Convergence Theory for a New Kind of Prediction Learning

Richard S. Sutton
GTE Laboratories Incorporated

This work concerns the learning problem of predicting the behavior of time series. For concreteness, consider a time series of the states passed through by an unknown ergodic Markov chain. When any of a special set of states S is entered, a scalar outcome z is emitted; its expected value depends on the special state entered. Whether or not an outcome is emitted, the current state is assumed to be explicitly known. We want an estimate e_i, for each state i, of the expected value of the *next* outcome, given that the process is in i.

We desire solutions that are 1) *incremental*, meaning that their memory and computational requirements do not increase with time, and 2) *efficient*, meaning that their space requirements and per-step computations are $O(n)$, where n is the number of states. We discuss four algorithms.

The *optimal algorithm* accumulates maximum likelihood estimates of the Markov process and the outcome process, and forms e_i that would be exact if these estimates were exactly correct. We call these e_i the *optimal estimates*. This algorithm is ruled out by the efficiency requirement, since it is $O(n^2)$ in memory and $O(n^3)$ in computation.

The *simple averaging algorithm* forms each e_i as the arithmetic average of the outcomes that have followed visits to i. The algorithm can be implemented incrementally and efficiently, yields unbiased estimates, and has well-understood convergence behavior. Although its estimates for a finite training series are not optimal, they do minimize the mean squared error (MSE) on the training data. The algorithm has two disadvantages: 1) it cannot be applied to non-stationary processes, and 2) it is difficult to extend to the important case in which the estimates are to be formed as a (typically linear) function of a real-valued *feature vector*.

The *recency-weighted averaging algorithm* overcomes both of these difficulties. Its estimates are averages of past outcomes that weight recent outcomes more than older ones. The algorithm starts with the e_i arbitrary. Then, for each occurrence of state i, e_i is updated by $e_i \leftarrow e_i + \alpha(z - e_i)$, where z is the actual next outcome and α, $0 < \alpha < 1$, is a learning-rate parameter. The estimates are not optimal, but they converge in the mean to the correct expected values, for sufficiently small α. If a finite series is presented repeatedly to the algorithm (with slight modifications for this case), then it converges to the same estimates as found by the simple averaging algorithm in one presentation. This algorithm, extended to handle the case of feature vectors, is widely used for pattern classification and prediction.

Finally, the new kind of prediction learning mentioned in the title is the family of *temporal-difference (TD) algorithms*. The simplest of these, called *TD(0)*, is only slightly different from the recency-weighted averaging algorithm. Whereas that algorithm adjusts each prediction to look more like the *outcome* that followed it, TD(0) adjusts each prediction to look more like the *prediction* that followed it. The e_i start arbitrary, and then, on each transition $i \rightarrow j \notin S$, e_i is updated $e_i \leftarrow e_i + \alpha(e_j - e_i)$, and, on each transition $i \rightarrow j \in S$, e_i is updated $e_i \leftarrow e_i + \alpha(z - e_i)$, where z is the outcome emitted upon arrival at j.

TD(0) converges in the mean to the correct expected values, and it is efficient and incremental. In fact, its estimates are usefully updated on *every* state transition, even on those on which an outcome is not emitted. It is also applicable to the non-stationary case and is easily extended to handle feature vectors. The estimates learned by TD(0) for a finite series presented once are not optimal, but in computational experiments they appear to be more accurate than those learned by the recency-weighted averaging algorithm. Moreover, the estimates learned by TD(0) *are* optimal for a finite series presented repeatedly. The relationship of TD(0) to the optimal algorithm is analogous to that of the recency-weighted averaging algorithm to the simple averaging algorithm.

Learning algorithms based on the TD idea have previously been used in Samuel's checker-playing program, in Holland's bucket brigade, in Barto, Sutton & Anderson's pole-balancing system, and in other learning systems studied by Witten, Booker, and Hampson. The TD idea has also been used in several models of animal learning phenomena. Surprisingly, there appears not to have been any previous studies of TD methods in the mathematical or engineering literatures.

The convergence, optimality, and computational results mentioned here are presented in the referenced article. This work provides a theoretical foundation for earlier TD studies and extends them in several directions, most notably by using TD algorithms to predict arbitrary quantities, not just evaluations. The proofs given are actually for the feature vector case, which subsumes that considered here. Major areas of application for TD algorithms are temporal pattern recognition such as speech recognition, the learning of evaluation functions, and learning control.

Reference

Sutton, R. S. "Learning to predict by the methods of temporal differences." *Machine Learning*, 1988, *3*, 9–44.