필기 데이터 인식을 위한 HMM 구조 최적화 기준에 대한 분석

박미나°, 하진영
강원대학교 컴퓨터정보통신공학과
jyha@kangwon.ac.kr, atom77@mail.kangwon.ac.kr

Analysis of HMM Topology Optimization Criteria for Handwriting Recognition

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

요 약

음성인식과 온라인 필기인식에서 우수한 성능을 보이는 은닉 마르코프(HMM)의 HMM의 구조는 휴리스틱 한 방법에 의해 결정되는 것이 일반적이기 때문에 최적의 모델을 선택하는데 어려움이 있다. 이에 본 논문에서는 HMM의 구조를 체계적인 방법으로 정합과 동시에 변별력의 단점을 개선 할 수 있는 방법으로 Anti-likelihood를 이용한 모델간의 변별력을 살펴보고 최적의 모델 선택 기준인 BIC와의 결합하여, 체계적이고 효율적인 최적 모델 선택이 가능한 방법론에 대해 연구하고 필기데이터에 대해 검증한 결과, 기존의 방법보다 파라미터의 수는 감소되고 인식률이 향상됨을 알 수 있다.

1. 서 론

필기인식에 대한 연구는 지난 30년 동안의 많은 연구의 결실로 실용화에 많이 근접하여 다수의 시제품과 상업용 시스템이 소개되었는데 상당수의 시스템들이 필기에 많은 제약을 두어 필기에 융통성을 두기 어렵고, 간혹 단어 단위의 모델을 생성한 후 단어 단위의 인식을 시도한 연구도 있었지만, 인식 대상 단어수가 커질 경우 단어 모델 생성의 어려움과 인식 속도의 문제로 인해 큰 관심을 끌지 못했다[1]. 그러므로 인식률 향상과 인식속도 향상, 메모리 문제 해결 등, 시스템의 최적화에 대한 연구가 절실히 필요하다.

지금까지 필기인식을 위한 다양한 방법론 등이 시도되어 왔는데 1980년대 말부터 현재까지 은닉 마르코프 모델(Hidden Markov Model: 이하 HMM)이 가장 우수한 성능을 보여주고 있다[2]. 은닉 마르코프 모델(HMM)이 이와 같이 널리 쓰이는 이유는 음성과 문자 등에서 발견되는 많은 변형들을 흡수할 수 있고, 시간에 따라 변해 가는 특성을 지닌 Data를 잘 모델링하며 파라미터의 수가 클수록 모델링을 잘 하기 때문이다[3].

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

이러한 HMM의 구조는 휴리스틱 한 방법에 의해 결정되는 것이 일반적이기 때문에 최적의 모델을 선택하는데 어려움이 있다. 이 HMM의 구조를 어떻게 최적으로 정해야 하는 가에는 몇 가지 휴리스틱 한 방법이 제안되었을 뿐 체계적인 방법론이 거의 없었다. 그 중 한 방법으로 모델 선택 시 가능한 한 적은 수의 인자를 갖는 것을 선호하는 Occam's razor principle에 기반을 둔 Bayesian Information Criterion(BIC)를 이용한 시도가 있었지만 모델의 인자 수를 줄여주는 데에는 성공했으나 인식률을 향상시키는 데에는 한계를 보였다[4].

이에 본 논문에서는 HMM의 구조를 체계적인 방법으로 정합과 동시에 변별력의 단점을 개선 할 수 있는 방법을 제안하고자 한다.

모델 선택 시 많이 사용되는 ML(Maximum Likelihood Criterion)과 BIC(Bayesian Information Criterion)와 비교 분석하고 HMM 적용 시 해당 모델과 해당 데이터에 기반한 Likelihood 뿐만 아니라 그 모델에 대한 다른 클래스 데이터의 Anti-likelihood를 이용한 데이터간의 변별력을 살펴보고 최적의 모델 선택 기준인 BIC와의 결합하여, 체계적이고 효율적인 최적 모델 선택이 가능한 방법론에 대해 연구하고 필기데이터에 대해 검증한 결과, 기존의 방법보다 파라미터의 수는 감소되고 인식률이 향상됨을 알 수 있다.

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. 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, 공분산행렬, 가중치라 하고, $\mu_i, \sum_i$ 그리고 w_i를 각각 특정 상태 S_i에 대한 평균 벡터, 공분산 행렬, mixture 가중치라고 정의한다[3].

3. 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.

$$BIC(M) = \log p(X|\theta_{ML}) - \alpha\frac{k}{2}\log N \quad (4.1)$$

위 식에서 BIC는 likelihood와 $\frac{k}{2}\log N$의 합인데, 후자는 모델 내의 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].

4. Anti-likelihood

HMM은 parameter의 수가 많아질수록 data에 대한 확률 값이 높아진다. 이와 같은 특성 때문에 HMM은 해당 클래스의 data나 다른 클래스의 data에 대해서도 높은 확률 값을 보인다. 이는 HMM구조의 모델에 대한 변별력 부족 때문이다[2]. 이에 본 논문에서는 다른 클래스의 data에 대해 확률의 증가를 억제하기 위해서 Anti-Likelihood 방법을 제안하고 그 방법과 BIC와의 결합으로 HMM topology를 최적화 하고자 한다. 해당 클래스 data에 대해서는 잘 모델링 되고 그 외의 data에 대해서는 잘 모델링 되지 않는 적당한 파라미터를 구하기 위해 Model criterion이 data에 대해 변별력을 가지는 Anti-Likelihood를 사용한다. Anti-Likelihood는 Likelihood와 Anti Criterion Likelihood의 차를 모델 선택 기준으로 사용하였다

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

(단, C는 해당 클래스 *data*)

5. BIC-Anti likelihood

Model의 최적 선택 기준인 BIC에 data의 변별력을 가진 Anti-likelihood를 결합함으로써 최적화된 parameter의 수와 인식률을 구한다.

BIC의 penalty 항에 Anti-likelihood를 합하고 임의의 α와 β를 적용한다.

$$BIC-Anti_c = \sum_{X\in C}\log p(X|\theta_{ML}) - \alpha\times(\beta\times\sum_{X'\notin C}\log p(X'|\theta_{ML}) + (1-\beta)\frac{k}{2}\log N) \quad (5.1)$$

위 (5.1)식에서 임의의 α는 BIC식의 penalty 항의 α와 같은 역할을 하게 된다. 이는 penalty항에 값을 증가시켜 인위적으로 parameter의 수를 줄여 준다. 이는 순수하게 parameter를 줄이는 것이 아니므로 총 parameter의 수가 감소하여도 큰 의의가 없다.

또한 β는 Anti likelihood의 값을 적용하는데 Anti의 값이 크게 적용이 되면 실험결과에서 볼 수 있듯이 parameter의 수는 β는 커질수록 감소함을 알 수 있는데 이것 또한 순수하게 parameter의 수를 최적화하거나 인식률을 증가시키는데는 아무런 의의가 없다. 그러므로 가능한 한 작은 값의 α와 β를 적용하여 parameter의 수를 감소 시키는 것이 중요하다.

BIC의 식에 Anti-likelihood Criterion을 적용하면 state의 수는 증가한다. 만일 state의 수가 감소하면 인식률이 낮아지게 되므로 state의 수는 전체 parameter의 수와 인식률에 있어서 감소하는 것보다는 증가하는 것이 인식률 향상에 영향을 미치게 된다. 그러므로 state의 수보다는 mixture의 수를 감소시켜 전체 parameter의 수를 줄여야 한다. 따라서 state수는 전체 parameter의 수를 최적화하는데 크게 영향을 미치지 않음을 알 수 있다.

6. 실험 및 결과 분석

6.1 실험데이타베이스

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. 이 Data Set(train_r01_v07)은 93개의 필기 data 특성을 포함하고 있으며 숫자, 영 대소문자, 부호 등과 같은 93개의 필기 data의 특성을 포함하고 있는 93개의 class로 분류하였다. 이 Data base를 세 개로 나누어 training data로 66,896개, Cross-validation data로 31,101개, 그리고 test data로 24,083개를 사용하였다.

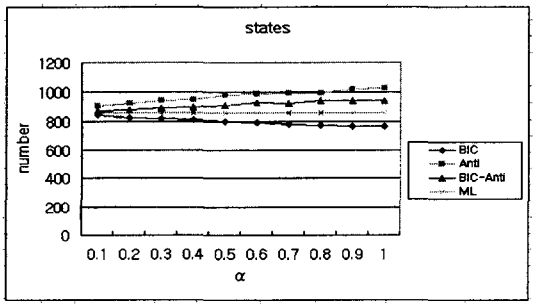
인식하기 전에 입력된 필기 데이타 신호를 시간에 따라 window를 움직여서 segmented, normalized, resampled 그리고 feature를 추출하였다.

PCA(Principal Component Analysis)를 이용하여 각 frame 별로 9개의 Dimension vector를 만들었다

6.2 Anti 데이터의 생성

대상 데이터에 해당하는 데이터를 제외한 나머지 데이터를 Merge하여 해당데이터에 대한 Anti 모델을 생성하였다. 하나의 모델은 해당 모델을 제외한 나머지 데이터가 모두 Merge되어, 모델의 크기가 본래의 데이터의 크기보다 상당히 커지게 되므로 본래의 Data양과의 균형을 맞추기 위해 본래의 데이터 양에 비례하게 랜덤한 데이터를 선별하여 적절한 양의 Anti 데이타를 생성하였다. 새로 생성된 클래스는 본래 자기 자신의 파일은 빼고 나머지 파일만 가지고 자기 자신의 파일을 만든다.

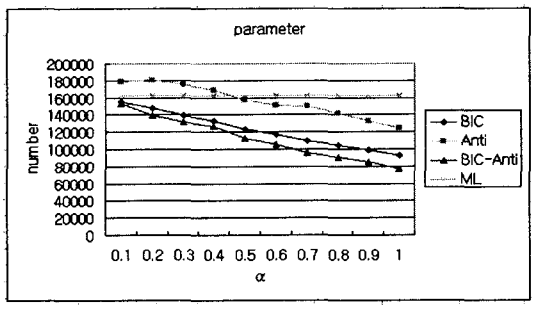
6.3 실험결과



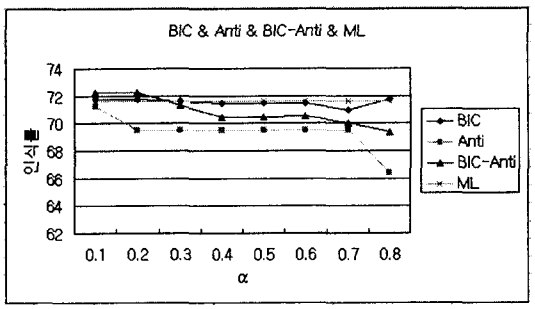
(a) BIC, Anti, BIC-Anti, ML의 state 수

위의 그림 (a)는 각 Model Selection Criteria의 state의 수를 측정한 결과 그래프이다. state의 수는 최적의 Model Selection Criterion인 BIC가 α가 증가할수록 감소하는데 이는 penalty의 영향이 크게 작용했음을 알 수 있다. 만일 BIC-Anti의 α가 0이라면 이 값은 BIC와 같게 되며 α와 β가 0라면 ML과 같은 값이 된다.

아래 그림(b)는각 Model selection criterion의 parameter 수의 값을 측정한 결과 그래프이다. parameter의 수는 본 논문에서 제안한 BIC와 Anti-likelihood의 결합인 BIC-Anti criterion이 가장 적은 결과를 보인다. α가 커질수록 그 값의 차이가 남을 알 수 있다 이는 BIC의 penalty 항에 α를 적용하여 그 값을 인위적으로 줄였기 때문에 나타나는 현상이다. 만일 state의 수와 마찬가지로 BIC-Anti의 α가 0 이라면 이 값은 BIC와 같게 되며 α와 β가 0라면 ML과 같은 값이 된다.



(b) BIC, Anti, BIC-Anti, ML의 parameter 수

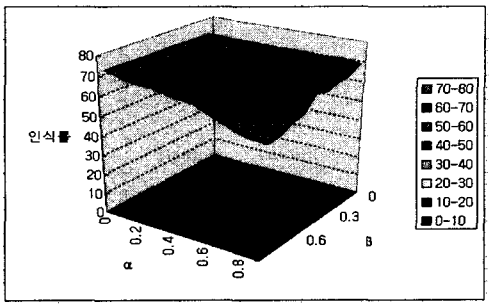


(C) ML, BIC, Anti, BIC-Anti 의 인식률

본 논문에서 제안한 방법 BIC-Anti의 인식률은 72.27%로 BIC 보다 0.7% 증가하였음을 알 수 있다. parameter의 수도 BIC보다 약 1% 감소한 153,306이 측정이 되었다. 그런데 그림(b)에서 보는 것처럼 α가 커지면 parameter의 수는 감소하지만 인식률도 또한 감소하는 것을 볼 수가 있다(그림(c)). 그러므로 BIC의 penalty 항의 값을 작게 사용하는 것이 parameter의 수의 감소뿐만 아니라 인식률 향상에도 영향을 미침을 알 수 있다. 또 β값을 크게 적용하게 되면 Anti likelihood의 값이 크게 적용되어 인식률 값은 72.3%로 증가하게 되나 parameter의 수도 크게 증가함을 볼 수가 있다.

아래 그림(d)는 α와 β를 적용시킨 인식률 결과 그래프이다. 그림에서 보는 것처럼 α와 β가 크게 적용될수록 인식률이 떨어짐을 알 수 있다. 비록 그림 (b), (c)에서처럼 α를 적용하여 parameter의 수를 감소시켰어도 이처럼 인식률이 떨어지면 parameter의 수를 줄이는 것은 큰 의의가 없으므로 parameter의 값이 감소하고 인식률이 증가하는 적절한 값을 찾아야 한다.

그러므로 parameter의 수를 감소하고 동시에 인식률을 증가시키는 α, β를 작은 값으로 적용하는 것이 좋음을 알 수 있다.



(d) α, β를 적용한 전체 인식률

7. 결론

본 논문에서는 모델 선택에 있어서 data의 변별력을 위해 Anti 모델을 적용하여 최적의 Model Selection Criterion인 BIC와 결합하여 BIC-Anti Model Selection Criterion을 제안한 결과 전체 parameter 수는 감소하였고 인식률이 증가하였음을 알 수 있었다. 그런데 UNIPEN 데이터 집합의 데이터 형식이 표준화되었어도 크기와 종류가 다양한데 비해 실험 대상모델에 대한 class를 하나만 만들어 실험하였고 문자의 대소문자를 구별하여 실험하였기 때문에 인식률이 약 72.3%에 지나지 않았다. 또한 본 논문에서는 feature vector의 수가 적게 사용되고 frame의 수가 작아서 state의 수에 대한 충분한 실험이 불가능하였다.

향후 과제로는 새로운 feature vector를 추출하고 해당 Data에 대한 class를 다양하게 만들어 실험하고자 한다.

참 고 문 헌

[1] 한국전자통신연구원 "펜을 이용한 문자/제스처 인식 시스템의 인식률 성능 향상을 위한 추가 개발에 관한 연구",1998.

[2] 박 미나, 하 진영, "HMM 모델링을 위한 HMM의 state수와 mixture 수 분석 " 한국 정보과학회 춘계 학술 논문 발표논문집, 2002.

[3] 하 진영, Alanin Biem, Jayashree Subrahmonia, 박 미나, "모델의 사전 확률 추정을 이용한 HMM 구조의 최적화", 한국정보과학회 추계 학술 논문 발표논문집, 2001.

[4] 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.

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

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

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

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