Biologically Inspired Cognitive Architectures 24 (2018) 1–20

Contents lists available at ScienceDirect

Biologically Inspired Cognitive Architectures

journal homepage: www.elsevier.com/locate/bica

Research article

Surprise-based learning of state representations

Thomas Joseph Collins*, Wei-Min Shen

Information Sciences Institute, University of Southern California, 4676 Admiralty Way, Marina Del Rey, CA, USA

ARTICLE INFO

Keywords:
Surprise-based learning
Active learning
Partially-observable Markov decision processes

ABSTRACT

There is an ever-increasing need for autonomous robots that are capable of adapting to and operating in challenging partially-observable and stochastic environments. Standard techniques for autonomous learning in such environments are often fundamentally reliant on human-engineered features, one of the most important of which is an *a priori* specification of the agent's *state space*. Designing an appropriate state space demands extensive domain knowledge, and even minor changes to the task or the agent might necessitate re-engineering. These limitations have given rise to end-to-end, *predictive* approaches, such as Predictive State Representations (PSRs) and our Stochastic Distinguishing Experiments (SDEs), that learn a representation of state encoded in the probabilities of key sequences of raw actions and observations (i.e., *experiments* the agent can perform). Discovering these experiments remains a key challenge, in part because existing techniques lack a formal relationship between predictive experiments and latent states in the agent's model of its environment. In this paper, we extend our SDE representation into a novel *hybrid latent-predictive* cognitive architecture in which each *latent* state is created and uniquely represented by the result of a *predictive* experiment that statistically distinguishes it from other states. We prove that deterministic environments and a useful subclass of POMDP environments can be *perfectly* represented with equivalent compactness by such models and provide an active algorithm for autonomously learning such models in unknown environments from experience based on the biologically-inspired notion of *surprises*. The agent begins using only its observations as a state space and *splits* those states into a hierarchy of additional latent states when it is *surprised* by high entropy resulting from repeatedly executing experiments that are automatically designed and selected to statistically disambiguate identical-looking states. We present experimental results demonstrating the feasibility of this learning procedure.

Introduction

There is an ever-increasing demand for autonomous robotic systems that are able to adapt to and operate in a range of challenging real-world environments. The environments in which such systems are to operate vary substantially, but almost all environments of interest exhibit high levels of *partial-observability* – meaning that the robot cannot directly sense all the salient aspects of its environment – and *stochasticity* – meaning that its actions and observations are *noisy*, *uncertain*, and *nondeterministic*. The combination of these factors makes learning particularly challenging, because it is difficult for the agent to disambiguate environmental noise from situations in which its model has failed to capture something important about the latent structure of its environment. Nevertheless, a host of biological life on Earth has successfully adapted to learning in the face of rampant partial-observability and stochasticity, which suggests that biologically-inspired methods might be a particularly powerful way to approach this problem.

In this paper, we make important theoretical progress by addressing a problem called *autonomous learning from the environment* (ALFE, first formulated in Shen, 1993a), in which an embodied agent is placed in an unknown discrete, partially-observable, stochastic environment and must build a *task-independent* model of the state and state dynamics of this environment from its experience, given the actions it can take and the observations it can make about its environment. The agent is *not* able to reset itself to a known state and has no prior knowledge about the number of underlying environment states or the expected results of executing its actions. The most distinguishing aspect of this problem is that the agent must decide simultaneously and autonomously both *what actions to take* in the absence of any external rewards and *how much experience* is sufficient to build a useful model.

In previous work, we introduced the Stochastic Distinguishing Experiments (SDEs, Collins & Shen, 2017) cognitive architecture as an approximate and purely *predictive* representation of state and state dynamics and provided a provably-convergent algorithm for actively learning an SDE model of unknown partially-observable and stochastic

* Corresponding author.
E-mail addresses: collinst@usc.edu (T.J. Collins), shen@isi.edu (W.-M. Shen).

https://doi.org/10.1016/j.bica.2018.04.005
Received 7 April 2018; Received in revised form 11 April 2018; Accepted 11 April 2018

environments based on the biologically-inspired notion of *surprises*. The resulting model could be used by the agent to approximate the history-dependent probability of any sequence of observations given any sequence of input actions. However, the connection between SDEs and the state space of the agent's environment model, the *representational capacity* of the SDE model (i.e., the types of environments SDE models could represent *perfectly*), and the applicability of existing POMDP planning and reinforcement learning techniques to SDE models were left as important future work.

In this paper, we extend our SDE representation into a novel *hybrid latent-predictive* cognitive architecture, called the *surprise-based partially-observable Markov decision process (sPOMDP)*, which is a partially-observable Markov decision process (POMDP, Kaelbling, Littman, & Cassandra, 1998) in which each *latent* state is uniquely represented by a maximally-probable *predictive* sequence of observations created by executing an associated Stochastic Distinguishing Experiment (SDE) from that state. Differences in these observation sequences can be used to statistically disambiguate identical-looking states. sPOMDPs, like other predictive models (e.g., Predictive State Representations, Littman & Sutton, 2002), are grounded in the raw actions and observations of the agent, enabling *end-to-end learning* that requires few, if any, human-engineered features; however, in contrast to other predictive models, sPOMDPs can also be used as traditional POMDPs, so state-of-the-art POMDP planning and reinforcement learning techniques can be applied straightforwardly.

We prove that Moore Machine (Moore, 1956) environments and a useful subclass of POMDP environments in which the agent experiences uniform noise around its most likely transitions and observations can be *perfectly* represented by sPOMDPs no larger than the *minimal* representations of these environments, thereby answering an important open theoretical question regarding the representational capacity of SDEs (Collins & Shen, 2017) and their precursors in the surprise-based learning literature (Shen, 1994). These theoretical results are used to develop a novel algorithm for the active, incremental learning of sPOMDP models from experience based on the biologically-inspired notion of *surprises*. The key idea is that the agent begins using only its observations as a state space and *splits* those states into a hierarchy of additional *latent* states when it is *surprised* by the high amount of entropy resulting from repeatedly executing the SDEs in its sPOMDP model. We provide experimental results demonstrating the feasibility of this learning procedure and compare its performance to the state-of-the-art purely predictive SDE model we presented in Collins and Shen (2017).

Related work

At a high level, this work is related to the problem of representation learning, which includes automatic feature extraction as a subproblem. Representation learning is currently dominated by research in probabilistic graphical models (Koller & Friedman, 2009) and deep neural networks (Goodfellow, Bengio, & Courville, 2016). These techniques have achieved tremendous success in signal processing, speech recognition, object recognition, artificial intelligence, natural language processing, etc. (Bengio, Courville, & Vincent, 2013). The success of deep reinforcement learning in playing Atari games above a human expert level (Mnih et al., 2015) and navigating complex, physics-based, 3D terrains (Heess et al., 2017) using raw pixel and sensor input are particularly powerful recent examples of what deep representations can accomplish. Nevertheless, the architecture and the number of hidden units and layers in deep neural network models continue to be important inputs that often require extensive manual tuning to yield high-quality results. In recurrent neural networks (RNNs) modeling dynamical systems, for example, the number of hidden units determines the number of underlying system states postulated (Goodfellow et al., 2016, Chapter 10), which is not available in most real-world situations without extensive human engineering. In contrast, the state space of an sPOMDP model grows nonparametrically with agent experience.

Similarly, Dynamic Bayesian Networks (DBNs), which include Hidden Markov Models (HMMs) as a special case (Koller & Friedman, 2009, Chapter 6), require an *a priori* specification of state variables, possible values, and their interconnections over time. Recent work such as the infinite Dynamic Bayesian Network (iDBN, Doshi, Wingate, Tenenbaum, & Roy, 2011) seeks to overcome this issue by using a Bayesian nonparametric model to place a prior over an unbounded number of possible DBN structures and using Markov Chain Monte Carlo (MCMC, Andrieu, De Freitas, Doucet, & Jordan, 2003) techniques to approximately infer the posterior over possible model structures given the observed data. This work differs fundamentally from sPOMDPs in that it does not consider the role of agent actions in the learning process.

A similar nonparametric model more closely related to the current work was proposed for learning a state representation of unknown POMDPs called the infinite Partially-observable Markov Decision Process (iPOMDP, Doshi-Velez, 2009; Doshi-Velez, Pfau, Wood, & Roy, 2015). One crucial difference between iPOMDPs and our sPOMDPs is that sPOMDPs do not require an external reward function for action selection: iPOMDPs require an expensive forward-looking search tree at every time step to select an approximately optimal next action based on a given reward function. This action selection routine also requires the offline solving of candidate POMDPs to estimate the Q values of various actions. In contrast, sPOMDP learning selects actions based on the amount of entropy resulting from the repeated execution of automatically designed and selected experiments, with all steps of the learning process executed in an online, incremental fashion.

Much of the research on POMDPs focuses on discrete POMDPs and is in the area of reinforcement learning (RL), where the goal is generally not to construct a state representation but rather to learn an optimal policy (and often the POMDP's dynamics). In almost every case, the agent is given a known state space. Traditional exact and approximate techniques for solving POMDPs for optimal policies include Murphy (2000), Pineau, Gordon, and Thrun (2003), and Roy, Gordon, and Thrun (2005). One of the most prominent exceptions is the family of instance-based (IB) RL methods (Liu, Jin, & She, 2016; McCallum, 1995, 1996; McCallum, Tesauro, Touretzky, & Leen, 1995), which memorize interactions with their environment and organize them into a suffix tree of actions and observations in a way that approximates an unknown discrete or continuous state space nonparametrically (in a task-specific way). Determining a memory size large enough to contain enough representative samples for good probability estimates can be challenging, and, in contrast to sPOMDP learning, a reward function is needed to guide the agent's actions.

There is a line of more recent work that sits at an interesting intersection between reinforcement learning, deep neural networks, and representation learning that also leverages mismatches between expected and actual sensory feedback in response to actions (similar in some ways to our notion of a *surprise*). This work can be broadly classified as reinforcement learning via *intrinsic motivation* (Barto, Singh, & Chentanez, 2004; Chentanez, Barto, & Singh, 2005; Mohamed & Rezende, 2015; Oudeyer & Kaplan, 2009; Oudeyer, Kaplan, & Hafner, 2007) and reinforcement learning via *artificial curiosity* (Frank, Leitner, Stollenga, Förster, & Schmidhuber, 2013; Pathak, Agrawal, Efros, & Darrell, 2017; Storck, Hochreiter, & Schmidhuber, 1995). The unifying theme of these approaches is that, since external rewards are often sparse in practice (if they exist at all), RL agents should use self-defined (intrinsic) or goal-independent metrics to evaluate their own performance and push themselves into novel situations in order to develop a hierarchy of skills that are broadly useful across a wide range of tasks. However, these approaches rely on a human-designed state space as input, making the problem being solved fundamentally different than that addressed in this work, in which the state space itself (and its dynamics) must be learned from experience.

Predictive state representations (PSRs, Littman & Sutton, 2002)