Regret Bounds for Learning State Representations in Reinforcement Learning

R. Ortner, M. Pirotta, A. Lazaric, R. Fruit, O. Maillard

Abstract

Selecting or designing the state representation is a well-known problem in RL. There are many approaches for feature extraction from high-dimensional observations. Not all these features describe the problem well or show Markovian dynamics.

We consider the feature RL problem: given a finite set $\Phi$ of models, we have to select online the appropriate model to solve the task. The learning efficiency is measured in terms of speed of convergence toward the optimal solution (i.e., regret). We introduce Ucb-Ms, an optimistic elimination algorithm that performs efficient exploration of the representations.

Example

Planar Navigation; set of state variables $\{x_0, y_0, \ldots, x_T, y_T, \phi_0, \phi_1, \phi_2, \ldots\}$ with state models: position, orientation, velocity.

What is the best representation for learning the optimal policy?

I) multiple redundant sensory measures
II) difficult also for experts to identify important variables

Settings

Online Learning

For time $t = 1, 2, \ldots$
- observe action $a_t \sim \pi$
- observe reward $r_t$ and observation $s_{t+1}$

Markov Decision Process (MDP)

$M = (S, A, r, p, \pi)$

Markov:

$P(s_{t+1}, r_t|s_t, a_t) = P(s_{t+1}|s_t, a_t)$

- Average reward $\rho^\pi(M) = \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} r_t(s_t, \pi(s_t))$
- Optimal policy $\pi^* = \arg\max_{\pi} \rho^\pi(M)$
- Diameter $D(M) = \max_{s, s'} \mathbb{E}[r_{t+1}(s_{t+1}, s')]$

$\phi$ can be an embedding from different neural networks and/or RNN.

A state-rep $\phi$ is Markov if it induces an MDP $M(\phi)$.

Problem

- Online learning
  - The learner has a finite set $\Phi$ of state-rep. models
  - At least one model $\phi \in \Phi$ is Markov

Goal

Find a policy that performs as well as $\pi^* \in \arg\max_{\pi} \rho^\pi(M(\phi))$,

$\pi^*, M(\phi)$ unknown

Why is it important?

- Difficult to know a priori if a representation is "reasonable"
- Automatically and quickly discard bad representation
- What is the best representation for learning the optimal policy?
- Learning directly online

Solution idea

quickly discard "bad" representations and keep following the optimal policy of models that perform well enough

Ucb-Ms Algorithm

For episodes $k = 1, 2, \ldots$
1. For each rep. $\phi \in \Phi$, compute optimistic policy $\pi_k\phi$

$(\hat{M}_k, \hat{\pi}_k) = \arg\max_{\phi \in \Phi} \rho^\phi(M(\phi))$

2. Choose best optimistic model

$\phi_k = \arg\max_{\phi \in \Phi} \rho^\phi(M(\phi))$

3. Execute best policy $\pi_k\phi_k$

Repeat until end of episode:

- Choose action $a_t \sim \hat{\pi}_k(s_t)$, get reward $r_t$ and observe next state $s_{t+1} = \phi_k(a_t, s_t) \in \Phi_k$;
- If $|t - t_k + 1| \geq \sqrt{\frac{AT}{D}}$

then $\hat{\pi}_k = \phi_k \cup \{s\}$ and terminate episode

Step 1) As UCRL2 [Jaksch et al., 2010], Ucb-Ms builds uncertainty about transitions and rewards for each $\phi \in \Phi$ (i.e., set of plausible MDPs).

Step 2) Optimism at the level of representations

Step 3) Discards models that are not achieving enough reward as "promised"

- Eq. 1 is a bound on the regret of a single episode of UCRL2
- An indication that the model is not Markov

Guarantees

With probability $1 - \delta$, the regret of Ucb-Ms using $D \geq D^*$ is

$R(T) \leq \text{const} \cdot D^* \sqrt{S_{\text{tr}} + \log(T) \cdot T}$

where $S_{\text{tr}} = \max_{\phi \in \Phi} S_{\text{tr}}$, and $S_{\text{tr}} = \sum_{\phi \in \Phi} S_{\text{tr}}$.

- Ucb-Ms adapts to most preferable model
- Needs the knowledge of upper-bound on the true diameter, i.e., $D \geq D^*$
- It reduces to UCRL2 when there is a single model: $D \leq D^*$
- No need to exactly identify the true model $\phi^*$ but if a non-Markovian model gives as much as reward of a Markovian one $
\Rightarrow$ no need of discarding it
- Improves regret compared to the state of the art (e.g., [Ortner et al., 2014])

Extensions

- Unknown diameter: $\Rightarrow$ doubling trick
  - if all the models are eliminated
  - double the estimate of the diameter

The only difference in the regret is an additional term $[|\Phi| \log(D)]$

- Effective size $S_{\Phi}$:
  - the entire state space can be covered with only $S_{\Phi}$ states using $\Phi$
  - Examples: Hierarchical structure

- regret bound scales with $S_{\Phi}$ rather than $S_{\text{tr}}$
  - $S_{\Phi} \gg S_{\text{tr}}$

- Unknown diameter and state size:
  - estimate directly the term $DS$
  - use similar doubling trick
  - the regret is upper-bounded by

$R(T) \leq \text{const} \cdot D^* \sqrt{S_{\text{tr}} + |\Phi| \log(DS_{\text{tr}}) \cdot T}$

References