Journal of Korea Multimedia Society Vol. 19, No. 1, January 2016(pp. 22-29)
http://dx.doi.org/10.9717/kmms.2016.19.1.022

불완전 시계열 데이터를 위한 이산 HMM 학습 알고리듬

신 봉 기†

Discrete HMM Training Algorithm for Incomplete Time Series Data

Bong-Kee Sin†

ABSTRACT

Hidden Markov Model is one of the most successful and popular tools for modeling real world sequential data. Real world signals come in a variety of shapes and variabilities, among which temporal and spectral ones are the prime targets that the HMM aims at. A new problem that is gaining increasing attention is characterizing missing observations in incomplete data sequences. They are incomplete in that there are holes or omitted measurements. The standard HMM algorithms have been developed for complete data with a measurements at each regular point in time. This paper presents a modified algorithm for a discrete HMM that allows substantial amount of omissions in the input sequence. Basically it is a variant of Baum-Welch which explicitly considers the case of isolated or a number of omissions in succession. The algorithm has been tested on online handwriting samples expressed in direction codes. An extensive set of experiments show that the HMM so modeled are highly flexible showing a consistent and robust performance regardless of the amount of omissions.

Key words: Hidden Markov Model, Training Algorithm, Discrete HMM, Incomplete Sequence Data

1. 서 론

은닉 마르코프 모형은 순차 데이터 또는 시계열 데이터의 모형으로 널리 활용되고 있으며 매우 폭넓게 활용되고 있고 시간이 갈수록 새로운 응용 사례가 나타나고 있다[1]. 1980년대 음성 인식에서 성공적으로 활용된 이후 다양한 패턴 인식 및 기계 학습 모형으로 응용 영역이 확대되어 왔다[2,3]. 최근 빅 데이터 붐을 타고 새로운 인기를 이어가고 있다.

오늘날 수많은 모바일 기기를 포함하여 다양한 센서로부터 들어오는 신호는 양은 많지만 개별 신호를 볼 때 누락되거나 은닉, 차폐 등으로 소실되는 경우가 많다[4]. 종종 그럼에도 불구하고 HMM으로 이러한 신호들을 모형화할 수 있어야 되는 경우가 많다. 누락 데이터에 대한 처리 문제는 이미 오랜 역사를 갖고 있다[5]. 음성인식에서도 오래 전부터 이와 같은 결측 데이터를 처리하기 위한 노력이 있어 왔다[4,5]. 음성의 경우 데이터 결손으로 인한 인식률의 급격한 저하는 참기가 어렵기 때문에 큰 주제거리가 되지는 않았지만 연구가 없지는 않았다[6]. 최근 GIS 응용의 일환으로 결측이 잦은 동적 위치 추적 문제에 HMM을 활용하는 사례도 있으며 앞으로 유사 사례가 크게 늘어날 것으로 예상된다[4].

HMM의 이론에 따르면 그래프 모형의 하나인

※ Corresponding Author : Bong-Kee Sin, Address: (608-737) Daeyon-dong, Nam-ku, Busan, TEL : +82-51-629-6256, FAX : +82-51-629-6261, E-mail : bkshin@pknu.ac.kr
Receipt date : Oct. 23, 2015, Approval date : Nov. 24, 2015

† Dept. of IT Convergence and Applications Engineering, Pukyong National University
※ This work was supported by the Pukyong National University Research Abroad Fund in 2011 (PS-2011-0304)

HMM은 어떤 변수가 있건 없건 간에 항상 추론이 가능하다. 따라서 일부의 관측 신호가 없더라도 추론에는 특별한 영향을 미치지 않는다. 있어야 할 변수가 없을 때 관심 변수를 추정 평가할 때 기대되는 확률의 정확성, 성능이 영향을 받을 뿐이다. 본 연구에서는 EM 기반의 표준의 학습 알고리듬을 최소한으로 수정하여 불완전 데이터에 대한 추론과 학습하는 알고리듬을 제시하고자 한다. 제안 방법의 E 단계는 전진 및 후진 확률 계산과 몇 가지의 사후확률 변수의 계산 공식를 변경함으로써 문제를 해결하였다. 연속 모형에도 벡터 단위의 결손이라는 가정을 만족할 때 이산 HMM의 학습 및 추론 방법을 그대로 적용할 수 있다.

본 논문의 구성은 다음과 같다. 2장에서 본 논문의 배경 이론이 되는 HMM에 관한 간략한 소개를 한 다음, 3장에서 제안 모형의 추론 및 학습 알고리듬을 전개 한다. 그리고 4장에서는 제안한 방법으로 실험한 결과를 다양한 각도에서 단계별로, 그리고 체계적으로 평가한 결과를 제시하며 성능의 변화를 해석한다. 마지막 5장에서 결론을 맺는다.

2. 은닉 마르코프 모형 이론

2.1 은닉 마르코프 모형 (HMM)의 정의

일차 마르코프 체인에 출력 과정을 얹은 형식의 은닉 마르코프 모형(HMM)은 입력 신호의 다양한 변형, 잡음 등에 뛰어난 모형화 능력이 있는 통계적 시계열 모형의 하나이다 [1/Rabiner+89]. 모형의 구조는 Fig. 1(a)와 같다. 마르코프 체인 $\{X_t\}$는 일련의 상태의 열로 정의하는 확률 과정인데 효율적 모형화를 위해 도입한 잠변수들이며 직접 관측되지 않는다. 그 대신 마르코프 체인의 확률 함수인 출력 과정 $\{Y_t\}$을 통해 간접적으로 드러낸다.

순차 데이터 $Y = Y_{1:T} = y_1 y_2 \cdots y_T$가 관측 되었을 때 HMM은 생성 확률을 다음과 같이 표현 한다.

$$P(Y|\lambda) = \sum_X P(Y, X|\lambda) \tag{1}$$

여기서 $X = X_{1:T} = X_1 X_2 \cdots X_T$는 Y를 생산하였을 가능성이 있는 임의의 마르코프 체인이다. 그런데 만약 관측 신호 중 t 번째의 신호 Y_t가 관측되지 않았다면 Fig. 1(b)와 같이 표현할 수 있다. 본 연구에서는 이때 변수 Y_t는 존재하며 그 값이 공($\varnothing$ 또는 널 null)이라고 두기로 한다. 그러면 불완전한 데이터의 경우에도 식 (1)을 그대로 적용할 수 있다. 다음 절에는 기존 학습 및 추론 알고리듬에서 달라지는 부분을 중심으로 기술하기로 한다. 이를 위해 그 전에 최소한의 HMM의 기초 내용을 소개하기로 한다.

HMM은 초기 상태 분포, 상태 천이 분포, 그리고 부호 출력 분포의 세 가지의 확률 모수로 기술할 수 있는 통계적 모형의 하나이다. $\lambda = (\pi, A, B)$로 표기한다. 유한한 N 개의 상태들의 집합을 $S = \{1, 2, \ldots, N\}$, 각 상태 $i \in S$ 에서 유한한 종류의 기호 또는 신호 $v \in V = \{1, 2, \ldots, M\}$를 만들어낸다고 하자. HMM의 세 가지 모수는 다음과 같이 정의 한다.

$$\pi = \{\pi_i : \pi_i = \Pr(X_1 = i),\ \forall i \in S\},\ \pi_i \geq 0,\ \sum_{i \in S} \pi_i = 1.$$

$$A = \{a_{ij} : a i_{ij} = \Pr(X_{t+1} = j | X_t = i),\ \forall i, j \in S\},$$

$$a_{ij} \geq 0,\ \sum_{j \in S} a_{ij} = 1.$$

$$B = \{b_i(v) : \Pr(Y_t = v | X_t = i),\ \forall i \in S,\ \forall v \in V\},$$

$$b_i(v) \geq 0,\ \sum_{v \in V} b_i(v) = 1.$$

HMM은 일련의 관측 데이터를 만들어 내는 생성 모형이다. 관측 순차 데이터 $Y = Y_{1:T} = Y_1 Y_2 \ldots Y_T$ 를 출력할 확률(우도)는 식 (1)과 같다. 여기서 X는 잠변수라고 하며 미지의 마르코프 체인 $X = X_1 X_2 \ldots X_T$을 나타낸다. Fig. 1(a)에 보인 HMM의 구조에 따르면 수평 및 수직 화살표로 표현되는 두 종류의 인과관계 $X_{t-1} \rightarrow X_t$와 $X_t \rightarrow Y_t$가 있다. 각각에는 일차 마르코프 가정과 조건부 독립 가정이 내포되어 있다.

Fig. 1. Graphical model representation. (a) Hidden Markov model, (b) Hidden Markov model with missing observations.

3. 제안 모형과 추론 알고리듬

3.1 불완전 데이터를 위한 HMM

표준 HMM의 이론은 모두 기본적으로 완전한 관측열을 중심으로 전개되어 있다[1]. 그러나 본 논문에서 제안하는 모형의 환경에서는 출력이 누락된 경우도 종종 있다고 가정한다. 출력이 없는 경우 공백 기호 ϕ로 두고 출력 과정을 다음과 같이 정의할 수 있다.

$$X_t \rightarrow Y_t \quad \in V \cup \varnothing \tag{2}$$

여기서 $\varnothing = \{\phi\}$이다. 이것은 단지 편리를 위한 것이다. $\varnothing$는 공백 기호가 있는 집합이며 집합론에서 다루는 공집합과 구별하기 바란다. 즉 관측 기호의 유무에 따라

$$P(X_t Y_t) = P(X_t)P(Y_t|X_t) \tag{3}$$

또는

$$P(X_t, Y_t = \varnothing) = P(X_t) \tag{4}$$

여기서

$$P(Y_t = \varnothing|X_t) = 1$$

와 같이 두 가지 경우로 구분하여 써야 하지만 확장된 출력의 집합 $V \cup \varnothing$을 도입하면 기존의 HMM 기술 방식을 그대로 쓸 수 있다.

3.2 추론 알고리듬

HMM의 전진 확률 변수는 표준 HMM과 다르지 않다.

$$\alpha_t(j) = \sum_i \alpha_{t-1}(i) a_{ij} b_j(y_t), \qquad \begin{matrix} j \in S \\ t = 1,\ldots, T. \end{matrix} \tag{5}$$

다만 관측 기호가 없는 경우에는 확률적 사건이 없으므로 다음과 같이 구분하면 된다.

$$b_j(y_t) = \begin{cases} \Pr(y_t|x_t) & y_t \neq \varnothing, \\ 1 & y_t = \varnothing. \end{cases} \tag{6}$$

모형의 평가 점수는 다음과 같다.

$$P = \sum_i \alpha_T(i) \tag{7}$$

유사하게 후진 확률은

$$\beta_t(i) = \sum_j a_{ij} b_j(y_t) \beta_{t+1}(j), \qquad \begin{matrix} i \in S \\ t = T-1,\ldots,1. \end{matrix} \tag{8}$$

전진 확률과 후진 확률의 경계 조건은 식 (6)에 따르는 표준 HMM과 같다.

3.3 학습 알고리듬

HMM은 학습 알고리듬은 유명한 EM 알고리듬의 하나로써 추정(E)과 최대화(M)라는 두 과정의 반복으로 구성된다[9,10]. 추정의 E 단계에서는 잠변수에 관한 모든 사건을 추정 한다. 학습 초기에 임의로 지정한 모수를 이용하여 다음과 같은 사후 확률 변수를 추정한다. 가장 기본적인 특정 t 시각의 상태 i의 추정 사후 확률은 표준 모형과 같다.

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{P} \tag{9}$$

반면 t 시각 상태 i에서 j로 천이하는 사건의 추정 확률은 관측 기호의 유무에 따라

$$\xi_t(i,j) = \frac{1}{P}\alpha_t(i) a_{ij} b_j(y_{t+1}) \beta_{t+1}(j) \tag{10}$$

또는

$$\xi_t(i,j) = \frac{1}{P}\alpha_t(i) a_{ij} \beta_{t+1}(j)$$

이다. 식 (6)을 이용하여 두 식을 결합하면 식 (9)로 통일할 수 있다. 마지막으로 t 시각 상태 i에서 기호를 관측할 확률은 다음과 같다.

$$\zeta_t(i) = \begin{cases} \dfrac{1}{P}\alpha_t(i)\beta_t(i) & y_t \neq \varnothing, \\ 0 & y_t = \varnothing. \end{cases} \tag{11}$$

관측이 없었으면 존재할 가능성이 없으므로 확률은 0이 된다.

이제 최적화의 M 단계이다. 세 종류의 사후 확률 변수를 모두 구하였으면 이전 단계의 임의적 모수보다 더 좋은 값을 구할 수 있다. 여기서 좋다는 것은 주어진 관측 데이터 Y에 대한 모형 λ의 우도(likelihood)가 더 크다는 것을 의미한다. EM 알고리듬으로 두 단계를 반복하면 매번 그 우도가 항상 증가한다는 것이 증명되어 있다. HMM의 세 모수의 새로운 값은 다음과 같다.

$$\widehat{\pi_i} = \gamma_1(i), \quad \forall i \in S \tag{12}$$

$$\widehat{a_{ij}} = \frac{\sum_{t=1}^{T-1}\sum \xi_t(i,j)}{\sum_{t=1}^{T-1}\gamma_{t(i)}}, \quad \forall i,j \in S \tag{13}$$

$$\hat{b}_j(k)=\frac{\sum_{t=1}^{T}\gamma_t(i)\delta_{y_t,k}}{\sum_{t=1}^{T}\gamma_t(i)},\quad \forall j\in S,\quad k\in V. \tag{14}$$

식 (14)의 k는 정의구역이 V인데 유의하기 바란다. 그리고 여기서 $\delta_{a,b}$ is Kronecker 델타 함수이다.

$$\delta_{a,b}=\begin{cases}1 & \text{if} \quad a=b,\\ 0 & \text{if} \quad a\neq b.\end{cases}$$

연속치를 출력하는 가우스 HMM의 경우에는 출력 모수 대신에 다음 공식을 사용하면 된다.

$$\hat{\mu}_i=\frac{\sum_{t=1}^{T}\gamma_t(i)x_t}{\sum_{t=1}^{T}\gamma_t(i)},\quad i\in S \tag{15}$$

$$\hat{\sigma}_i=\frac{\sum_{t=1}^{T}\gamma_t(i)(x_t-\hat{\mu}_i)^2}{\sum_{t=1}^{T}\gamma_t(i)},\quad i\in S \tag{16}$$

4. 실험 결과 및 고찰

4.1 실험 데이터

불완전 순차 데이터를 모형화하는 이산 HMM의 평가를 위하여 와콤 태블릿/전자펜 장치를 사용하였으며 총 216 표본의 숫자 필기를 수집하였다. 각 숫자별로 15~25개의 표본을 획득하였다. 본 실험의 주목적은 다양한 비율로 손상된 시계열 데이터를 얼마나 잘 모형화 하며 성능의 변화가 어떤지를 살펴보는데 있다.

필기 데이터는 펜의 궤적을 작은 단위 선분으로 나누고 각각의 진행 방향으로 16 종류의 코드로 표현하였다. 원 입력 데이터는 손상이 없으므로 $D^{(0)}$로 부르기로 한다. $D^{(0)}$의 각 필기의 임의의 위치에 10% 정도가 손상되어 기호가 사라진 데이터를 $D^{(10)}$라고 한다. 비슷하여 20%~50%가 손상된 데이터 집합을 차례로 $D^{(20)}$, ..., $D^{(50)}$이라고 한다. $D^{(0)}$의 깨끗한 원필기는 Fig. 2의 최상단의 숫자와 같다. 그림은 점진적으로 손상이 많은 $D^{(10)}$, $D^{(20)}$, ..., $D^{(50)}$의 필기 예도 보여준다. $D^{(0)}$ 필기는 방향 코드를 그대로 그림으로 가시화한 것이지만 이하의 손상된 필기는 중간의 기호들이 사라졌기 때문에 Fig. 2와 같이 그릴 수 없다. 단지 참고를 위해 손상되지 않은 코드를 손상전의 위치에 그대로 그렸을 뿐이다. 순차 데이터로서는

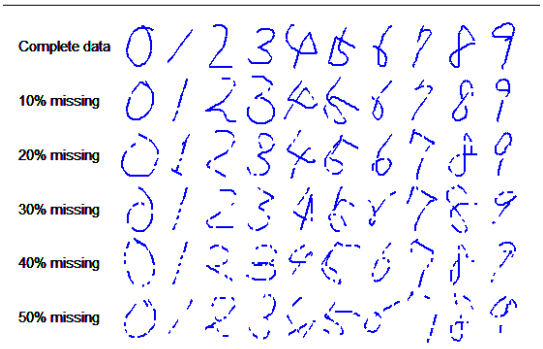


Fig. 2. Samples from six data sets. From the top: $D^{(0)}$, $D^{(10)}$, $D^{(20)}$, $D^{(30)}$, $D^{(40)}$, $D^{(50)}$.

10%, 20% 등의 비례적 손상이지만 필기의 모양은 크게 달라진다는 점을 이해하여야 한다. 따라서 이미 손상된 데이터로부터 모양을 정확하게 복구하는 것은 불가능하다.

$D^{(0)}$를 제외한 다른 손상 데이터는 학습 데이터와 시험 데이터를 별도로 생성하였다. 기본적으로 모두 $D^{(0)}$에서 유래한 것이며 손상 비율도 같지만 손상 위치가 다른 모의 실험 데이터를 사용하였다.

4.2 교차 검증

HMM은 시계열 패턴 인식으로 널리 활용된다. 숫자 필기 인식을 위하여 교차 검증법을 적용하기로 한다. Fig. 2의 각의 데이터에서 90%로 훈련한 다음 나머지 10%로 평가한다. 각 데이터 별로 총 10 번의 검증 평가를 실시하였다. 모형은 일반적인 어고딕(ergodic) 모형과 선형적 순서를 표현하는 선형 사슬 모양의 모형의 두 가지로 평가하였고 모두 상태수를 $N=7$로 하였다. 결과는 Table 1과 같다.

모든 데이터에 대해 선형 모형의 성능이 높다. 필기 신호의 특성상 선형적 특징이 엄격하게 적용되어야 하기 때문이라고 설명할 수 있다. 각 데이터로 학습된 모형은 비슷한 비율의 손상 데이터에 대해서 거의 유사한 수준의 성능을 보여준다. 이는 상당히 놀라운 발견이다. 손상으로 인하여 관측 데이터가 줄어들면 손상 비율에 비례하여 식별 능력이 줄어들 것이라고 기대하였지만 그 예상이 크게 빗나간 결과를 얻었다. Table 1은 각 경우에 대해 세 번의 평가를 하였고 성능의 중간 값만 보인 표이다. Fig. 3은 세 번의 성능을 모두 보인 그래프이다. HMM의 모수를

Table 1. Cross-validation results with two types of models ($M^{(n)}$) from six data sets ($D^{(n)}$)

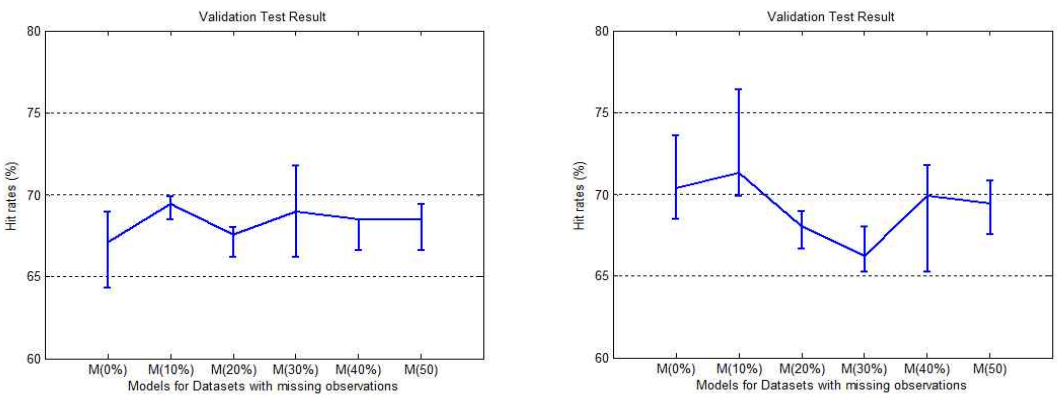
Train set / Model type	$D^{(0)}$ ($M^{(0)}$)	$D^{(10)}$ ($M^{(10)}$)	$D^{(20)}$ ($M^{(20)}$)	$D^{(30)}$ ($M^{(30)}$)	$D^{(40)}$ ($M^{(40)}$)	$D^{(50)}$ ($M^{(50)}$)
Ergodic model	67.13	69.44	67.59	68.98	68.52	68.52
Linear Model	70.37	71.30	68.06	66.20	69.91	69.44

임의로 초기화 하였기 때문에 매번 성능의 변동이 있지만 통계적으로 유의할 만한 추세를 보이지는 않는다.

4.3 다양한 시험 데이터에 대한 성능

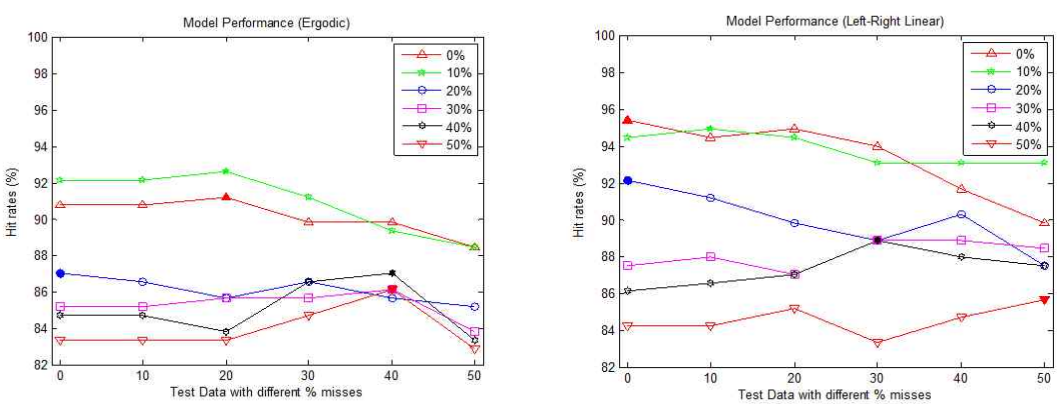
만약 훈련 데이터와 같은 정도의 손상이 있는 동일한 시험 조건이 아니라면 어떨까? 실제 입력 데이터는 손상이 없거나 손상이 오히려 더 많은 상황이라면 성능의 변화는 어떤지 알아보기로 한다. 상태수는 여전히 모두 7로 두었다.

Fig. 4(a)는 어고딕 모형의 성능 그래프이다. 각 곡선은 특성 비율로 손상된 데이터로 훈련한 모형인데 수평축의 다양한 손상 데이터에 대해 변화는 있지만 크게 감소하는 추세를 보이지는 않는다. 즉 다양



(a) (b)

Fig. 3. Results of cross validation tests with three repetitions for each case. (a) ergodic models, and (b) linear models.



(a) (b)

Fig. 4. Test results on data sets with different rates of missing observations. Each curve represents a model and the test sets (horizontal axis) are separately prepared and thus differ from the training sets. (a) Ergodic models, (b) linear models.

한 변화를 모두 수용하는 강건한 모형화 능력을 보여준다. Fig. 4(b)는 선형 모형의 결과이다. 상대적으로 손상이 적은 데이터로 훈련된 모형은 시험 데이터의 손상이 클수록 성능이 떨어지는 경향을 읽을 수 있다. 하지만 오른 쪽으로 갈수록 모든 모형의 성능이 수렴하는 모습을 보여 준다. 저품질의 데이터로 훈련한 모형은 손상이 덜한 데이터에 대해서 성능이 높지 않지만 품질이 떨어질수록 성능이 비슷하거나 오히려 상승하는 양상이다.

4.4 모형 크기에 따른 성능 변화

지금까지는 상태수를 $N=7$로 고정하였지만 상태수의 변화에 따라 더 높은 성능을 올릴 수도 있다. 상태수가 커지만 모형의 복잡도가 커지기 때문에 과도한 학습(overfitting)을 하게 되고 일반화 성능이 떨어질 수 있다.

Fig. 5는 HMM의 상태수에 따라 성능이 변하는 추세를 분석한 결과를 도시한 것이다. Fig. 5(a)는 무손상의 원자료로 학습한 어고딕 모형의 성능 변화이다. 상태가 늘면 일관되게 성능이 향상됨을 알 수 있다. Fig. 5(b)는 20% 정도의 결손이 있는 데이터 $D^{(20)}$로 훈련한 모형 $M^{(20)}$의 성능 변화이다. 전체적으로 $M^{(0)}$에 비하여 다소 저하되기는 하였지만 유사한 성능 변화의 추세를 보인다. Fig. 5(c)와 (d)는 선형 모형으로 유사한 실험을 한 결과이다. 어고딕 모형과 달리 $N=6$, 9 및 $N=5$, 8에서 두 개의 정점을 갖는다. 어고딕 모형과 선형 모형 모두 $N=9$ 부근에서 정점을 찍고 이후 성능이 하락하는 경향을 보인다.

4.5 모형 크기와 데이터에 따른 성능 분석

마지막 실험에서는 모델의 크기(N)과 시험 데이터 손상 정도에 따라 모형의 성능 변화 추이를 살펴본다. 특정 학습 데이터에 대한 HMM은 N이 클수록 성능이 높지만 반면 시험 데이터의 손상이 심할수록 식별이 어려워 질 거라고 예상할 수 있다. Fig. 6은 그런 점을 확인하는 시험 결과를 나타낸 것이다. 각 픽셀의 밝기는 숫자 인식률을 나타낸다. 비슷한 인식률을 나타내는 등고선을 보면 예상대로 N이 클수록 (윗 방향), 데이터의 손상이 적을수록(왼쪽 방향) 성능이 높은 경향을 보여준다. Fig. 6(a)와 (b)는 각각 $D^{(0)}$와 $D^{(20)}$으로 학습한 어고딕 모형의 결과이다. Fig. 6(c)와 (d)는 같은 데이터로 학습한 선형 모형의 결과이다. 후자의 경우 상태수가 각각 6과 9, 5와 8일 때 성능의 정점을 보여준다. 이와 같은 정점은 순차 데이터의 특징에 따른 결과이다. 데이터의 평균 길이와 패턴의 복잡도 등에 따라 결정된다.

Fig. 7은 유사한 시험을 모든 훈련 데이터 $D^{(0)}$, ..., $D^{(50)}$으로 제작한 모형 $M^{(0)}$, ..., $M^{(50)}$ 모두의 성능 변화의 모습을 보인 것이다. 오른 쪽으로 갈수록 어두워지며(성능 저하) 모형간의 인식률 차이가 없이 균등해지는 현상을 보여준다. 어고딕 구조든 선형 구조든 같은 추세를 보여준다. 놀라운 점은 상태수가 크

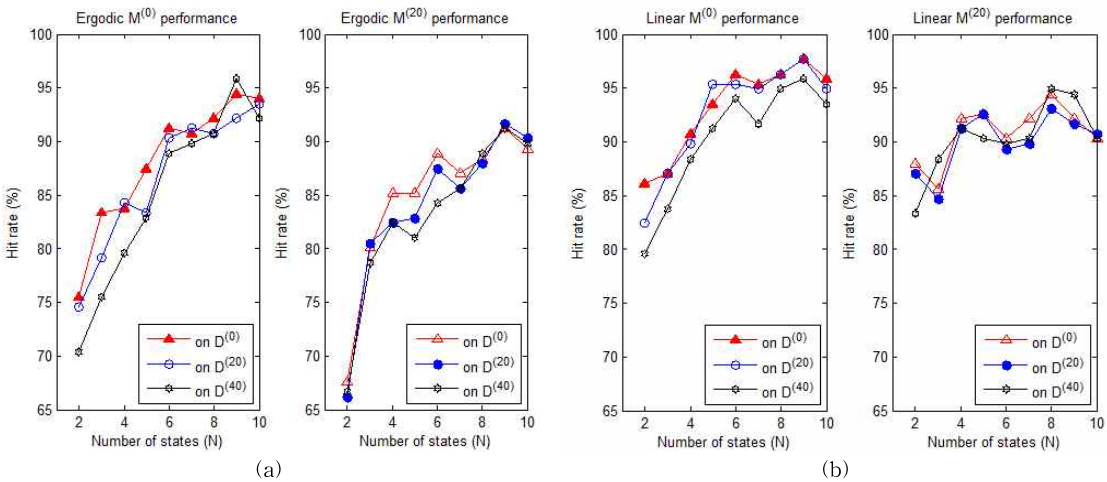


Fig. 5. Performance curve as a function the number of states N. (a) Ergodic model $M^{(0)}$ trained with $D^{(0)}$, (b) ergodic model $M^{(20)}$ using $D^{(20)}$, (c) linear model $M^{(0)}$ for $D^{(0)}$, and (d) linear model $M^{(20)}$ for $D^{(20)}$.

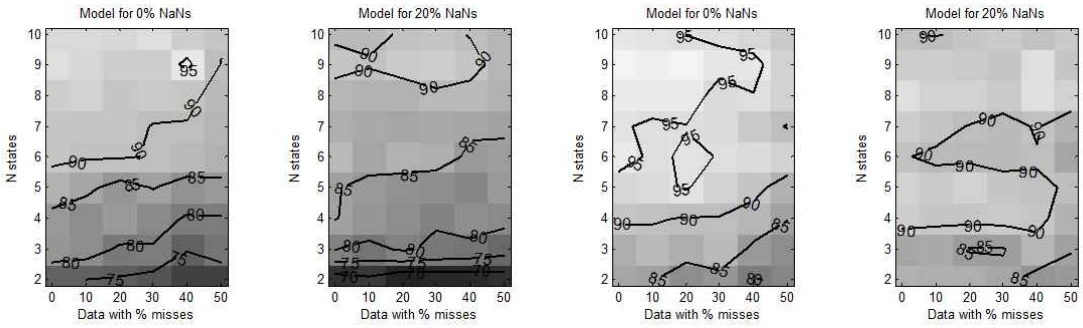


Fig. 6. Model performance over varying model size (N) and the degree of missing observations. (a) Ergodic model $M^{(0)}$'s performance, (b) ergodic $M^{(20)}$'s, (c) linear $M^{(0)}$'s, and (d) linear $M^{(20)}$'s.

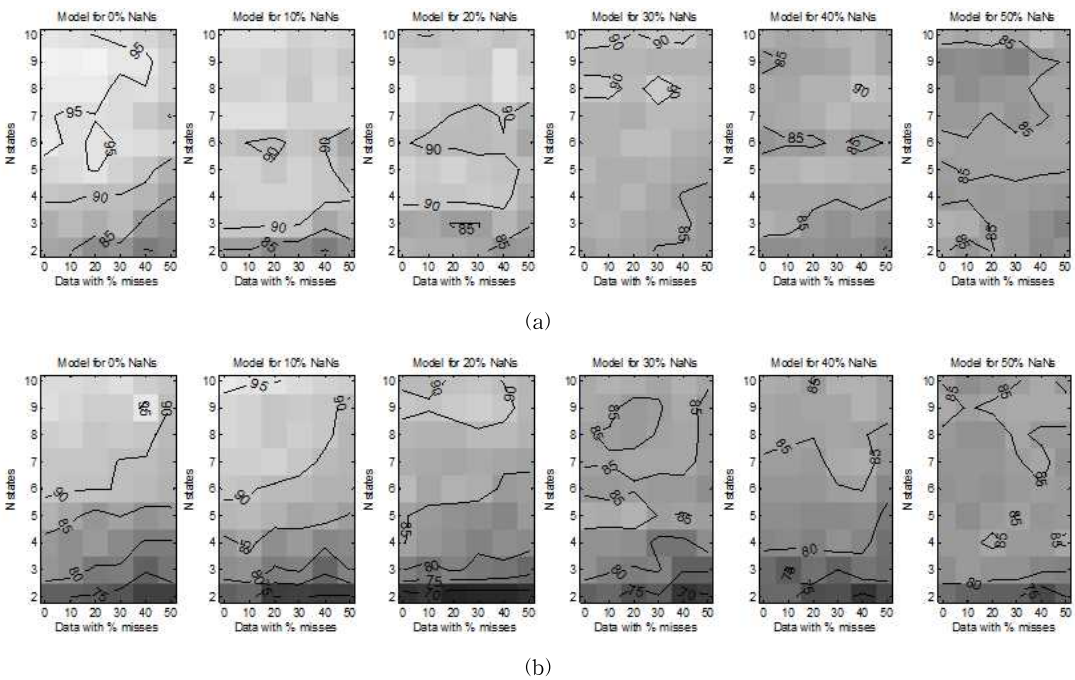


Fig. 7. The full version of Fig. 6, about the model performance over varying model size (vertical axis) and the data quality (horizontal axis). (a) Ergodic models' performance, and (b) linear models' performance.

지 않아도, 그리고 시험 데이터의 손상 정도가 상당한 정도(대략 40%)까지는 성능의 저하가 크지 않다는 점이다.

5. 결 론

본 논문에서는 순차적 관측 데이터에 손상이 있거나 관측 자체가 불완전한 경우에도 HMM을 응용할 수 있는 추론 및 학습 알고리듬을 제시하였다. 실험 결과 어떤 상황에서 훈련 하였든 완전 또는 불완전 데이터를 평가할 수 있음을 보여 주었다. 이때 성능의 저하가 크지 않다는 것을 확인할 수 있었다. 그 성능 저하가 거의 데이터 손상 자체로 인한 영향이며 HMM의 모형화 능력은 상당한 손상 범위 내에서 살아있음을 확인할 수 있었다. 손상 외에 보편적으로 관측이 누락되는 현상은 오늘날의 다양한 빅 데이터에서 관찰된다. 향후 본 연구의 결과는 이와 같은 실제 데이터에 널리 활용할 수 있을 것이다.

REFERENCE

[1] L. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE*, Vol. 77, No. 2, pp. 257-286, 1989.

[2] B.-K. Sin, "Recognizing Hand Digit Gestures Using Stochastic Models," *Journal of Korea Multimedia Society*, Vol. 11, No. 6, pp. 807-815, 2008. 6.

[3] B.-K. Sin and K.-R. Kwon, "Feature Space Analysis of Human Gait Dynamics in Single View Video," *Journal of Korea Multimedia Society*, Vol. 13, No. 12, pp. 1778-1785, 2010. 12.

[4] F. Torre, D. Pitchford, P. Brown, and L. Terveen, "Matching GPS Traces to (Possibly) Incomplete Map Data: Bridging Map Building and Map Matching," *Proceeding of ACM SIGSPATIAL GIS'12*, pp. 546-549, 2012.

[5] R. Little and D.B. Rubin, *Statistical Analysis with Missing Data*, John Wiley & Sons, Inc., NY, USA, p. 408, 1987.

[6] M. Cooke, P. Green, and M. Crawford, "Handling Missing Data in Speech Recognition," *Proceeding of International Conference on Spoken Language Processing*, pp. 1555-1558, 1994.

[7] A.C. Morris, M.P. Cooke, and P.D. Green, "Some Solution to the Missing Feature Problem in Data Classification, with Application to Noise Robust ASR," *Proceeding of International Conference on Acoustics, Speech and Signal Processing*, Vol. 2, pp. 737-740, 1998.

[8] S. Parveen and P.D. Green, "Speech Recognition with Missing Data using Recurrent Neural Network," in *Advances in Neural Information Processing Systems, 14*, pp. 1189-1195, 2001.

[9] L.E. Baum and J.A. Eagon, "An Inequality with Applications to Statistical Estimation for Probabilistic Functions of Markov Processes and to a Model for Ecology," *Bulletin of the American Mathematical Society*, Vol. 73, No. 3, pp. 360-363, 1967.

[10] A.P. Dempster, N.M. Rubin, and D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society, Series B*, Vol. 39, No. 1, pp. 1-38, 1977.

신 봉 기

1985년 서울대학교 공과대학 자원공학과 학사 졸업.
1987년 한국과학기술원 전산학과 석사 졸업.
1995년 한국과학기술원 전산학과 박사 졸업.

1987년~1999년 한국통신 SW연구소 선임연구원
1999년~현재 부경대학교 IT융합응용공학과 교수
관심분야 : 패턴인식, 기계학습, 컴퓨터 시각, 인공지능.