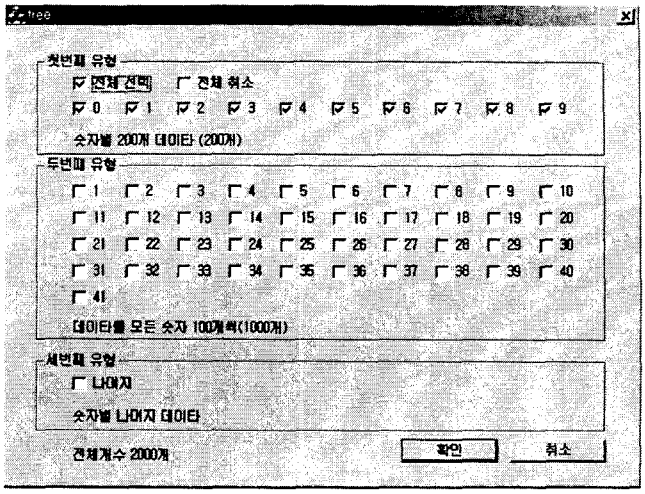

그림 4. 시물레이션 프로그램의 데이타 로딩부

열화나 해상도에 따른 데이터 소실을 보상하기 위하여, 모든 점에 대한 확장을 한번 수행하여 중심점을 추출하였다.

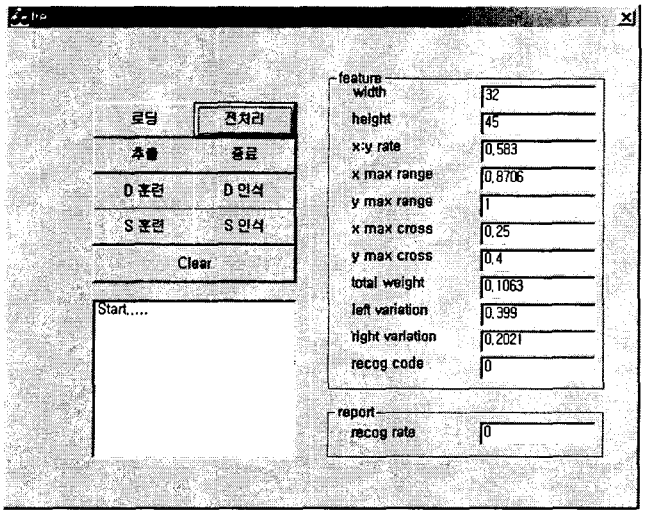

그림 5. 시물레이션 프로그램의 전처리 과정

그림 5는 원 데이타에서 중심점을 추출한 데이타이다. 전처리가 완료된 데이타에서 인식에 사용할 특징을 추출한다. 실험에서는 넓이에 대한 높이 대비, 최대 가로 부피율, 최대 세로 부피율, 최대 가로 교차점 수, 최대 세로 교차점 수, 전체 데이터 분포율, 우측 변화율, 좌측 변화율로 8개의 특징을 추출하여 인식에 사용한다. 각 특징은 프로세싱의 효율을 위해 모든 특징을 0~1 사이로 평준화하였다. 전처리 과정을 거친 데이타의 특징을 실험에서는 ATM과 신경망의 입력으로 사용한다. 본 논문에서 제안한 ATM 알고리즘의 검증을 위해 신경망의 한 부류인 SOFM(self-organizing feature maps)을 이용한다. ATM에 사용되는 특징의 우선순위 결정과 SOFM에서 입력

데이타와 히든 계층의 노드사이의 가중치 결정에 유전 알고리즘을 사용한다. 유전 알고리즘으로 최적화된 가중치는 SOFM의 가중치로 바로 적용하며, 최적화된 가중치를 정렬하여 정렬된 수치에 따라 특징의 우선 순위로 적용한다. 입력된 데이터에 대한 ATM 훈련은 루프가 필요하지 않으므로, 단 1회의 훈련으로 맵이 생성되었으며 입력 데이터에 대한 인식률은 100%의 결과를 나타내었다. SOFM은 루프가 완료된 이후, 훈련시의 인식 결과에 따라 루프 회수, SOFM의 맵 크기, 노드 조정에 필요한 α 값에 대하여 수동으로 조절이 필요하다. 여기서 α값은 다음과 같은 조정식에서 사용되는 인자이다. 식 2에서 r은 현재 반복 횟수이며, S_r은 전체 반복 회수이다.

$$f(x) = \frac{1}{\frac{S_r}{S_r - r}} \times \alpha \qquad \text{(식 2)}$$

다음에 설명되는 표 1, 2는 SOFM의 반복 횟수=1000, SOFM의 맵 크기 = 10×10, α값 = 0.911의 기본값을 근거로 인자를 조정한 인식 결과와 훈련 시간을 표현한 것이다.

표 1. SOFM 크기에 따른 인식 결과

SOFM 크기	10×10	20×20	30×30	40×40	50×50	60×60
인식률 (%)	33.50	56.15	73.30	88.45	92.00	95.75
학습 시간	00:05:57	00:15:22	00:38:22	01:01:44	01:36:59	05:38:06

표 2. SOFM 반복 회수에 따른 인식 결과

SOFM 의 α 값	0.0911	0.511	0.911	1.000
인식률(%)	82.00	91.75	92.00	91.85

표 3. 기존 신경망 알고리즘과 비교 평가

Methods	Recog [%]	Error [%]	Reject [%]	Reliability [%]
Legault [3]	93.90	1.60	4.50	98.32
Krzyzak [4]	94.85	5.15	0.00	94.85
Suen [5]	93.05	0.00	6.95	100.0
SOFM	92.35	3.15	3.50	96.70
ATM	98.95	0.00	1.05	100.0

표 3은 기존 인식 알고리즘 성능과 ATM을 비교한 결과이다. Suen은 일반적인 특징과 인식 알고리즘을 사용하였고, Legault는 패턴 정보에서 윤곽선 정보와 파라미터 특징을 이용하여 인식에 사용하였으며, Krzyzak은 패턴 분류를 위해 기존의 오류 역전파 알고리즘을 개선하였다. 여기서 신뢰도(Reliability)는 다음 식과 같다.

$$f(x) = \frac{r(x)}{r(x)+e} \times 100\ [\%] \qquad \text{(식 3)}$$

식 3에서, $r(x)$는 인식률이며, e는 에러율을 나타낸다. 제안된 ATM의 평균 인식률은 올바로 인식된 테스트 데이터 값을 전체 테스트 데이터 개수로 나눈 값에 대한 백분율이다.

Ⅳ. 결 론

본 논문에서 제안한 ATM은 기존의 신경망과 비교해서 입력되는 데이터의 순서에 관계없이 동적인 맵을 1회의 훈련과정으로 구성이 가능하여 시간을 절약할 수 있으며, 새로운 입력 데이터가 추가되는 경우 맵을 동적으로 확장 가능하기 때문에 적응성이 뛰어나다. 또한, 초기맵을 생성하지 않고, 입력되는 데이터에 따라 동적으로 맵이 확장되므로, 초기맵 생성에 따른 인식의 영향을 받지 않으며, 사용되지 않는 노드는 확장되지 않으므로 메모리 낭비가 없다. 입력되는 데이터의 특성에 따라 입력이 되어, 훈련 과정에서는 인식률이 100%에 이른다. 그리고, 데이타의 특징에 대한 중요도를 유전자 알고리즘을 통한 특징 우선순위로 고려한다. 일반적인 신경망은 입력 데이터에 대해 많은 시간의 실험과 인자 조정이 필요하지만, ATM은 입력 데이타에 따른 실시간 동적맵 구성이 가능하므로 일반적인 인식 문제가 발생하는 경우 최적의 특징을 갖는 맵을 구성한 인식 시스템을 구성할 수 있게 된다.

참 고 문 헌

[1] L. D. Davis, The Handbook of Genetic Algorithms, Van Nostrand Reinhold, 1991.

[2] 하성욱, 권기항, 강대성, "유전 목 지도의 동적확장", 정보과학회논문지:소프트웨어 및 응용 제29권 제6호 pp. 386-395 2002.6.

[3] R. Legault and C. Y. Suen, "Contour Tracking and Parametric Approximations for the Digitized Patterns," Computer Vision and Shape Recognition : Singapore, pp. 225-240, 1989.

[4] A. Krzyzak, W. Dai and C. Y. Suen, "Unconstrained Handwritten Character Classification using Modified Backpropagation Model," In Proc. 1st Int. Workshop on Frontiers in Handwriting Recgnition, pp. 155-166, 1990.

[5] C. Y. Seun, C. Nadal, R. Legault, T. A. Mai and L. Lam, "Computer Recognition of Unconstrained Handwritten Numerals, Proceeding of the IEEE, Vol. 80, No. 7, pp. 1162-1180, 1992.