Building Relational World Models for Reinforcement Learning

Lisa Torrey, Trevor Walker, Jude Shavlik
{ltorrey,twalker,shavlik}@cs.wisc.edu
Computer Science Department
University of Wisconsin – Madison

Richard Maclin
rmaclin@d.umn.edu
Computer Science Department
University of Minnesota - Duluth

This research is supported by DARPA IP TO under contract FA8650-06-C-7606.

Abstract
Many reinforcement learning domains are highly relational. While traditional temporal-difference methods can be applied to these domains, they are limited in their capacity to exploit the relational nature of the domain. Our algorithm, AMBIL, constructs relational world models in the form of relational Markov decision processes (MDP). AMBIL works backwards from collections of high-reward states, utilizing inductive logic programming to learn their preimage, logical definitions of the region of state space that leads to the high-reward states via some action. These learned preimages are chained together to form an MDP that abstractly represents the domain. AMBIL estimates the reward and transition probabilities of this MDP from past experience. Since our MDPs are small, AMBIL uses value iteration to quickly estimate the Q-values of each action in the induced states and determines a policy. AMBIL is able to employ complex background knowledge and supports relational representations. Empirical evaluation on both synthetic domains and a sub-task of the RoboCup soccer domain shows significant performance gains compared to standard Q-learning.

Q-Learning vs AMBIL for Reinforcement Learning

Traditional Q-Learning Process
- Initialize random starting policy
- Play games using current policy
- Calculate Q-values
- Generate next policy

Q-values are calculated using standard Bellman backups, based upon the observed rewards, next state, and previous policy.

Over time, the policy will be refined, with a controlling the learning rate.

AMBIL Learning Process
- Initialize random starting policy
- Play games using current policy
- Build and solve world model
- Use solved world model as next policy

Building a world model completely replaces the Q-value learning mechanism from traditional Q-learning. AMBIL partitions state space into a relational Markov Decision Process (MDP). Each abstract state of the MDP contains estimates for the expected rewards, transition probabilities, and Q-values for each action.

Reinforcement Learning

Why use ILP?
Action Abstraction
- Simpler actions learned as single concept
- Reduces the number of concepts to learn
- Increases data available for each concept

Object Abstraction
- Objects in domains can be abstracted into classes
- Opponent classes

Learning Domain
2-on-1 Breakaway Soccer

Object: blue team attempts to shoot goal in head amount of time.
Learn: player will have to move, pass, and shoot at appropriate times.
Reinforcement: +1 Goal = Other outcome.

Maclin et al. (2004) demonstrated RL can be used on this task.

Results

Q-Learning vs AMBIL

Collect examples states by playing

Select concept to learn

Learn concept using ILP

Create MDP States

Each MDP becomes an MDP state.
For each transition in an MDP state, the preimage and episodes are evaluated and stored in a relational Markov Decision Process (MDP).

A state score is calculated using Linkage backups.

Building a Relation World Model

Select relational concepts

Learn concept and extend model

Final model

This is the final model for the example.
A special "uncovered" state provides a valid set for example states that couldn’t be covered. The score for the final model is the sum of the linkages and states.

Building a Relation World Model

Select relational concepts

Learn concept and extend model

Final model

This is the final model for the example.
A special “uncovered” state provides a valid set for example states that couldn’t be covered. The score for the final model is the sum of the linkages and states.

Building a Relation World Model

Select relational concepts

Learn concept and extend model

Final model

This is the final model for the example.
A special “uncovered” state provides a valid set for example states that couldn’t be covered. The score for the final model is the sum of the linkages and states.