Microsoft

Interacting with Reinforcement Learning

Michael L. Littman
Columbia University
Embedded Adaptable Systems

More and more computers interact with us in the real world. They need to adapt to us. How do they find out what to do?

Explicit programming (end-user friendly). Explicit reward delivery.

Learn to connect language and behavior. Learn from demonstrations.

Apology: I am focusing on work, but there are many people doing excellent work in this area.
Programming Via Programming

Experimenting with user-friendly programming languages like Scratch and IFTTT for device programming.

- If a photo is taken, then send an email to mlittman@cs.brown.edu
- If a new photo is taken, then turn on colored lights when I've lost weight!

- Forever if Device On?
  - broadcast Turn On Fan
  - repeat until not Device On?
  - broadcast Turn On Lamp
  - wait 0.5 secs
  - broadcast Turn Off Lamp
  - wait 0.5 secs
  - broadcast Turn Off Fan

Clapper

Most obnoxious alarm ever

Restricted interface
Learning in End-User Programming

Specifying sensing events can be hard. Supervised learning...

when visitor in my office..

Specifying strategies can be hard. Reinforcement learning...

keep temperature around 72 without making it too dry or using too much energy..

opportunities for integrating learned conditions and behaviors into end-user programming languages
Programming via Rewards

reward function : reinforcement-learning algorithm

Source code : compiler  the thing that specifies desired behavior

Getting a good reward function is like programming, it is likely to be hard for end users.

What about delivering rewards “by hand”. Essentially train programs like training a dog.

See TAMER by Knox et al. (Ten to Watch!)
Exploring RL To Human Reward Delivery

...algorithms developed for static reward functions do not match how humans use feedback.

...a communication channel.

Figure 4. The average number of actions required to successfully train the base Bayesian, Sim-TAMER, and Sim-COBOT algorithms to a policy that is 50%, 75%, or 100% correct, and until the participants decide the learning is complete.

Figure 7. Exit survey responses to the question "The robot was able to solve the task in a way that was satisfactory to me" - Agree includes responses "agree" and "strongly agree", while disagree includes "disagree" and "strongly disagree".
A (provably!) larger class of environmental dynamics can be learned efficiently from demonstration than pure autonomous exploration.
Apprenticeship Learning About Multiple Intentions

H. C. Wang-Vroman
Computer Science, Rutgers University, 110 Frelinghuysen Rd, Piscataway, NJ 08854 USA

VUKOSI@CS.RUTGERS.EDU

K. Ramanian
Computer, Georgia Institute of Technology, 801 Atlantic Dr., Atlanta, GA 30332 USA

K. KAUSUBBU@GATECH.EDU

D. A. Littman
Computer Science, Rutgers University, 110 Frelinghuysen Rd, Piscataway, NJ 08854 USA

MLITTMAN@CS.RUTGERS.EDU

Figure 1. A single trajectory from start to goal.

Figure 2. Reward computed using

Figure 7. Simulated Highway Car

Figure 8. Value of the computed policy of length of driving trajectories for three learning about multiple intentions.

Left: Grid world showing the start states (grey), (G), puddles and three sample trajectories. Posterior probabilities of the three trajectories between Expert 1.

Figure 6. Average reward for Student trajectory for EM approach with varying IRL/AL components.
Verbal Commands As Rewards

Top verbal instructions to behavior using paired examples text and trajectories.

Push the star to the green room. = reward for in(star,room1)

Learning to Interpret Natural Language Instructions

Glashan* and Monica Babeș-Vroman+ and Kevin Winner* and Ruoyuan Gao+ Adjogah* and Marie desJardins* and Michael Littman+ and Smaranda Muresan#

Department of Computer Science and Electrical Engineering, University of Maryland, Baltimore County
+ Computer Science Department, Rutgers University
# School of Communication and Information, Rutgers University

Figure 1: Training data for two tasks: 1) Taking the star to the green room (left) and 2) Going to the green room (right). Each training instance is a demonstration of a verbal instruction, shown below the demonstration on the right.