Machine Learning, 22, 7–9 (1996)

Introduction

LESLIE PACK KAELBLING lpk@cs.brown.edu
Computer Science Department, Brown University, Providence, RI, USA 02912-1910

This is the second special issue of *Machine Learning* on the subject of reinforcement learning. The first, edited by Richard Sutton in 1992, marked the development of reinforcement learning into a major component of the machine learning field. Since then, the area has expanded further, accounting for a significant proportion of the papers at the annual *International Conference on Machine Learning* and attracting many new researchers.

As the field grows, its boundaries become, perhaps necessarily, more diffuse. People are increasingly confused about what reinforcement learning *is*. I will take advantage of this guest editorial to outline one general conception of the field, and then to very briefly survey the current state of the art, including the papers in this issue.

1. What is reinforcement learning?

It is useful to think of reinforcement learning as a class of problems, rather than as a set of techniques. As Sutton said in the introduction to his special issue, "Reinforcement learning is the learning of a mapping from situations to actions so as to maximize a scalar reward or reinforcement signal." It is distinguished from supervised learning by the lack of a teacher specifying examples of the desired mapping and by the problem of maximizing reward over an extended period of time.

The most common techniques for solving reinforcement-learning problems are based on dynamic programming, developed in the operations research community (see texts by Puterman (1994) or Bertsekas (1995) for excellent overviews). They are based on the idea that an estimate of the utility of a state can be improved by looking ahead and using estimates of the utility of successor states. This is the basis of the temporal difference (TD) techniques (Sutton, 1988, Watkins, 1989). There are other methods, however, based on direct optimization of a policy (the paper by Moriarty and Miikkulainen in this issue illustrates the use of genetic algorithms for this purpose) or of the value function (the paper by Bradtke and Barto in this issue appplies least-squares methods), and it is important to consider these and others not yet invented, when referring to "reinforcement-learning techniques."

Robot control problems, such as navigation, pole-balancing, or juggling, are the canonical reinforcement-learning problems; but reinforcement-learning problems occur in many other situations. A particularly interesting set of applications for reinforcement learning occur in symbol-level learning (Dietterich, 1986). Tesauro's TD-Gammon system (Tesauro, 1995) is an example of one kind of symbol-level reinforcement learning. The system knows a complete model of backgammon initially and so could, in principle, simply compute the optimal strategy. However, this computation is intractable,

so the model is used to generate experience, and a strategy is learned from that experience. What results is an extremely good approximate solution, which is focused, by the training experience, on the most typical situations of the game. Another example of symbol-level reinforcement learning is Zhang and Dietterich's scheduling system (Zhang & Dietterich, 1995). In this case, the problem of learning search-control rules in a problem solver is modeled as a reinforcement-learning problem; this model is much more appropriate than the typical explanation-based learning model, in which successful traces are thought of as providing instances for a supervised learning method (Dietterich & Flann, 1995).

2. State of the art

The problem of learning from temporally-delayed reward is becoming very well understood. The convergence of temporal-difference (TD) algorithms for Markovian problems with table look-up or sparse representations has been strongly established (Dayan & Sejnowksi, 1994, Tsitsiklis, 1994). The paper by Tsitsiklis and Van Roy provides convergence results for feature-based representations; results like these are crucial for scaling reinforcement-learning to large problems. Most convergence results for TD methods rely on the assumption that the underlying environment is Markovian; the paper by Schapire and Warmuth shows that, even for environments that are arbitrarily non-Markovian, a slight variant of the standard TD(λ) algorithm performs nearly as well as the best linear estimate of the value function.

One oft-heard complaint about the TD and Q-learning algorithms is that they are slow to propagate rewards through the state space. Two of the papers in this issue address this problem with traces. The paper by Singh and Sutton considers a new trace mechanism for TD and shows that it has some theoretical and empirical advantages over the standard mechanism. The technical note by Peng and Williams develops the use, suggested by Watkins, of traces in Q-learning.

Nearly all of the formal results for TD algorithms use the expected infinite-horizon discounted model of optimality; in this issue, two additional cases are considered. Mahadevan's paper explores the problem of finding policies that are optimal in the average case, considering both model-free and model-based methods. Heger's paper addresses dynamic programming and reinforcement learning in the minimax case, in which the agent should choose actions to optimize the worst possible result.

The problem of exploration in unknown environments is a crucial one for reinforcement learning. Although exploration is well-understood for the extremely simple case of k-armed bandit problems, this understanding does not extend to the exploration of more general environments. The paper by Koenig and Simmons considers the special case of exploration in multi-state environments with goals; it shows that the problem of finding the goal even once is potentially intractable, but that simple changes in representation can have a large impact on the complexity of the problem.

A crucial problem in reinforcement learning, as in other kinds of learning, is that of finding and using bias. Bias is especially crucial in reinforcement learning, because it plays a dual role: in addition to allowing appropriate generalizations to be made, it can

guide the initial exploration in such a way that useful experience is gathered. The paper by Maclin and Shavlik allows humans to provide bias in the form of "advice" to their reinforcement-learning system; this advice is added to a neural-network representation of the value function and can be adjusted based on the agent's experience.

The papers in this issue represent a great deal of progress on problems of reinforcement learning. There is still, of course, a great deal of work remaining to be done. In particular, there are still important questions of scaling up, of exploration in general environments, of other kinds of bias, and of learning control policies with internal state. These problems, as well as others, are the subject of much current research. The future of reinforcement learning is exciting and challenging, and I hope you find this issue informative and inspiring.

References

Bertsekas, Dimitri P., (1995). *Dynamic Programming and Optimal Control.* Athena Scientific, Belmont, Massachusetts. Volumes 1 and 2.

Dayan, Peter & Sejnowski, Terrence J., (1994). TD(λ) converges with probability 1. *Machine Learning*, 14(3).

Dietterich, Thomas G., (1986). Learning at the knowledge level. *Machine Learning*, 1(3):287–315.

Dietterich, Thomas G. & Flann, Nicholas S., (1995). Explanation-based learning and reinforcement learning: A unified view. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 176–184, Tahoe City, California. Morgan Kaufmann.

Puterman, Martin L., (1994). *Markov Decision Processes.* John Wiley & Sons, New York.

Sutton, Richard S., (1988). Learning to predict by the method of temporal differences. *Machine Learning*, 3(1):9–44.

Tesauro, Gerald, (1995). Temporal difference learning and TD-Gammon. *Communications of the ACM*, pages 58–67.

Tsitsiklis, John N., (1994). Asynchronous stochastic approximation and Q-learning. *Machine Learning*, 16(3).

Watkins, C. J. C. H., (1989). *Learning from Delayed Rewards.* PhD thesis, King's College, Cambridge.

Zhang, Wei & Dietterich, Thomas G., (1995). A reinforcement learning approach to job-shop scheduling. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pages 1114–1120, Montreal, Canada. Morgan Kaufmann.