Solving Hidden-Semi-Markov-Mode Markov Decision Problems

Emmanuel Hadoux, Aurélie Beynier, and Paul Weng

Sorbonne Universités, UPMC Univ Paris 06, UMR 7606, LIP6, Paris, France
firstname.surname@lip6.fr

Abstract. Hidden-Mode Markov Decision Processes (HM-MDPs) were proposed to represent sequential decision-making problems in non-stationary environments that evolve according to a Markov chain. We introduce in this paper Hidden-Semi-Markov-Mode Markov Decision Process es (HS3MDPs), a generalization of HM-MDPs to the more realistic case of non-stationary environments evolving according to a semi-Markov chain. Like HM-MDPs, HS3MDPs form a subclass of Partially Observable Markov Decision Processes. Therefore, large instances of HS3MDPs (and HM-MDPs) can be solved using an online algorithm, the Partially Observable Monte Carlo Planning (POMCP) algorithm, based on Monte Carlo Tree Search exploiting particle filters for belief state approximation. We propose a first adaptation of POMCP to solve HS3MDPs more efficiently by exploiting their structure. Our empirical results show that the first adapted POMCP reaches higher cumulative rewards than the original algorithm. However, in larger instances, POMCP may run out of particles. To solve this issue, we propose a second adaptation of POMCP, replacing particle filters by exact representations of beliefs. Our empirical results indicate that this new version reaches high cumulative rewards faster than the former adapted POMCP and still remains efficient even for large problems.

1 Introduction

Markov Decision Processes (MDPs) provide a general formal framework for sequential decision-making under uncertainty. They have proved to be powerful for solving many planning problems [14]. However, MDPs run under the assumption that the environment is stationary, i.e., the transition function and/or the reward function do not evolve through time. In many real-world applications, this assumption does not hold and the sources of non-stationarity are diverse. For instance, the environment may change due to external events. In finance, when investing on the stock market, a financial crisis changes the dynamics of stock prices. Another example of non-stationary environment concerns multi-agent systems. Indeed, from the viewpoint of one agent, a change of behavior (e.g., due to learning) of another agent may affect the environment of the first agent.

U. Straccia and A. Calì (Eds.): SUM 2014, LNAI 8720, pp. 176–189, 2014.

Planning in a non-stationary environment is a difficult problem to tackle in the general case. We focus instead on a subclass of problems where non-stationary environments evolve according to a small number of non-observable modes, which are modeled as MDPs and represent different possible dynamics and rewards of that environment. An example of problem belonging to this subclass is that of elevator control [6] where the environment can typically be in three modes: morning rush-hour, late-afternoon rush-hour and non-rush-hour. Planning in such non-stationary environments has already been studied in the MDP framework [8] and in the reinforcement learning framework [7,9,15]. In all those works, non-stationary environments are represented with multiple modes. The model of Hidden-Mode Markov Decision Processes (HM-MDPs) proposed by Choi et al.[8] formalizes this idea. HM-MDPs constitute a subclass of Partially Observable MDPs. In HM-MDPs, the environmental changes are described by a Markov chain and thus occur at each decision step. However, we argue that this assumption is not always realistic. Indeed, in the elevator problem for instance, allowing, even with a small probability, the environment to be able to change between different rush modes at every move of the elevator is debatable.

In this paper, we propose a natural extension of HM-MDPs, called Hidden-Semi-Markov-Mode Markov Decision Process (HS3MDP), where the non-stationary environment evolves according to a semi-Markov chain. This new model is to hidden semi-Markov models [17] what HM-MDPs are to hidden Makov models. In HS3MDPs, when the environment stochastically changes to a new mode, it stays in that mode during a stochastically drawn duration. While HM-MDPs assume that environmental changes follow a geometric law, this assumption is relaxed in HS3MDPs.

In order to solve large-sized HS3MDPs, we exploit the Partially Observable Monte Carlo Planning (POMCP) algorithm [16], an online algorithm proposed for approximately solving POMDPs, based on Monte Carlo Tree Search and particle filters for belief state approximation. We present two improvements of POMCP for solving HS3MDPs more efficiently. The first adaptation exploits the special structure of HS3MDPs and the second furthermore represents belief states exactly instead of using particle filters. Finally, we experimentally validate those algorithms showing their effectiveness on a diverse range of domains.

In Sect. 2, we recall the necessary notations and definitions. Then, in Sect. 3, we introduce our new model. In Sect. 4, we present two adapted algorithms for solving HS3MDPs. Experimental results are presented in Sect. 5. Finally, we conclude in Sect. 6.

2 Background

Markov Decision Process. A *Markov Decision Process* (MDP) [14] is defined by $\langle \mathbf{S}, \mathbf{A}, T, R \rangle$ where $\mathbf{S}$ is a finite set of states, $\mathbf{A}$ is a finite set of actions, $T(s, a, s')$ is the probability of reaching state s' from s after executing action a and $R(s, a) \in \mathbb{R}$ is the immediate reward obtained after performing action a in s. A *policy* π is a sequence $(\delta_0, \delta_1, \ldots, \delta_t, \ldots)$ of *decision rules* such as each