Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence

Multiagent Stochastic Planning With Bayesian Policy Recognition

Alessandro Panella
Department of Computer Science
University of Illinois at Chicago
Chicago, IL 60607
apanel2@uic.edu

Introduction

A rational, autonomous decision maker operating in a *partially observable*, *multiagent* setting must accurately predict the actions of other entities. For this purpose, some kind of model of the other agents needs to be maintained and updated. One option is to consider *intentional* models, and simulate their decision making process. For instance, a POMDP-based agent might consider other decision makers to be POMDP-based themselves, and maintain a probability distribution over the collection of POMDP specifications it considers possible, as in the case of interactive POMDPs (Gmytrasiewicz and Doshi 2005). In general, this approach involves maintaining a *probability distribution over all possible agents' specifications*, including their own beliefs and structure. Given the complexity of such space, this is clearly an impractical task, even without considering the additional complication that the other agents might themselves model other entities, including our agent; this recursion gives rise to an *infinite hierarchy of nested beliefs*.

An alternative is to consider *subintentional* models, which provide a predictive distribution over other agents' actions without explicitly simulating their decision-making algorithm. Within this class, stochastic finite state controllers (FSCs) provide a good tradeoff between expressivity and compactness. An FSC is a collection of nodes, each associated with a probability distribution over the agent's actions, and transition probabilities for each possible observation. Each node can be thought of as an abstract memory state of the modeled agent, a statistic that summarizes its past history. Learning accurate FSC models of other agents from their observed behavior yields an implicit *marginalization over the set of possible structures*, particularly their reward functions, when predicting their actions. It also has the benefit of *flattening the belief hierarchy*, since all the information the modeled agents use to take their decisions is implicitly contained in the inferred FSCs, regardless of their actual mental process.

From a Bayesian learning standpoint, let $m \in M$ be an FSC, and D be the observed data about the behavior of an agent being modeled. We want to compute the posterior probability $p(m|D) \propto p(D|m)p(m)$. The complexity (i.e. size) of an FSC that accurately "explains" the behavior of a modeled agent is not known *a priori*, as it depends on the unknown agent's specifications and its decision making process. Therefore, the prior $p(m)$ is over a class M of models of *unknown, unbounded size*. Recent developments in the field of Bayesian nonparametric methods offer theoretical and computational support for dealing with this requirement. In particular, priors based on the Dirichlet process have proven to be suitable to tackle a wide range of similar problems (Hjort et al. 2010).

My thesis work proposes a novel framework for autonomous planning in multiagent, partially observable domains with weak knowledge about the other agents' structure. To correctly predict their actions, FSCs are learned from their behavior using nonparametric Bayesian methods.

Progress to Date

To date, I have developed a Bayesian method for the posterior inference of an agent's FSC, given a sample of its interactions with the environment. While it is realistically not assumed that the observations received by the agent being modeled are observable by the modeling agent, the state of the world is considered observable, as well as the agent's actions. The input to the algorithm is therefore a trajectory $D = (s_{1:T}, a_{1:T})$ of state/action pairs.. Building upon the hierarchical Dirichlet process hidden Markov model (Teh et al. 2006), I have derived a collection of models with varying assumptions about the knowledge of the modeled agent's structure. Its reward function is always unknown, whereas the observation function (i.e., the probability $p(w|a, s)$ of receiving an observation $\omega \in \Omega$ for each action/state pair) can be: *a*) known; *b*) unknown, but with known cardinality $|\Omega|$; *c*) completely unknown. In the latter two cases the observation model needs to be inferred, in addition to the FSC itself. Particularly, in the case where $|\Omega|$ is unknown, a second hierarchical Dirichlet process prior needs to be used to deal with the unknown number of the agent's observations. As a result, the hypothesis space is dramatically expanded, yet the experimental findings show good accuracy, validating the proposed methodology. Inference is performed by a blocked Gibbs sampling algorithm that generates samples from the posterior distribution over FSCs.

Figure 1 reports preliminary results obtained on four test domains: three instances of the Tiger Problem, with varying

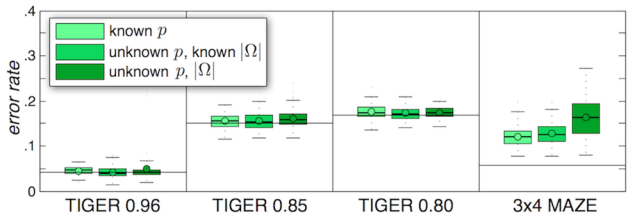

Figure 1: Action prediction error. The horizontal lines show the best achievable accuracy. The y axis is clipped at 0.4.

sensor accuracy (0.96, 0.85, 0.8), and a navigation problem in a 4x3 maze. Results are shown for each of the three assumptions regarding the modeled agent's observation function, with statistics gathered over 100 runs. The training sequences were generated from the FSCs obtained by solving the corresponding POMDP (3, 5, 7, and 44 nodes respectively.) The reported error is the proportion of incorrect action predictions along a test sequence, where each prediction is sampled from the distribution $p(a_{t+1}|s_{1:t}, a_{1:t})$ computed from the inferred FSC at each timestep. The horizontal lines represent the expected least possible error in each case, obtained by making prediction using the *actual* FSC of the agent being modeled.

In addition to multiagent settings, the proposed methodology can be used in the context of *apprenticeship learning*, where the goal is to generate a policy that closely imitates the behavior of an expert, acquired through demonstrations. I plan to submit a paper describing my progress to date to the 2013 edition of the Conference on Neural Information Processing Systems.

Proposed Work

Multiagent Planning and Online FSC Learning

In order to be used when acting in multiagent settings, the learned models must be included in a planning algorithm. In the case of POMDPs, the set of nodes of the FSC is added to the set of states of the world of the POMDP specification. Since the dynamics of the two sets are conditionally independent given the modeled agent's observation, they can be treated separately, rather than combined explicitly as in a Cartesian product, so that the growth in complexity is attenuated. I plan to integrate FSC models in both offline and online POMDP algorithms.

The batch learning method described above can be used in a scenario where an observing agent passively acquires knowledge about oblivious entities, and subsequently exploits the learned models to act optimally. Nevertheless, other agents might recursively be modeling our agent themselves. In order to take this into consideration, the agents' FSCs must be inferred *while the interaction takes place*. With learning happening online, the modeling agent's own actions influence the behavior of the other entities, and hence in turn the learning process itself. I plan to investigate the resulting dynamics as one of the contributions of this work. Regarding the technique to be adopted for online FSC learning, I am exploring the use of an online version of the Gibbs sampler and online variational inference (Bryant and Sudderth 2012).

Reward Function Inference

An additional objective of my thesis work is to learn information about the interacting agents' reward function. Inferring preferences is valuable because it generates *transferable knowledge* that can be applied to different domains of interaction, and it can be used to build *descriptive models* of human and animal behavior. Reverse-engineering the reward function of decision-theoretic agents from their behavior is known as *inverse reinforcement learning* in the context of fully observable MDPs. For POMDPs, existing approaches assume that either the agent's FSC is known, or its belief trajectory is observable (Choi and Kim 2011). My objective is to develop a methodology that does not make such assumptions, since they are often unrealistic. Two approaches will be investigated: (*a*) infer the reward function from the FSC after it is learned, and (*b*) learn FSC and reward function simultaneously. Combining the latter with online FSC learning would lead to a *preference elicitation* methodology.

Summary and Timeline

The existing and expected contributions of my thesis can be summarized as:

- Develop models and algorithms for offline and online learning of finite state controllers from observing the behavior of other agents.
- Integrate such methodologies in planning algorithms.
- Building on these results, investigate techniques for the inference of other agents' preferences and generation of transferable knowledge.

As a tentative timeline, I expect to design an online inference algorithm by July 2013, along with the integration in a POMDP planner. The remaining part, and tackling emerging challenges and new directions, will be addressed in the months to follow.

References

Bryant, M., and Sudderth, E. 2012. Truly nonparametric online variational inference for hierarchical dirichlet processes. In *Advances in Neural Information Processing Systems 25*, 27082716.

Choi, J., and Kim, K.-E. 2011. Inverse reinforcement learning in partially observable environments. *Journal of Machine Learning Research* 12:691730.

Gmytrasiewicz, P. J., and Doshi, P. 2005. A framework for sequential planning in multi-agent settings. *J. Artif. Int. Res.* 24(1):4979.

Hjort, N. L.; Holmes, C.; Mller, P.; and Walker, S. G., eds. 2010. *Bayesian Nonparametrics*. Cambridge University Press.

Teh, Y. W.; Jordan, M. I.; Beal, M. J.; and Blei, D. M. 2006. Hierarchical dirichlet processes. *Journal of the American Statistical Association* 101:1566–1581.