Available online at www.sciencedirect.com
ScienceDirect
Mathematics and Computers in Simulation 137 (2017) 350–365
MATHEMATICS AND COMPUTERS IN SIMULATION
www.elsevier.com/locate/matcom

Original articles

The evaluation problem in discrete semi-hidden Markov models

J.F. Gómez-Lopera[a], J. Martínez-Aroza[b,*], R. Román-Roldán[a], R. Román-Gálvez[a], D. Blanco-Navarro[a]

[a] Dept. Física Aplicada, University of Granada, Avenida Fuente nueva s.n. 18071, Granada, Spain
[b] Dept. Matemática Aplicada, University of Granada, Avenida Fuente nueva s.n. 18071, Granada, Spain

Received 29 October 2015; received in revised form 22 September 2016; accepted 13 December 2016
Available online 29 December 2016

Abstract

This paper is devoted to discrete semi-hidden Markov models (*SHMM*), which are related to the well-known hidden Markov models (*HMM*). In particular, the *HMM* associated to an *SHMM* is defined, and the forward algorithm for solving the evaluation problem in *SHMM*s is introduced. Experiments show that in a set of randomly generated sequences with different *SHMM*s, the maximum value for the conditional probability of each sequence being generated by the model most frequently matches the model that generated the sequence. Something similar happens to associated *HMM*s, suggesting that the *HMM* associated to a given *SHMM* shows a certain affinity to this, which is higher than other *HMM*s.

Keywords: Semi-hidden Markov models; Hidden Markov models; Symbolic run sequences; Evaluation problem; Forward algorithm

1. Introduction

A hidden Markov model (hereinafter *HMM*) is a powerful statistical tool to characterize discrete-time series. In this context the system being modelled is assumed to be a Markov process with unobserved (hence hidden) states.

Initially proposed by L.E. Baum and others [1–5], *HMM*s are widely used in science and engineering in many areas such as speech recognition, optical character recognition, machine translation, computer vision, finance and economics, social sciences, etc. *HMM*s are especially known for their application in temporal pattern recognition such as speech, handwriting, gesture recognition, part-of-speech tagging, musical-score following and bioinformatics [8–13,15].

Discrete semi-hidden Markov models (hereinafter *SHMM*), recently introduced in [14], are a new kind of stochastic models related with *HMM*s. Strictly speaking, an *SHMM* is not a Markov model, since it does not verify the "Markov property" [7], that is a characteristic of memoryless sources. A source driven by an *SHMM* changes its state depending not on the current state but on the last already emitted symbols, unlike *HMM*s. The number of last emitted symbols determining the new state can be called the *memory size* of the source.

* Corresponding author.
E-mail address: jmaroza@ugr.es (J. Martínez-Aroza).

http://dx.doi.org/10.1016/j.matcom.2016.12.002

Given a symbolic sequence emitted by an *HMM* source, determining the sequence of states run by the source is, in general, impossible. However, given a symbolic sequence generated by an *SHMM* source, then it is possible to know the sequence of states if we know the state changing positions [14]. This is why an *SHMM* is not completely hidden, and it is called semi-hidden.

There are certain symbolic sequences in which long and frequent runs can appear, either of a unique repeated symbol, or also compositional runs (constant frequencies for each symbol along a fragment of the sequence). Furthermore, there can be cases in which the emitting source shows a high inertia (high resistance to change the current state). *SHMM* are intended to modelize the behaviour of symbolic sequences showing memory and inertia, such as a temporal series of temperatures in a short period of time. The temperature of an object in a given instant is not expected to be very different than that in the previous instant, which means that the system evolves with some kind of memory and inertia.

This paper is organized as follows: in Section 2, formal descriptions of *HMM* and *SHMM* are included; in Section 3 a new concept, the *HMM* associated to a given *SHMM*, is introduced and illustrated; in Section 4, the forward algorithm for *HMM*s is described[1]; in Section 5, an analogous efficient new algorithm is introduced to solve the same problem as in the previous section, but related to a given *SHMM* source; Section 6 is devoted to presenting a set of experiments to show the performance of the forward algorithm to discriminate between the corresponding generating models and their associated ones.

2. Component elements of discrete *HMM* and *SHMM*

Although both kind of stochastic models have been already defined, definitions of *HMM* and *SHMM* are given in this Section. Since *HMM*s are well known and widely studied in literature [1–5,13], no more comments are included here. Concerning *SHMM*s, a brief explanation of the meaning and effect of each *SHMM* parameter on the generated sequences is included at the end of this Section.

2.1. Elements of an HMM

The elements of a discrete *HMM* are the following [13]:

- The alphabet $V = \{v_1, \dots, v_m\}$, with m symbols, $m \in \mathbb{N}$.
- The set $\{1, 2, \dots, n\}$ of n hidden states, $n \in \mathbb{N}$.
- The $n \times n$ right stochastic transition matrix among states $A = \{a_{ij}\},\ 1 \le i, j \le n,\ 0 \le a_{ij} \le 1,\ \sum_{j=1}^{n} a_{ij} = 1\ \forall i$. Each matrix element $a_{ij} = P[q_{t+1} = j | q_t = i]$ is the probability of changing from state i to j at any position t when generating a sequence, being q_t the current state at position t in the sequence. As in [13], we label the individual states as $\{1, 2, \dots, n\}$.
- The $n \times m$ right stochastic matrix of probability distributions of symbols emission in each state $B = \{b_{jk}\},\ 0 \le b_{jk} \le 1,\ 1 \le j \le n,\ 1 \le k \le m,\ \sum_{k=1}^{m} b_{jk} = 1\ \forall j$.
- The initial-state probability distribution $\Pi = \{\pi_i\},\ i = 1, 2, \dots, n$.

Note that in the definition of matrices A and B, a numerical order in the model states and alphabet symbols is assumed. In the specialized literature, an *HMM* is denoted by the 3-tuple $\lambda = \{A, B, \Pi\}$. A source that emits symbols driven by an *HMM* is called an *HMM* source.

2.2. Elements of a discrete SHMM

A discrete *SHMM* is a stochastic model that operates by switching among different states and generating symbols of an alphabet in a similar way to an *HMM*. It has been presented in [14] and its main features are included here for the benefit of the reader. Although they are not Markov models (do not verify the "Markov property"), these models are related to *HMM* ones, and are intended to generate and analyse symbolic sequences containing frequent runs.[2] The elements of an *SHMM* model are the following:

[1] The forward algorithm is a procedure to efficiently compute the conditional probability $P[O|\lambda]$ of generating a particular sequence O, given the model λ.

[2] In this context we consider that a run is a segment of statistically constant composition along its length, including the degenerate case of a chain of one repeated symbol.