Balancing Exploration Ratio in Reinforcement Learning

Ozcan, Ozkan

http://hdl.handle.net/10945/44365
Balancing Exploration and Exploitation Ratio in Reinforcement Learning

Ozkan Ozcan (1stLT/ TuAF)
oozcan@nps.edu
Outline

• Introduction
• Abstract
• Background
• Approaches
• Study
  – Methods
  – Analysis
• Results
• Future Work
Introduction

• Reinforcement Learning is a method by which agents can learn how to maximize rewards from their environment by managing exploration and exploitation strategies and reinforcing actions that led to a reward in the past.

• Balancing the ratio of exploration and exploitation is an important problem in reinforcement learning. The agent can choose to explore its environment and try new actions in search for better ones to be adopted in the future, or exploit already tested actions and adopt them.
• Using a strategy of pure exploration or pure exploitation cannot give the best results. The agent must find a balance between exploration and exploitation to obtain the best utility.
Abstract

• Methods to control this ratio in a manner that mimic human behavior are required for use in cognitive social simulations, which seek to constrain agent learning mechanisms in a manner similar to that observed in human cognition.

• In my thesis I used traditional and novel methods for adjusting the exploration and exploitation ratio of agents in cognitive social simulation.
Abstract

• There are different algorithms in Reinforcement Learning, such as Q-learning which can be combined with different action selection techniques such as Boltzmann selection. In my thesis I focused particularly on Q-learning algorithm and Boltzmann selection.
• Reinforcement Learning (RL)
  – Every round of the simulation each agent will be faced with a choice of possible actions.
  – The agent pairs their current state, with the possible actions called State-Action pairs.
  – Each State-Action pair will have an associated Q-value, initially determined to be equal to the default utility and later updated by the reinforcement algorithm.
  – Using the Q-values of all the actions for the agent’s current state, the agent will apply a selection algorithm to make a choice.
Environments

- Two-Armed Bandit
- Grid World (GW)
- Cultural Geography Model (CG)
Approaches

• Temperature (Exploitation-Exploration Ratio) – Controls how likely the agent is to select an action with a higher utility. Higher temp means more likely to explore, lower temp means more likely to exploit.

• I examined two methods for dynamically controlling the ratio of exploration and exploitation.
Study

- Temperature will be varied dynamically
  - Technique 1: Time Based Search then Converge
    - Start with high temp and cool over time so the agents are exploratory initially and then exploitive after some period of time.
  - Technique 2: Aggregate Utility Driven Exploration
    - Same as above except now the agents will cool based on current utility.
Technique 1: Time Based Search then Converge

\[ \tau_{\text{new}} = \frac{\tau_{\text{Initial}}}{1 + \frac{t}{t_{\text{Exploit}}}} \]

When we decrease Exploit Start Time agent is behaving greedier in a short time. By making analyze on this technique we can define what is the best exploit time in a given scenario length.
Technique 1: Two Armed Bandit

Expected Utility vs. Exploit Start Time

Scenario Length
- Range 100
- Range 200
- Range 300
- Range 400
- Range 500
- Range 600
- Range 700
- Range 800
- Range 900
- Range 1000

Expected Utility
- <= 150
- <= 200
- <= 250
- <= 300
- <= 350
- <= 400
- <= 450
- <= 500
- <= 550
- > 550
Technique 1: Grid World Example

**Expected Utility vs. Exploit Start Time**

![Graph showing Expected Utility vs. Exploit Start Time with different scenario lengths](image)

**Scenario Length**
- Range 500
- Range 600
- Range 700
- Range 1000
- Range 1200

**Legend for Expected Utility**
- <= 0.600
- <= 0.800
- <= 1.000
- <= 1.200
- <= 1.400
- <= 1.600
- <= 1.800
- > 1.800
Technique 1: Cultural Geography Model

![Expected Utility- Action 4 vs. Exploit Start Time](image)

- Expected Utility vs. Exploit Start Time
- Scenario Length
  - Range 150
  - Range 350
  - Range 540

![Expected Utility- Action 4](image)

- Expected Utility for Action 4
- Range Values:
  - <= -2.100
  - <= -2.000
  - <= -1.900
  - <= -1.800
  - <= -1.700
  - <= -1.600
  - <= -1.500
  - <= -1.400
  - > -1.400
Technique 2: Aggregate Utility Driven Exploration

\[ \tau_{\text{new}} = \frac{\tau_{\text{Initial}}}{1 + \frac{U(t)_{\text{aggregate}}}{U_{\text{acceptable}}}} \]

This technique is based on current utility of agent. When CurrentUtility is low, Temperature is getting high, and probability to select that action is also getting high.
Technique 2: Two Armed Bandit

Expected Utility vs. Exploit Utility

Scenario Length
- Range 100
- Range 200
- Range 300
- Range 400
- Range 500
- Range 600
- Range 700
- Range 800
- Range 900
- Range 1000

Expected Utility
- <= 150
- <= 200
- <= 250
- <= 300
- <= 350
- <= 400
- <= 450
- <= 500
- <= 550
- > 550

Exploit Utility
- 0
- 1
- 2
- 3

Range
- 0
- 100
- 200
- 300
- 400
- 500
- 600
- 700
- 800
- 900
- 1000
Technique 2: Grid World Example

Expected Utility vs. Exploit Utility In Different Ranges

Scenario Length
- Range 500
- Range 1000
- Range 1500
- Range 2000
- Range 2500
- Range 3000
- Range 3500
- Range 4000
- Range 4500
Technique 2: Cultural Geography Model

Expected Utility vs. Exploit Utility

Expected Utility-Action 1
- <= -0.850
- <= -0.800
- <= -0.750
- <= -0.700
- <= -0.650
- <= -0.600
- <= -0.550
- <= -0.500
- <= -0.450
- > -0.450
Results

In dynamic environments by having dynamic Temperature we can get better utilities.

Expected Utility has a relationship with scenario length and dynamic Temperature.

By figuring out this relationship we can say that for this simulation and for this scenario length if you run your simulation with this Temperature schedule you get the best utility and you can make your agent learn faster.
Future Work

The next step to adapting this model to more closely match human behavior will be to develop a technique that is:

• A combination of time based and utility based learning
• Allows for an increase in temperature due to agitation of the agent such as a change in perceived happiness
Questions ?