Integrated Architecture for Learning, Planning and Reacting Based on Approximating Dynamic Programming.

Monica Patel (260728093)

31-March-2017

McGill University

Synopsis of the Paper

- The paper presents two architecture of DYNA.
- DYNA-PI: based on dynamic programming policy iteration method.
- DYNA-Q: Based on Watkins's Q-learning, a new kind of reinforcement learning.

Understanding DYNA

Different ways to approach a Decision Making Problem:

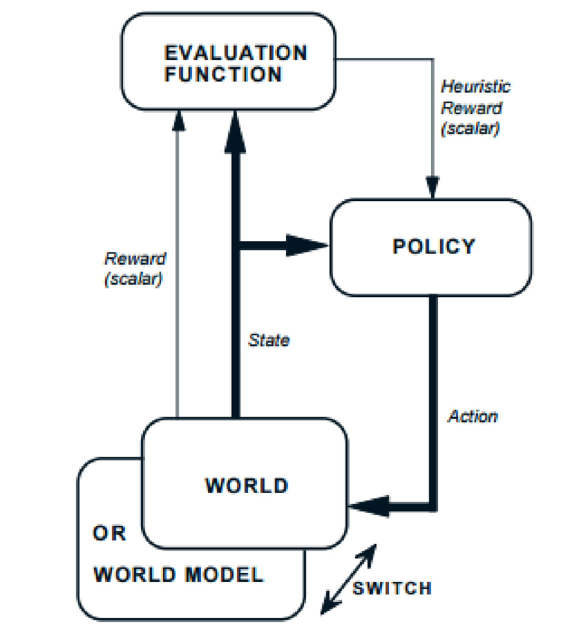
- Deciding best action given current Goal and world model.
- Plan in advance and compile its results into a set of rapid reactions, or situation-action rules, which can be used in real time decision making.
- To learn a good set of reactions by trial and error, this eliminates dependence on world model.

Dyna architectures are those that learn a world model online while using approximations to DP to learn and plan optimal behavior.

DYNA-PI

- DYNA-PI is based on approximating a Dynamic Programming method known as Policy Iteration.
- DYNA-PI included interaction of four components:
    1. Policy
    2. world (Actual)
    3. World Model
    4. Evaluation Function
- The evaluation function the policy and the world model are each updated by separate learning processes.

DYNA-PI ARCHITECHTURE



Understanding DYNA-PI

- DYNA-PI is reactive system for a fixed policy.
- A policy in DYNA-PI is in a sense plan that is conditioned on current input.
- There is switching between real world and world model.
- The planning process consist of shallow searches, each of one step layer. This ultimately produces same result as conventional search algorithm.
- DYNA-PI is a Monte Carlo or stochastic approximation variant of policy iteration in which the world model is only sampled not examined directly.
- To summarize, actual world can directly be used for sampling instead of world model (as in reinforcement learning algo) DYNA-PI accumulates results of both real world learning and planning on hypothetical model.

DYNA-Q: Dyna by Q-learning

- Only one fundamental memory structure is maintained in Q-learning. For each state action pair the Q value is stored.
- Thus in DYNA-Q only one data structure is maintained as compared to DYNA-PI
- But Q learning require one extra complexity step to calculate policy from the Q-values.
- But One advantage of Q-learning is that it requires no special adjustments if the action selection during hypothetical experience is different from the current policy.
- To encourage exploration in the method each state-action pair is given an exploration bonus proportional to this uncertainty measure.

GRID WORLD

Implementation using grid world can be seen in the video.

Limitations

- The example taken for implementation have countable small number of state action so the table can be maintained for such problem.
- For larger problem generalization will be needed.
- Explicit knowledge of the world space is assumed.
- Search control in Dyna boils down to the decision of whether to consider hypothetical or real experiences and of picking the order in which to consider hypothetical experiences, more sophisticated methods can be used for same.
- Paper conclusion is that: it is not necessary to choose between planning systems, reactive systems and learning systems they can be integrated.