Learning Robot Optimal Trajectories Online Using Inverse Reinforcement Learning

Author: Shupeng

Supervisor: Iason Batzianoulis, Prof. Aude Billard.

INTRODUCTION

Let's imagine an individual with limited capabilities sitting on a wheelchair, which is equipped with a robot arm to assist him/her and a vision system. The robot arm can execute reach-to-grasp motion. With the identification of obstacle by the vision system, we use the modulated dynamical system to help the robot arm to avoid the obstacle. To take into account the user preference, we use IRL combined with the dynamical system. Inverse reinforcement learning could model human behavior and enable Imitation Learning by learning the reward function. Reward function provides the most succinct and transferable definition of the task. The feedback from the user is limited due to the noise of the biomedical signals. We are using limited information and adapt rapidly to user's feedback.

BACKGROUND INFORMATION AND THEORY

To move the robot arm end effector from starting point to the target, we use a P controller. The obstacle avoidance motion is achieved by modulates the original dynamical system (ḟ = Mḟ + C f + U). The modulation is based on the definition of the ellipsoid. The interesting parameters for our case is reactivity p and safety margin n. By varying these two parameters, different obstacle avoidance motions can be generated. Reinforcement learning is about with given environment and the reward function to compute the reward-maximizing behavior (or say the optimal policy). In contrast, Inverse reinforcement learns the reward function with observed behavior provided, and usually, we use the reward to compute the optimal behavior.

We classify current IRL algorithms into two categories: Maximum margin formulation and maximum entropy formulation.

The Maximum Margin Formulation

The Maximum Entropy Formulation

 inverse Reinforcement Learning

 Methods

We have two scenarios: the first one is fixed parameters rho and eta, the second one is dynamically reconfigured parameters. In the fixed parameters scenario, the user input is binary: optimal or close to optimal and others. The use labels those obstacle avoidance motion as negative if either too close or too far away from the obstacle. The algorithm we proposed is the maximum margin in the obstacle. The algorithm we proposed is the maximum margin in the interest parameter space. The learning is the iteration of the gradient of log likelihood w.r.t the weights of reward. For the fixed parameters scenario, the user input is binary: optimal or not.

In the future, it would be interesting to replace the mouse with the EEG signal.

DISCUSSION AND CONCLUSION

We combine the dynamic system with IRL, so to incorporate advantages of both dynamical system and IRL such as perturbation resistance and reward learning capability. The reward function is useful, like interpreting the human decision and generalizing. One potential usage related to our specific topic is correct the position of the obstacle.

In the future, it would be interesting to replace the mouse with the EEG signal.