Available online at www.sciencedirect.com
ScienceDirect
Mathematics and Computers in Simulation 118 (2015) 320–328
MATHEMATICS AND COMPUTERS IN SIMULATION
www.elsevier.com/locate/matcom

Original articles

Semi-hidden Markov models for generation and analysis of sequences

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

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

Received 18 March 2014; received in revised form 31 July 2014; accepted 20 November 2014
Available online 12 December 2014

Abstract

In this work a new kind of stochastic model is presented, the semi-hidden Markov model (*SHMM*). The proposed model is related to the hidden Markov model (*HMM*), and it is called semi-hidden because generated sequences need less information than *HMM* sequences to infer the succession of states run by the source.

The main feature of *SHMM* is that they work with statistical memory, i.e. the symbol's emission probability distribution on the current state of the emitting source depends on a number of symbols already emitted in the previous state. The proposed model is useful for the generation and analysis of processes and symbolic sequences containing runs.

Keywords: Hidden Markov models, *HMM*; Generation of symbolic sequences; Symbolic run sequences

1. Introduction

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

Initially proposed by L.E. Baum and others [3,2,5,4,1] *HMM* are widely used in science, engineering and many other areas like speech recognition, optical character recognition, machine translation, computer vision, finance and economics, social sciences, etc. [12]. *HMM* 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 [9,8,7,10,13].

2. Hidden Markov models *(HMM)*

A Markov chain (a first order *HMM*) is a discrete-time random process which generates a sequence of symbols with the Markov propriety: the probability of switching to any particular state only depends on the current state. Consequently, in a Markov chain the state is directly visible to the observer, and therefore the state transition probabilities

* Corresponding author.
E-mail address: jfgomez@ugr.es (J.F. Gómez-Lopera).

http://dx.doi.org/10.1016/j.matcom.2014.11.009

are the only parameters. In a *HMM*, the state is not directly visible, but only the output (dependent on the state). This model can be seen as a double stochastic process, in which every state is characterized by two probability distributions: a probability distribution over the possible output symbols and a probability distribution that controls the change of state. Therefore the sequence of symbols generated by a *HMM* gives some information about the sequence of states, although not all information. In this context, 'hidden' refers to the state sequence followed by the source, not to the parameters of the model. In fact, even if the model parameters are known, the model still remains 'hidden' in the sense that the state sequence controlling the output is, in general, unknown.

The elements of a discrete *HMM* are the following [12]:

- The alphabet $V = \{v_1, \ldots, v_m\}$, with m symbols.
- The set of n hidden states $S = \{s_1, \ldots, s_n\}$.
- The $n \times n$ left stochastic transition matrix among states $A = \{a_{ij}\}$, $1 \leq i, j \leq n$, $0 \leq a_{ij} \leq 1$, $\sum_{i=1}^{n} a_{ij} = 1$.
- The $m \times n$ matrix of distributions of symbols' emission probability in each state $B = \{b_{ij}\}$, $i = 1, \ldots, m$, $j = 1, \ldots, n$, $0 \leq b_{ij} \leq 1$, $\sum_{i=1}^{m} b_{ij} = 1$.
- The initial-state probability distribution $\Pi = \{\pi_i\}$, $i = 1, 2, \ldots, m$.

A *HMM* is denoted by $\lambda = \{A, B, \Pi\}$. The three fundamentals problems to solve in the field of *HMM* are the following:

- Evaluation problem: given an observed sequence $O = \{o_1, \ldots, o_T\}$ and a model $\lambda = \{A, B, \Pi\}$, evaluate the conditional probability $P(O|\lambda)$ of O being generated by λ. This problem can be solved by means of the forward algorithm [12].
- Decoding problem: given an observed sequence $O = \{o_1, \ldots, o_T\}$ and a model $\lambda = \{A, B, \Pi\}$, obtain the 'optimal' sequence of states $Q = \{q_1, \ldots, q_T\}$ (in the sense that it best explain the observations). This problem can be solved by means of the Viterbi algorithm [12].
- Training problem: given an observed sequence $O = \{o_1, \ldots, o_T\}$, adjust the model parameters $\lambda = \{A, B, \Pi\}$ to maximize $P(O|\lambda)$. This is a problem difficult to solve since it has no analytical solution. Some usual algorithms to approximate the solution are Dempster [6] and Levinson [11].

3. Semi-hidden Markov models

Semi-Hidden Markov models (hereinafter *SHMM*) are proposed as a modification of *HMM*. The main feature of this kind of models is the inclusion of statistical inertia, which allows the generation and analysis of symbolic sequences containing runs. In this way, each new state is determined by the sequence of z last symbols generated. In addition, the lower the value of z, the lower the number of possible states.

3.1. Formal definition of a SHMM

The elements of a discrete *SHMM* are the following:

- The alphabet $V = \{v_1, \ldots, v_m\}$, with m symbols.
- The stationary probability distribution $P^* = \{p_1^*, \ldots, p_m^*\}$. It is used to compute the probability distribution of every state.
- The length of statistical inertia $z \in \mathbb{N}$, $z \geq 1$.
- The weight of statistical inertia $w \in \mathbb{R}$, $0 \leq w \leq 1$.
- The probability of changing the state $\pi \in \mathbb{R}$, $0 \leq \pi \leq 1$.

A source driven by a given *SHMM* $\lambda = \{P^*, z, w, \pi\}$ generates a sequence of symbols $O = \{o_1, o_2, \ldots, o_T\}$ according to the following procedure:

1. Generate z random symbols following the stationary probability distribution $P = P^*$, as previous symbols not belonging to the sequence. Let F be the relative frequency distribution of these z symbols.
2. Set the initial state as the corresponding to the probability distribution
$$P = w \cdot F + (1 - w) \cdot P^* \tag{1}$$
3. Generate a symbol of the alphabet according to P and output to the sequence.