Balancing Exploration and Exploitation in Agent Learning

Ozkan Ozcan, Jonathan Alt, and Christian J. Darken
MOVES Institute, Naval Postgraduate School, Monterey, California, USA
{oozcan, jkalt, cjdarken} @ nps.edu

Abstract

The issue of controlling the ratio of exploration and exploitation in agent learning in dynamic environments provides a continuing challenge in the application of agent learning techniques. Methods to control this ratio in a manner that mimics human behavior are required for use in the representation of human behavior, which seek to constrain agent learning mechanisms in a manner similar to that observed in human cognition. This paper describes the use of two novel methods for adjusting the exploration and exploitation ratio of agents using a Cultural Geography (CG) Model.

Introduction

Balancing the ratio of exploration and exploitation is an important problem in reinforcement learning [1]. If you examine the relationship between agent and the environment in reinforcement learning, agent has two action selections in its environment: exploration and exploitation. 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 just one of the actions, neither exploration nor exploitation can give best results. Because of that the agent must find a balancing between exploration and exploitation to obtain the best value.

In this paper, we propose two approaches based on Boltzmann Selection to change the ratio between exploration and exploitation and examine the behavior of agents in the CG Model.

Reinforcement Learning

The Boltzmann technique assigns probabilities to all actions corresponding to their expected values, based on the value of the temperature parameter, τ. A high τ leads to exploratory behavior while a low τ leads to greedy behavior [2]. If the probability is higher it means its expected value is higher, and it is most likely to be taken. The probability is measured by following formula:

$$P_i = \frac{e^{\frac{U_i}{\tau}}}{\sum_j e^{\frac{U_j}{\tau}}}$$

The exploration and exploitation problem using this function becomes one of dynamically setting the temperature parameter in a manner that allows the agent to learn something about the environment, while eventually taking advantage of this information.

Cultural Geography Model

The Cultural Geography (CG) model is a government-owned, open-source agent-based model designed to address the behavioral response of civilian populations in conflict environments [3]. Agents within the CG Model select their action according to a constant temperature setting over the course of a model run. To enhance the functionality of agents in selecting their actions and to get more realistic results with better utilities we changed this constant to a dynamic parameter which depends on time in Time Based Selection and on utility in Aggregate Utility Selection.

Time Based Search then Converge

Inspired by the search then converge class of algorithms, this method requires that the modeler have knowledge regarding the environment in order to specify the half-life of the temperature parameter. The general form of the algorithm is shown below,

$$\tau_{new} = \frac{\tau_{Initial}}{1 + \frac{t}{t_{Exploit}}}$$

, where t is the current simulation time and $t_{Exploit}$ is the specified transition point from exploration to exploitation, equivalent to the half-life of the initial temperature.

Aggregate Utility Driven Exploration

The second method still requires the user to know something about the environment that the agent will operate in. In this case rather than an arbitrary transition

time from exploratory to greedy behavior, the user is required to know something about the reward structure of the environment. Keeping the same general form as the time based algorithm, this approach requires a user specified acceptable utility. The general form of the algorithm is shown below,

$$\tau_{new} = \frac{\tau_{Initial}}{1 + \frac{u(t)_{aggregate}}{u_{acceptable}}}$$

, where the aggregate utility at simulation time t is divided by the user specified acceptable level of utility. In dynamic environment where the aggregate utility varies over time this algorithm results in greedy behavior when an acceptable level of behavior is reached, but provides the agent the opportunity to shift back into exploratory mode should the aggregate utility drop below the threshold, due to discounting or other effects from the environment.

Approach

This section provides computational experiments, 26 replications of three different scenario lengths, to explore the two algorithms performance in CG model. For the Time Based Search Then Converge technique exploit start time is examined at every 10 steps starting from 1 to 750. For the second technique exploit utility value was changed at every 0.1 point starting from 0.1 to 4. Both techniques focused on mean expected utilities of 26 replications.

Time Based Search then Converge Results

Recall for this case that the parameter of interest to the modeler is the $t_{Exploit}$, functionally equivalent to the half-life of the initial temperature. In order to better understand the relationship between $t_{Exploit}$ and scenario length an experiment to that systematically varied these two parameters was constructed and executed for the CG Model.

Fitting a statistical model to the results indicated a significant linear relationship between scenario length and expected utility, as would be expected. A non-linear relationship was identified between $t_{Exploit}$ and expected utility.

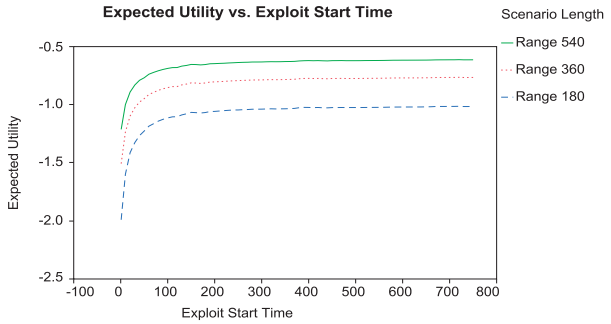

Figure 1. *Expected utility as a function of* ***tExploit*** *for 3 different scenario lengths for CG model.*

Aggregate Utility Driven Exploration Results

Recall for this case that the parameter of interest to the modeler is the $u_{acceptable}$, the threshold that the agent's aggregate utility must achieve prior to the agent's behavior becoming greedy.

Similar non-linear results were observed in fitting a regression model. Although $u_{acceptable}$ accounted for less of the variance, Rsquare=0.85 fitted with a fifth order polynomial, in the response than the same term in the time based algorithm, the agent had better utilities for each scenario lengths.

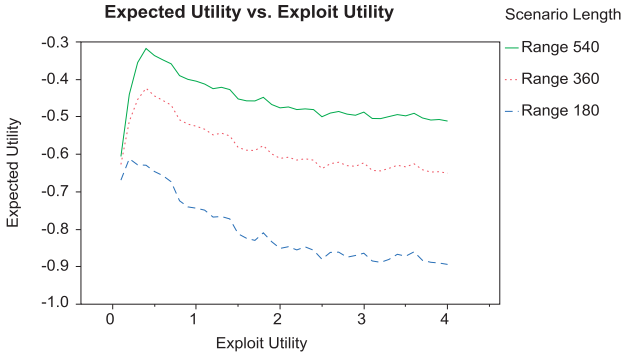

Figure 2. *Expected utility as a function of* $\boldsymbol{u_{acceptable}}$ *for 3 different scenario lengths for CG model.*

Future Work

This research demonstrated the application of two novel approaches to controlling the level of exploration and exploitation in reinforcement learning. These early results are promising as simple approaches requiring minimal knowledge of the environment to ascertain their initial setting. Future work will apply these algorithms to more complex environments, with the intended application in the representation of human behavior within modeling and simulation.

References

[1] R. Sutton and A. Barto, 1998. Reinforcement Learning: An Introduction. Cambridge: MIT Press.

[2] S. Russel and P. Norvig, 2003. Artificial intelligence: A Modern Approach, Second. Prentice Hall.

[3] Alt, J., Jackson, L., Hudak, D., and Lieberman, S. 2009. The Cultural Geography Model: Evaluating the Impact of Tactical Operational Outcomes on a Civilian Population in an Irregular Warfare Environment. Journal of Defense Modeling and Simulation: Applications, Methodology, Technology, 6(4) 185-199.