Online Discovery of Feature Dependencies

Alborz Geramifard, Finale Doshi, Joshua Redding, Nicholas Roy, Jonathan P. How

ICML 2011

Review notes by Xuejun Liao

October 3, 2011
A Markov Decision Process (MDP) (Sutton & Barto, 1998) is a tuple defined by \((S, A, P_{ss'}, R_{ss'}, \gamma)\) where \(S\) is a finite set of states, \(A\) is a set of actions, \(P_{ss'}^{a}\) is the probability of getting to state \(s'\) by taking action \(a\) in state \(s\), \(R_{ss'}^{a}\) is the corresponding reward, and \(\gamma \in [0, 1)\) is a discount factor balancing current and future rewards.
Given policy $\pi : S \rightarrow A$, the action-value function is

$$Q^\pi(s, a) = \mathbb{E}_\pi \left[ \sum_{t=0}^{\infty} \gamma^t r_t \mid a_0 = a, s_0 = s \right]$$

The value function is governed by Bellman equation

$$Q^\pi(s, a) = R(s, a) + \gamma \sum_{s'} P^a_{ss'} Q^\pi(s', \pi(s'))$$
Temporal Difference (TD) Learning

Given a trajectory \((s_0, a_0, r_0, s_1, a_1, r_1, \cdots, s_t, a_t, r_t, s_{t+1})\), define TD error

\[
\delta_t = r_t + \gamma \max_{a_{t+1}} Q^\pi(s_{t+1}, a_{t+1}) - Q^\pi(s_t, a_t)
\]

and a TD learning rule updates

\[
Q^\pi(s_t, a_t) \leftarrow Q^\pi(s_t, a_t) + \alpha_t \delta_t
\]

where \(\alpha_t > 0\) is a learning rate.
The value function can approximated as

$$Q^\pi(s, a) = \sum_i w_i \phi_i(s, a)$$

where $\phi_i : S \times A \rightarrow \{0, 1\}$ maps states and actions to features (binary as considered here).

Omiting action $a$ for simplicity, the set $\{i : \phi_i(s) = 1\}$ contains the features (indices) that are active for state $s$. 
Basic Idea of Online Feature Discovery

Start from an initial set of binary features.
Expand features by adding conjunctions of existing features where large errors persist.
For any state, prune away the active features that are contained in larger conjunction sets.

Figure 1. A snapshot of iFDD: Initial features are circles, conjunctive features are rectangles. The relevance $\psi_f$ of a potential feature $f$ is the filled part of the rectangle. Potential features are discovered if their relevance $\psi$ reaches the discovery threshold $\xi$. 

- Online Discovery of Feature Dependencies

A. Geramifard, F. Doshi, J. Redding, N. Roy, J. P. How
Algorithm 1: Discover

Input: $\phi(s), \delta_t, \xi, F, \psi$
Output: $F, \psi$

1. foreach $(g, h) \in \{(i, j) | \phi_i(s) \phi_j(s) = 1\}$ do
   2. $f \leftarrow g \land h$
   3. if $f \notin F$ then
      4. $\psi_f \leftarrow \psi_f + |\delta_t|$
      5. if $\psi_f > \xi$ then
         6. $F \leftarrow F \cup f$
Online Discovery of Feature Dependencies

Algorithm 2: Generate Feature Vector ($\phi$)

Input: $\phi^0(s), F$
Output: $\phi(s)$

1. $\phi(s) \leftarrow \emptyset$
2. $activeInitialFeatures \leftarrow \{i | \phi^0_i(s) = 1\}$
3. $Candidates \leftarrow \varnothing(activeInitialFeatures)$ *sorted*
4. while $activeInitialFeatures \neq \emptyset$ do
   5. $f \leftarrow Candidates.next()$
   6. if $f \in F$ then
      7. $activeInitialFeatures \leftarrow activeInitialFeatures - f$
      8. $\phi_f(s) \leftarrow 1$
   9. return $\phi(s)$
Theorem
Given initial features and a fixed policy that turns the underlying MDP into an ergodic Markov chain, iFDD-TD is guaranteed to discover all possible feature conjunctions or converge to a point where the TD error is identically zero with probability one.

Corollary
If at each step of iFDD-TD the policy changes but still induces an ergodic Markov chain (e.g., via-greedy or Boltzmann exploration), then iFDD-TD will explore all reachable features or converge to a point where the TD error is identically zero with probability one.
Corollary 3.3 With probability one, iFDD-TD converges to a weight vector $\theta$ and feature matrix $\Phi$, where the approximated value function error, as originally shown by Tsitsiklis and Van Roy (1997) for a fixed set of linear bases, is bounded by:

$$
\left\| \Phi \theta - V^* \right\|_D \leq \frac{1}{1 - \gamma} \left\| \Pi V^* - V^* \right\|_D,
$$

where $D_{|S| \times |S|}$ is a diagonal matrix with the stationary distribution along its diagonal, $\Pi = \Phi_\infty (\Phi_\infty^T D \Phi_\infty)^{-1} \Phi_\infty^T D$, and $\| \cdot \|$ stands for the weighted Euclidean norm.
The state is a 12-dimensional vector of remaining fuel, location, motor sensor status and camera sensor status for each of the three UAVs for a total of approximately 150 million state-action pairs.
Online Discovery of Feature Dependencies

Initial features were the fuel and position of each UAV, the communication UAV mode (move or perch), and the rescue status at each node. The total state-action pairs exceeded 200 million.
Online Discovery of Feature Dependencies

Results (1/3): performance

(c) Persistent Surveillance

(d) Rescue Mission

A. Geramifard, F. Doshi, J. Redding, N. Roy, J. P. How
Figure 4. Average final feature counts. ATC and SDM, even using more features, performed poorly on high-dimensional examples. The black bar depicts the total number of state-action pairs.
Table 1. The final performance with 95% confidence intervals of iFDD and random expansion with equal number of features.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Expansion Scheme</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Random</td>
</tr>
<tr>
<td>Inverted Pendulum</td>
<td>2953 ± 30</td>
</tr>
<tr>
<td>BlocksWorld</td>
<td>−0.80 ± 0.06</td>
</tr>
<tr>
<td>Persistent Surveillance</td>
<td>174 ± 44</td>
</tr>
<tr>
<td>Rescue Mission</td>
<td>10 ± .74</td>
</tr>
</tbody>
</table>