Using human reinforcement learning models to improve robots as teachers

Sayanti Roy, Emily Kieson, Charles Abramson and Christopher Crick
Oklahoma State University
Stillwater, Oklahoma
sayanti.roy@okstate.edu,kieson@okstate.edu,charles.abramson@okstate.edu,chriscrick@cs.okstate.edu

ABSTRACT

Robotic teaching has not received nearly as much research attention as robotic learning. In this research, we used the humanoid robot Baxter to provide feedback and positive reinforcement to human participants attempting to achieve a complex task. Our robot autonomously casts the teaching problem as one that invokes the exploration/exploitation tradeoff to understand the cognitive strategy of its human partner and develop an effective motivational approach. We compare our learned reinforcement model with a baseline non-reinforcement approach and with a random reinforcer.

1 INTRODUCTION

Human cognition is complex, hidden, and often difficult to interpret. A robot's teaching strategy should be more effective if the robot possesses some understanding of its human student's mindset. In this research, we employ the humanoid robot Baxter to act as a facilitator during an individual's learning process by motivating them extrinsically. When people who are less motivated to persevere with a difficult task can receive some positive reinforcer to overcome their challenge, their learning rate is expected to increase [1]. We not only explore the effective teaching strategies available to a robot, but also try to identify which positive reinforcements are are most effective at motivating a particular human student's learning style. We divide the subject pool into three groups, which received no reinforcements, random reinforcements, or learned reinforcements respectively. We compare the number of mistakes made in each category against each other. We found that subjects in the learned group made comparatively fewer mistakes. We also discovered that the robot's regret, in a machine learning sense, strongly correlates with the probability that a test subject makes more versus fewer mistakes.

2 RELATED WORK

Thomaz [2] explored how feedback influences future learning processes through a motivational channel which successfully improves a robot's learning behaviour. Roy [3, 4] discussed effective mutual robot learning and teaching using semantic labels and shared conceptual hierarchies. Leite [5] performed experiments where a robot attempted to discover individually appropriate supportive behaviours based on interactions with children over time.

3 TECHNICAL DESCRIPTION

When Baxter is trying to motivate an individual, it provides a positive reinforcer [6]. Four different reinforcer were available: a reward, verbal encouragement, a motivational gesture, and none. Initially, the robot assigns a uniform prior across its potential reinforcement behaviours. When a subject is given a particular reinforcement, the robot evaluates his performance on the immediately following subtask, and reweighs its reinforcement strategy appropriately:

$$S_t = \{\nu\phi s^+_{t-1}, \frac{\nu}{|S|-1}(1-\phi)s_{t-1}\forall s \in S_{s\neq s^+}\} \quad (1)$$

where S_t is the weight distribution over all reinforcement strategies at time t, s^+_{t-1} is the particular reinforcement strategy chosen at time $t-1$, ϕ is 1 if the subject successfully completed the subtask immediately following the previous reinforcer, and 0 otherwise, and ν is a learning rate parameter (empirically set to 0.03).

After several interactions, the robot can conclude which particular reinforcements are inducing the candidate to perform well. To evaluate how well the robot can assess its own performance as a good teacher, we correlated its regret with the number of mistakes made by its students. If the regret correlates well with mistakes made, Baxter's own self-assessment is reliable, and its representation of reinforcement strategies is appropriate to the teaching task. Regret is defined as the difference between the reinforcement strategy selected and the reinforcer with maximum weight.

$$R = s_{max} - s^+ \quad (2)$$

The reinforcer that Baxter has assigned maximum weight is the one that the robot has decided is the most appropriate strategy for the human learner with which it is currently interacting. Low regret means that the robot is exploiting this knowledge to increase the learning rate, while if it selects a different reinforcement strategy it is exploring to discover more about the student's individual motivational receptiveness.

HRI '18 Companion, March 5–8, 2018, Chicago, IL, USA

ACM ISBN 978-1-4503-5615-2/18/03.
https://doi.org/10.1145/3173386.3177074

4 EXPERIMENTAL PROCEDURE

$n = 110$ participants are recruited for the experiment. The no reinforcer group contains $n = 35$, the random reinforcer group contains $n = 22$ and the learned model group contains $n = 53$ participants. Subjects in each group were given three tasks to perform in ascending order of difficulty. The robot teaches them to recreate different patterns with markers. Depending on the test condition, when its students made mistakes, the robot either provided no motivation, chose a random strategy, or attempted to learn and use the best reinforcement strategy for the particular student.

4.1 Performance Evaluation

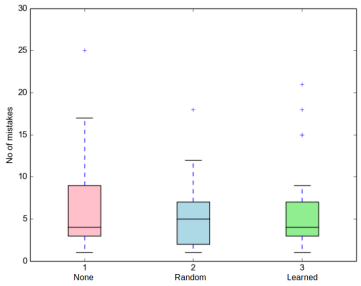

Figure 1: Performance of participants

Fig. 1 shows the number of mistakes made by participants in the three conditions. Although the median performance between the subjects that received no reinforcement and those whose preferences the robot learned and exploited is similar, the range of the mistakes differ. More than a quarter of the participants in the group without motivational feedback made more mistakes than almost anyone in the learned group. Fewer mistakes were made by the worst performers in the learned group than any other. The group receiving random reinforcement performed at an intermediate level compared to those that received no reinforcement and those whose reinforcement was individually learned and tailored by the robot. The reinforcement strategy is considered to be working for a participant when the participant starts making fewer mistakes with same kind of reinforcer, and this also leads to a lower computed regret for the robot. Positive reinforcement has a salutary effect on human learning, and a robot that can learn an appropriate reinforcement strategy will be more successful at teaching complex tasks.

4.2 Regret Analysis

As mentioned in Section 3, regret is calculated from the interplay between the subject's performance and the robot's attempts to understand and characterize the most appropriate reinforcement strategy. Fig. 2 shows the correlation between the number of mistakes made by the human participants and the total regret felt by the robot. The figure illustrates a very strong $r = 0.88$ relationship between the variables. Thus, the robot's computed regret and the test subjects' mistakes are positively and tightly correlated, which demonstrates that the robot's own exploration and exploitation in the reinforcement learning space is very appropriate for understanding human responses to the various available reinforcement strategies.

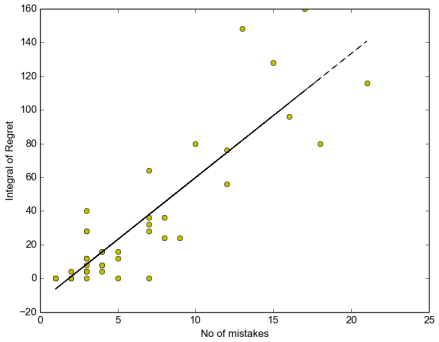

Figure 2: Regret analysis of the robot

5 CONCLUSION AND FUTURE WORK

In this work, we develop a reinforcement learning model for robotic teaching where the robot both attempts to learn an action sequence that leads to high reward (understood as successful human learning) and represents a human's own learning process as a reinforcement process as well, providing appropriate rewards to motivate better performance. In our work, both humans and robots use reinforcement learning techniques to learn from one another, sharing strategies, knowledge and decision making processes. In our future work we would like to jointly determine and characterize the mental models of the robots and humans as both learners and teachers. Robots should be able to serve as teachers and collaborators to help their human partners.

This work was supported by NSF award #1527828 (NRI: Collaborative Goal and Policy Learning from Human Operators of Construction Co-Robots).

REFERENCES

[1] M. Csikszentmihalyi and M. Wong, *Motivation and Academic Achievement: The Effects of Personality Traits and the Quality of Experience*. Springer Netherlands, 2014, pp. 437–465.

[2] A. L. Thomaz and C. Breazeal, "Teachable robots: Understanding human behavior to build more effective robot learners," *Artificial Intelligence*, vol. 172, pp. 716–737, 2008.

[3] S. Roy, E. Kieson, C. Abramson, and C. Crick, "Semantic structure for robotic teaching and learning," in *Proceedings of the 26th IEEE International Symposium on Robotics and Human Interactive Communication (RO-MAN)*, 2017.

[4] S. Roy, H. Maske, G. Chowdhary, and C. Crick, "Teaching and learning using semantic labels," in *Proceedings of the 12th ACM/IEEE Conference on Human-Robot Interaction (HRI)*, 2017.

[5] I. Leite, G. Castellano, A. Pereira, C. Martinho, and A. Paiva, "Long-term interactions with empathic robots: Evaluating perceived support in children," in *Proceedings of the 4th International Conference on Social Robotics (ICSR)*, 2012.

[6] K. Wheldall and F. Merrett, *Positive teaching: The behavioral approach*. Routledge, 2017.