Hidden Markov Model 과 Genetic Algorithm을 이용한 온라인 문자인식에 관한 연구

홍 영 표, 장 춘 서
금오공과대학교 컴퓨터공학부
전화 : 054-467-4418 / 핸드폰 : 011-548-1381

On-Line Character Recognition using Hidden Markov Model and Genetic Algorithm

Young-Pyo Hong, Choon-Seo Jang
School. of Computer Engineering, Kumoh National University of Technology
E-mail : yphong@cespc1.kumoh.ac.kr

Abstract

HMM(Hidden Markov Model)은 시간적인 정보를 토대로 하는 수학적인 방법으로서 문자인식에 많이 사용되어지고 있다. 그런데 HMM이 적용되고자 하는 문제에서 사용되어지는 상태 수와 HMM에서 사용되어지는 parameter들은 처음에 결정되는 값들에 의해서 상당히 많은 영향을 받게 된다. 따라서 한글의 특성을 이용한 HMM의 상태 수를 결정한 후 결정되어진 각각의 HMM parameter들을 Genetic Algorithm을 이용하였다. Genetic Algorithm은 매개변수 최적화 문제에 대하여 자연의 진화 원리를 모방한 알고리즘으로 선택, 교배, 돌연변이 연산을 이용하여 최적의 개체를 구하게 된다. 여기서는 HMM에서의 Viterbi Algorithm을 적합도 검사에 사용하였다.

I. 서론

앞으로 PDA 사용 환경이 점차 증가하게 되면서 입력 수단으로 키보드를 제외한 다른 수단의 필요성을 느끼게 됨으로써 문자인식에 관한 연구가 활발히 이루어지게 되었다. 본 연구에서는 문자인식을 행함에 있어서 한글의 특성을 이용하여 HMM(Hidden Markov Model)의 상태를 결정한 후 결정되어진 각각의 HMM을 Genetic Algorithm을 이용하여 parameter를 최적화 한 후 하나의 한글 인식기를 구현하여 학습결과를 확인 하고자 한다.

온라인 문자인식에서는 시간에 따른 필기 상태를 이용하여 인식하는 방법이 적합하기 때문에 확률 통계적 방법 중에서 HMM이 많이 이용되어지고 있다. 최근에는 HMM 외에 여러 가지 방법들을 혼합하는 시도가 많이 이루어지고 있다. 여기에서는 HMM과 Genetic Algorithm을 이용하여 학습한 후 문자를 인식 하고자 한다. 일반적으로 HMM은 다음과 같이 정의한다.

1) 초기상태 확률분포

$\pi = \{\pi i\}$: 이것은 다음 [식 1]에 정의될 매개변수 A의 특별한 경우로 볼수 있으며, 초기 t=1 에서의 상태 확률분포를 나타낸다.

$$\pi i = P(q1=Si), \quad 1 \le i \le N, \quad \text{[식 1]}$$

$$\sum_i \pi i = 1$$

2) 상태전이 확률분포

A = {aij} : 마르코프 모델의 모든 상태 전이 확률을 나타내는 벡터 매개변수로, 그 값은 다음과 같이 정의된다.

$$aij = P(qt+1 = Sj \mid qt = Si), \quad 1 \le i,j \le N, \quad \text{[식 2]}$$

aij는 임의 시각에 상태 Si에 있다가 다음 순간에 상태 Sj로 전이할 확률을 말한다. 만약 aij > 0이면 Si에서 Sj로 한번의 상태 전이로 갈 수 있다는 뜻이다. 그리고

이들 매개 변수는 다음의 조건을 만족한다.

$$\sum_j aij = 1 \qquad [식\ 3]$$

3) 관측 심볼 확률 분포

B = {bi(k)}는 다음 [식 4]과 같이 정의된다.

$bi(k) = P(vk \mid Si) \qquad 1 \le i \le N,\ 1 \le k \le M$ [식 4]

bi(k)는 시각 t에 상태 Si에서 k번째 심볼 vk를 관찰하게 될 확률을 말하며 다음 [식 5]과 같은 확률 조건을 만족한다.

$$\sum_k bi(k) = 1 \qquad 1 \le i \le N \qquad [식\ 5]$$

그리고 이밖에 모델의 기본 구조를 정의하는 상태의 수 N 과 관측 심볼의 집합 V = {v1, v2, ... , vM}가 있다. 그러나 HMM은 위에서 설명한 모델의 확률적 특성을 기술하는 매개변수 λ로 표현하는 것이 보통이다.

$$\lambda = (A, B, \pi)$$

HMM은 시간적인 정보를 토대로 특정 사건의 일어날 확률을 예측하는 수학적 모델이므로 온라인 상에서 발생하는 순차적인 문자 데이터를 입력으로 사용하기에 매우 적합하다. HMM의 구성 요소와 원리를 보면 다음과 같다.

HMM의 기본적인 형태는 다음 [그림 1]과 같다.

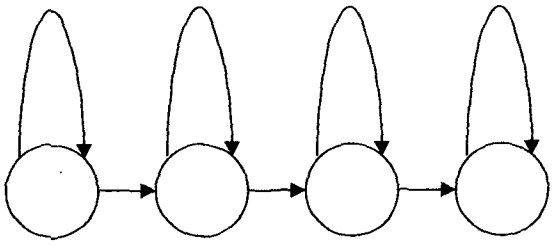

그림 1. left-to-right simple HMM

HMM은 그 명칭에서 알 수 있듯이 모델링 방식의 하나이다. 실제로, HMM은 이산 심볼열(스트링)로 샘플링된 신호를 해독하는 문제에 통계적 모델을 적용하기 위한 기법의 하나로 출현한 것이었다. 이러한 모델링 방식에서는 우선 정보(또는 신호)의 불확실성 근원에 대해서 적절한 모델을 설정하게 된다. 그 다음에는 모델 매개변수의 값 추정, 주어진 입력에 대한 모델 평가 그리고 주어진 입력을 가장 최적으로 표현해주는 모델 안에서의 경로를 찾는 세 가지 문제가 나타난다. 이미 각 문제에 대하여 효율적인 방법이 개발되어 널리 이용되고 있으며, 이런 효율적인 방법들은 HMM의 실용성을 높여준다.

Genetic Algorithm 또는 진화적 알고리즘은 자연세계의 진화현상에 기반한 계산 모델이다. 진화알고리즘은 풀고자 하는 문제에 대한 가능한 해들을 정해진 형태의 염색체로 표현한 다음, 이 염색체들의 집단 중에서 해결해야할 문제에 적합한 정도에 따라 적합도가 계산되며, 각 염색체들은 적합도에 비례해서 다음 세대에 재생산될 확률을 가진다. Genetic Algorithm에는 선택, 교배, 돌연변이 등의 세 가지 기본적인 연산이 있다. 이런 세가지 연산들을 이용해서 점차적으로 새로운 세대로의 진행을 하면서 최적화된 값을 구하게 된다.

Genetic Algorithm의 기본적인 흐름도는 다음 [그림 2]와 같다.

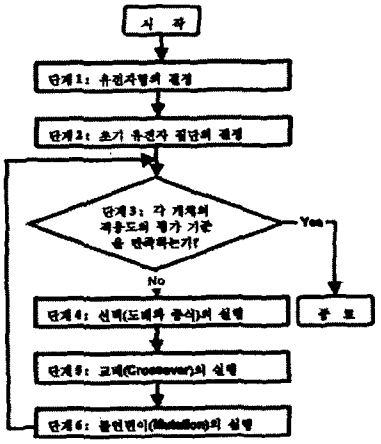

그림 2. Genetic Algorithm의 흐름도

Genetic Algorithm의 장점으로서 다음과 같은 점을 들 수 있다.

■ 복수 개의 개체 사이의 상호 협력에 의한 해의 탐색 - genetic algorithm은 복수의 개체 사이에서 선택이나 교배 등의 유전적 조작에 의해서 상호 협력적으로 해의 탐색을 수행하고 있다. 따라서, 단순한 병렬적 해의 탐색과 비교하여 보다 좋은 해를 발견하기 쉽다.

■ 번거로운 미분 연산 등이 불필요 - 뉴럴 네트워크, 특히 백프로파게이션 알고리즘 등에서는 평가함수의 미분값을 필요로 하였다. genetic algorithm에서는 현재 적응도를 분별할 수 있으면 되기 때문에 알고리즘이 단순하고, 평가 함수가 불연속인 경우에도 적용이 가능하다.

한편 다음과 같은 문제점이 있다.

■ 대상으로 하는 문제를 genetic algorithm으로 해결하기 위한 일반적인 방법이 없다. 특히문제의 유전자형으로의 표현(코딩)은 설계자의 숙달로 되어 있다.

■ 개체수, 선택 방법이나 교배법의 결정, 돌연변이의

비율 등 파라미터의 수가 많다.

II. HMM과 Genetic Algorithm을 이용한 문자인식

2.1 HMM의 상태수 결정과 염색체 구조

입력되는 데이터로부터 먼저 전처리를 수행한다. 각 획의 시작점에서부터 끝점까지의 각 점 사이의 각도의 변화에 따라 특징 점을 추출하게된다. 한글의 경우 각 획 사이의 각도의 변환이 90도 이상이 되는 경우가 많은데 이러한 방법을 이용하여 각각의 자소에 대한 HMM의 상태수를 결정할 수 있게 된다. 이러한 방법으로 HMM의 상태 수를 자동적으로 결정한 후 염색체들의 초기 상태를 생성한다.

각각의 염색체에서 가져야 되는 조건으로 첫 번째로는 상태 전이 확률 a_{ij} 가 다음의 조건을 가져야 한다.

$$1 = \sum_{j=0}^{5} a_{ij} \text{ where } i = 1, \ldots, N \quad [식\ 6]$$

두번째로 관측 확률 b_{ik} 에서도 마찬가지로 다음의 조건에 만족하여야 한다.

$$1 = \sum_{k=1}^{8} b_{ik} \text{ where } i = 1, \ldots, M \quad [식\ 7]$$

(상태수가 5이고 관측심볼의 수가 8이라고 가정)따라서 하나의 HMM이 하나의 염색체로 바뀌게 되면 다음과 같은 [그림 3]의 염색체 구조를 가지게 된다.

a_{11}	a_{12}	.....	a_{55}	b_{11}	b_{12}	b_{13}	.....	b_{57}	b_{58}

그림 3. HMM을 염색체 구조로 표현

2.2 Genetic Algorithm functions

초기의 Genetic Algorithm는 적합도에 단순히 비례하는 선택방법을 사용하였는데, 적합도 사이에 분산이 큰 초기에 적합도가 높은 소수의 개체들과 이들의 자손들이 개체군 내에서 빠르게 증식되기 때문에 다양한 공간을 탐색하지 못하고 조기에 수렴하게 된다. 따라서 개체를 선택하는 강도를 일정하게 해주는 sigma scaling 방법을 이용하여 선택을 수행하였다. sigma scaling 방법은 시간 t 에서 염색체 p가 다음세대에 선택될 기대값을 아래의 [식 8]으로 표현한다.

$$S(\lambda p, t) = 1 + \frac{ft(\lambda p) - ft(\lambda)}{2\sigma t(\lambda)}, \text{ if } \sigma t(\lambda) \neq 0$$

$$= 1 \quad , \text{ if } \sigma t(\lambda) = 0$$

[식 8] 유전자 알고리즘 에서의 sigma scaling 식
시간 t에서 적합도의 평균, 적합도의 표준편차 이용.

두 염색체 사이의 교배는 일 점 교배와 이점 교배를 사용하였다. 위의 그림에 나와 있는 대로 하나의 염색체는 두 부분으로 이루어져있다. 따라서 전이 확률 부분에서는 일 점 교배를 행하고 관측 확률 부분에서는 이 점 교배를 행한다. 교배 연산을 수행한 후에 위의 [식 6],[식 7]에 맞도록 해주어야 한다.

돌연변이 연산에서도 위의 교배 연산과 마찬가지로 전이 확률의 한 점과 관측 확률의 두 점에 대해서 돌연변이 연산을 수행한다. 돌연변이 연산을 수행한 후에도 역시 위의 [식 6],[식 7]에 맞도록 해주어야 한다.

HMM에 적용된 Viterbi algorithm은 주어진 입력열 O에 대한 최적의 경로를 계산하는 것이다. Viterbi algorithm에 따르면 각 상태의 확률은 경로상의 여러 가능성 중 최적 또는 최대 값으로 결정되는데 이는 어떤 시점에서 현재 노드 까지 오는 경로 중 최적인 것 하나만을 기억한다는 것이다. 실제로 여러 경로를 끝까지 확장하여 보았을 때 도중에 버려진 경로들은 각 시점에서 선택한 최적 경로보다 낮은 확률값을 갖는다. 따라서 적합도를 평가하기 위해서 Viterbi algorithm을 이용하여 각각의 염색체들에게서 구한 최적 경로의 값들을 전체 염색체들의 최적 경로의 합으로 나누어 준 값을 적합도 값으로 한다. 다음의 [식 4]를 이용해서 적합도를 구한다.

Step 1
Initialization: $\delta_1(i) = \pi_i b_i(O_1), \ 1 \le i \le N$
Step 2
Recursion: For $2 \le t \le T, \ 1 \le j \le N$

$$\delta_t(i) = \max_{1 \le i \le N} [\delta_{t-1}(i)\ a_{ij}] * b_j(O_t)$$

Step 3
Termination - Maximum probability, argument and calculate fitness:

$$Pr = \max_{1 \le i \le N} [\delta_T(i)]$$

$$f = Pr(n) \Big/ \sum_{n=1}^{\text{all gene}} Pr(n)$$

[식 9] 적합도 검사를 위한 식
위의 [식 9]에 나와 있듯이 염색체의 초기화가 끝난후

단계2 에서 모든 경로들에 대한 합의 최적값을 구한다. 그 최적값을 단계3 모든 염색체의 확률값으로 나누어 준 값을 해당 염색체의 최적값으로 한다.

III. 실험 및 결론

3.1 실험 결과

훈련 및 인식실험에 사용된 데이터는 각각의 자소들이다. 실험에서 훈련에 사용된 자소들은 각각 40개씩을 사용하였다. 그리고 인식 실험에 사용하기 위해서 40개씩의 데이터를 이용하였다. 모든 데이터들은 실험을 하기 위해 전처리를 거친 후 사용하였다. 다음 표에서 결과를 확인해볼수 있다.

	Hidden Markov Model	Genetic Algorithm + Hidden Markov Model
초성	84%	91%
중성	93%	95%
종성	81%	89%

중성의 경우는 단조로운 패턴으로 인해 인식률이 좋은 경우가 많았지만 종성의 경우 복잡한 패턴들이 존재하기 때문에 인식률의 저하가 생겨났다.

3.2 결론

일반적으로 Hidden Markov Model을 이용하는 경우 모델에 대한 매개 변수를 추정한 결과에 따라 인식률의 차이가 난다. 이러한 모델의 매개 변수를 추정하기 위해서 Genetic Algorithm을 사용하였다. 복잡한 탐색 공간에서 해를 찾아내는 문제에 간단하게 적용될 수 있으므로 기존의 방법보다 구현이 간단하고, 적합도를 나타내는 함수를 변형하면서 자신이 원하는 모델을 쉽게 만들 수 있다. 하지만 확률적인 방법에 의존하기 때문에 잘못된 방향으로 진화할 확률도 가지고 있다. 이와 같은 같은 단점은 여러번의 훈련을 거친후 최적의 모델을 선택하는 방법으로 해결할 수 있다. 또한 자소가 아닌 문자 또는 문장으로서의 인식기를 구현하기 위해서는 각각의 자소에 해당하는 염색체들의 연결에 대한 방법이 필요하다.

참고문헌(또는 Reference)

[1] 하진영, HMM 네트워크 기반의 한글 인식기를 위한 구조 특성열의 적용, 제 10회 한글 및 한국어 정보처리 학술대회, 1998

[2] 하진영,신봉기 , 온라인 한글 인식을 위한 HMM 상태 수의 최적화, 정보처리학회, 1998

[3] 김찬우, 퍼지 방향 코드를 이용한 온라인 한글 인식, 1995

[4] 이재준, 은닉 마르코프 모델을 이용한 한/영 혼용 필기의 온라인 인식

[5] Melanie Mitchell, An Introduction to Genetic Algorithms,

[6] C.W.Chau, Optimization of HMM by a Genetic Algorithm, 1997 IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, 1997

[7] Tomio Takara, Isolated Word Recognition Using The HMM Structure Selected by the Genetic Algorithm