모델의 사전 확률 추정을 이용한 HMM 구조의 최적화

하진영*0, Alain Biem**, Jayashree Subrahmonia**, 박미나*
*강원대학교 컴퓨터· 정보통신공학과
**Pen Technologies Group, IBM T.J. Watson Research Center
jyha@kangwon.ac.kr, {biem, jays}@us.ibm.com, atom77@mirae.kangwon.ac.kr

HMM Topology Optimization using Model Prior Estimation

Jin-Young Ha*0 Alain Biem** Jayashree Subrahmonia** Mina Park*
*Dept. of Computer Engineering, Kangwon National University
**Pen Technologies Group, IBM T.J. Watson Research Center

요 약

본 논문은 온라인 문자 인식을 연속 밀도 HMM의 구조의 최적화 문제를 다룬다. 최적이란 최소한의 모델 파라미터를 사용하여 최소한의 오류를 허용하는 것이라고 정의할 수 있다. 본 연구에서는 HMM 구조의 최적화를 위해 Bayesian 모델 선택 방법론을 사용한다. 먼저 잘 알려진 BIC(Bayesian Information Criterion)을 적용해보고, 그것을 HMM의 복잡한 구조에 적합하도록 본 논문에서 제안한 HBIC(HMM-Oriented BIC)와 비교해본다. BIC는 모델의 사전 확률 분포를 추정하지 않고 다변량 정규분포라고 가정하는데 비해, HBIC는 모델의 각 파라미터로부터 사전 확률을 추정한 후 그것들을 사용함으로써 더 좋은 결과를 얻도록 한다. 실험 결과 BIC와 HBIC 둘 다 기존 방법보다 모델의 파라미터 수를 현저히 감소시킴을 확인했고, HBIC가 BIC에 비해 더 적은 수의 파라미터를 사용해도 비슷한 인식률을 얻을 수 있었다.

1. 서 론[1)]

은닉 마르코프 모델 (HMM)은 음성 인식과 온라인 필기 인식에서 우수한 성능을 보여 왔다. 이러한 HMM의 성공은 가변 길이의 시계열에 대한 높은 모델링 능력과 EM-알고리즘과 같이 주어진 모델 구조에 맞춰 모델 파라미터를 재추정할 수 있는 강력한 훈련 알고리즘에 기인한 바가 크다. 온라인 필기 인식에서 많이 사용되는 HMM은 left-to-right HMM으로 모델 구조는 상태 수와 상태 당 mixture 수, 그리고 전이 확률에 의해 결정된다. 기존 연구에서 이러한 HMM 구조는 휴리스틱한 방법에 의해 결정되는 것이 일반적이었기 때문에 최적 모델을 선택하는데 어려움이 있었다. 여기서 최적 모델이란 최소한의 모델 파라미터로서 최소의 오류를 허용하는 것을 의미한다. 최근 보급이 확산되고 있는 PDA와 휴대용 무선 장치에서의 필기 인식을 위해 시스템의 메모리 용량과 인식 성능의 균형에 직접적인 영향을 미치는 HMM 구조의 최적화는 시급하고도 중요한 이슈로 부각되고 있다.

HMM 구조의 최적화를 위한 기존 연구는 다양한 방법으로 진행되어 왔다. 높은 점유를 갖는 상태부터 순차적으로 분할해서 점차 상태 수를 증가시키는 방법[1]과 Dirichlet 사전 확률에 기반한 사후 확률을 사용하여 복잡한 구조로부터 점차 구조를 감소시켜 나가는 방법이 있었다[2]. 또한 최대확률 기준을 이용하거나 BIC를 이용한 연구가 있었다[3,4].

본 논문에서는 HMM 구조의 최적화를 모델 선택의 문제로 보았다. 즉, 가능한 후보 모델들로부터 최적 모델을 선택하는 것이다. 이 방법은 구조 추정에 있어서 단순한 알고리즘을 적용할 수 있고, 정보 이론과 베이지안 추론의 장점을 살릴 수 있다. 더욱이 제한된 용량, 다시 말하면 작은 장치 플랫폼을 위한 모델을 선택하거나, 좀 더 강력한 성능을 가진 시스템을 위한 모델을 선택하거나 할 때 극히 유용하게 사용될 수 있다.

Occam의 razor 원리는 가능한 후보 집합 중에서 작은 모델을 선택하는 것을 선호하는 것으로 모델 선택에 대한 기본 철학을 제공해준다. 모델 선택에 대한 Bayesian 방법은 자연스럽게 이러한 원리를 구현할 수 있게 해준다. 하지만 Schwarz의 BIC (Bayesian Information Criteria) 등과 같이 이미 제안된 모델 선택 기준은 동질의 파라미터를 갖는 통계적으로 잘 행동하는 모델을 가정하고 있어서 연속 밀도 HMM 등과 같이 복잡한 구조에는 적합하지 않다[5].

본 논문에서 제안하는 모델 선택 기준인 HBIC (HMM-oriented BIC)는 파라미터의 유형에 따라 다른 확률 밀도를 추정하여 사전 모델 확률 (a priori model probability)로 사용한다. 이러한 확률 밀도 추정은 훈련 데이터를 사용한다. 제안하는 HBIC를 BIC 및 기존 연구와 비교한다.

1) 본 연구는 한국과학재단의 해외 Post-Doc. 프로그램의 지원을 받았음을 밝힙니다.

2. 은닉 마르코프 모델 (Hidden Markov Model)

은닉 마르코프 모델은 유한 상태 기계로, 그 안에 있는 상태는 은닉되어 있고, 다만 출력열이 관측될 뿐이다. 상태 간 전이 확률은 마르코프 프로세스를 따르며 출력 확률은 각 상태에 지정되어 있다. HMM은 다음과 같이 $\{S, A, B\}$로 표현될 수 있다.

- $S=\{ S_1,..., S_Q \}$, 총 상태 수가 Q개인 HMM 상태의 집합.
- $A=[a_{ij}]$, 상태 전이 확률 행렬.
- $B=\{b_i(x)\}$, 출력 확률 집합으로 $b_i(x)$는 다음과 같이 정의된 상태 S_i 에 연관된 확률이다.

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

여기에서 $N(x,\mu_{il}, \Sigma_{il})$ 는 정규분포이고, μ_{il}은 i-번째 상태의 l-번째 mixture의 평균이고, Σ_{il}은 공분산, ω_{il}는 가중치이다. 각 mixture에서는 D-차원의 특징벡터가 있고, 모든 상태에서 L개의 mixture가 있다고 가정한다.

- μ,Σ,ω 를 각각 모델 전체에 대한 평균 벡터, 공분산 행렬, mixture 가중치라 하고, μ_i, Σ_i, 그리고 ω_i를 각각 특정 상태 S_i에 대한 평균 벡터, 공분산 행렬, mixture 가중치라고 정의한다.

2.1 HMM 구조

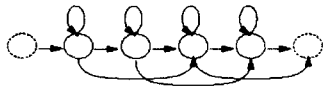

그림 1. 6-상태 HMM

온라인 필기 인식 시스템은 일반적으로 left-to-right HMM을 사용한다. 그림 1은 본 논문에서 사용한 연속 밀도 left-to-right HMM 구조의 예이다. 두 종류의 상태가 있는데, 출력 확률이 연관된 상태(실선으로 표시)와 출력 확률이 연관되지 않은 상태(점선으로 표시)로 나눠진다. 출력 확률이 연관되지 않은 상태는 시작 상태와 종료 상태뿐이다. HMM은 구조 M과 주어진 M에 대한 파라미터 θ로 특징 지워질 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 $\theta=\{A,\mu,\Sigma,\omega\}$ 의 합집합으로 볼 수 있다.

3. 베이지안 모델 선택 (Bayesian Model Selection)

베이지안 구조(Bayesian Framework)에서는 모델 선택이 다음과 같은 구조 M을 선택함으로서 이루어진다.

$$\begin{aligned}\hat{M} &= \arg\max_M P(M|X) \\ &= \arg\max_M P(M)P(X|M)\end{aligned} \quad (2)$$

최적 모델의 선택은 주어진 데이터 X에 대해 가장 높은 결합 확률값 $P(M,X)=P(M)P(X|M)$을 갖는 구조를 선택하는 것이다. 최적 구조는 특정한 구조에 대한 선호도를 나타내는 모델 구조의 사전 확률 $P(M)$과 주어진 데이터 X에 대한 모델 구조 M의 확률의 곱을 계산함으로써 찾을 수 있다. 후자를 증거라고도 부른다. 사전 확률과 증거를 둘 다 사용하는 것은 동등하게 가능성 있는 모델이 주어졌을 때 단순한 모델을 선호하는 Occam의 razor 원칙을 구현하는 것이다[5,6].

베이지안 모델 선택에서의 통상적인 방법은 모델 구조에 대한 사전 확률 $P(M)$을 무시하는 것이다. 다시 말하면 모든 구조에 대해 동일한 사전 확률 분포를 가정한다. 그리고 증거인 $P(X|M)$만을 모델 선택의 유일한 기준으로 삼는다. $P(X|M)$는 다음 수식과 같이 모든 파라미터 집합에 대한 적분을 계산함으로써 구해진다.

$$p(X|M)=\int p(X|M,\theta)\,p(\theta|M)\,d\theta \quad (3)$$

$$=\int g(\theta)\,d\theta \quad (4)$$

베이지안 적분에서의 주요 문제는 식 (3)의 적분을 계산하는 것이다. 이 적분은 구조가 복잡할 경우 거의 계산 불가능하게 되고, 수치해석적 방법에 의해 계산하거나 근사치를 구하는 방법을 사용해야 한다. 본 논문에서는 후자의 방법을 택하였다.

3.1 라플라스 근사 (Laplacian Approximation)

적분 근사에 대한 라플라스의 방법은 식 (3)의 적분을 계산하는데 널리 사용되어 왔다[5,7]. 다음 함수

$$g(\theta)=p(X|M,\theta)\,p(\theta|M) \quad (5)$$

가 가장 가능성 있는 파라미터 집합 θ_{MP}에서 피크를 보인다고 가정하면 증거 $p(X|M)$는 이 함수의 최대치 주위에서의 테일러 확장 (Taylor Expansion)에 의해 다음과 같은 다루기 쉬운 형태로 근사될 수 있다.

$$\begin{aligned}p(X|M) \approx\ & p(X|M,\theta_{MP}) \\ & p(\theta_{MP}|M)\,(2\pi)^{\frac{k}{2}}\det(A)^{-\frac{1}{2}}\end{aligned} \quad (6)$$

여기에서 k는 모델의 자유 파라미터 개수이고 $A=-\nabla^2\log P(\theta|X,M)\,|_{\theta=\theta_{MP}}$ 이다.

N이 커짐에 따라 $\det(A)$는 $N^k\det(I)$에 근접한다. I는 하나의 관측에 대한 Fisher 정보 행렬이고, N은 데이터 집합의 크기이다. 함수 $g(\theta)$가 확률 항 $p(X|\theta)$에 의해 크게 좌우되기 때문에 θ_{MP}는 최대 우도 추정치 θ_{ML}에 의해 근사 될 수 있다. 이러한 조건과 함께 식 (6)에 log를 취하면 증거는 다음과 같이 근사 된다.

$$\begin{aligned}\log p(X|M) \approx\ & \log p(X|\theta_{ML}) + \log p(\theta_{ML}|M) \\ & + \frac{k}{2}\log(2\pi) - \frac{k}{2}\log N - \frac{1}{2}\log(\det(I))\end{aligned} \quad (7)$$

3.2 BIC(Bayesian Information Criterion)

BIC는 식 (7)에 대한 점근 근사(Asymtotic Approximation)를 가정한다. Central Limit 정리에 의해, 파라미터의 사전 확률 $p(\theta_{ML}|M)$은 평균 θ_{ML}과 공분산 I^{-1}을 갖는 다변량 정규 밀도로 간주될 수 있다. 이러한 조건은 다음과 같이 정의된 널리 알려진 베이지안 정보 기준으로 인도한다.

$$BIC(M)=\log p(X|\theta_{ML})-\frac{k}{2}\log N \quad (8)$$

위 식에서 BIC는 우도와 $\frac{k}{2}\log N$의 합인데, 후자는 모델 내의 파라미터 개수에 대한 페널티(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.

4. HBIC (HMM-Oriented BIC)

HMM 상황에 알맞은 선택 기준을 도출하기 위해 BIC로부터 출발하여 다음과 같은 과정을 거친다. 첫째, 모델 구조에 대한 사전 확률 $P(M)$에 대해 설명한다. 둘째, 식 (7)에서 Fisher의 정보 행렬 항인 $\log(\det(I))$이 단일 관측열에 의존하고 대규모 데이터 집합에서는 $\log(N)$ 항에 의해 좌우되기 때문에 무시될 수 있다. 모델 구조 사전 확률 $P(M)$을 사용하여, HBIC는 다음과 같이 정의된다.

$$\begin{aligned}HBIC(M) =\ & \log p(X|\theta_{ML}) + \log p(\theta_{ML}|M) \\ & + \frac{k}{2}\log(2\pi) - \frac{k}{2}\log N + \frac{1}{2}\log P(M) \\ =\ & BIC(M) + \log p(M) \\ & + \log p(\theta_{ML}|M) + \frac{k}{2}(2\pi)\end{aligned} \quad (9)$$

HBIC는 BIC 선택 기준에다 HMM의 복잡한 구조를 반영하는데 보다 적합한 항의 합으로 구성된다. $P(M)$을 계산하는데 HMM 구조를 이용함으로써, HBIC는 HMM 구조 추정에 보다 적합한 선택 기준이 될 수 있다.

4.1 $P(M)$의 추정

상태의 개수 Q는 상태 당 mixture의 개수 L과 독립적이라는 가정을 한다. 또한 모델 내의 모든 상태는 동일한 mixture 개수를 갖는다고 가정한다. 그러면 다음 식이 성립된다.

$$P(M)=P(Q)P(L) \quad (10)$$

모델 구조 사전 확률 $P(M)$은 $P(Q)$와 $P(L)$를 따로 추정함

으로써 얻어진다.

4.2 Estimatig $p(\theta_{ML} \mid M)$

HMM 구조가 선택된 후, 다음 단계는 $p(\theta_{ML}|M)$을 추정하는 것이다. 전이 확률 행렬, 가중치, 평균, 공분산이 서로 통계적으로 독립적이라고 가정하면 다음 식을 얻을 수 있다.

$$\begin{aligned}\log p(\theta_{ML}|M) &= \log p(A|M) + \log p(\omega|M) \\ &\quad + \log p(\mu \mid M) + \log p(\Sigma \mid M) \\ &= \sum_i \log p(a_i \mid M) + \sum_i \log p(\omega_i \mid M) \\ &\quad + \sum_i \log p(\mu_i \mid M) + \sum_i \log p(\Sigma_i \mid M)\end{aligned} \tag{11}$$

위 식을 이용하면 사전 확률 $p(\theta_{ML}|M)$의 추정은 각 유형별로 사전 확률을 추정한 후 그것을 모두 합하면 얻을 수 있다.

5. 실험 및 결과 분석

5.1 과제 및 데이터베이스

본 논문에서는 UNIPEN 데이터를 실험 대상으로 삼았다. UNIPEN은 온라인 문자 인식에 관계된 세계 각국의 대학, 연구소, 기업 등 다양한 기관들이 공통의 파일 표준을 만들어 필기 데이터를 모아 놓은 것인데, 그 중 train_r01_v07을 사용하였다. UNIPEN 데이터에는 숫자, 영대소문자, 부호 등이 있지만, 숫자만을 실험 대상으로 하였다. 382명의 필기자로부터 수집한 9436개의 샘플을 훈련 집합으로 사용했고, 131 필기자로부터 얻은 3296개의 샘플을 교차 검증 집합으로 사용했다. 그리고 121 필기자로부터 획득한 3221개의 샘플을 테스트 데이터로 사용했다.

숫자만을 실험 대상으로 했기 때문에 모두 10개의 클래스가 있고, 필기 형태가 서로 상이한 것은 별도의 allograph로 만들어 총 25개의 allograph를 생성하였다. 필기 데이터를 먼저 크기 정규화한 후 각 획에서 특징점을 찾고 특징점 사이의 데이터에 대해 특징을 추출하여 PCA(Principal Component Analysis) 과정을 거쳐 총 9 차원 벡터를 생성했다.

5.2 사전 확률 추정 결과

HBIC를 구현하는 데에는 모델 구조와 구조에 따르는 파라미터를 추정하는 것이 요구된다. 본 논문에서는 사전 확률 분포를 추정함에 있어서 훈련 데이터를 다음과 같이 사용하였다.

5.2.1 모델 구조

상태 수에 대한 사전 확률 분포 $P(Q)$를 추정할 때, 각 allograph에 대한 훈련 집합에서 입력 프레임 수의 평균을 구해 평균에 피크가 되는 Beta 분포를 $P(Q)$로 사용하였다. 그 이유는 기존 연구에서 HMM의 상태 수의 추정에 입력 프레임 수의 평균 또는 최빈수를 사용하여 어느 정도의 성능을 보이는 데에 착안하였다. Mixture 개수에 대한 사전 확률 분포는 이미 훈련된 모델 중에서 최대 우도 기준에 의해 선택한 HMM에서 mixture 개수를 추출하여 그것에 가장 근사한 Beta 분포를 찾았다. 본 실험에서 확률 분포 유형은 정규 분포를 포함하여 총 7개를 고려 대상으로 삼았다.

5.2.2 모델 파라미터

HMM 파라미터 유형별로 5.2.1절에서 언급한 7개의 확률 분포를 대상으로 가장 근사하는 확률 분포를 찾았다. 확률 분포 차이 계산은 Anderson-Darling 테스트를 사용하였다[8].

- 전이 확률 : 다음 상태로의 전이에는 Beta 분포가, 셀프 루프와 한 상태를 건너뛰는 전이에 대해서는 Gamma 분포가 적합함을 알 수 있었다.
- 평균 : 삼각 분포, 정규 분포, Gamma 분포, Logistic 분포, Weibull 분포 등 벡터의 각 차원에 대해 다양한 확률 분포로 근사되었다.
- 공분산 : 공분산 행렬은 대각 행렬만 허용하게 한 후 확률 분포 근사를 한 결과 모든 차원에서 Gamma 분포가 가장 가까움을 알 수 있었다.
- Mixture 가중치 : Beta 분포로 근사되었다.

5.3 인식 결과

표 1은 각 모델 선택 기준에 따른 인식률 및 전체 모델의 상태 수, 전체 모델의 총 파라미터 수를 보여준다. 인식률의 차이는 크지 않았지만, BIC와 HBIC 둘 다 최대 우도 방법이나 입력 프레임의 평균치로 상태 수를 결정하는 휴리스틱 방법보다 파라미터 수가 현저히 감소했음을 알 수 있었다. 또한 HBIC가 BIC보다 파라미터 수의 감소가 현격함을 알 수 있었다. 하지만 HBIC가 지나치게 파라미터 수를 감소시켜 다른 기준에 비해 인식률이 약간 저하되는 현상을 보였다.

표 1. 인식률 비교

모델 선택 기준	인식률(%)	상태 수	파라미터 수
최대 우도 (ML)	91.77	207	35140
입력 프레임 평균	91.83	221	37861
BIC	91.93	185	20596
HBIC	91.06	182	15761

6. 결 론

본 논문에서는 베이지안 정보 기준 BIC와 그것을 HMM의 복잡한 구조에 맞도록 사전 확률 추정을 통해 개선한 HBIC에 대해 필기 문자를 대상으로 실험하였다. HBIC가 BIC에 비해 모델 파라미터 수의 감소에는 더 효과적이었지만, 다소 인식률이 저하되는 문제가 남아 있다. 보다 정확한 사전 확률의 추정 및 다양한 데이터 집합에 대해 실험해 보는 것이 향후 연구 과제로 남아 있다.

참 고 문 헌

[1] H.Singer and M. Ostendorf, "Maximum likelihood successive state splitting", in *ICASSP*, pp.601-604, 1996.

[2] Andress Stolcke and stephen Omohundro, "Hidden Markov Model induction by bayesian model merging", in Advances in NIPS, vol. 5, pp.11-18, 1993.

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

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

[5] G. Schwarz, "Estimation the dimension of m model," Ann.Statist., vol.6, no2, pp. 461-464, 1978.

[6] A. Barron, J. Rissanen, and B. Yu, "The minimum description length principle in coding and modeling," IEEE Trans. Inform. Theory, vol. 44, pp. 2743-2760, 1998.

[7] J.Olivier and Baxter R,"MML and Bayesianism" Similarities and differences (Introduction to minimum encoding inference - Part II", Technical Report 206, Monash University, Australia, 1994.

[8] M.A.Stephens, "EDF statistics for goodness of fit and some comparisons", Journal of the American Statistical Association, vol 69. pp.730-737, 1974.