HMM 모델링을 위한 HMM의 State수와 Mixture수 분석

박미나[0], 하진영
강원대학교 컴퓨터정보통신공학과
jyha@kangwon.ac.kr, atom77@mirae.kangwon.ac.kr

Analysis of the Number of States and Mixtures for HMM Modeling

Mi-Na Park[0], Jin-Young Ha
Dept. of Computer Engineering, Kangwon National University

요 약

본 논문에서는 음성인식과 온라인 필기인식에서 우수한 성능을 보이는 은닉 마르코프(HMM)의 모델링의 문제점을 다룬다. HMM은 파라미터의 수가 클수록 자기 데이터에 대해 잘 모델링하는 특징으로 음성인식과 온라인 필기인식등에서 많이 쓰이고 있다. 그러나 그러한 특징으로 인해 해당 클래스 데이터가 아닌 다른 클래스 데이터에 대해서도 파라미터의 수가 클수록 잘 모델링하는 단점이 나타났다. 이에 본 연구에서는 대상 데이터를 분석하여 state의 수와 mixture의 수를 조정하여 가장 적절한 HMM의 구조의 파라미터를 구하는 가능성을 본다.

1. 서 론

은닉 마르코프 모델(HMM)이 이와 같이 널리 쓰이는 이유는 음성과 문자등에서 발견되는 많은 변형들을 흡수할 수 있고, 시간에 따라 변해가는 특성을 지니는 Data를 잘 모델링하며 파라미터의 수가 클수록 잘 모델링 하기 때문이다[1].

이러한 HMM의 성공은 가변 길이의 시계열에 대한 높은 모델링 능력과 EM-알고리즘과 같이 주어진 모델 구조에 맞춰 모델 파라미터를 재 추정할 수 있는 강력한 훈련알고리즘에 기인한 바가 크다.

온라인 필기인식에서 많이 사용되는 HMM은 left-to-right HMM으로 모델 구조는 state 수와 state당 mixture 수, 그리고 전이 확률에 의해 결정된다.

이러한 HMM의 구조는 휴리스틱한 방법에 의해 결정되는 것이 일반적이기 때문에 최적의 모델을 선택하는데 어려움이 있다.

HMM 구조의 최적화를 위한 연구는 다양한 방법으로 진행되어 왔다. 높은 점유를 갖는 상태부터 순차적으로 분할해서 점차 상태 수를 증가시키는 방법이나 Dirichlet 사전 확률에 기반한 사후 확률을 사용하는 복잡한 구조로부터 점차 구조를 감소시켜나가는 방법이 있었다. 또한 최대확률 기준을 이용하거나 BIC를 이용한 연구가 있었다[1][4][5].

이에 본 연구에서는 HMM의 구조를 최적화하기 위해 HMM 적용 시 해당 모델과 해당 데이터에 기반한 likelihood 뿐만 아니라 그 모델에 대한 다른 클래스의 데이터에 대한 확률 값인 Anti-likelihood를 이용하여 데이터를 분석한다.

2. 은닉 마르코프 모델(Hidden Markov Model)

은닉 마르코프 모델(HMM)은 유한개(N)의 노드 $S=\{S_1,\cdots,S_n\}$와 각 상태 사이를 방향성 있게 연결하는 전이하는 집합으로 구성된 네트워크로 정의된다.[2] HMM은 유한상태 기계로 그 안에 있는 상태는 은닉되어 있고 단지 출력열이 관측되며 출력확률은 각 상태에 지정되어 있다. HMM은 다음과 같이 {S,A,B}로 표현될 수 있다.

· $S=\{S_1,\cdots,S_n\}$, 총 상태수가 N개인 HMM 상태의 집합.
· $A=[a_{ij}]$, 상태 전이 확률 행렬.
· $B=\{b_i(x)\}$, 출력확률 집합으로 $b_i(x)$는 다음과 같이 정의된 상태 S_i에 연관된 확률이다.

$$b_i(x)=\sum_{l=1}^{L} w_{il}\ N(x,\mu_{il},\sum_{il})$$

$N(x,\mu_{il},\sum_{il})$는 정규분포이고 μ_{il}는 i-번째상태의 l-번째 mixture의 평균이고 $\sum_{il}$은 공분산, w_{il}는 가중치이다. 각 mixture에서는 D-차원의 feature vector가 있고 모든 상태에서 L개의 mixture가 있다고 가정한다.

· $\mu, \sum, w$ 를 각각 모델 전체에 대한 평균벡터, 공분산 행렬, 가중치라 하고, $\mu_i, \sum_i$ 그리고 w_i를 각각 특정 상태 S_i에 대한 평균 벡터, 공분산 행렬, mixture 가중치라고 정의한다.

2.1 HMM의 구조

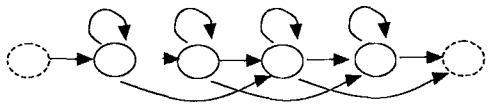


그림 1. 6- 상태 HMM

위의 그림1은 본 논문에서 사용한 연속밀도 HMM 구조의 예이다. HMM의 구조 M과 주어진 M에 대한 파라미터 θ로 특징 지워 질 수 있다. 상태 간 전이 구조는 고정되어 있기 때문에 M은 상태 수와 상태 당 mixture 수로 유일하게 결정된다.

따라서 모델을 모델 구조 $M = \{Q, L\}$과 파라미터 $\theta = \{A, \mu, \sum, w\}$의 합집합으로 볼 수 있다[1].

3. Anti-likelihood

HMM은 파라미터의 수가 많아질수록 해당 데이터에 대한 인식률이 상당히 높다.

이에 다른 클래스 데이터에 대해서도 파라미터의 수가 많아질수록 인식률 또한 높게 나오는 결과를 얻었다.

그러므로 다른 데이터에 대한 변별력이 어렵다.

그래서 본 연구에서는 해당 클래스 데이터에 대해선 잘 모델링 되고 그 외의 데이터에 대해선 잘 모델링 되지 않는 적당한 파라미터를 구하기 위해, 해당 클래스 데이터외의 다른 데이터에 대해서 실험하였다.

해당모델에 대한 다른 클래스란 해당모델이름으로 자기 자신외의 다른 데이터를 묶어 만든 해당 클래스외의 데이터의 집합을 말한다. 이 클래스의 데이터와 해당모델 데이터와의 차를 이용한다.

$$Anti_c = \underset{[X \in C]}{\log p(X|\theta_{ML})} - \underset{[X' \notin C]}{\log p(X'|\theta_{ML})}$$

단, C는해당클래스데이타

4. BIC(Bayesian Informaition Criterion)

Central Limit 정리에 의해, 파라미터의 사전 확률 $p(\theta_{ML}|M)$은 평균 θ_{ML}과 공분산 Γ^{-1}을 갖는 다변량 정규밀도로 간주될 수 있다. 이러한 조건은 다음과 같이 정의된 널리 알려진 베이지안 정보기준으로 인도한다.

$$BIC(M) = \log p(X|\theta_{ML}) - \frac{k}{2}\log N \qquad (1)$$

위 식에서 BIC는 likelihood와 $\frac{k}{2}\log N$의 합인데, 후자는 모델내의 파라미터 개수에 대한 패널티(penalty) 항 또는 로그 사전 확률로 볼 수 있다. 여기에서 사전 확률은 자유 파라미터 개수에만 제한되고, 모델을 정의하는 각각의 파라미터 유형에 따라 별도의 고려를 하지 않는다. HMM에는 동질적이지 못한 파라미터 집합이 존재하기 때문에 이러한 제한점은 부적절하다[1].

5. 실험 및 결과 분석

5.1 실험데이타베이스

온라인 문자인식에 관계된 세계 각국의 대학, 연구소, 기업 등 다양한 기관들이 공통의 파일 표준을 만들어 필기 데이터를 모아 놓은 UNIPEN 데이터를 실험 대상으로 삼았다[1].

해당 클래스의 데이터와 해당 클래스외의 다른 데이터를 모아 그 차를 이용하여 실험하였다.

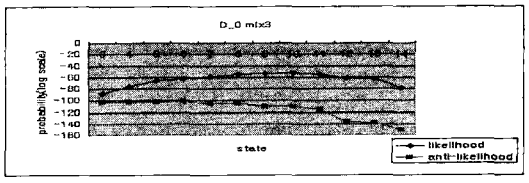
5.1.1 Anti 데이터의 생성

대상데이터에 해당하는 데이터를 제외한 나머지 데이터를 merge하여 해당데이터에 대한 Anti 모델을 생성하였다. 그 모델의 크기가 상당히 커지는 결과에 의해 그 데이터양에 비례하게 랜덤한 데이터를 선별하여 적절한 양의 Anti 데이타를 생성하여 실험하였다.

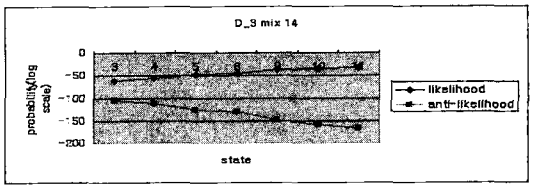
5.2 실험결과

해당데이터에 대해서는 고정 mixture의 수에 state수가 증가할수록 인식 결과가 좋은 반면 다른 데이터에 대한 결과는 좋지 않은 결과를 보인다.

그런데 일정 state수에 대해 mixture의 수가 증가할수록 해당 데이터에 대한 인식결과와 다른 데이터에 대한 인식결과는 모두 좋아지는 현상을 보였다.



(a)



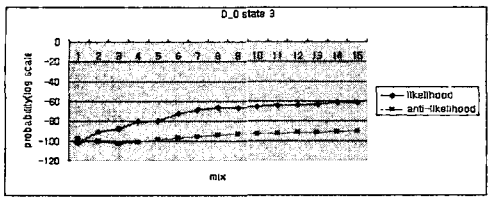
(b)

그림 2. (a)model D_0 mix 3, (b)model D_3 mix14

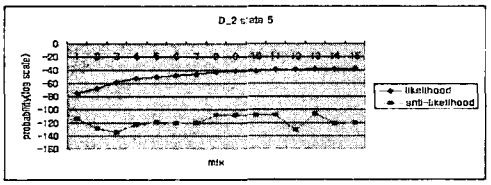
위의 그림2는 각각의 모델에 대해 mixture의 수를 (a) 3, (b)14(으)로 고정하고 state의 수를 증가했을 경우 결과를 그래프로 나타낸 결과이다.

해당 데이터에 대한 state의 수가 증가할수록 y좌표

확률 값이 증가하지만 다른 데이터에 대한 확률 값은 감소되는 것을 확인 할 수 있다.



(a)



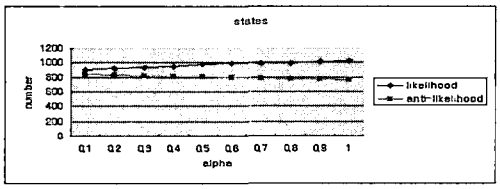
(b)

그림 3. (a)model D_0 state 3, (b)model D_2 state 5

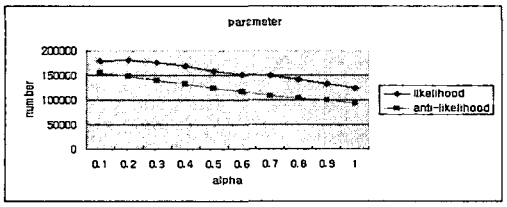
위의 그림3은 각각의 모델에 대해 state의 수를 (a) 3, (b) 5(으)로 고정하고 state의 수를 증가했을 경우 결과를 그래프로 나타낸 결과이다. 해당 데이터에 대한 mixture의 수가 증가할수록 y좌표 확률값이 증가한다. 그런데 다른 데이터에 대한 확률값도 또한 증가하는 현상을 확인 할 수 있다.

HMM은 특징은 파라미터의 수가 많을수록 데이터에 대한 모델링을 잘 한다는 특징을 가지고 있다.

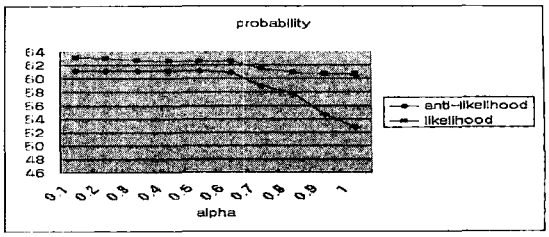
그런데 해당데이터가 아닌 다른 데이터에 대해서도 파라미터의 수가 많을수록 모델링을 잘 한다는 문제점을 가지고 있다.



(a)



(b)



(c)

그림 4. alpha값을 적용했을 경우 (a)states의 수 parameter의 수(b), 확률 값(c)

위의 그림4 (a),(b),(c)는 alpha값을 적용하였을 경우, states의 수와 parameter의 수, 확률 값을 보인 것이다.

alpha의 값이 커질수록 states의 수는 증가하고 parameter의 수는 낮아지고 전체적인 인식률은 낮아지기는 하나 높은 인식률을 보이는 것을 알 수 있다.

HMM의 구조를 최적화하기 위해 데이터를 분석한 결과 실험결과에서 보이는 것처럼 mixture의 수가 증가할 때 해당 데이터나 다른 데이터 모두 높은 인식률을 보였으며 state의 수가 증가할 때 해당데이터는 높은 인식률을 보이며 다른 데이터에 대해서는 낮은 인식률을 보이는 결과를 얻었다.

이후엔 Bayesian Information Criterion(BIC)과의 적절한 결합으로 HMM 구조 모델링의 최적의 파라미터를 구하고자 한다.

참 고 문 헌

[1] 하진영, 박미나, Alanin Biem, Jayashree Subrahmonia " 모델의 사전 확률 추정을 이용한 HMM 구조의 최적화 ", 한국정보과학회 추계 학술논문 발표논문집, 2001.

[2] 한국정자통신연구원 "펜을 이용한 문자/제스처 인식 시스템의 인식률 성능 향상을 위한 추가 개발에 관한 연구", pp.7-8,1998.

[3] LAWRENCE R. RABINER, FELLOW, IEEE " A Tutorial on Hidden Markov Models and Selected Applications in speech Recognition", 1989.

[4] 하진영, 신봉기," 온라인 한글 인식을 위한 HMM 상태 수의 최적화", 한국정보과학회 추계 학술발표논문집, pp.372-374,1998.

[5] J.Li, A. Biem and J. Subrahmonia, " HMM Topology Optimization for Handwriting Recognition", ICASSP, 2001.

[6] H.Singer and M, ostendorf,"Maximum likelihood successive state splitting", in ICASSp, pp.601-604, 1996.

[7] Andress Stolcke and stephen Omohundro, 'Hidden Markov Model induction by bayesian model mergint", in Advances in NIPS, vol5. pp.11-18, 1993.