Robots Learning from Human Teachers

Andrea L. Thomaz
Georgia Institute of Technology

Reinforcement Learning & Decision Making
June 9, 2015
Socially Intelligent Machines Lab

Georgia Tech

Robotics @ GT & Intelligent Machines
Social Robot

A robot who's functional goal involves interacting with people in an environment designed for humans.
Many robot successes are in structured environments, program once and repeat

Humans and human environments are dynamic, pre-programming controllers is not an option
Our goal is to enable robots to function in new or changing environments by interacting with end-users to learn what to do.
• Flexible manufacturing co-workers
• Easily trainable to **whatever task is needed now** not programmed once and deployed for months
• Service robots in a home

• Tasks will be different in every home and only the end-user knows what the robot should do
**Problem:** build a policy or model of the task to be performed in a given environment: $S \times A$
Interactive Machine Learning

Algorithms need to use the information that people are **good at providing**
Interactive Machine Learning

**Demonstrations**

\[ S_1, a_1 \quad S_2, a_2 \quad \ldots \quad S_n, a_n \]

**Critique**

\[ S_1, a_1 \quad S_2, a_2 \quad \ldots \quad S_n, a_n \]

bad \quad good\quad good
Interactive Machine Learning

Demonstrations

Critique

\[
S_1, a_1 \quad S_2, a_2 \quad \ldots \quad S_n, a_n
\]
Interactive Reinforcement Learning

Agent

Environment

\[
Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + \alpha_t(s_t, a_t) \cdot \left( R_{t+1} + \gamma \max_a Q_t(s_{t+1}, a) - Q_t(s_t, a_t) \right)
\]
What is the best way to incorporate human feedback here?
Reward Shaping

- Is this something people can do?
- **Sophie’s Kitchen:** experiments testing this basic formulation of reward shaping

Thomaz & Breazeal, JAI 2008
Sophie’s Kitchen Results

• People aren’t very good at reward shaping: gave “feedback” about the future, positive bias in rewards

• **Changing interface and algorithm helps:** guidance for future directed input, undo on negative feedback

• Learns in 49% fewer trials
• 38% fewer ending failure
• Exploring 50% fewer states
Don’t convert “that’s right/wrong!” to a number, people are talking directly about the current policy and how to change it.

Policy Shaping instead of Reward Shaping

Griffin et al., NIPS 2013
Human’s Feedback Policy $\pi_F$

Human gives right/wrong labels to (s,a) pairs seen, calculate their policy given this history of feedback.

Probability that the human is **correct**

$$\pi_F(s, a) \propto C^{\Delta_{s,a}} (1 - C) \sum_{j \neq a} \Delta_{s,j}$$

- number of times **this action** was labeled correct
- number of times **some other action** was labeled correct
Agent’s Policy  $\pi_R$

Using Bayesian Q-Learning the agent’s policy is a probability distribution as well.

$Q(s, a)$ distribution

$\mathcal{N}(\mu_{s,a}, \tau_{s,a}) \sim \text{Normal} - \text{Gamma}(\mu_{0,s,a}, \lambda_{s,a}, \alpha_{s,a}, \beta_{s,a})$

$\mathcal{N}(\mu_{s,a}, \tau_{s,a})$ is sampled 100 times to obtain $\pi_R(s, a)$
Policy Shaping: ADVISE

- The probability that an action is optimal in a given state is the combination of these two distributions.
- Leads to a natural arbitration between listening to the human vs. self-exploration.

\[ \pi(s, a) \propto \pi_F(s, a) \times \pi_R(s, a) \]
Experiments

- We simulated human feedback with an oracle policy, testing sensitivity to **Likelihood** and **Correctness**

- **Baselines:** state of the art Interactive RL approaches that come close to policy shaping but still use human input as a reward
  - Action Biasing:
  - Control Sharing:
  - Reward Shaping:
Experiments

• We simulated human feedback with an oracle policy, testing sensitivity to Likelihood and Correctness

• **Baselines:** state of the art Interactive RL approaches that come close to policy shaping but still use human input as a reward

  • **Action Biasing:** $\text{argmax}_a \hat{Q}(s, a) + B[s, a] \cdot H[s, a]$
  
  • **Control Sharing:** $P(a = \text{argmax}_a H[s, a]) = \text{min}(B[s, a], 1.0)$
  
  • **Reward Shaping:** $R'(s, a) \leftarrow R(s, a) + B[s, a] \cdot H[s, a]$
## Experiments

### Ideal Case ($L = 1.0, C = 1.0$) vs Reduced Feedback ($L = 0.1, C = 1.0$)

<table>
<thead>
<tr>
<th>Action Biasing</th>
<th>Control Sharing</th>
<th>Reward Shaping</th>
<th>Advise</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pac-Man</td>
<td>Frogger</td>
<td>Pac-Man</td>
<td>Frogger</td>
</tr>
<tr>
<td>0.58 ± 0.02</td>
<td>0.16 ± 0.05</td>
<td>0.16 ± 0.04</td>
<td>0.04 ± 0.06</td>
</tr>
<tr>
<td>0.34 ± 0.03</td>
<td>0.07 ± 0.06</td>
<td>0.01 ± 0.12</td>
<td>0.02 ± 0.07</td>
</tr>
<tr>
<td>0.54 ± 0.02</td>
<td>0.11 ± 0.07</td>
<td>0.14 ± 0.04</td>
<td>0.03 ± 0.07</td>
</tr>
<tr>
<td>0.77 ± 0.02</td>
<td>0.45 ± 0.04</td>
<td>0.21 ± 0.05</td>
<td>0.16 ± 0.06</td>
</tr>
</tbody>
</table>

### Reduced Consistency ($L = 1.0, C = 0.55$) vs Moderate Case ($L = 0.5, C = 0.8$)

<table>
<thead>
<tr>
<th>Action Biasing</th>
<th>Control Sharing</th>
<th>Reward Shaping</th>
<th>Advise</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pac-Man</td>
<td>Frogger</td>
<td>Pac-Man</td>
<td>Frogger</td>
</tr>
<tr>
<td>-0.33 ± 0.17</td>
<td>0.05 ± 0.06</td>
<td>0.25 ± 0.04</td>
<td>0.09 ± 0.06</td>
</tr>
<tr>
<td>-2.87 ± 0.12</td>
<td>-0.32 ± 0.13</td>
<td>-0.18 ± 0.19</td>
<td>0.01 ± 0.07</td>
</tr>
<tr>
<td>-0.47 ± 0.30</td>
<td>0 ± 0.08</td>
<td>0.17 ± 0.12</td>
<td>0.05 ± 0.07</td>
</tr>
<tr>
<td>-0.01 ± 0.11</td>
<td>0.02 ± 0.07</td>
<td>0.13 ± 0.08</td>
<td>0.22 ± 0.06</td>
</tr>
</tbody>
</table>
Experiments

**Frogger**

*Ideal: L = 1, C = 1*

**Pac-Man**

*Moderate: L = .5, C = .8*
Interactive Machine Learning

Demonstrations

\[ S, S, \ldots, S \]

Critique

\[ S_1, a_1, S_2, a_2, \ldots, S_n, a_n \]

• Reward Shaping needs modification to be effective

• Better idea is Policy Shaping, don’t convert feedback to a reward number at all
Interactive Machine Learning

**Demonstrations**

$S_1, a_1 \quad S_2, a_2 \quad \ldots \quad S_n, a_n$

**Critique**

s \quad s \quad \ldots \quad s

- How can we efficiently get good demo input?
- What should we really learn from these demos?
Interactive Machine Learning

Demonstrations

\[ S_1, a_1 \quad S_2, a_2 \quad \ldots \quad S_n, a_n \]

Critique

s  s  \ldots  s

bad  good  good

• How can we efficiently get good demo input?
• What should we really learn from these demos?
Trajectory Demonstrations

- Typical input method for LfD, all state recorded from start to end.
- Many existing algorithms (GMM+GMR[1], SEDS[2], DMP[3])

Trajectory Demonstrations

A typical bad trajectory demo
Keyframe Demonstrations

- Sparse trajectory, a sequence of poses
- Learning instead of playback (e.g., animation, factory automation)
Keyframe Demonstrations
Trajectory vs. Keyframe Demonstrations

Experiment — 34 people teaching various skills to Simon in both Trajectory and Keyframe interactions

Akgun et al., HRI 2012

(a) Insert  (b) Stack  (c) Touch  (d) Close
(e) Salute  (f) Beckon  (g) Raise  (h) Throw
Trajectory vs. Keyframe Demonstrations

- People were **positive toward both** interaction modes
- Providing full, error-free, trajectory demonstration was hard and **required a very long practice** session
Keyframe-based Learning from Demonstration

Start like this

Then go here

Then go here

Finish like this

Akgun et al., J. Soc. Robotics 2012
Keyframe-based Learning from Demonstration
Keyframe-based Learning from Demonstration

**Optimal model** captures as much variance as possible along each keyframe cluster dimension
Embodied Queries for Robot Learners

- Label Queries
  “Was this correct?”

- Demo Queries
  “Can you show me an example like this?”

- Feature Queries
  “Does this feature matter for the task?”

Cakmak & Thomaz, HRI 2012
Generating Questions

LABEL QUERIES

Can I do this?
LABEL QUERIES
Generating Questions

FEATURE QUERIES

Do I have to have this rotation at the end?
FEATURE QUERIES
Interactive Machine Learning

Demonstrations

\[ S_1, a_1 \quad S_2, a_2 \quad \ldots \quad S_n, a_n \]

Critique

\[ s \quad s \quad \ldots \quad s \]

- How can we efficiently get good demo input?
- **What should we really learn from these demos?**
Learning **Goals from Demonstration**

- **Start like this**
- **Then go here**
- **Then go here**
- **Finish like this**
Visual Goals

Start like this

Then go here

Then go here

Finish like this
Haptic Goals

ATI mini40 F/T plates
Pose Constraint Goals

Start like this

Then go here

Then go here

Finish like this
Learning **Goals** from Demonstration

Start like this

Then go here

Then go here

Then go here

Finish like this

Start like this
Learning **Goals from Demonstration**

Start like this

Then go here

Then go here

Finish like this

Start like this

Then go here
Most manipulation goals aren’t only visual...
Can we build a goal model of what an action-object pair feels like?
Learning Haptic Affordances

Object-Action Demo

Human-Guided Exploration

Affordance Model
Experiment

- We tested the approach with eight different object-directed actions:
  
  - Insertion
  - Scooping
  - Pour/Shake Cheese
  - Hammering
  - Close Box
  - Cap bottle
Action Demonstration
Human Guided Exploration
Affordance Modeling

- We build an HMM from the F/T trajectories, one for the success and one for near-miss.
- Recognizing a new trajectory looks at the relative likelihood of these two models.

10 Successful Trajectories

10 Failed (near-miss) Trajectories
Results: Online Affordance Testing

• Seven object-action testing scenarios, each executed 5 times
Results: Online Affordance Testing

- Seven object-action testing scenarios, each executed 5 times
- Correctly identified 6 of 7
Results: Online Affordance Testing
Results: Online Affordance Testing
Learning **Goals** from Demonstration

Start like this

Then go here

Then go here

Then go here

Finish like this

Start like this
Best of Both: Keyframe and Trajectory

The keyframe demo highlights the salient parts of the skill,

The trajectory demo shows the dynamics at these salient points,
Best of Both: Keyframe and Trajectory

Geometric  +  Dynamic
Best of Both: Keyframe and Trajectory

\[ G = \{g_1, \ldots, g_n\} \]
\[ D = \{d_1, \ldots, d_m\} \]
\[ C = \{c_1, \ldots, c_n\} \]

\( c_i \) is pose+velocity for each keyframe

this is our input to a planner (CBiRRT)
Keyframe Demo for Geometric Constraint
Trajectory Demo for Dynamic Constraint

Slow

Fast
Resulting planned motion with constraints

Slow

Fast
Interactive Machine Learning

Demonstrations

$s_1, a_1, s_2, a_2, \ldots, s_n, a_n$

Critique

$\text{bad} \quad \text{good} \quad \text{good}$

• How can we efficiently get good demo input?
• **What should we really learn from these demos?**
Robots Learning from Human Teachers

Algorithms need to use the information that people are good at providing.
Socially Intelligent Machines Lab

http://www.gatech.edu/social-machines/

Baris Akgun
Maya Cakmak (Phd 2012)
Vivian Chu
Tesca Fitzgerald
Crystal Chao (PhD 2015)
Karl Jiang (MS 2012)
Jaewook Yoo (MS 2012)
Kalesha Bullard
Eric Huang
Yannick Schorocker
Michael Gielniak (PhD 2011)
Questions?

Andrea L. Thomaz
Georgia Institute of Technology
http://www.gatech.edu/social-machines/

Sponsors: NSF, ONR, DARPA, Honda, GM