Artificial Intelligence 174 (2010) 215–243

Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint

Hidden semi-Markov models

Shun-Zheng Yu

Department of Electronics and Communication Engineering, Sun Yat-Sen University, Guangzhou 510275, PR China

ARTICLE INFO

Article history:
Received 14 April 2009
Available online 17 November 2009

Keywords:
Hidden Markov model (HMM)
Hidden semi-Markov model (HSMM)
Explicit duration HMM
Variable duration HMM
Forward–backward (FB) algorithm
Viterbi algorithm

ABSTRACT

As an extension to the popular hidden Markov model (HMM), a hidden semi-Markov model (HSMM) allows the underlying stochastic process to be a semi-Markov chain. Each state has variable duration and a number of observations being produced while in the state. This makes it suitable for use in a wider range of applications. Its forward–backward algorithms can be used to estimate/update the model parameters, determine the predicted, filtered and smoothed probabilities, evaluate goodness of an observation sequence fitting to the model, and find the best state sequence of the underlying stochastic process. Since the HSMM was initially introduced in 1980 for machine recognition of speech, it has been applied in thirty scientific and engineering areas, such as speech recognition/synthesis, human activity recognition/prediction, handwriting recognition, functional MRI brain mapping, and network anomaly detection. There are about three hundred papers published in the literature. An overview of HSMMs is presented in this paper, including modelling, inference, estimation, implementation and applications. It first provides a unified description of various HSMMs and discusses the general issues behind them. The boundary conditions of HSMM are extended. Then the conventional models, including the explicit duration, variable transition, and residential time of HSMM, are discussed. Various duration distributions and observation models are presented. Finally, the paper draws an outline of the applications.

1. Introduction (History)

A hidden Markov model (HMM) is defined as a doubly stochastic process. The underlying stochastic process is a discrete-time finite-state homogeneous Markov chain. The state sequence is not observable and so is called hidden. It influences another stochastic process that produces a sequence of observations. An excellent tutorial of HMMs can be found in Rabiner [150], a theoretic overview of HMMs can be found in Ephraim and Merhav [57] and a discussion on learning and inference in HMMs in understanding of Bayesian networks is presented in Ghahramani [66]. The HMMs are an important class of models that are successful in many application areas. However, due to the non-zero probability of self-transition of a non-absorbing state, the state duration of an HMM is implicitly a geometric distribution. This makes the HMM has limitations in some applications.

As an extension of the HMM, a hidden semi-Markov model (HSMM) is traditionally defined by allowing the underlying process to be a semi-Markov chain. Each state has a variable duration, which is associated with the number of observations produced while in the state. The HSMM is also called "explicit duration HMM" [60,150], "variable-duration HMM" [107, 155,150], "HMM with explicit duration" [124], "hidden semi-Markov model" [126], generalized HMM [94], segmental HMM [157] and segment model [135,136] in the literature, depending on their assumptions and their application areas.

E-mail address: syu@mail.sysu.edu.cn.

0004-3702/$ – see front matter
doi:10.1016/j.artint.2009.11.011

The first approach to hidden semi-Markov model was proposed by Ferguson [60], which is partially included in the survey paper by Rabiner [150]. This approach is called the explicit duration HMM in contrast to the implicit duration of the HMM. It assumes that the state duration is generally distributed depending on the current state of the underlying semi-Markov process. It also assumes the "conditional independence" of outputs. Levinson [107] replaced the probability mass functions of duration with continuous probability density functions to form a continuously variable duration HMM. As Ferguson [60] pointed out, an HSMM can be realized in the HMM framework in which both the state and its sojourn time since entering the state are taken as a complex HMM state. This idea was exploited in 1991 by a 2-vector HMM [93] and a duration-dependent state transition model [179]. Since then, similar approaches were proposed in many applications. They are called in different names such as inhomogeneous HMM [151], non-stationary HMM [164], and recently triplet Markov chains [144]. These approaches, however, have the common problem of computational complexity in some applications. A more efficient algorithm was proposed in 2003 by Yu and Kobayashi [199], in which the forward–backward variables are defined using the notion of a state together with its remaining sojourn (or residual life) time. This makes the algorithm practical in many applications.

The HSMM has been successfully applied in many areas. The most successful application is in speech recognition. The first application of HSMM in this area was made by Ferguson [60]. Since then, there have been more than one hundred such papers published in the literature. It is the application of HSMM in speech recognition that enriches the theory of HSMM and develops many algorithms for HSMM.

Since the beginning of 1990's, the HSMM started being applied in many other areas such as electrocardiograph (ECG) [174], printed text recognition [4] or handwritten word recognition [95], recognition of human genes in DNA [94], language identification [118], ground target tracking [88], document image comparison and classification at the spatial layout level [81], etc.

In recent years from 2000 to present, the HSMM has been obtained more and more attentions from vast application areas such as change-point/end-point detection for semi-conductor manufacturing [64], protein structure prediction [162], mobility tracking in cellular networks [197], analysis of branching and flowering patterns in plants [69], rain events time series model [159], brain functional MRI sequence analysis [58], satellite propagation channel modelling [112], Internet traffic modelling [198], event recognition in videos [79], speech synthesis [204,125], image segmentation [98], semantic learning for a mobile robot [167], anomaly detection for network security [201], symbolic plan recognition [54], terrain modelling [185], adaptive cumulative sum test for change detection in non-invasive mean blood pressure trend [193], equipment prognosis [14], financial time series modelling [22], remote sensing [147], classification of music [113], and prediction of particulate matter in the air [52], etc.

The rest of the paper is organized as follows: Section 2 is the major part of this paper that defines a unified HSMM and addresses important issues related to inference, estimation and implementation. Section 3 then presents three conventional HSMMs that are applied vastly in practice. Section 4 discusses the specific modelling issues, regarding duration distributions, observation distributions, variants of HSMMs, and the relationship to the conventional HMM. Finally, Section 5 highlights major applications of HSMMs and concludes the paper in Section 6.

2. Hidden semi-Markov model

This section provides a unified description of HSMMs. A general HSMM is defined without specific assumptions on the state transitions, duration distributions and observation distributions. Then the important issues related to inference, estimation and implementation of the HSMM are discussed. A general expression of the explicit-duration HMMs and segment HMMs can be found in Murphy [126], and a unified view of the segment HMMs can be found in Ostendorf et al. [136]. Detailed review for the conventional HMM can be found in the tutorial by Rabiner [150], the overview by Ephraim and Merhav [57], the Bayesian networks-based discussion by Ghahramani [66], and the book by Cappe et al. [29].

2.1. General model

A hidden semi-Markov model (HSMM) is an extension of HMM by allowing the underlying process to be a semi-Markov chain with a variable *duration* or *sojourn time* for each state. Therefore, in addition to the notation defined for the HMM, the duration d of a given state is explicitly defined for the HSMM. State duration is a random variable and assumes an integer value in the set $\mathcal{D} = \{1, 2, \ldots, D\}$. The important difference between HMM and HSMM is that one observation per state is assumed in HMM while in HSMM each state can emit a sequence of observations. The number of observations produced while in state i is determined by the length of time spent in state i, i.e., the duration d. Now we provide a unified description of HSMMs.

Assume a discrete-time Markov chain with the set of (hidden) states $\mathcal{S} = \{1, \ldots, M\}$. The state sequence is denoted by $S_{1:T} \triangleq S_1, \ldots, S_T$, where $S_t \in \mathcal{S}$ is the state at time t. A realization of $S_{1:T}$ is denoted as $s_{1:T}$. For simplicity of notation in the following sections, we denote:

- $S_{t_1:t_2} = i$ – state i that the system stays in during the period from t_1 to t_2. In other words, it means $S_{t_1} = i$, $S_{t_1+1} = i, \ldots,$ and $S_{t_2} = i$. Note that the previous state S_{t_1-1} and the next state S_{t_2+1} may or may not be i.