ARTICLES PUBLISHED ELSEWHERE

TEMPORAL DIFFERENCE LEARNING *AND TD-GAMMON*

by G. Tesauro[1]

Yorktown Heights, NY, USA

Communications of the ACM, Vol. 38, No. 3, pp. 58-68.

We provide an abstract, selectively using the author's formulations:
"The article presents a game-learning program called TD-GAMMON. TD-GAMMON is a neural network that trains itself to be an evaluation function for the game of backgammon by playing against itself and learning from the outcome. It was not developed to surpass all previous computer programs in backgammon; rather, its purpose was to explore some new ideas and approaches to traditional problems in reinforcement learning.

The basic paradigm of reinforcement learning is as follows: the learning agent observes an input state or input pattern, it produces an output signal (most commonly thought of as an "action" or "control signal"), and then it receives a scalar "reward" or "reinforcement" feedback signal from the environment indicating how good or bad its output was. The goal of learning is to generate the optimal actions leading to maximal reward, which often is delayed (i.e., is given at the end of a long sequence of inputs and outputs). The learner then has to solve the *temporal credit-assignment* problem, i.e., it must figure out how to apportion credit and blame to each of the inputs and outputs leading to the final reward signal.

The reinforcement-learning paradigm has held great intuitive appeal and has attracted considerable interest for many years because of the notion of the learner being able to learn on its own, without the aid of an intelligent "teacher".

Unfortunately, despite the considerable attention devoted to reinforcement learning, so far there have been few practical successes in solving large-scale, complex, real-world problems. The crux was that, with delay, the temporal credit assignment aspect of the problem has remained extremely difficult. Many traditional approaches have also been limited to learning lookup tables or linear evaluation functions, neither of which seem adequate for handling many classes of real-world problems.

Two major recent developments promise to overcome these traditional limitations. One of these is the developments of a wide variety of novel nonlinear function approximation schemes, such as decision trees, localized basis functions, spline-fitting schemes and multilayer perceptions, which appear to be capable of learning complex nonlinear functions of their inputs.

The second development is a class of methods for approaching the temporal credit-assignment problem which have been termed "Temporal Difference" (or simply TD) learning methods. The basic idea of TD methods is that the learning is based on the difference between temporally successive predictions. In other words, the goal of learning is to make the learner's current prediction for the current input pattern more closely match the next prediction at the next time step. The most recent of these TD methods is an algorithm for training multi-layer neural networks called TD(λ).

When added to an already strong backgammon program, NEUROGAMMON, it greatly improved the latter's playing strength. The implementation has shown that, in a time span easily achieved on supercomputers, learning is indeed possible starting from random scores and random weights in the evaluation function."

The author also remarks that the results of TD-GAMMON have already modified human players' notions of the best move in some specific game positions, where conventional wisdom has turned out to be false.

---

[1] IBM Thomas J. Watson Center, P.O. Box 704, Yorktown Heights, NY, 10598. Email: tesauro@watson.ibm.com.