HBIC와 BIC_Anti 기준을 이용한 HMM 구조의 최적화

(HMM Topology Optimization using HBIC and BIC_Anti Criteria)

박 미 나 † 하 진 영 ††
(Mi-Na Park) (Jin-Young Ha)

요 약 본 논문에서는 연속 밀도 HMM 구조의 최적화 문제를 다룬다. HMM 구조의 최적화를 위해 여러 연구가 있었는데, 그 중에서도 잘 알려진 BIC(Bayesian Information Criterion)등과 같이 이미 제안된 모델 선택 기준은 동질의 파라미터를 갖는 데이타에 대해 통계적으로 잘 행동하는 모델을 가정하고 있어서 연속 밀도 HMM 등과 같이 복잡한 파라미터를 갖는 구조에는 적합하지 않고, 파라미터 수를 줄이는 데는 어느 정도 효과가 있었으나 인식률 향상에 있어서는 한계를 보였다.

이에 본 논문에서는 HMM의 파라미터 유형에 따라 별도의 확률 밀도를 추정하여 사전 모델 확률(a priori model probability)로 사용하는 모델 선택 기준인 HBIC(HMM-oriented BIC)를 제안했다. 또한 HMM의 변별력을 높이기 위해 변별력 특성을 갖는 안티확률을 BIC와 결합한 새로운 모델 선택 기준인 BIC_Anti를 제안했다. 제안한 모델 선택 기준의 유용성을 검증하기 위해 온라인 필기 데이타를 대상으로 실험하여 기존의 연구와 비교하였다. 그 결과 제안한 HBIC와 BIC_Anti 모델 선택 기준을 사용하는 것이 BIC를 사용하는 것보다 더 적은 파라미터 수로도 향상된 인식률을 얻을 수 있음을 확인했다.

키워드 : 은닉마르코프모델, HBIC, 안티확률, BIC_Anti, 구조최적화

Abstract This paper concerns continuous density HMM topology optimization. There have been several researches for HMM topology optimization. BIC (Bayesian Information Criterion) is one of the well known optimization criteria, which assumes statistically well behaved homogeneous model parameters. HMMs, however, are composed of several different kind of parameters to accommodate complex topology, thus BIC's assumption does not hold true for HMMs. Even though BIC reduced the total number of parameters of HMMs, it could not improve the recognition rates.

In this paper, we proposed two new model selection criteria, HBIC (HMM-oriented BIC) and BIC_Anti. The former is proposed to improve BIC by estimating model priors separately. The latter is to combine BIC and anti-likelihood to accelerate discrimination power of HMMs. We performed some comparative research on couple of model selection criteria for online handwriting data recognition. We got better recognition results with fewer number of parameters.

Key words : HMM, BIC, HBIC, Anti-likelihood, BIC_Anti, Topology optimization

1. 서 론

필기인식을 위해 시도되어 온 다양한 방법론 중에서 1980년대 말부터 현재까지 은닉 마르코프 모델(Hidden Markov Model: 이하 HMM)은 시간에 따라 변해 가는 특성을 지닌 데이타에 대한 높은 모델링 능력과 파라미터 수가 많을수록 모델링이 잘되는 특성 때문에 음성인식과 온라인 필기 인식에서 우수한 성능을 보여 왔다 [1,2].

지금까지 소개된 실용화 단계의 시스템들은 필기에 많은 제약을 두어 필기에 융통성을 두기 어렵고, 인식 대상의 단어수가 커질 경우 단어 모델 생성의 어려움과 인식 속도의 문제로 인해 큰 관심을 끌지 못했다. 따라서 인식률과 인식속도 향상, 메모리 문제 해결 등 시스

· 본 논문은 한국과학재단의 박사후해외연수지원사업과 강원대학교 BK21 사업단의 지원을 받았음을 밝힙니다.
† 비 회 원 : 강원대학교 컴퓨터정보통신공학과
mnpark@kwnu.kangwon.ac.kr
†† 정 회 원 : 강원대학교 전기전자정보통신공학부 교수
jyha@cc.kangwon.ac.kr
논문접수 : 2002년 5월 16일
심사완료 : 2003년 4월 9일

템의 최적화에 대한 연구가 절실히 필요하다.

온라인 필기인식에서 많이 사용되는 HMM은 left-to-right HMM으로 이 모델 구조는 상태 수와 상태 당 믹스쳐 수, 그리고 상태 사이의 전이 확률에 의해 결정된다. 이러한 HMM의 구조는 휴리스틱 방법에 의해 결정되는 것이 일반적이기 때문에 최적의 모델을 선택하는데 어려움이 있다. 여기서 최적 모델이란 최소한의 모델 파라미터로서 최소의 오류를 허용하는 것을 의미한다. HMM 구조의 최적화를 위한 기존 연구는 다양한 방법으로 진행되어 왔다[3-9]. 높은 점유를 갖는 상태부터 순차적으로 분할해서 점차 상태 수를 증가시키는 방법[3]과 Dirichlet 사전 확률에 기반 한 사후 확률을 사용하여 복잡한 구조로부터 점차 구조를 감소시켜 나가는 방법이 있었다[4]. 또한 최대 확률 기준을 이용하거나 베이지안 정보기준(Bayesian Information Criterion: 이하 BIC)를 이용한 연구가 있었다[5,6].

모델 선택 시 가능한 한 적은 수의 인자를 갖는 것을 선호하는 Occam's Razor Principle에 기반을 둔 BIC를 이용한 연구는 모델의 파라미터 수를 줄이는 데에는 성공했으나 인식률 향상에서 한계를 보였다[10].

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[11].

Occam's Razor Principle은 가능한 후보 집합 중에서 작은 모델을 선택하는 것을 선호하는 것으로 모델 선택에 대한 기본 철학을 제공해준다. 모델 선택에 대한 베이지안 방법은 자연스럽게 이러한 원리를 구현할 수 있게 해준다. 하지만 Schwarz의 BIC등과 같이 이미 제안된 모델 선택 기준은, 동질의 파라미터를 갖는 통계적으로 잘 행동하는 모델을 가정하고 있어서, 연속 밀도 HMM 등과 같이 복잡한 구조에는 적합하지 않다[12].

본 논문에서는 파라미터의 유형에 따라 각각의 확률 밀도를 추정하여 사전 모델 확률(*a priori* model probability)로 사용하는 모델 선택 기준인 HBIC (HMM-oriented BIC)와 기존의 모델 선택 기준인 BIC와 데이타간의 변별력 특성을 가진 안티확률을 기반으로 한 모델 선택 기준 BIC_Anti를 필기 데이타를 대상으로 실험하여 기존의 연구와 비교한다.

2. 은닉 마르코프 모델(Hidden Markov Model)

은닉마르코프 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. HMM은 유한상태 기계로 그 안에 있는 상태는 은닉되어 있고 단지 출력 열이 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-번째 상태의 l-번째 믹스쳐의 평균이고, $\sum_{il}$은 공분산, w_{il}는 가중치이다. 각 믹스쳐에서는 D-차원의 특징 벡터가 있고, 모든 상태에서 L개의 믹스쳐가 있다고 가정한다.

$\mu, \sum, w$를 각각 모델 전체에 대한 평균벡터, 공분산 행렬, 가중치라 하고, μ_i, $\sum_i$, 그리고 w_i를 각각 특정 상태 S_i에 대한 평균 벡터, 공분산 행렬, 믹스쳐 가중치라고 정의한다[11].

2.1 HMM 구조

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 사이의 전이 구조는 고정되어 있기 때문에 M은 상태 수와 상태 당 믹스쳐 수로 유일하게 결정된다. 따라서 모델을 모델 구조 $M=\{Q,L\}$과 파라미터 $\theta=\{A,\mu,\Sigma,\omega\}$의 합집합으로 볼 수 있다[13].

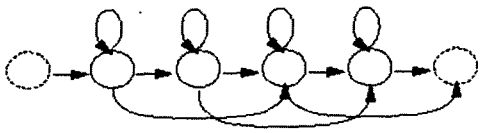


그림 1 6-상태 HMM

2.2 HTK(Hidden Markov Model Toolkit)

은닉 마르코프 툴킷(Hidden Markov ToolKit : 이하 HTK)는 HMM의 훈련과 인식 실험을 위한 HMM 기

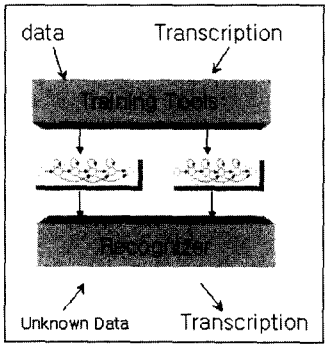


그림 2 HTK의 two stage

반의 음성 인식 처리 툴이다[14]. HTK는 1989년 Cambridge University Engineering Department의 Steve Young이 처음 버전을 만들었으며 1995년 Entropic Cambridge Research Laboratory(ECRL)가 설립되면서 상용화되기 시작했다. 현재 버전은 3.2까지 나왔으며 본 논문에서 사용된 HTK version은 3.0이다.

HTK는 훈련 툴과 인식기 두 단계로 나누어 데이타를 처리한다. 훈련 툴은 훈련 데이타를 이용하여 HMM 집합의 파라미터를 추정하며, 인식기는 알려지지 않은 데이타를 인식하는 단계이다. HTK을 이용하여 데이타를 훈련하고 파라미터 수와 인식률을 추정한다.

3. 관련연구

3.1 ML(Maximum Likelihood) 기준

최대 우도(Maximum Likelihood)의 일반적인 방법은 파라미터 공간 A에 속하는 파라미터 a에 속한 집합 S 안의 관찰된 임의의 변수 X의 밀도 함수를 $f(X|a)$로 정의한다. 여기서 우도(likelihood)함수 L은

$$L(a|X)=f(X|a) \quad a \in A \text{ and } X \in S$$

로 정의할 수 있다. 최대 우도방법은 우도 함수 $L(a|X)$의 값이 최대가 되는 파라미터 a의 평균값 $\mu(x)$를 구하는 것이다. $\mu(x)$은 파라미터 a의 최대 우도 추정이 된다. 이 방법은 직관적인 방법으로 관찰된 데이타를 가장 사실적으로 만드는 파라미터의 값을 찾는 것이다. $L(a|X)$의 최대 값은, 자연로그함수 ln이 증가할 때 이 값이 존재한다면 $\ln[L(a|X)]$의 최대 값이다. $\ln[L(a|X)]$ 함수는 로그 우도(log likelihood)함수라고 불린다. 이 함수는 많은 경우에 우도 함수보다 더 쉽게 연산 할 수 있기 때문에 많이 쓰인다. 이 방법은 학습데이타의 양이 충분하다는 조건 하에서 다른 어느 추정방법 못지않은 성능을 보인다.

3.2 베이지안 모델 선택 (Bayesian Model Selection)

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

$$\begin{aligned}\hat{M} &= \arg\max_{M} P(M|X) \\ &= \arg\max_{M} P(M)P(X|M)\end{aligned} \tag{1}$$

최적 모델의 선택은 주어진 데이타 X에 대해 가장 높은 결합 확률 값 $P(M,X)=P(M)P(X|M)$을 갖는 구조를 선택하는 것이다. 최적 구조는 특정한 구조에 대한 선호도를 나타내는 모델 구조의 사전 확률 $P(M)$과 주어진 데이타 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 Occam's Razor Principle을 구현하는 것이다[15].

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[13]. $P(X|M)$는 다음 수식과 같이 모든 파라미터 집합에 대한 적분을 계산함으로써 구해진다.

$$p(X|M)=\int p(X|M,\theta)\,p(\theta|M)\,d\theta \tag{2}$$

$$=\int g(\theta)\,d\theta \tag{3}$$

베이지안 적분에서의 주요 문제는 식 (2)의 적분을 계산하는 것이다. 이 적분은 구조가 복잡할 경우 거의 계산 불가능하게 되고, 수치 해석적 방법[16]에 의해 계산하거나 근사치를 구하는 방법을 사용해야 한다. 본 논문에서는 후자의 방법을 택하였다.

3.2.1 라플라스 근사 (Laplacian Approximation)

적분 근사에 대한 라플라스의 방법은 식 (2)의 적분을 계산하는데 널리 사용되어 왔다[6,17]. 다음 함수

$$g(\theta)=p(X|M,\theta)\,p(\theta|M) \tag{4}$$

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.

$$p(X|M) \approx p(X|M,\theta_{MP})\,p(\theta_{MP}|M)\,(2\pi)^{\frac{k}{2}}\det(A)^{-\frac{1}{2}} \tag{5}$$

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 데이타 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 근사될 수 있다. 이러한 조건과 함께 식 (5)에 로그(log)를 취하면 증거는 다음과 같이 근사된다.

$$\log p(X|M) \approx \log p(X|\theta_{ML}) + \log p(\theta_{ML}|M) + \frac{k}{2}\log(2\pi) - \frac{k}{2}\log N - \frac{1}{2}\log(\det(\mathrm{I})) \quad (6)$$

3.2.2 BIC(Bayesian Information Criterion)

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

$$BIC(M) = \log p(X|\theta_{ML}) - \alpha\frac{k}{2}\log N \quad (7)$$

위 식에서 BIC는 우도와 $\frac{k}{2}\log N$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].

4. HMM 특성을 이용한 모델 선택 기준

4.1 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 $\log(\det(\mathrm{I}))$이 단일 관측 열에 의존하고 대규모 데이타 집합에서는 $\log(N)$ 항에 의해 좌우되기 때문에 무시될 수 있다. 모델 구조 사전 확률 $P(M)$을 사용하여, HBIC는 다음과 같이 정의된다.

$$\begin{aligned} HBIC(M) &= \log p(X \mid \theta_{ML}) + \log p(\theta_{ML} \mid M) \\ &\quad + \frac{k}{2}\log(2\pi) - \frac{k}{2}\log N + \frac{1}{2}\log P(M) \\ &= BIC(M) + \log p(M) \\ &\quad + \log p(\theta_{ML} \mid M) + \frac{k}{2}(2\pi) \end{aligned} \quad (8)$$

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.

4.1.1 $P(M)$의 추정

상태의 개수 Q는 상태 당 믹스처 개수 L과 독립적이라는 가정을 한다. 또한 모델 내의 모든 상태는 동일한 믹스처 개수를 갖는다고 가정한다. 그러면 다음과 같이 식이 성립된다.

$$P(M) = P(Q)P(L) \quad (9)$$

모델 구조 사전 확률 $P(M)$은 $P(Q)$와 $P(L)$를 따로 추정함으로써 얻어진다.

4.1.2 $p(\theta_{ML} \mid M)$의 추정

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.

$$\begin{aligned} \log p(\theta_{ML}|M) &= \log p(A|M) + \log p(\omega|M) \\ &\quad + \log p(\mu \mid M) + \log p(\Sigma \mid M) \\ &= \sum_i \log p(a_i \mid M) + \sum_i \log p(\omega_i \mid M) \\ &\quad + \sum_i \log p(\mu_i \mid M) + \sum_i \log p(\Sigma_i \mid M) \end{aligned} \quad (10)$$

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

4.2 안티 확률 추정

HMM은 파라미터의 수가 많아질수록 해당 클래스 데이타에 대한 확률 값은 높아진다. 이에 HMM은 해당 클래스 이외에 다른 클래스 데이타에 대해서는 어떤 확률 값을 갖는지, 데이타에 대해 변별력을 갖고 있는지 알아보기 위해 해당 클래스 이외에 다른 클래스 데이타에 대해서 실험한다.

해당 클래스 외의 다른 클래스 데이타의 집합은 해당 데이타를 제외한 나머지 클래스의 데이타를 모두 한 데이타로 병합한다. 이 때 그 크기가 너무 크게 되므로 이를 본래의 해당 클래스 데이타의 크기와 비례하도록, 한 클래스 당 데이타를 랜덤하게 선별하여 적절한 크기의 안티 데이타를 만든다.

안티 데이타의 확률은

$$Likelihood_{Anti} = \sum_{X' \notin C} \log p(X' \mid \theta_{ML}) \quad (11)$$

이 된다.

본 논문에서는 다른 데이타에 대해 증가하는 확률을 억제하는 방법으로 안티 확률 방법을 제안하였다. 해당 클래스 데이타에 대해서는 잘 모델링되고 그 외의 데이타에 대해서는 모델링되지 않는 적당한 파라미터 수를 구하기 위해서 데이타에 대해 변별력을 가지는 안티 확률 기준은 자기 자신의 데이타에 대한 확률과 안티 데이타에 대한 기준 확률의 차를 모델 선택 기준으로 사용하였다.

$$\begin{aligned} Anti_c &= \sum_{X \in C} \log p(X|\theta_{ML}) \\ &\quad - \alpha \times \sum_{X' \notin C} \log p(X'|\theta_{ML}) \end{aligned} \quad (12)$$

(단, C는 해당클래스)

전체 파라미터 수(N)는 다음과 같이 표현된다.

$$N = (Ns-2)(3+Nm(2\times 9+1)) - 1 \quad (13)$$

(N : Total Number of Parameters

Ns : Number of states

Nm : Number of mixtures)

위 (13)식에서 상태 수와 믹스쳐 수를 조정하여 전체 파라미터 수를 기존의 모델 선택 기준보다 최소화한다. 그런데 이 안티 기준은 실험에 나타난 것처럼 상태 수는 증가하고 믹스쳐 수 감소를 선호하므로 상태 수 보다는 믹스쳐 수를 조정하여 전체 파라미터 수를 감소한다.

4.3 BIC_Anti 확률 기준

안티 확률 기준은 데이타에 대해 변별력을 이용하여 기존의 모델 선택 기준 중에 파라미터의 수를 줄이는데 성공적인 연구 결과를 보인 BIC와 결합하여 좀더 향상된 인식률을 구하고자 한다.

BIC의 패널티 항에 안티 확률 기준을 합하고 임의의 α와 β를 적용한다.

$$BIC\_Anti_c = \sum_{X \in C} \log p(X \mid \theta_{ML}) - \alpha \times (\beta \times \sum_{X \notin C} \log p(X' \mid \theta_{ML}) + (1-\beta)\frac{k}{2}\log N) \quad (14)$$

위 (12)식에서 임의의 α는 BIC식의 패널티 항의 α와 같은 역할을 하게 된다. 패널티 항에 값이 증가하면 파라미터 수가 줄어들고 인식률도 저하된다. 그러므로 패널티 값을 조정하여 총 파라미터 수가 감소되는 것은 큰 의의가 없다.

또한 β는 안티 확률기준이 적용되는데 안티 기준 값이 증가할수록 안티 기준의 특성 때문에 인식률이 저하되므로 이것 또한 파라미터의 수를 최적화하거나 인식률을 증가시키는 데는 아무런 의의가 없다. 따라서 파라미터 수를 감소시키고 인식률의 향상에 영향을 주기 위해 가능한 작은 값의 α와 β를 적용해야 한다.

4.4 모델 선택 처리 절차

실험에서 사용할 데이타 집합을 HMM 훈련을 위해 훈련 데이타 집합과 교차검증을 위한 held-out 집합, 테스트를 위한 테스트집합으로 가정하였다. 모델 선택 절차는 첫 번째로 훈련 데이타 집합으로 다양한 HMMs 구조를 상태 수와 상태 당 믹스쳐 수를 변화시키면서 훈련시킨다. 그리고 held-out 데이타 집합을 사용하여 훈련 된 HMM의 우도를 계산한다. 그리고 각 선택 기준에서 가장 좋은 HMM 구조를 선택한다.

5. 실험 및 결과

5.1 실험 데이타베이스

본 논문에서는 UNIPEN 데이타를 실험 대상으로 삼았다. UNIPEN은 온라인 문자 인식에 관계된 세계 각국의 대학, 연구소, 기업 등 다양한 기관들이 공통의 파일 표준을 만들어 필기 데이타를 모아 놓은 것으로써, 그 중 train_r01_v07을 사용하였다.

이 데이타 집합t(train_r01_v07)은 93개의 필기 데이타 특성을 포함하고 있으며 숫자, 영 대소문자, 부호 등과 같은 93개의 필기 데이타의 특성을 포함하고 있는 93개의 클래스로 분류하였다. 이 데이타 기준을 세 개로 나누어 훈련 데이타로 66,896개, 교차 검증 데이타로 31,101개, 그리고 테스트 데이타로 24,083개를 사용하였다. 필기 데이타를 먼저 크기 정규화한 후에 각 획에서 특징점을 찾고 특징점 사이의 데이타에 대해 특징을 추출하여 PCA(Principal Component Analysis) 알고리즘을 이용하여 총 9차원 벡터를 생성하였다.

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 데이타를 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. 믹스쳐 개수에 대한 사전 확률 분포는 이미 훈련된 모델 중에서 최대 우도 기준에 의해 선택한 HMM에서 믹스쳐 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

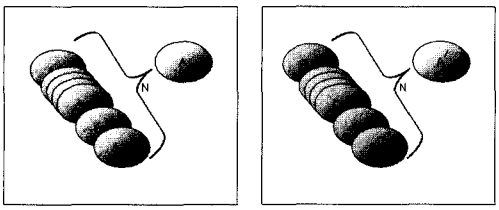
HMM 파라미터 유형별로 5.2.1절에서 언급한 7개의 확률 분포를 대상으로 가장 근사하는 확률 분포를 찾았다. 확률 분포 차이 계산은 Anderson-Darling 테스트를 사용하였다[18].

- 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.
- 믹스쳐 가중치 : Beta 분포로 근사되었다.

5.3 안티 데이타 생성

대상 데이타에 해당하는 데이타를 제외한 나머지 데이타를 병합하여 해당데이타에 대한 안티모델을 생성하였다.

하나의 안티모델은 대상 클래스를 제외한 나머지 데이타를 모두 병합한다. 모델의 크기가 본래의 데이타의 크기보다 일반적으로 커지게 되므로 대상 모델의 데이타양과의 균형을 맞추기 위해 대상 데이타양과 비례하게 데이타를 무작위로 추출하여 적절한 양의 안티 데이타를 생성한다. 새로 생성된 클래스는 대상 데이타 자기 자신의 모델은 제외하고 나머지 모델만 가지고 자기 자신의 모델을 만든다.



(i) 대상 데이타 (ii) Anti 데이타

그림 3 데이타

5.4 실험결과

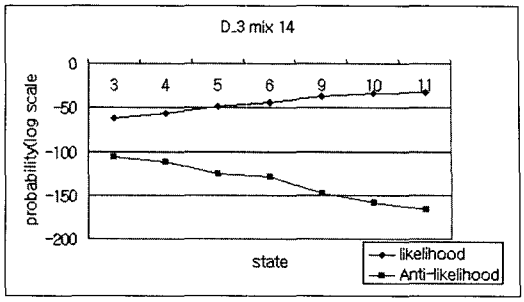
그림 4는 자기 자신의 데이타에 대한 확률과 자기 자신외의 데이타에 대한 확률을 믹스쳐와 상태 수에 따른 변화를 나타낸 그래프이다. 그림 4의 a)는 모델에 대해 믹스쳐 수를 일정하게 하고 상태수를 단계별로 증가시켜 실험한 결과그래프로 해당클래스 데이타에 대한 모델 확률 값은 증가하고 다른 클래스 데이타에 대한 모델 확률 값은 감소한다.

이는 해당 데이타에 대한 확률 값이 증가하고 다른 데이타에 대한 확률 값은 감소하므로 이는 해당데이타에 대한 변별력을 갖고 있음을 알 수 있으며, 그래프에서 보는 것처럼 상태의 수는 증가할수록 모델 확률이 좋음을 알 수 있었다.

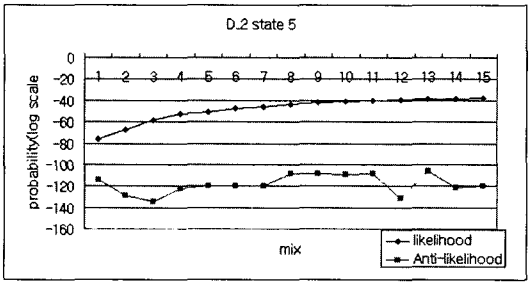
그림 b), c)는 모델에 대해 상태 수를 일정하게 하고 믹스쳐 수를 단계별로 증가시켜 실험한 결과그래프로 믹스쳐 수가 증가할수록 해당 클래스 데이타에 대한 모델 확률 값과 다른 클래스의 데이타에 대한 모델 확률 값이 모두 증가한다. 이것은 모든 데이타에 대해 좋은 확률 값을 가지므로 데이타에 대한 변별력을 가질 수 없으며 믹스쳐 수를 무조건 증가시키는 것은 모델의 변별력에 큰 도움이 되지 않는다는 것을 알 수 있었다.

그림 5는 각 모델 선택 기준에 대해 α에 따른 파라미터 수를 측정한 결과 그래프이다.

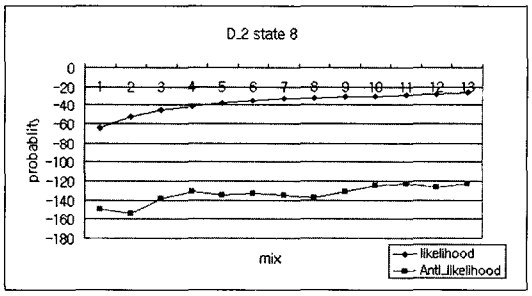
α가 0.1일 때, 본 논문에서 제안한 방법 BIC_Anti기준의 파라미터 수는 ML보다 약 6% 감소하였으며, BIC보다 약 1%감소하였고 사전확률을 이용한 HBIC의 파라미터 수는 ML보다 약 5%감소하였고 BIC 보다는 0.6% 감소하였다.



a) 모델 D_3 mix4, 상태 수에 따른 확률



b) 모델 D_2 state5, mix에 따른 확률



c) 모델 D_2 state8, mix에 따른 확률

그림 4 상태수와 믹스쳐 수에 따른 Anti 확률(log scale)

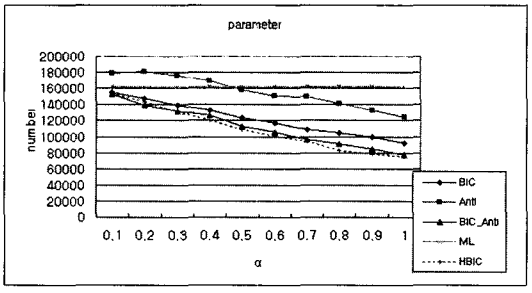
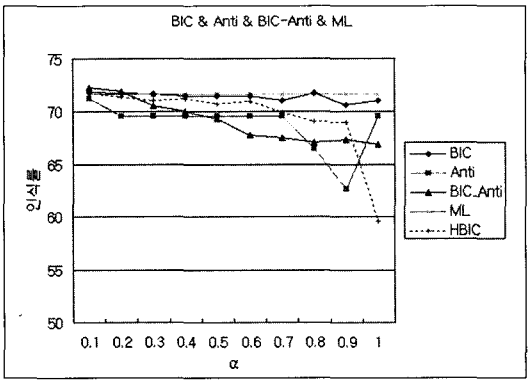


그림 5 BIC, Anti, BIC_Anti, ML, HBIC의 파라미터 수

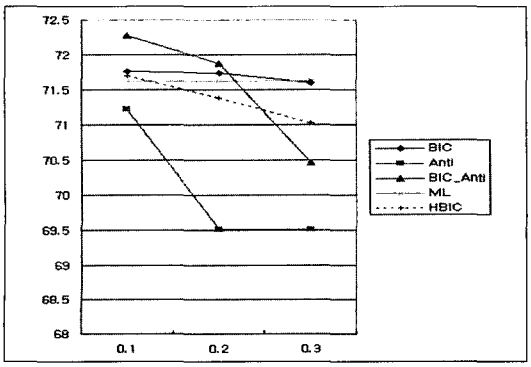
파라미터 수는 본 논문에서 제안한 BIC와 안티 확률 기준의 결합인 BIC_Anti 모델 선택 기준이 가장 적은 결과를 보인다. α가 커질수록 그 수가 차이가 나는 것은 BIC의 패널티 항에 α를 적용하였기 때문이다. α가

커지면 HBIC 모델 선택 기준의 파라미터 수가 더 적어지는 현상을 볼 수 있는데 이는 BIC의 α와 같은 의미로 강제적으로 파라미터 수를 줄이게 되는 것이므로 그다지 큰 의의가 없다. α가 증가할수록 파라미터 수가 감소하고(그림 5), 인식률(그림 6)도 감소하는 것을 볼 수가 있다. 따라서 BIC의 패널티 값이 파라미터 수의 감소뿐만 아니라 인식률 향상에 영향을 주는데 패널티 값을 너무 많이 주게 되면 파라미터 수가 많이 감소하지만 인식률도 감소하게 되므로 패널티 값은 가능한 작게 적용시켜 파라미터 수를 감소시키고 인식률은 감소되지 않도록 해야 한다.

그림 7은 BIC_Anti에 α와 β를 적용시킨 인식률 결과 그래프이다. 결과 그림에서 보면 α와 β가 커질수록 인식률이 떨어진다. 비록 그림 6, 7에서처럼 α를 증가시켜 파라미터 수가 감소되었어도 이처럼 인식률이 떨어지면 파라미터 수를 줄이는 것이 의의가 없으므로 파라미터 수가 감소하고 인식률이 증가하는 적당한 값을 찾아야 한다. 또 β값은 안티 기준 값의 적용 비율이 되는데 안티기준을 크게 적용하였을 경우 그 파라미터 수가 감소되지만 그 인식률 또한 감소하게 되므로 파라미터 수를 감소하고 동시에 인식률을 증가시키는 적당한 값을 찾아야 한다.



a) 전체 인식률



b) 인식률부분 확대

그림 6 ML, BIC, Anti, BIC_Anti의 인식률

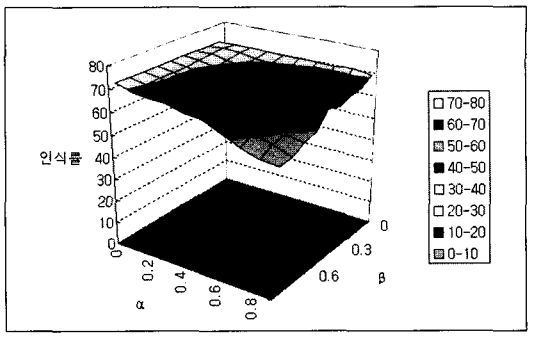


그림 7 BIC_Anti α, β를 적용한 인식률

α를 최소 값으로 적용하였을 때 각각의 모델 선택 기준에서 파라미터 수와 인식률을 비교하여 아래 표 1과 같이 정리하였다. BIC_Anti의 α, β의 값이 각각 0.1일 때 BIC_Anti기준의 파라미터 수는 ML보다 약 6% 감소하였으며, BIC보다 약 1%감소하였고 사전확률을 이용한 HBIC는 ML보다 약 5.43%감소하였고 BIC 보다는 0.68% 감소하였다. 인식률은 ML은 71.63%, BIC는 71.76%, BIC_Anti는 72.27%로 BIC_Anti가 ML보다는 약 0.64% 증가하였고 BIC 보다는 약 0.5% 증가하였다. 또 BIC_Anti의 α, β 값이 각각 0.1, 0.2일 때 파라미터 수는 ML보다 약 2.74% 감소하였고 , BIC보다 약 1.9% 감소하였으며 인식률은 ML보다 약 0.67% 증가하였고 BIC보다 약 0.54%증가하였다.

α, β 값이 각각 0.2, 0.1 일 때 파라미터 수는 ML은 162,239, BIC는 154,923, BIC_Anti는 139,208로 BIC_Anti가 ML보다는 약 16.5% 감소하였고, BIC보다는 약 10%감소하였다. 인식률은 ML은 71.63%, BIC는 71.76%, BIC_Anti는 71.88%로 ML보다는 약 0.25% 증가하였고, BIC보다는 약 0.11% 증가하였다.

표 1 α = 0.1 일 때 각 모델 선택 기준 비교

모델선택기준		파라미터 수	인식률(%)
ML		162,239	71.63
BIC	α=0.1	154,923	71.76
Anti(α = 0.1)		178,834	69.51
HBIC(α = 0.1)		153,876	71.70
BIC_Anti	α = 0.1, β = 0.1	153,306	72.27
	α = 0.1, β = 0.2	157,908	72.30
	α = 0.2, β = 01	139,208	71.88
	α = 0.3, β = 0.1	132,397	70.46

표 2 모델 선택 기준 간의 증·감 비교

모델선택기준		파라미터 수(%)		인식률(%)	
		ML	BIC	ML	BIC
HBIC(α = 0.1)		5%↓	0.6%↓	0.7%↑	0.6%↓
BIC_Anti	α = 0.1 β = 0.1	6%↓	1%↓	0.64%↑	0.5%↑
	α = 0.1 β = 0.2	2.74%↓	1.9%↓	0.67%↑	0.54%↑
	α = 0.2 β = 01	16.5%↓	10%↓	0.25%↑	0.11%↑

HBIC 모델 선택 기준은 파라미터 수가 153,876으로 ML보다 5.43%감소하였으며 BIC보다 0.68% 감소하였다. 인식률은 71.70%로 ML보다 0.07% 향상되었고 BIC보다는 약 0.06% 감소되었다.

본 논문에서 제안된 BIC_Anti 모델 선택 기준과 HBIC 모델 선택 기준은 기존의 선택 기준인 BIC보다 파라미터 수는 작게는 1%에서 크게는 16% 줄어들었고 인식률도 향상된 결과를 보였으며 사전 확률을 이용한 HBIC는 지나치게 파라미터 수를 감소시켜 다른 기준에 비해 인식률이 약간 저하되는 현상을 보였다.

본 논문의 실험결과에 의하면 BIC_Anti 전체 패널티는 0.2, 그리고 안티 기준은 0.1로 적용하였을 때 BIC보다 파라미터 수가 10% 감소하고 인식률은 0.17%증가하였다.

6. 결론 및 향후 과제

본 논문에서는 파라미터의 유형에 따라 다른 확률 밀도를 추정하여 사전 모델 확률(a priori model probability)로 사용하는 모델 선택 기준인 HBIC와, 파라미터의 수가 많을수록 모델링을 잘되는 특성을 가진 HMM에 모델간의 변별력을 주기 위해 BIC_Anti 기준을 제안하고 온라인 필기 데이타 집합인 UNIPEN 데이타에 대해 기존 모델 선택 기준인 BIC와 비교 실험하였다. 실험 결과 제안한 HBIC와 BIC_Anti 모델 선택 기준은 적은 수의 파라미터를 사용해도 BIC의 인식률을 상회하는 결과는 얻는데 성공하였다.

향후 연구 과제는 보다 다양한 데이타 집합에 대해 실험하여 인식률 향상 폭을 넓힘과 동시에 보다 최적화된 파라미터 수를 구하는 것이다.

참 고 문 헌

[1] 박미나, 하진영, "HMM 모델링을 위한 HMM의 State 수와 Mixture 수 분석", 한국정보과학회 춘계 학술발표논문집, 제29권 제1호, pp. 658-660, 2002년 4월.

[2] 하진영, A. Biem, J. Subrahmonia, 박미나, "모델의 사전 확률 추정을 통한 HMM 구조의 최적화", 한국정보과학회 추계 학술발표논문집, 제28권 제2호, pp. 325-327, 2001년 10월.

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

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

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

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

[7] S. Chen and P. S. Gopalakrishnan. "Clustering via the Bayesian Information criterion with applications in speech recognition," ICASSP, 2:645-649, 1998.

[8] Raymond C.Vasko, Jr., Amro EL-Jaroudi J.Robert, Boston "An Algorithm To determine hidden Markov Model Topology," IEEE, 3577-3579. 1996.

[9] Raymond C.Vasko, Jr., Amro EL-Jaroudi J.Robert, Boston "Application of Hidden Markov Model Topology Estimation To Repetitive Lifting Data," IEEE, pp. 4073-4076. 1997.

[10] A. Biem, J.-Y. Ha and J. Subrahmonia, "A Bayesian Model Selection Criterion for HMM Topology Optimization," ICASSP 2002. Orlando, Florida, U.S.A., pp. I-989 - I-992, May, 2002.

[11] Jin Young Ha, Alain Biem and Jayasheree Subrahmonia "Use of Model Prior for HMM Topology Optimization," The 4th Korea-China Joint Symposium on Information Technology for Oriental Language Processing and Pattern Recognition, Nov. 16-17, 2001.

[12] Schwarz, "Estimating the dimension of a model," Ann, Statist., Vol. 6, No. 2, pp.461-464, 1978.

[13] MacKay, "Bayesian interpolation," Neural computation, Vol. 4, No. 3, pp. 415-447, 1992.

[14] Steve Young, Dan KerShaw, Julian Odell, Dave ollason, Valtcho Valthchev, Phil Woodland, *The HTK Book(for HTK version 3.0)*, Microsoft corporation, 1999.

[15] 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.

[16] E. Kass and A. Raftery, "Bayes factors" Technical Report 254, University of Washington, Department of Statistics, 1994.

[17] 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.

[18] M.A. Stephens, "EDF statistics for goodness of fit and some comparisons," Journal of the American

Statistical Association, vol. 69. pp.730-737, 1974.

박 미 나

1999년 8월 강원대학교 컴퓨터공학과(학사). 2002년 8월 강원대학교 컴퓨터정보통신공학과(석사). 현재 동 대학원 박사과정 중, 관심분야는 패턴인식, 음성인식 등

하 진 영

1987년 2월 서울대학교 컴퓨터공학과(학사). 1989년 2월 한국과학기술원 전산학과(석사). 1994년 2월 한국과학기술원 전산학과(박사). 1994년 3월~1997년 2월 (주) 핸디소프트 기술연구소. 1997년 3월 ~현재 강원대학교 전기전자정보통신공학부 교수. 2000년 7월~2001년 7월 IBM T.J. Watson Research Center 방문 연구원. 관심분야는 패턴인식, HCI 등