Research report

Kaushik Roy

October 2019

1 Current form of paper and suggested improvements

1.1 Q value bootstrapping

Gather samples from trajectories and then bootstrap Q value estimates from successive states. Fit Q value approximator to these bootstrapped estimates. However, if number of state action pairs too little, repeated bootstrapping actually takes the policy captured away from the true Q value policy.
To tackle this problem to some extent, we might use a replay buffer to store uncorrelated states to prevent a biased Q value approximation in favor of the bootstrapped Q value estimates from the samples. This problem is exacerbated in relational domains as the biasing happens because of high dimensional ground states that aren't actually all that different in their lifted representation. Hence, a replay buffer is a good bet.

1.2 Assuming the existence of many good trajectories for this to work

If we could somehow only gather a few high quality trajectories in that accurately follow the true underlying policy, we could exploit the structure in the domain to generalize to the rest of the state space. That is, similar states execute similar actions according to the true underlying policy.
How to capture similarity between states? We can do supervised metric learning from the high quality sampled trajectories. After this for any new state the policy expected can be a bayesian model average of 'k' nearest neighboring states. This can be nearest neighbor regression or prediction based on whether they are used for fitted value iteration, policy gradient methods or in imitation learning settings.

1.2.1 Metric learning

We seek to find a mapping x to Lx where, L is a linear map. Of course, distance metric properties have to hold for distances after this mapping takes place. We

could just as easily instead figure out a non linear map $\phi(x)$ by using gradient boosting. This optimization problem is non convex however and can therefore be affected by poor initialization.

1.2.2 Advice

To get a better map estimate of the bayesian posterior, we can use knowledge in conjunction with metric learning. The knowledge can be added as a multivariate gaussian prior during the learning phase for imitation learning or Harsha's way for Q value estimation for either fitted value iteration or policy gradient methods.
If the covariance matrix is diagonal then it reduces to a radial basis function network for modelling the advise prior. We can of course force the prior to prefer the model to be high recall as we might desire this, treating the occurrence of a state action pair in a trajectory as a relatively 'rare' event in a structurally complex domain.
After our previous discussion, I believe I now have a full bayesian formulation to incorporate advice information of various kinds into the prior to obtain a map estimate using bayesian model averaging. For Harsha's regression advice, it is possible to arrive at that formulation from a bayesian perspective where she calculates the fixed point estimate of the map solution instead of the expected parameter value according to the posterior distribution which bayesian model averaging will allow us to achieve.

2 Forays

As already detailed, thinking about this problem had me foray into relational metric learning and different kinds of advice modelled as adding a prior to form a bayesian posterior from which the map estimate can be found.

2.1 Relational support vector machines

This can also interestingly lead to a relational Support vector machine similar to the kind presented in F Takahashi's paper titled tree based SVM's.

2.2 Energy minimization classifier

An energy minimzation classifier using the learned metric using large margin nearest neighbor methods can also be done in relational settings. Having a non probablistic relational classifier might be really useful in situations where the prediction simply cannot afford the risk of an associated probability score.

2.3 Advice in distance based classifiers

Advice in these classifiers can follow ideas similar to Gautam's adviceptron paper to change the polytopes that bound the different classes. A clause is

equivalent to defining a polytope in the space so this shouldn't be too big a problem. But, have to give this way more thought.

2.4 POMDPs

Pomdps can be thought of as solving bayesian RL, where each MDP has a prior probability. In other words, each policy described by the MDP has a prior probability. The key difference is that we only observe the observations emitted by states in the MDP and have to figure out the probability of being in each state in the MDP. This can of course be done by the EM algorithm augmented with the learned metric and advice. This can help the sensitivity of the EM algorithm to faulty initialization. Thus model free RL in a pomdp domain would involve bayesian posterior map estimation over an observation space where the emission probabilities are learned from data. The transition probabilities can be empirically adjusted for with sample trajectories.