Proceedings of the Twenty-Ninth International
Florida Artificial Intelligence Research Society Conference

Reward from Demonstration in Interactive Reinforcement Learning

Syed Ali Raza, Benjamin Johnston, and Mary-Anne Williams
Centre for Quantum Computation and Intelligent Systems
University of Technology, Sydney, Australia
syed.a.raza@student.uts.edu.au, {benjamin.johnston, mary-anne.williams}@uts.edu.au

Abstract

In reinforcement learning (RL), reward shaping is used to show the desirable behavior by assigning positive or negative reward for learner's preceding action. However, for reward shaping through human-generated rewards, an important aspect is to make it approachable to humans. Typically, a human teacher's role requires being watchful of agent's action to assign judgmental feedback based on prior knowledge. It can be a mentally tough and unpleasant exercise especially for lengthy teaching sessions. We present a method, Shaping from Interactive Demonstrations (SfID), which instead of judgmental reward takes action label from human. Therefore, it simplifies the teacher's role to demonstrating the action to select from a state. We compare SfID with a standard reward shaping approach on Sokoban domain. The results show the competitiveness of SfID with the standard reward shaping.

Introduction

Designing an autonomous agent which can learn to act in its environment is a major challenge in Artificial Intelligence (AI). As robots become more widespread and pervasive in society, ordinary people should be able to design robot's behavior in ordinary circumstances. For a naïve user, interactively transferring knowledge is an intuitive solution to robot programming. This is, generally, referred to as 'Learning from Demonstrations' (LfD).

The goal of learning in a social setting is to minimize the anti-social aspects while increasing the learning performance. Asking a human teacher for numerous and repeated demonstrations for an idle and nonresponsive robot (like mannequin), as required in some LfD methods (Argall et al. 2009), is anti-social. Instead, humans like an interactive learner which responds as the learning progresses. Reinforcement learning (Sutton & Barto 1998) provides a rich framework for interactive learning. Specifically, reward shaping method (Ng, Harada & Russell 1999) in RL allows the teacher to interactively give judgmental feedback on learner's behavior.

Reward shaping is a state-of-the-art technique to speed-up RL. Nevertheless, the engagement of a human teacher adds another dimension to the reward shaping research. For a wider adoption of reward shaping in the design of social robots, learning should occur through natural teaching style. In this paper, we have altered the standard reward shaping to take human input as demonstration instead of reward signal. The resulted method of reward shaping offers to teach by acting naturally from a state. Therefore, it can potentially substitute the learning from human-generated numerical rewards. We heuristically measure the numerical reward signal from the demonstrated behavior by matching the agent's policy with the teacher's policy. We show that the policy learned by the proposed method is compatible to the policy learned through standard reward shaping.

Similar approaches of using online human feedback in RL have been proposed in the past. Thomaz & Breazeal (2006) proposed a modified RL method to allow human teacher to provide reward signal for the last action as well as guidance for the future action selection. Suay & Chernova (2011) studied it further and observed that positive effects of guidance increases with state space size. Knox & Stone (2009) introduced a framework called TAMER which learns a predictive model of human reward to use it in place of the reward functions of standard RL methods. In another approach, Griffith et al. (2013) introduced 'policy shaping' to formalize human feedback as policy advice. Among recent works, Loftin et al. (2015) used the models of trainer strategy to improve learning performance and De la Cruz et al. (2015) studied the use of online crowd to speedup RL.

Reinforcement Learning Background

The interaction of a RL agent with its environment, at discrete time steps $t = 1,2,3, \ldots$, is modeled within the frame-

work of Markov decision process (MDP). An MDP M is specified by the tuple $M = (S, A, T, R, \gamma, D)$. $S = \{s_1, s_2, \dots, s_n\}$ is a set of states that agent can perceive and $A = \{a_1, a_2, \dots, a_n\}$ is the set of actions that agent can perform. T is a transition function $T(s'|s, a): S \times A \times S \rightarrow [0,1]$. Upon taking an action a_t from a state s_t the agent moves to state s_{t+1} and receives an environmental reward r defined by the reward function $R(s, a): S \times A \rightarrow \mathbb{R}$. $\gamma \in [0,1]$, the discount factor, is used to discount future reward values. $D \subseteq S$ is the subset of states used to get the start state distribution. In Q-learning, an update is made to the action-value function $Q: S \times A \rightarrow \mathbb{R}$ at each time step as follow,

$$Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + \alpha_t \delta_t \quad (1)$$

$$\text{Where, } \delta_t = r_{t+1} + \gamma max_a Q_t(s_{t+1}, a) - Q_t(s_t, a_t) \quad (2)$$

The term δ_t is referred to as temporal-difference error. $\alpha_t \in (0,1)$ is the learning rate. The agent uses a cy, $\pi: S \rightarrow A$, indicating the action to select from the current state as $\pi(s) = argmax_a\, Q(s, a)$.

Interactive Reinforcement Learning

The RL provides an ideal framework for step-by-step or sequential decision making. In the interactive reinforcement learning, the teacher can interact with the agent while the actual learning in underway and can influence or evaluate the learner's decision making. Due to the inherent exploratory nature of reinforcement learner, it can act by itself at any decision step, even in absence of teacher interaction. For example, one may use random policy, selecting random actions, or a fixed policy, selecting fixed action from every state, etc. This self-acting ability can enhance the social experience of the interaction as compared to an idle learner which only records demonstrations. From this perspective, it can be argued that the role of the teacher is to guide learning instead of teaching a task from scratch. The teacher's input can guide learner to avoid unnecessary exploration which can cause damage to the robot. In addition, it adds teacher's preference in the learned policy. Note that, unlike standard RL where the reward function and therefore the optimal policy is fixed, in interactive RL the final policy is a teacher's preferred policy.

Interactive Reward Shaping

In RL, environmental rewards are generally assigned at the completion of the task or on achieving a sub-goal. Therefore, these rewards are sparse and their effect on the intermediate action-selections may start to occur after numerous episodes of exploration. To expedite learning, immediate rewards are used to indicate the desirability of the recent action. It shapes the policy locally such that it leads to the accomplishment of goal. An applicable solution to reward shaping is by allowing a human teacher to show the rewarding rules through interaction with the agent, known as interactive reward shaping.

In interactive reward shaping, also referred to as interactive shaping, a human closely observes the agent and evaluates its last action in the context of domain knowledge. The teacher, then, maps the usefulness of the recent action in longer run to a positive or negative value. In practice, the repeated labeling becomes a tedious task. A teacher is usually engaged in extensive sessions of rewarding which may depend upon various factors such as domain's search space, exploration method, and initial policy. Thus, due to lengthy and judgmental nature of feedback it becomes a mentally exhaustive exercise. However, our main contribution is towards eliminating the judgmental nature of feedback.

Formulation & Algorithm

In a typical reward shaping setting, when the agent samples a new experience, it might receive a reward from MDP's underlying reward function, $R(s, a)$, and an additional shaping reward, $H(s, a)$. Thus, the resulting shaping reward $R_s(s, a)$ is,

$$R_s(s, a) = R(s, a) + H(s, a) \quad (3)$$

Therefore, Eq. 2 becomes,

$$\delta_t = (r_{t+1} + h_{t+1}) + \gamma max_a Q_t(s_{t+1}, a) - Q_t(s_t, a_t) \quad (4)$$

We distinguish the human-teacher-generated reward by denoting it as $H^{\pm}(s, a)$, which is produced by visually observing agent's last state-action and assigning a signed numerical reward to it.

Algorithm 1: Shaping from Interactive Rewards

1: **While** learning **Do**
2: $\boldsymbol{a}$ = get action with highest $\boldsymbol{Q(s, a)}$
3: Execute $\boldsymbol{a}$, and transition to next state $\boldsymbol{s_{t+1}}$, and sense environmental reward $\boldsymbol{r}$
(Wait for human reward, $\boldsymbol{h}$)
4: Update $\boldsymbol{Q(s, a)}$ using eq. 1 and 4
5: **End While**

Algorithm 1 gives the basic steps of the interactive reward shaping process adopted in this research. The process starts with greedy action selection using Q-values. After taking action the agent waits for a fixed amount of time to receive human reward. Finally, the Q-values are updated and the same process repeats in the next time step. The algorithm uses a random initial policy which selects a random action from a previously unvisited state.

Shaping from Interactive Demonstrations

We propose a method to derive shaping reward function from interactive demonstration called Shaping from Interactive Demonstrations (SfID). It is designed after the popu-

lar interactive reward shaping but differs in how the teacher provides input to the learner. The teacher only observes the state and labels the action to choose from it. The teaching method of direct action labelling (demonstration) eliminates the judgemental evaluation task of interactive reward shaping. In addition, it is more explicit as compared to rewarding in eliciting teachers preferred policy. For the policy derivation, SfID follows the learning mechanism of interactive reward shaping.

The use of demonstrations in SfID is different from that of a typical LfD method. In a typical LfD, the teacher fully controls actions executed and the demonstrations are recorded as a chain of state action pairs taking the agent from the start state and leading towards goal state. In SfID, the action selection process is under agent's control and the teacher's demonstration only contributes towards computing a reward signal for the preferred policy. Therefore, agent's initial policy may not necessarily obey the teacher's policy but later converges towards it.

Formulation & Algorithm

In an interactive learning setting, the human teacher observes agent's current state and indicates the action to select from it as,

$$\pi_t^h(s) = a_b \tag{5}$$

The above equation gives the human teacher's policy $\pi^h: S \rightarrow A$. $a_b \in A$ is defined as the best action to perform from state s as per teacher's knowledge. Note that t in subscript represents the human teacher's policy might change over time.

A reward function is derived from the teacher's policy through a mapping $\Omega: A \times A \rightarrow \mathbb{R}$,

$$H^d(s,a) = \Omega(\pi_t^h(s), \pi(s)) \tag{6}$$

Therefore, the problem is to define a mapping Ω to get the reward function $H^d(s,a)$ from $\pi_t^h(s)$ to substitute $H(s,a)$ in Eq. 3. We empirically show that, the policy learned by using $H^d(s,a)$ is no less than the policy learned by $H^{\pm}(s,a)$ i.e. $\pi_{H^d(s,a)} \cong \pi_{H^{\pm}(s,a)}$. The mapping function Ω is given by the following equation,

$$\Omega\left(\pi_t^h(s), \pi(s)\right) = \begin{cases} 1, & \pi_t^h(s) = \pi(s) \\ -1, & otherwise \end{cases}$$

The above equation is based on a simple heuristic. A fixed positive reward is assigned if the agent's action matches with the prescribed action otherwise a negative reward is assigned. Algorithm 2 details the complete process of SfID. The process is similar to the one described in Algorithm 1 except that the agent waits for a fixed time span for the teacher to provide demonstration (action label). Afterwards, it computes the reward from the demonstration and executes the greedy action.

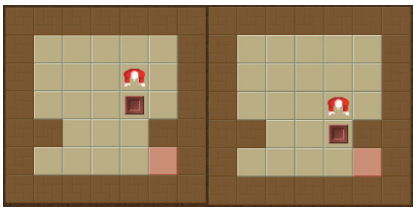

Figure 1. From left to right it shows the result of a down action and push.

Algorithm 2: Shaping from Interactive Demonstrations

While learning Do
  $\boldsymbol{a}$ = get action with highest $\boldsymbol{Q}(\boldsymbol{s},\boldsymbol{a})$
  (Wait for the demonstration, $\boldsymbol{a_b}$)
If $\boldsymbol{a_b}$ provided **Then**
  $\boldsymbol{h}$ = $\boldsymbol{\Omega}(\boldsymbol{a_b},\boldsymbol{a})$
Else
  $\boldsymbol{h}$ = 0
End If
execute $\boldsymbol{a}$, and transition to $\boldsymbol{s_{t+1}}$, and sense environmental reward $\boldsymbol{r}$
update $\boldsymbol{Q}(\boldsymbol{s},\boldsymbol{a})$ using eq. 1, 3, 4 and 6.
End While

Experiments

Sokoban is a 2d grid-world game where each cell can be a wall or free cell. Free cell can be occupied by either player or a box. The player can choose from four actions: left, right, up, and down. The player's task is to push each box using four actions and drive it to the goal position without letting the box get stuck, deadlock. Since the player does not have pull action, a box can get stuck if it cannot be derived to destination without pulling it. At the start of game player and each box are positioned at a fixed location. Figure 1 shows the domain and an example of gameplay. The solution to Sokoban maze has been shown as PSPACE complete (Culberson 1999). We have simplified it to avoid frequent deadlocks. Typically, at the start of the game player and each box are positioned at a fixed location. Instead, we have used a state distribution for both box and player for their start positions.

In our implementation of the Sokoban, we have designed an interface to allow the human teacher to provide reward and demonstration using keys. To show the effectiveness of the proposed method we have conducted experiments using two human teachers of completely different skill levels. Subject 1 was an expert in robots and machine learning and has previous experience of playing Sokoban. Subject 2 was a naïve user who never played Sokoban and had no experience in machine learning. Conducting an extensive user study with a large variety of teachers is planned as an extension of this work. The criterion to stop teaching is when the agent acts according to teacher's preferred policy. Even if the teacher's preferred policy was

learned earlier, the teacher was asked to keep observing the policy of the agent for a number of episodes to ensure that no further improvement is needed. The interface included an option to set the time span to wait for reward or demonstration. The teachers were asked to use it according to their need at different stages of learning. We set RL parameters as: $\alpha = 0.1$, and $\gamma = 0.9$. The experiments were run for 30 episodes for each subject.

Results

The results of the experiments are summarized in Figure 2. These are the results for the performance during the learning. Interactive reinforcement learning and SfID are compared using four criterions; absolute error (Figure 2(a)), number of steps taken to reach goal (Figure 2(b)), the cumulative reward gained in each episode (Figure 2(c)), and the number of inputs per episode from the teacher (Figure 2(d)). The graphs for subject 1 show that in each of the criterion the SfID successfully matched the performance of interactive reward shaping. However, the graphs of subject 2 reflect that the learning performance using SfID is better than interactive reward shaping. The worst learning performance was achieved by interactive reward shaping by a naïve user. Thus, these results support our claim that SfID can at least achieve the learning performance of interactive reward shaping. In addition, for a naïve user SfID can be more efficient than interactive reward shaping. We also recorded the time used to teach the policy. The expert teacher taught using both the methods in almost same amount of time. However, for a naïve subject SfID helped to teach policy in less time. In addition, a common feedback from both the subjects is that teaching using SfID was less onerous than interactive reward shaping.

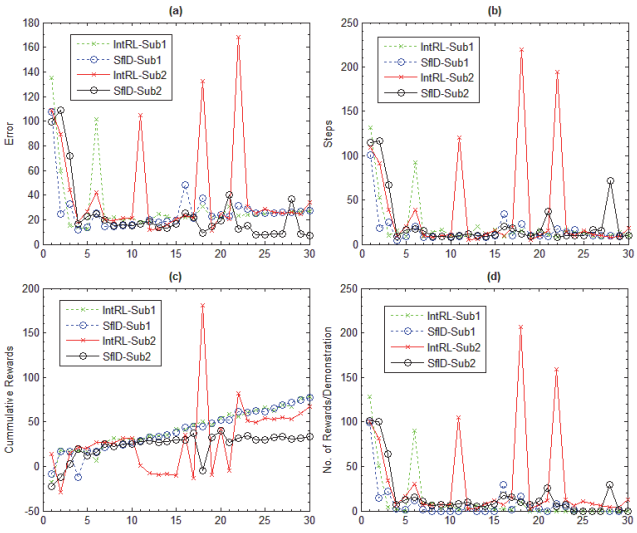

Figure 2. Comparison of SfID with interactive reward shaping. IntRL is a short for interactive reward shaping. The horizontal axis in all the graphs shows the number of episodes. (a) Absolute temporal difference error as given by Eq. 2. (b) The steps taken in each episode. (c) Cumulative discounted reward earned over the episodes. (d)Tthe count of teacher's input in each episode, (reward signal in interactive reward shaping and demonstration in SfID)

Conclusion

In this research, we have shown the competitiveness of utilizing interactive demonstration for the sake of reward shaping. A human teacher interactively teaches the policy to the learner by providing the demonstration of how to act from a state. The demonstration is interpreted as a reward signal and incorporated in the reward shaping process. The results obtained suggest that the performance of SfID either matches or outperforms that of the traditional method of teaching policy via reward shaping. However, the major advantage of SfID over interactive reward shaping is the ease it provides in the teaching method.

References

Argall, B.D., Chernova, S., Veloso, M. & Browning, B. 2009, 'A survey of robot learning from demonstration', Robotics and Autonomous Systems, vol. 57, no. 5, pp. 469-83.

Culberson, J. 1999, 'Sokoban is PSPACE-complete', Proceedings in Informatics, vol. 4, pp. 65-76.

De la Cruz, G.V., Peng, B., Lasecki, W.S. & Taylor, M.E. 2015, 'Towards Integrating Real-Time Crowd Advice with Reinforcement Learning', Proceedings of the 20th International Conference on Intelligent User Interfaces Companion, ACM, pp. 17-20.

Griffith, S., Subramanian, K., Scholz, J., Isbell, C. & Thomaz, A.L. 2013, 'Policy Shaping: Integrating Human Feedback with Reinforcement Learning', paper presented to the Advances in Neural Information Processing Systems 26.

Knox, W.B. & Stone, P. 2009, 'Interactively shaping agents via human reinforcement: the TAMER framework', paper presented to the Proceedings of the fifth international conference on Knowledge capture, Redondo Beach, California, USA.

Loftin, R., Peng, B., MacGlashan, J., Littman, M.L., Taylor, M.E., Huang, J. & Roberts, D.L. 2015, 'Learning behaviors via human-delivered discrete feedback: modeling implicit feedback strategies to speed up learning', Autonomous Agents and Multi-Agent Systems, pp. 1-30.

Ng, A. Y., Harada, D., and Russell, S. 1999. Policy invariance under reward transformations: Theory and application to reward shaping. ICML.

Suay, H.B. & Chernova, S. 2011, 'Effect of human guidance and state space size on Interactive Reinforcement Learning', RO-MAN, 2011 IEEE, pp. 1-6.

Sutton, R.S. & Barto, A.G. 1998, Reinforcement learning: An introduction, vol. 1, MIT press Cambridge.

Thomaz, A.L. & Breazeal, C. 2006, 'Reinforcement learning with human teachers: evidence of feedback and guidance with implications for learning performance', paper presented to the Proceedings of the 21st national conference on Artificial intelligence - Volume 1, Boston, Massachusetts.