Handling Context Dependence with Dual Hidden Markov Model

Réjean Plamondon
École Polytechnique de Montréal, Montréal, H3C 3A7
rejean.plamondon@polymtl.ca
Xiaolin Li
CADlink Technology Corp., 2440 Don Reid Drive, Suite 100, Ottawa K1H 1E1
xli@cadlink.com
Chunhua Feng
Laboratoire Scribens, École Polytechnique de Montréal, Montréal H3C 3A7
feng@scribens.polymtl.ca

Keywords - Hidden Markov model, first order partial context dependence, dual hidden Markov model, observation evaluation, multiple observation training

Abstract

Hidden Markov models (HMMs) are stoc hastic models capable of statistical learning and classification. They have been used in speech recognition and handwriting recognition because of their great adaptability and versatility. However, as these models deal with the state transition information in just one direction, they do not fully explore the bidirectional context dependence contained in a random sequence. The state transition information in the other direction is missing from observation evaluation and model training. This paper presents a dual hidden Markov model (DHMM) whic h gives a solution to the above problem. The model is based on the first order partial context dependence assumption. With this assumption, a DHMM is relativ ely simple such that it contains two HMMs that share a common sym bol emission matrix. One of the HMMs deals with the backward context dependece (forward state transition) while the other deals with the forward context dependence (backward state transition). The bidirectional context information is then integrated into observation evaluation and model training. The observation evaluation is solved using Baum's forward-backward procedure, and the DHMM m ultiple observation training is solved using a maxim um likelihood estimation method. The deriv ed training equations guarantee not only the convergence but also the adaptability of the training process.

1 Introduction

Hidden Markov models (HMMs) are stoc hastic models which were introduced and studied in the late 1960s and early 1970s [1, 2, 3, 4, 5]. These models have been explored by many researchers since then, see for example [6, 7, 11, 12, 13, 14, 15, 16] because they have a rich mathematical structure. For many years HMMs ha ve been used in speech recognition [6, 7, 8, 9, 10, 11]. More recently they have also been proposed for handwriting recognition [17, 18, 19, 20, 21] as they are adaptive to random sequential signals and capable of statistical learning and classification.

A hidden Markov process is a doubly stochastic process: an underlying process which is hidden from observation, and an observable process which is determined by the underlying process. The underlying process is characterized by a state transition probability distribution, where a current state is hidden from observation and depends only on previous state(s). On the other hand, the observable process is characterized by a symbol emission probability distribution, where a current symbol depends on the current state transition, or simply the current state.

Indeed, HMMs ha ve a great adaptability in handling sequential signals. On the other hand, if one review these models from con text dependence point of view, one can see that these models utilize the context dependence in just forward state transition, and the context dependence in backward state transition is not explored. As a result, the bidirectional context information contained in a random sequence is missing from observation evaluation and model training.

This paper presents a dual hidden Markov model (DHMM) whic h gives a solution to the above problem. The model is based on the first order partial context dependence assumption and is capable of handling the

bidirectional context information. With this assumption, a DHMM is relatively simple such that it contains two HMMs that share a common symbol emission matrix. One of the HMMs deals with the backward context dependence (forward state transition) while the other deals with the forward context dependence (backward state transition). The bidirectional context information is then integrated into observation evaluation and model training.

The remainder of this paper is organized as follows. Section 2 outlines the dual hidden Markov model. Section 3 presents a solution to the observation evaluation problem using both the forward state transition and the backward state transition. Section 4 describes a method for DHMM multiple observation training. Finally, Section 5 concludes this paper.

2 Dual Hidden Markov Model

2.1 The hidden process

Let $q(t)$ be a discrete stochastic hidden process such that

$$q(t) = q_t \in S, \quad t = 1, 2, \cdots, T \tag{1}$$

where

$$S = \{S_1, S_2, \cdots, S_N\} \tag{2}$$

is a set of states, and $\forall t$, q_t satisfies:

$$\begin{aligned} &0 \le P(q_t = S_i) \le 1 \\ &\sum_{i=1}^{N} P(q_t = S_i) = 1 \end{aligned} \tag{3}$$

Thus, $q(t)$ forms a random source. From this source, we can get many different state sequences and any of them can be expressed as:

$$Q = q_1 q_2 \cdots q_T, \quad q_t \in S, \quad 1 \le t \le T \tag{4}$$

Let us consider the property of $q(t)$. Within a state sequence Q, the first order bidirectional context dependence of a current state q_t is rather complicated:

$$P(q_t | q_{t-1}, q_{t+1}) = \frac{P(q_t, q_{t-1}, q_{t+1})}{P(q_{t-1}, q_{t+1})} \tag{5}$$

And it is not feasible to construct a context dependence model corresponding to the above conditional probability distribution. For simplicity, we assume that $q(t)$ satisfies a first order partial dependence property with respect to both the backward and the forward direction:

$$\begin{aligned} P(q_t | q_{t-1}, q_{t-2}, \cdots, q_1) &= P(q_t | q_{t-1}), \quad 1 < t \le T \\ P(q_t | q_{t+1}, q_{t+2}, \cdots, q_T) &= P(q_t | q_{t+1}), \quad 1 \le t < T \end{aligned} \tag{6}$$

We also assume that the state transition is time aligned such that

$$\begin{aligned} P(q_{t+\tau} = S_j | q_{t+\tau-1} = S_i) &= P(q_t = S_j | q_{t-1} = S_i) \\ P(q_{t-\tau} = S_j | q_{t-\tau+1} = S_i) &= P(q_t = S_j | q_{t+1} = S_i) \end{aligned} \tag{7}$$

where τ is an integer representing an appropriate time shift.

The above process $q(t)$ is hidden from observation, i.e. one can not assess $q(t)$ directly but one can assess another discrete stochastic process that depends on this underlying process.

2.2 The observable process

Let $o(t)$ be a discrete stochastic process such that

$$o(t) = o_t \in V, \quad t = 1, 2, \cdots, T \tag{8}$$

where

$$V = \{V_1, V_2, \cdots, V_M\} \tag{9}$$

is a set of observation symbols, and $\forall t$, o_t satisfies:

$$\begin{aligned} &0 \le P(o_t = V_j) \le 1 \\ &\sum_{j=1}^{M} P(o_t = V_j) = 1 \end{aligned} \tag{10}$$

Thus, any observation sequence can be expressed as:

$$O = o_1 o_2 \cdots o_T, \quad o_t \in V, \quad 1 \le t \le T \tag{11}$$

Now consider the property of $o(t)$ and notice that it depends on the underlying process $q(t)$ only. For simplicity, we assume that $o(t)$ satisfies the following symbol emission property:

$$P(o_t | X) = \begin{cases} P(o_t | q_t), & q_t \in X \\ P(o_t), & q_t \notin X \end{cases} \tag{12}$$

where X is a set of random events and $o_t \notin X$[1].

2.3 Dual properties

Let us consider the probability of a state sequence Q given all the above assumptions. Denoting the model as λ, we have

$$\begin{aligned} P(Q|\lambda) &= P(q_1|\lambda) P(q_2|q_1, \lambda) \cdots P(q_T | q_{T-1}, \lambda) \\ P(Q|\lambda) &= P(q_T|\lambda) P(q_{T-1}|q_T, \lambda) \cdots P(q_1 | q_2, \lambda) \end{aligned} \tag{13}$$

Thus, we can use a dual pair of HMMs $\lambda^{(f)}$ and $\lambda^{(b)}$ to represent λ:

$$\lambda = (\lambda^{(f)}, \lambda^{(b)}) \tag{14}$$

[1] If we let $o_t \in X$, in the case that $X = \{o_t\}$, we will have $P(o_t | o_t) = 1$.

Using this notation, we have

$$\begin{aligned} P(Q|\lambda) &= P(q_1|\lambda^{(f)}) \cdots P(q_T|q_{T-1}, \lambda^{(f)}) \\ P(Q|\lambda) &= P(q_T|\lambda^{(b)}) \cdots P(q_1|q_2, \lambda^{(b)}) \end{aligned} \tag{15}$$

and the probability of a state sequence Q given a model λ becomes

$$P(Q|\lambda) = \frac{1}{2}\left[P(Q|\lambda^{(f)}) + P(Q|\lambda^{(b)})\right] \tag{16}$$

2.4 Model elements

Based on previous discussion, we can construct a dual hidden Markov model using the following elements:

- *forward state transition matrix*

$$A^{(f)} = \{a_{ij}^{(f)}\} \tag{17}$$

where

$$\begin{aligned} a_{ij}^{(f)} &= P(q_t = S_j | q_{t-1} = S_i) \\ \sum_{j=1}^{N} a_{ij}^{(f)} &= 1 \end{aligned} \tag{18}$$

- *backward state transition matrix*

$$A^{(b)} = \{a_{ij}^{(b)}\} \tag{19}$$

where

$$\begin{aligned} a_{ij}^{(b)} &= P(q_t = S_j | q_{t+1} = S_i) \\ \sum_{j=1}^{N} a_{ij}^{(b)} &= 1 \end{aligned} \tag{20}$$

- *symbol emission matrix*

$$B = \{b_j(k)\} \tag{21}$$

where

$$\begin{aligned} b_j(k) &= P(o_t = V_k | q_t = S_j) \\ \sum_{k=1}^{M} b_j(k) &= 1 \end{aligned} \tag{22}$$

- *initial state distribution matrix*

$$\pi^{(f)} = \{\pi_i^{(f)}\} \tag{23}$$

where

$$\begin{aligned} \pi_i^{(f)} &= P(q_1 = S_i) \\ \sum_{i=1}^{N} \pi_i^{(f)} &= 1 \end{aligned} \tag{24}$$

- *final state distribution matrix*

$$\pi^{(b)} = \{\pi_i^{(b)}\} \tag{25}$$

where

$$\begin{aligned} \pi_i^{(b)} &= P(q_T = S_i) \\ \sum_{i=1}^{N} \pi_i^{(b)} &= 1 \end{aligned} \tag{26}$$

Using the above elements, the dual pair of HMMs involved in this description can be expressed as

$$\begin{aligned} \lambda^{(f)} &= (A^{(f)}, B, \pi^{(f)}) \\ \lambda^{(b)} &= (A^{(b)}, B, \pi^{(b)}) \end{aligned} \tag{27}$$

2.5 Model types

Similar to a hidden Markov model, $\lambda^{(f)}$ and $\lambda^{(b)}$ can be classified into either ergodic model or unidirectional (left-right) model in the light of the constrait on their state transition, where the former type has full state transition while the latter type has only unidirectional state transition. In the light of its component models, a dual hidden Markov model can be classified into one of the four types:

1. ergodic-ergodic
2. ergodic-unidirectional
3. unidirectional-ergodic
4. unidirectional-unidirectional

3 Observation Evaluation

Based on previous discussion, given an observation sequence O and a model λ, we have

$$\begin{aligned} P(O|\lambda) &= \sum_{Q} P(O|Q,\lambda) P(Q|\lambda) \\ &= \frac{1}{2}[\sum_{Q} P(O|Q,\lambda) P(Q|\lambda^{(f)}) + \\ &\qquad \sum_{Q} P(O|Q,\lambda) P(Q|\lambda^{(b)})] \\ &= \frac{1}{2}[P(O|\lambda^{(f)}) + P(O|\lambda^{(b)})] \end{aligned} \tag{28}$$

Associating Baum's forward and backward variables with the forward direction (see Figure 1), the forward component probability can be computed as:

$$P(O|\lambda^{(f)}) = \begin{cases} \sum_{i=1}^{N} \alpha_t^{(f)}(i) \beta_t^{(f)}(i), & \forall t \\ \sum_{i=1}^{N} \alpha_T^{(f)}(i), & t = T \end{cases} \tag{29}$$

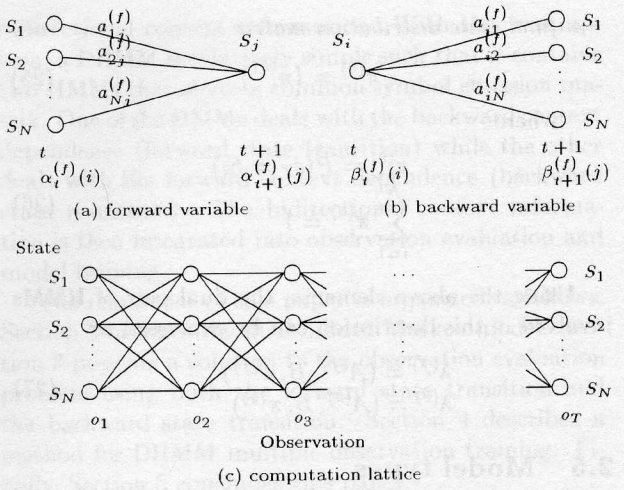

Figure 1: Forward-backward procedure in forward state transition

where $\alpha_t^{(f)}(i)$ and $\beta_t^{(f)}(i)$ are the forwad variable and the backward variable associated with the forward direction, respectively:

$$\alpha_t^{(f)}(i) = P(o_1 o_2 \cdots o_t, q_t = S_i | \lambda^{(f)}) \tag{30}$$

$$\beta_t^{(f)}(i) = P(o_{t+1} o_{t+2} \cdots o_T | q_t = S_i, \lambda^{(f)}) \tag{31}$$

and $\alpha_t^{(f)}(i)$ and $\beta_t^{(f)}(i)$ can be solved inductively, see [9] for more details.

The backward component probability can be computed in a symmetric manner by associating Baum's forward and backward variables with the backward direction. We drop the details here for brevity.

From the above one can see that the computational complexity of this procedure is $O(TN^2)$.

4 Multiple Observation Training

Let us consider a set of observation sequences from a pattern class:

$$\mathbf{O} = \{O^{(1)}, O^{(2)}, \cdots, O^{(K)}\} \tag{32}$$

where

$$O^{(k)} = o_1^{(k)} o_2^{(k)} \cdots o_{T_k}^{(k)}, \quad 1 \le k \le K \tag{33}$$

are individual observation sequences. Assuming that $O^{(k)}$ are independent each other, the multiple observation probability under the model can be expressed as:

$$P(\mathbf{O}|\lambda) = \prod_{k=1}^{K} P(O^{(k)}|\lambda) \tag{34}$$

Using the maximum likelihood estimate method, we have the following training equations for the forward component model:

1. forward state transition probability:

$$\bar{a}_{mn}^{(f)} = \frac{\sum_{k=1}^{K} \nu_k^{(f)} \sum_{t=1}^{T_k-1} \xi_t^{(f)(k)}(m,n)}{\sum_{k=1}^{K} \nu_k^{(f)} \sum_{t=1}^{T_k-1} \gamma_t^{(f)(k)}(m)} \tag{35}$$

where $1 \le m, n \le N$.

2. symbol emission probability:

$$\bar{b}_n(m) = \frac{\sum_{k=1}^{K} [u_k^{(f)} + u_k^{(b)}]}{\sum_{k=1}^{K} [v_k^{(f)} + v_k^{(b)}]} \tag{36}$$

where $1 \le n \le N$, $1 \le m \le M$, and

$$\begin{aligned} u_k^{(f)} &= \nu_k^{(f)} \sum_{t=1, o_t^{(k)}=v_m}^{T_k} \gamma_t^{(f)(k)}(n) \\ u_k^{(b)} &= \nu_k^{(b)} \sum_{t=T_k, o_t^{(k)}=v_m}^{1} \gamma_t^{(b)(k)}(n) \\ v_k^{(f)} &= \nu_k^{(f)} \sum_{t=1}^{T_k} \gamma_t^{(f)(k)}(n) \\ v_k^{(b)} &= \nu_k^{(b)} \sum_{t=T_k}^{1} \gamma_t^{(b)(k)}(n) \end{aligned} \tag{37}$$

3. initial state probability:

$$\bar{\pi}_n^{(f)} = \frac{\sum_{k=1}^{K} \nu_k^{(f)} \gamma_1^{(f)(k)}(n)}{\sum_{k=1}^{K} \nu_k^{(f)}} \tag{38}$$

where $1 \le n \le N$.

In the above equations, the coefficient $\nu_k^{(f)}$ is given by:

$$\nu_k^{(f)} = \frac{P(O^{(k)}|\lambda^{(f)})}{P(O^{(k)}|\lambda)} \tag{39}$$

and $\xi_t^{(f)(k)}(m,n)$ and $\gamma_t^{(f)(k)}(m)$ are joint event and state variable associated with $O^{(k)}$ in forward state

transition, respectively. $\xi_t^{(f)(k)}(m,n)$ and $\gamma_t^{(f)(k)}(m)$ are defined as:

$$\begin{aligned}\xi_t^{(f)(k)}(m,n) &= P(q_t = S_m, q_{t+1} = S_n | O^{(k)}, \lambda^{(f)}) \\ &= \frac{\alpha_t^{(f)(k)}(m) a_{mn}^{(f)} b_n(o_{t+1}^{(k)}) \beta_{t+1}^{(f)(k)}(n)}{P(O^{(k)}|\lambda^{(f)})} \end{aligned} \quad (40)$$

$$\begin{aligned}\gamma_t^{(f)(k)}(m) &= P(q_t = S_m | O^{(k)}, \lambda^{(f)}) \\ &= \sum_{n=1}^{N} \xi_t^{(f)(k)}(m,n) \end{aligned} \quad (41)$$

And $\nu_k^{(b)}$ and $\gamma_t^{(b)(k)}(m)$ are associated with $O^{(k)}$ in backward state transition and hence can be defined symmetrically.

The training of the backward component model is symmetric and hence we drop the details for brevity.

5 Conclusions

In this paper, we have presented a dual hidden Markov model (DHMM) capable of handling bidirectional context dependence. The DHMM contains two hidden Markov models (HMMs) that share a common symbol emission matrix. One of the HMMs deals with the backward context dependence (forward state transition) while the other deals with the forward context dependence (backward state transition). The bidirectional information is then integrated into state sequence evaluation.

Based on this integration, the observation evaluation problem has been solved using Baum's forward-backward procedure. The DHMM multiple observation training equations have also been derived using the maximum likelihood estimation method. These equations always guarantee the convergence of the training process.

The DHMM has the advantage over the conventional HMM in that that it explores the context dependence bidirectionally while it does not increase the computational complexity. Furthermore, the concept of bidirectional context dependence can be generalized into the concept of multidirectional context dependence and hence it can help to explore more complicated cases such as a hidden Markov random field (HMRF).

Acknowledgments

This research work was supported by NSERC Grants OGP00915 and STP 0192785 and FCAR Grant IL-0004 to R. Plamondon.

References

[1] L.E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite state Markov chains". *The Annals of Mathematical Statistics*, Vol.37, 1554-1563 (1966)

[2] L.E. Baum and J.A. Egon, "An inequality with applications to statistical estimation for probabilistic functions of a Markov process and to a model for ecology". *Bull. Amer. Meteorol Soc.*, Vol.73, 360-363 (1967)

[3] L.E. Baum and G.R. Sell, "Growth functions for transformations on manifolds". *Pac. J. Math.*, Vol.27, No.2, 211-227 (1968)

[4] L.E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains". *The Annals of Mathematical Statistics*, Vol.41, No.1, 164–171 (1970)

[5] L.E. Baum, "An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes". *Inequalities*, Vol.3, 1-8 (1970)

[6] S.E. Levinson, L.R. Rabiner and M.M. Sondhi, "An introduction to the application of the theory of probabilistic functions of Markov process to automatic speech recognition". *Bell System Technical Journal*, Vol.62, No.4, 1035-1074 (1983)

[7] L. R. Bahl, F. Jelinek and R. L. Mercer, "A maximum likelihood approach to continuous speech recognition". *IEEE Trans. Pattern Anal. Machine Intell.* Vol. PAMI-5, 179-190 (1983)

[8] L. R. Rabiner and S. E. Levinson, "A speaker-independent, syntax-directed, connected word recognition system based on hidden Markov models and level building". *IEEE trans. Acoust. Speech Signal Processing*, Vol. ASSP 33, No. 3, 561-573 (1985)

[9] Lawrence R. Rabiner, "A Tutorial on hidden Markov models and selected applications in speech recognition". *Proceedings of the IEEE*, 77(2). 257-286 (1989)

[10] Kai-Fu Lee, "Automatic Speech Recognition - the Development of SPHINX System". *Kluwer Academic Publishers* (1989)

[11] L. Rabiner and B.H. Juang, "Fundamentals of Speech Recognition". Prentice Hall, Englewood Cliffs, N.J., 1993

[12] Makoto Iwayama, Nitin Indurkhya and Hiroshi Motoda, "A new algorithm for automatic configuration of hidden Markov models". *Proc. 4th International Workshop on Algorithmic Learning Theory (ALT'93)*, 237–250 (1993)

[13] A. Kaltenmeier, T. Caesar, J.M. Gloger and E. Mandler, "Sophisticated topology of hidden Markov models for cursive script recognition". *Proceedings of the 2nd International Conference on Document Analysis and Recognition*, 139-142 (1993)

[14] P. Baldi and Y. Chauvim, "Smooth on-line learning algorithms for hidden Markov models". *Neural Computation*, Vol.6, 307–318 (1994)

[15] S.B. Cho and J.H. Kim, "An HMM/MLP architecture for sequence recognition". *Neural Computation*, Vol.7, 358–369 (1995)

[16] J. Dai, "Robust Estimation of HMM parameters using fuzzy vector quantization and Parzen's window". *Pattern Recognition*, Vol.28, No.1, 53–57 (1995)

[17] A. Kundu, Y. He and P. Bahl, "Recognition of handwritten word: first and second order hidden Markov model based approach". *Pattern Recognition*, Vol.22, No.3, 283–297 (1989)

[18] S.R. Veltman and R. Prasad, "Hidden Markov models applied to on-line handwritten isolated character recognition". *IEEE Trans. Image Processing*, Vol.3, No.3, 314–318 (1994)

[19] E.J. Bellegarda, J.R. Bellegarda, D. Nahamoo and K.S. Nathan, "A fast statistical mixture algorithm for on-line handwriting recognition". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.16, No.12, 1227–1233 (1994)

[20] G. Rigoll, A. Kosmala, J. Rottland, and Ch. Neukirchen, "A Comparison between continuous and discrete density hidden Markov models for cursive handwriting recognition". *Proceedings of ICPR'96*, 205–209 (1996)

[21] J. Hu, M.K. Brown and W.Turin, "HMM based on-line handwriting recognition". *IEEE transactions on Pattern Analysis and Machine Intelligence*, Vol.18, No.10, 1039-1045 (1996)