Machine Learning, 8, 225-227 (1992)

Introduction: The Challenge of Reinforcement Learning

Reinforcement learning is the learning of a mapping from situations to actions so as to maximize a scalar reward or reinforcement signal. The learner is not told which action to take, as in most forms of machine learning, but instead must discover which actions yield the highest reward by trying them. In the most interesting and challenging cases, actions may affect not only the immediate's reward, but also the next situation, and through that all subsequent rewards. These two characteristics—trial-and-error search and delayed reward—are the two most important distinguishing features of reinforcement learning.

Reinforcement learning is both a new and very old topic in AI. The term appears to have been coined by Minsky (1961), and independently in control theory by Waltz and Fu (1965). The earliest machine learning research now viewed as directly relevant was Samuel's (1959) checker player, which used temporal-difference learning to manage delayed reward much as it is used today. Of course learning and reinforcement have been studied in psychology for almost a century, and that work has had a very strong impact on the AI/engineering work. One could in fact consider all of reinforcement learning to be simply the reverse engineering of certain psychological learning processes (e.g., operant conditioning and secondary reinforcement.)[1]

Despite the early papers mentioned above, reinforcement learning was largely forgotten in the late 1960s and the 1970s. Not until the early 1980s did it gradually become an active and identifiable area of machine learning research (Barto, et al., 1981, 1983; see also Hampson, 1983). Research in genetic algorithms and classifier systems, initiated by John Holland (1975, 1986), has also been an influential part of reinforcement learning research, as has learning automata theory (see Narendra & Thathachar, 1974). Most recently, Chris Watkins (1989) and Paul Werbos (1987), among others, have invigorated theoretical research in reinforcement learning by linking it to optimal control theory and dynamic programming.

The seven articles of this special issue are representative of the excellent reinforcement learning research ongoing today. Some are theoretical, some empirical. Most of them use some form of connectionist network as part of their learning method.[2] The article by Williams introduces a gradient theory of reinforcement learning analogous to that available for connectionist supervised learning. Whereas Williams' theory treats the case of immediate reward, the article by Tesauro focusses on delayed reward. Tesauro compares temporal-difference and supervised-learning approaches to learning to play backgammon. Among other surprising results, his temporal-difference program learns to play significantly better than the previous world-champion program and as well as expert human players.

Closely related to temporal-difference learning is Q-learning (Watkins, 1989), currently the most well-understood and widely-used reinforcement learning algorithm. The technical note by Watkins and Dayan presents for the first time a complete proof of the convergence of Q-learning, a landmark result in reinforcement learning theory. The papers by Lin and by Singh take up the broader challenge of extending and scaling Q-learning and other simple reinforcement-learning methods so that they are applicable to larger and harder tasks. In one of the largest systematic comparisons of learning methods, Lin demonstrates significantly accelerated learning using novel methods for teaching-by-example and re-using prior experience. Singh's article opens up an important new direction in which to extend reinforcement learning methods—structuring them to permit transfer from simple tasks to larger, composite tasks. The next paper, by Dayan, uses Q-learning techniques to extend the theory of temporal-difference learning methods and weaken their reliance on a Markov-world assumption. Finally, Millan and Torras's paper on path-finding is noteworthy for its use of continuous rather than discrete state and action spaces. This is the first work that I know of to combine continuous actions with temporal-difference learning.

It gives me great pleasure to have assembled and to present this set of papers. These works constitute an excellent introduction to modern reinforcement learning research, but they are by no means complete. I would be remiss if I did not mention at least some of the other ongoing reinforcement-learning work, including that by Barto, et al. (1991) on dynamic programming, by Whitehead and Ballard (1991) on active perception, by Mahadevan and Connell (1990) on Q-learning in robots, and by Booker (1988) and Grefensteete, et al. (1990) on reinforcement learning in genetic systems. Many more interesting papers can be found in the proceedings of recent machine learning meetings. An excellent tutorial introduction to reinforcement learning remains to be written, but the best choices for a place to start are either the theses by Kaelbling (1990) or Watkins (1989), or else the early papers by Barto, et al. (1981).

Part of the appeal of reinforcement learning is that it is in a sense the whole AI problem in a microcosm. The task is that of an autonomous learning agent interacting with its world to achieve a goal. The framework permits the simplifications necessary in order to make progress, while at the sime time including and highlighting cases that are clearly beyond our current capabilities, cases that we will not be able to solve effectively until many key problems of learning and representation have been solved. That is the challenge of reinforcement learning.

Richard S. Sutton
GTE Laboratories Incorporated
Waltham, MA 02254
(SUTTON@GTE.COM)

Notes

1. Psychologists do not use exactly the term "reinforcement learning," so we can feel free to use it, as I do here, to refer exclusively to the engineering enterprise.
2. This is typical, but by no means necessary, as is shown by reinforcement learning research that instead uses genetic algorithms (e.g., Grefenstette, et al., 1990).

References

Barto, A.G. Bradtke, S.J. & Singh, S.P. (1991). *Real-time learning and control using asynchronous dynamic programming* (Technical Report 91-57). Amherst, MA: University of Massachusetts, Computer Science Department.

Barto, A.G. & Sutton, R.S. (1981). Landmark learning: An illustration of associative search. *Biological Cybernetics, 42,* 1-8.

Barto, A.G., Sutton, R.S. & Anderson, C.W. (1983). Neuronlike elements that can solve difficult learning control problems. *IEEE Trans. on Systems, Man, and Cybernetics, SMC-13,* 834-846.

Barto, A.G., Sutton, R.S. & Brouwer, P.S. (1981). Associative search network: A reinforcement learning associative memory. *Biological Cybernetics, 40,* 201-211.

Booker, L.B. (1988). Classifier systems that learn world models. *Machine Learning, 3,* 161-192.

Grefenstette, J.J., Ramsey, C.L. & Schultz, A.C. (1990). Learning sequential decision rules using simulation models and competition. *Machine Learning, 5,* 355-382.

Hampson, S.E. (1983). *A neural model of adaptive behavior.* Ph.D. dissertation, Dept. of Information and Computer Science, Univ. of Calif., Irvine (Technical Report #213). A revised edition appeared as *Connectionist Problem Solving,* Boston: Birkhäuser, 1990.

Holland, J.H. (1975). *Adaptation in natural and artificial systems.* Ann Arbor, MI: Univ. of Michigan Press.

Holland, J.H. (1986). Escaping brittleness: The possibilities of general-purpose learning algorithms applied to parallel rule-based systems. In: R.S. Michalski, J.G. Carbonell, & T.M. Mitchell (Eds.), *Machine learning, An artificial intelligence approach, Volume II,* 593-623, Los Altos, CA: Morgan Kaufman.

Kaelbling, L.P. (1990). *Learning in embedded systems.* Ph.D. dissertation, Computer Science Dept, Stanford University.

Mahadevan, S. & Connell, J. (1990). Automatic programming of behavior-based robots using reinforcement learning. IBM technical report. To appear in *Artificial Intelligence.*

Minsky, M.L. (1961). Steps toward artificial intelligence. *Proceedings IRE, 49,* 8-30. Reprinted in E.A. Feigenbaum & J. Feldman (Eds.), *Computers and Thought,* 406-450, New York: McGraw-Hill, 1963.

Narendra, K.S. & Thathachar, M.A.L. (1974). Learning automata—a survey. *IEEE Transactions on Systems, Man, and Cybernetics, 4,* 323-334. (Or see their textbook, *Learning Automata: An Introduction,* Englewood Cliffs, NJ: Prentice Hall, 1989.)

Samuel, A.L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal on Research and Development, 3,* 210-229. Reprinted in E.A. Feigenbaum & J. Feldman (Eds.), *Computers and Thought,* 71-105, New York: McGraw-Hill, 1963.

Waltz, M.D. & Fu, K.S. (1965). A heuristic approach to reinforcement learning control systems. *IEEE Transactions on Automatic Control, AC-10,* 390-398.

Watkins, C.J.C.H. (1989). *Learning with delayed rewards.* Ph.D. dissertation, Psychology Department, Cambridge University.

Werbos, P.J. (1987). Building and understanding adaptive systems: A statistical/numerical approach to factory automation and brain research. *IEEE Transactions on Systems, Man and Cybernetics,* Jan-Feb.

Whitehead, S.D. & Ballard, D.H. (1991). Learning to perceive and act by trial and error. *Machine Learning, 7,* 45-84.