필기 데이터 인식을 위한 이산 HMM과 연속 확률밀도 HMM에서의 HMM구조 최적화 기준 분석

박미나O, 하진영
강원대학교 컴퓨터정보통신공학과
mnpark@kangwon.ac.kr, jyha@kangwon.ac.kr

Analysis of HMM Topology Criteria on Discrete HMM and Continuous-Density HMM for Handwriting Recognition

Mi-Na PARKO, Jin-Young HA
[1]Department of Computer and Information Communications Engineering,
Kangwon National University, Chuncheon, Kangwon, 200-701, Republic of Korea.

요 약

은닉 마르코프(HMM)의 HMM의 구조 최적을 위한 모델 선택 방법에 많은 방법들이 연구되어지고 있다. HMM의 구조를 어떻게 최적으로 정해야 하는 가에 대해 HMM의 구조를 체계적인 방법으로 정함과 동시에 변별력의 단점을 개선 할 수 있는 방법으로 Anti-likelihood(ALC1)를 제안하였고 이를 모델 선택 기준인 BIC와의 결합(ALC2)하여 필기 데이터에 대해 실험한 결과 기존의 방법보다 파라미터의 수는 감소되고 인식률이 향상됨을 알 수 있었다. 이를 Discrete HMM에도 적용하여 제안된 ALC2가 HMM 구조를 최적화 하는 모델 선택 기준임을 Continuous-Density HMM과 비교하여 실험 검증 한다.

1. 서 론

지난 30년간 음성·필기인식에 대한 연구로는 지난 30년 동안의 많은 연구의 결실로 실용화에 많이 근접하여 다수의 시제품과 상업용 시스템이 소개되었으나 인식률 향상과 인식속도 향상, 메모리 문제 해결 등, 시스템의 최적화에 대한 연구의 필요에 의해 지금까지 필기인식을 위한 다양한 방법론 등이 시도되어 왔다[1]. 1980년대 말부터 현재까지 은닉 마르코프 모델(Hidden Markov Model: 이하 HMM)이 가장 우수한 성능을 보여주고 있다[2]. 은닉 마르코프 모델(HMM)는 음성과 문자 등에서 발견되는 많은 변형들을 흡수하고, 시간에 따라 변해가는 특성의 Data를 잘 모델링하기 때문이다[3].

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

이러한 HMM의 구조를 어떻게 최적으로 정해야 하는 가에는 몇 가지 휴리스틱 한 방법이 제안되었다. 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]. HMM의 구조를 체계적인 방법으로 정함과 동시에 변별력의 단점을 개선 할 수 있는 방법으로 HMM 적용 시 해당 모델과 해당 데이터에 기반한 Likelihood 뿐만 아니라 그 모델에 대한 다른 클래스 데이터의 Anti-likelihood를 제안하였고 이를 BIC 모델 선택 기준과 결합하여 HMM의 구조를 체계적인 방법으로 정함과 동시에 변별력의 단점을 개선하여 필기데이터에 대해 기존의 방법보다 파라미터의 수는 감소되고 인식률이 향상된 결과를 보였다.

이에 본 논문에서는 후자의 방법을 Discrete HMM에 대해서도 실험하여 선택 기준 방법에 대해 Continuous-Density HMM과 비교 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. 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.

* 이 논문은 2005년도 강원대학교 두뇌 한국 21 사업에 의하여 지원되었음.

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

$N(x, \mu_{il}, \sum_{il})$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. 모델 선택 기준(Model Selection Criteria)

3.1 ML(Maximum Likelihood) Criterion

최대 확률의 일반적인 방법은 파라미터 공간 A에 속하는 파라미터 a에 속한 집합 S안의 관찰된 임의의 변수 X의 밀도 함수를 $f(X|a)$로 정의한다. 여기서 확률 함수 L은

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

로 정의 할 수 있다. 최대 확률 방법은 확률 함수 $L(a|X)$의 값이 최대가 되는 파라미터 a의 평균값 $\mu(x)$을 구하는 것이다. $\mu(x)$은 파라미터 a의 최대 확률 추정이 된다. 이 방법은 직관적인 방법으로 관찰된 데이터를 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 로그확률함수라고 불린다. 이 함수는 많은 경우에 확률함수보다 더 쉽게 연산 할 수 있기 때문에 많이 쓰인다. 이 방법은 학습 데이터의 양이 충분하다는 조건 하에서 다른 어느 추정방법 못지않은 성능을 보인다.

3.2 BIC(Bayesian Informaition Criterion)

Central Limit 정리에 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 베이지안 정보기준으로 인도한다.

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

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].

3.3 ALC1(Anti-likelihood Criterion)

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.

$$ALC1 = logp(X|\theta_{ML}) - logp(X'|\theta_{ML}) \qquad [X \in C]\ \ [X' \notin C] \qquad (2)$$

(단, C는 해당 클래스 $data$)

3.4 ALC2(BIC-Anti likelihood Criterion)

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.

$$ALC2 = \sum_{X \in C} \log p(X|\theta_{ML}) - \alpha \times (\times \sum_{X \notin C} \log p(X|\theta_{ML})\ \beta)\frac{k}{2}\log N \qquad (3)$$

위 (3)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 증가 시키는 데는 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.

4. 실험 및 결과 분석

4.1 실험데이타베이스

4.1.1 Data

본 논문에서는 UNIPEN Data Base 와 KAIST Data Base를 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 있다. 본 논문에서는 Continuous-Density HMM 과 Discrete HMM과 비교 실험하기 위해 숫자만을 실험 대상으로 하였다. 또 문자 전체에 대해 실험하여 그 인식률에 대해서도 알아보았다. 먼저 숫자 Data에 대하여 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.

UNIPEN 전체 Data에 대해서는 훈련데이타 66,896 개, 교차검증 데이타 31,101개, 그리고 테스트 데이타로 24,083개를 사용하였다.

Discrete HMM을 실험하기 위해 KAIST의 필기 숫자 Data Base를 사용하였다. 총 23명의 필기자로부터 수집한 4064개의 샘플을 실험 데이타로 삼았다. 이중 2229개의 샘플을 훈련데이타로 1092개의 샘플을 교차검증집합으로 743개의 샘플을 테스트 데이타로 사용했다.

실험하기 위한 Toolkit으로 HTK를 사용하였으며 실험 data를 HTK에서 제공되는 Vector Quantization 파일의 형태로 수정하여 훈련하였다.

4.1.2 Anti 데이터

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.

4.2 실험 결과

다음은 Continuous-Density HMM과 Discrete HMM의 실험 결과이다.

Table 1. a) Continuous -Density HMM Recognition Results(숫자만)

Model selection criteria	Total number of parameters	Recognition Rates(%)
ML	351,40	91.77
BIC	20,596	91.93

b) Continuous -Density HMM Recognition Results(숫자와 문자)

Model selection Criteria		Total number of parameters	Recognition Rates(%)
ML		162,239	71.63
BIC	α=0.1	154,923	71.76
ALC1(α = 0.1)		178,834	69.51
ALC2i	α = 0.1 β = 0.2	157,908	72.30

Table 2. Discrete HMM Recognition Results(숫자)

Model selection criteria		Total number of parameters	Recognition Rates(%)
ML		3,740	92.60
BIC	α =0.1	3,434	93.27
ALC1	α =0.1	1,734	87.35
ALC2	α = 0.1 b = 0.1	612	94.21

5. 결론

본 논문에서는 모델 선택에 있어서 모델의 변별력을 위해 ALC1 Model Selection Criterion을 최적의 Model Selection Criterion인 BIC와 결합한 ALC2 Model Selection Criterion에 Discrete HMM을 적용하여 UNIPEN과 KAIST의 숫자 데이타에 대해 실험한 결과 Continuos-Density HMM를 실험하였을 때와 마찬가지로 기존의 방법보다 전체 parameter 수는 감소하였고 인식률이 증가하는것을 보였다. UNIPEN data 집합의 크기와 종류가 다양하여 데이타의 특징을 찾는데 어려움을 갖고 있으나 숫자에 대해서는 비교적 높은 인식률의 결과를 보였다. 그러나 숫자와 문자전체 데이타에 대해서는 실험 대상모델에 대한 class를 하나만 만들어 실험하였고 문자의 대소문자를 구별하여 실험하였기 때문에 UNIPEN 데이타의 특징으로 인해 저조한 인식률을 보였다.

앞으로 KAIST Data에 대해서도 숫자외의 다른 문자를 포함하여 그 결과를 비교하여 보고 UNIPEN 데이타의 일반화 작업을 통해 그 인식률을 향상시키고자한다.

참 고 문 헌

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

[2] 박미나, 하진영, "HMM 모델링을 위한 HMM의 state수와 mixture 수 분석 " 한국 정보과학회 춘계 학술 논문 발표논문집, 2002.

[3] 하진영, Alanin Biem, Jayashree Subrahmonia, 박미나, "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] 한국전자통신연구원 "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 merging", in Advances in NIPS, vol5. pp.11-18, 1993.