Towards Student/Teacher Learning in Sequential Decision Tasks

(Extended Abstract)

Lisa Torrey
St. Lawrence University
ltorrey@stlawu.edu

Matthew E. Taylor
Lafayette College
taylorm@lafayette.edu

Categories and Subject Descriptors

I.2.6 [**Learning**]: Miscellaneous

General Terms

Algorithms, Performance

Keywords

Reinforcement Learning, Inter-agent teaching, Transfer Learning

1. INTRODUCTION

Significant advances have been made in allowing agents to learn, both autonomously and with human guidance. However, less attention has been paid to the question of how agents could best teach each other. For instance, an existing robot in a factory should be able to instruct a newly arriving robot, even if it is from a different manufacturer, has a different knowledge representation, or is not optimal itself.

This work investigates teaching methods in sequential decision tasks. In particular, we consider a reinforcement learning student-agent that must learn from 1) autonomous exploration of the environment and 2) the guidance of another teacher-agent. In order to minimize inter-operability requirements, the teacher and student are presumed not to know each others' internal workings; teachers can only help students by suggesting actions. Furthermore, the teacher may have limited expertise in the student's task and should be careful not to over-advise the student. Our primary question: how should the teacher decide when to give advice?

This teaching context is related to the more well-studied problem of *transfer learning* [5], in which an agent uses knowledge from a source task to aid its learning in a target task, but differs in that we do not assume agents can directly access each others' internal knowledge. Another related area is *learning from experts* [1, 3], where agents may imitate experts or ask for their advice. Our approach differs because control is given to the teacher, rather than the student, and we focus on non-expert teachers. Our hope is that this paper enables and inspires the agents community to develop further methods by which agents can teach other agents.

Appears in: *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2012)*, Conitzer, Winikoff, Padgham, and van der Hoek (eds.), 4-8 June 2012, Valencia, Spain.

2. METHODS

Our methods build upon Probabilistic Policy Reuse [2] (PPR), which allows an agent to learn a task faster by taking advantage of an existing policy. The PPR method changes the action selection step of model-free RL methods. With probability ψ, the agent exploits an old policy; the rest of the time, it uses normal ϵ-greedy action selection. The value of ψ decays over time according to a decay rate v so that the agent makes less use of old policies as it improves its own. We use this method for teaching by letting ψ be the teacher's probability of giving advice.

However, for a teacher with limited expertise, PPR may not be the optimal teaching algorithm. A PPR teacher provides action advice with a global probability ψ that is uniform across all states. If the teacher is more confident in some states than others, it makes more sense for advice probabilities to be higher in some states than others. We therefore propose a new approach that uses *confidence measures* to make advice probabilities state-specific.

2.1 Confidence Measure

In our proposed approach, agents need to be able to estimate their confidence in a state. Because we assume limited expertise, agents may not have much data to work with. To allow meaningful estimates with limited data, we introduce a confidence measure called *update counting*. The update-count of a state indicates how many times a non-zero Q-value update has been made there. This measure has a straightforward tabular implementation in discrete settings, but it is also adaptable to continuous settings through tile coding. Update-counts can be associated with tiles; the update-count of a state is then the sum of the update-counts of its component tiles.

2.2 Advice Probabilities

Because our proposed approach builds upon PPR, our teachers compute a probability of giving advice in a state. We believe there are several desirable properties for advice probabilities. First, they should be higher in states where the teacher is more confident, so that teachers give advice in proportion to their expertise. Second, they should be capped at the ψ of the PPR algorithm, so that the confidence-based teaching method integrates cleanly with the PPR framework. Third, they should decay over time, so that teachers gradually decrease their guidance as the student learns.

We now propose an algorithm that computes a probability of giving advice $p(s)$ with the above properties. Let $c_t(s)$ and $c_s(s)$ represent the teacher and student confidences (i.e., update-counts) in that state, respectively.

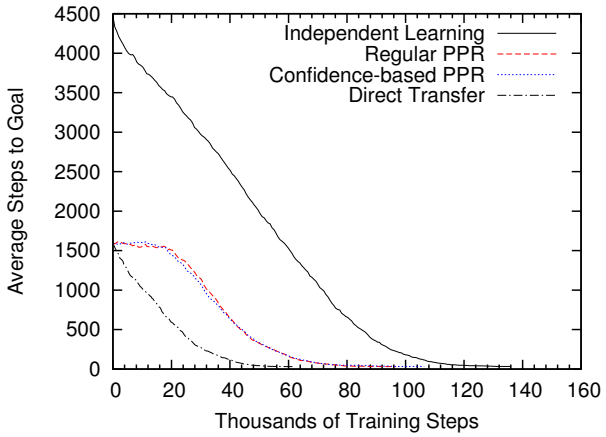

Figure 1: Q-learning agents in a maze

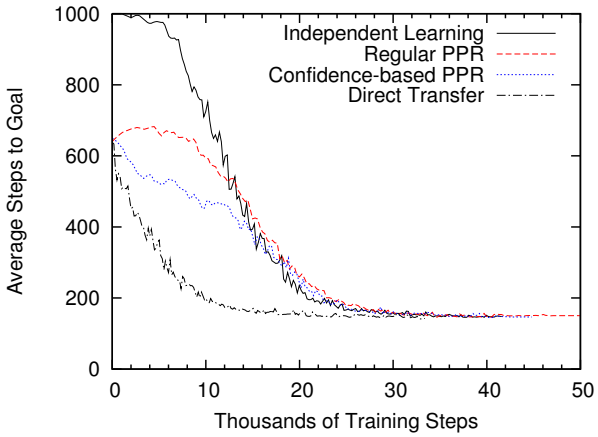

Figure 2: Sarsa learning students in mountain car

The *confidence-based PPR* algorithm computes:

$$p(s) = \begin{cases} 0 & \text{if } c_t(s) < 1 \\ \min\left(1 - \frac{c_s(s)}{c_t(s)+d}, \psi\right) & \text{if } c_t(s) \geq 1 \end{cases}$$

The first condition states that a teacher should never give advice if it has no knowledge about a state. The second dictates that the advice probability $p(s)$ depends on the relationship between the student's confidence and the teacher's confidence. In states where the teacher has higher confidence than the student, it gives advice with a higher probability, up to a maximum of ψ. As the student's confidence in a state grows, the teacher gives advice with lower probability. As the delay parameter d increases, teachers require students to reach higher levels of confidence before they stop giving advice.

3. EVALUATION

To evaluate this novel teaching algorithm, we perform teaching experiments in two domains. One is a discrete 20×20 maze, fully described in the earlier Ask-For-Help work [1], in which our teachers learn via standard tabular Q-learning. The other is mountain car, a benchmark continuous domain [4], in which our teachers learn via Sarsa with tile coding.

To produce teachers with limited expertise, we do not allow them to train until their policies converge. Instead, we train teachers for only 20 episodes. Each teacher then gives advice to students using regular PPR or confidence-based PPR. We average 100 teacher-student pairs for each experiment. Lower bounds for students are represented by *independent* agents, who learn without teachers. Upper bounds are represented by *direct-transfer* agents, who copy the teacher's entire Q-function, which our students do not have access to.

Figure 1 shows some results from the maze. The most effective parameter settings here are $v = 0.99$ and $d = 0$. As expected, students with teachers outperform students without teachers. The two teaching algorithms perform comparably in this domain. Confidence-based PPR does not outperform regular PPR because this domain lacks critical decision points, which means teachers can be less discerning about giving advice. However, we did find that it achieves comparable performance using two orders of magnitude less advice.

Figure 2 shows some results from mountain car. The most effective parameter settings here are $v = 0.99$ and $d = 100$. Both teaching methods again speed up learning, but in this domain confidence-based PPR outperforms regular PPR. Differences in the areas under these learning curves are statistically significant at the 95% confidence level.

There are several potential reasons that confidence-based PPR outperforms regular PPR in this domain. Mountain car may have some critical decision points, where the relative values of student and teacher confidences can play important roles in advice decisions. The tile coding in mountain car provides state-space generalization, which can cause student confidence to grow quickly in some sets of states. Using confidence-based PPR, teachers are able to back off quickly in these states, while still giving advice in less common states.

4. FUTURE WORK AND CONCLUSIONS

This paper contributes an initial study of algorithms that agents can use to teach each other in sequential decision tasks. We assume a broadly applicable setting, in which teachers and students interact only through action advice and in which teachers can have limited expertise.

There are many potential directions for future work in this area. For instance, teachers could explicitly reason about the expense of communication versus the expected gain, which would be appropriate in domains where communication has a non-zero cost. There could also be multiple teachers, with different areas of expertise, who must coordinate with each other.

5. REFERENCES

[1] J. A. Clouse. *On integrating apprentice learning and reinforcement learning*. PhD thesis, University of Massachusetts, 1996.

[2] F. Fernandez and M. Veloso. Probabilistic policy reuse in a reinforcement learning agent. In *Proceedings of the 5th International Conference on Autonomous Agents and Multiagent Systems*, 2006.

[3] B. Price and C. Boutilier. Accelerating reinforcement learning through implicit imitation. *Journal of Artificial Intelligence Research*, 19:569–629, 2003.

[4] S. Singh and R. S. Sutton. Reinforcement learning with replacing eligibility traces. *Machine Learning*, 22:123–158, 1996.

[5] M. E. Taylor and P. Stone. Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research*, 10(1):1633–1685, 2009.