Reinforcement Learning for Adaptive Cognitive Orthotics

Matthew R. Rudary and **Satinder Singh** and **Martha E. Pollack***
Computer Science and Engineering
University of Michigan
Ann Arbor, MI 48109
{mrudary,baveja,pollackm}@umich.edu

Abstract

Reminder systems support people with impaired prospective memory and/or executive function, by providing them with reminders of their functional daily activities. We integrate temporal constraint reasoning with reinforcement learning (RL) to build an *adaptive* reminder system and in a simulated environment demonstrate that it can personalize to a user and adapt to both short- and long-term changes. In addition to advancing the application domain, our integrated algorithm contributes to research on temporal constraint reasoning by showing how RL can select an optimal policy from amongst a set of temporally consistent ones, and it contributes to the work on RL by showing how temporal constraint reasoning can be used to dramatically reduce the space of actions from which an RL agent needs to learn.

Introduction

Reinforcement learning (RL) has been successfully applied to a number of problems in control and operations research, but there have been relatively few applications to the design of human-computer interaction (HCI) systems; notable exceptions are Singh *et al.* (2002) and Roy, Pineau, & Thrun (2000). We describe the use of RL and temporal constraint reasoning to induce an effective interface for a *cognitive orthotic* system—a system intended to support people with impaired memory and/or executive function, by providing suitable reminders of functional daily activities. The goal of such systems is to increase the autonomy of cognitively impaired persons, allowing them to be more self-sufficient and/or to maintain self-sufficiency longer. These systems must have interfaces that are extremely intuitive and straightforward, and hence the timing and content of the interactions must be carefully considered. Moreover, because people differ from one another in many regards, and because even an individual user will change over time—particularly if she has progressive cognitive decline—the interactions must be *personalized* to the needs of the user, and *adaptive* to both short- and long-term changes in those needs. We modify Autominder (Pollack *et al.* 2003), a schedule-management system that models and maintains status information about the user's plan of daily activities, processes information obtained from sensors to infer whether and when activities have been performed, and reasons about discrepancies between the user's plan and what she has been observed doing and determines what reminders to issue. In the version of Autominder currently being used in field tests, a hand-crafted reminder strategy is used; we make this strategy adaptive.

We have therefore adopted an approach of learning effective strategies for interacting with the user of a cognitive orthotic system. Specifically, we use RL to induce an interaction policy, i.e., a function from features of the current state to interface actions, including if and when to issue a reminder to perform a certain activity. From the perspective of RL, there is at least one rather unusual and interesting challenge in building adaptive cognitive orthotic systems. In general in RL systems, the set of actions available in every state is either fixed or quite easy to determine. In contrast, in our application, determining the set of actions available in the current state is itself an NP-hard problem. At any point in time, the set of legitimate actions depends on the history of the user's activities so far as well as on the details of the user's daily plan, which in general will contain a number of complex temporal constraints; extracting these actions from the plan is computationally hard. Although in principle we could specify that the fixed set of actions for every state is the collection of all possible reminders that the system might take at any time during the day, in practice this approach is highly inefficient. We therefore integrate two powerful technologies: constraint-based temporal reasoning, which employs powerful heuristics and pruning strategies to efficiently determine what actions are legitimate in the current state, and RL to learn from experience which of the legitimate actions is optimal there.

In a series of experiments with a simulated user and environment, we demonstrate that our approach results in a personalized and adaptive cognitive orthotic system. In addition to our contribution to the application domain, our integrated learning algorithm also contributes to research on temporal constraint reasoning by showing how RL can be used to select an optimal policy from among temporally consistent ones, and it contributes to the work on RL by showing how temporal constraint reasoning can be used to dramatically reduce the space of actions from which an RL agent needs to learn.

*This research was supported by a grant from the Intel Research Council. Rudary, Singh, & Pollack (2004) is a longer version of this extended abstract.

RL-Based System Architecture

To address the limitations of a hand-crafted reminder strategy, we employ RL to infer an optimal interaction policy for each user of the cognitive orthotic system. For the most part, our agent conforms to a standard RL architecture (e.g. Sutton & Barto, 1998): the agent observes its environment (the user) with sensors and performs actions within its environment (by issuing reminders). It also receives a reward signal[1] to drive its behavior. However, we depart from the standard architecture by introducing an action proposer.

At the start of a day, the system is given the user's plan, i.e., a record of all activities the user is supposed to perform, along with constraints on the times of their performance. The action proposer has to compute which activities, if any, the user can do at each time step while still allowing the remaining activities to be done without violating any constraints (that are not already violated). This is a challenging task and we adapt Autominder's plan manager to this end. The plan manager reasons about the plan and the user's activities, and can be used to determine a set of legally executable user actions. The action proposer transforms these user actions into a set of allowable reminders, and adds a do-nothing action to the set. These actions become candidates for the agent, which then selects and executes one.

Experiments

We performed experiments using a simulator to model both a user and the sensors that detect that user's activities. We designed the simulator so that we could vary how often each activity is forgotten, when activities are executed if remembered, and how the user responds to reminders. The sensors detect when each activity is started and finished.

In each experiment, we used function approximation-based Q-learning (Watkins 1989) to train a separate linear neural network for each activity in the user's plan, plus one network for the do-nothing action. The inputs to these networks are features extracted from the user's plan and its state of execution. These experiments and their results are described in detail by Rudary, Singh, & Pollack (2004).

One of these experiments, whose results are shown in Figure 1, showed that the adaptive Autominder is able to adapt to long-term changes in the cognitive ability of a user. The bottom half of the figure shows the probability that the user forgets each of the four activities in a simple plan, over the course of 250 days. This profile may be seen in a patient who, for example, has a mild stroke at day 50, and thereafter enters a period of steady cognitive decline. In this experiment, the agent was trained every ten days using the prior 50 days of experience, and the agent acted according to an ϵ-greedy policy (where it chooses the action that currently looks best with 90% probability and a random action the rest of the time). The top half of Figure 1 shows the reward obtained each day, averaged over 10 runs of the experiment

[1]Our two goals drive the choice of reward function. We wish to maximize compliance, so we give positive reward (+2.0) every time the user correctly completes an activity. We also want to minimize dependence on the orthotic, so we charge a fixed cost (-0.6) for each reminder issued.

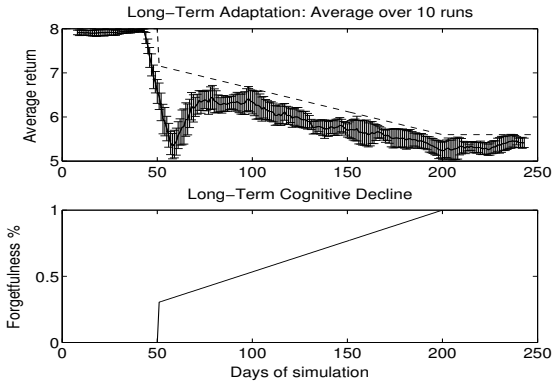

Figure 1: Results for the long-term adaptation experiment.

and then smoothed by averaging together the results of 15 consecutive days. The dashed line shows the expected value of the optimal policy; adaptive Autominder is somewhat below this because of its ϵ-greedy strategy. However, we see that the agent adapts to the changing behavior readily.

Conclusion

Our experiments show that a combination of RL and temporal constraint reasoning can produce a cognitive orthotic system that is personalized and adaptive to both short- and long-term changes in a user. The next step in this research trajectory is to deploy this adaptive Autominder system in field studies with real users. One issue that must first be addressed is the length of time required to produce a reasonable policy; in some of our experiments, as much as 30 days of data were required. Another interesting area of continued work involves generalizing the interaction policy we learn from a particular plan to other plans for the same user.

References

Pollack, M. E.; Brown, L.; Colbry, D.; McCarthy, C.; Peintner, B.; Ramakrishnan, S.; and Tsamardinos, I. 2003. Autominder: An intelligent cognitive orthotic system for people with memory impairment. *Robotics and Autonomous Systems* 44(3-4):273–282.

Roy, N.; Pineau, J.; and Thrun, S. 2000. Spoken dialogue management for robots. In *Proceedings of the 38th Annual Meeting of the Assn. for Computational Linguistics*.

Rudary, M.; Singh, S.; and Pollack, M. E. 2004. Adaptive cognitive orthotics: Combining reinforcement learning and constraint-based temporal reasoning. In *The Twenty-First International Conference on Machine Learning*.

Singh, S.; Litman, D.; Kearns, M.; and Walker, M. 2002. Optimizing dialogue management with reinforcement learning: Experiments with the njfun system. *Journal of Artificial Intelligence Research* 16:105–133.

Sutton, R. S., and Barto, A. G. 1998. *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.

Watkins, C. J. C. H. 1989. *Learning from Delayed Rewards*. Ph.D. Dissertation, King's College, Cambridge.