Research

Reward Schemes for Learning Systems

The credit assignment problem for rule based systems in delayed payoff situations has been formalized in a *Conceptual Model* in which the environment of the system is a finite automaton. A reward scheme has been exhibited that avoids detrimental biases even when eligibility sets overlap. The demonstration that such biases are avoided is an analogue of the proof of Fisher's fundamental theorem of natural selection. This result applies to what in Classifier Systems is called Profit Sharing.

To help provide a conceptual testbed for reward schemes, a new two component cascade decomposition of finite automata has been obtained. Sometimes current payoff is due to a system action taken a long time ago. The question is, how long ago might such an action have occurred? This question can be usefully re-formulated if the environment is decomposed using the cascade decomposition. In the re-formulation it is often possible to ignore all but the first component of the decomposition.

Main Present Work

My article "The bucket Brigade is not Genetic" explains that the Bucket Brigade (like sensible learning systems) directly rewards the system for achieving subgoals, whereas biological evolutionary systems (like Genetic Algorithms, the Pitt approach, and Profit Sharing) do not. It says the difference is fundamental. The article was intended to provoke a counter argument, possibly from me, for if the article is right then there is little the evolutionary process can tell us about learning, and the whole programme that gave rise to the Genetic Algorithms community is flawed. A counter argument must rely on a shift in viewpoint. It has been long in coming. Since 2005 I have been tackling the counter argument directly and making new analogies between sequential and parallel adaptive systems. What follows is a rough sketch of the plot line and state of play. (Most definitions, limitations, and caveats are omitted.)

Classifier Systems and Genetic Algorithms formalism gives us a way of using the Conceptual Model to examine the issues. Reproduction is supposed to increase the payoff per time unit. The rate of increase I call the climb rate. The climb rate depends on the reproductive rates. I define climb rate formally, assuming no noise. A system must estimate the appropriate (allele) reproductive rates, but these estimates are subject to sampling error noise. To prevent premature convergence, we want the standard deviation of this noise to be low. Thus we want the standard deviation of the (genotype) reproductive rates to be low, although the standard deviation of the (genotype) values needs to be high. The ratio of the climb rate to the standard deviation of the reproductive rates I call the *Effectiveness Ratio*. It is independent of the learning rate coefficient and it needs to be high. Under certain assumptions, it is highest when reproductive rate is proportional to value (what Holland calls "fitness proportional selection").

In more general situations, we can increase the Effectiveness Ratio by increasing sampling rate, because this decreases the noise. In Classifier Systems, a rule (classifier) in a firing sequence is analogous to an allele in a genotype. We can implicitly increase sampling rate by estimating reproductive rates not with a Profit Sharing scheme but with a subgoal reward scheme like the Bucket Brigade. But this is at the cost of introducing additional biases that can decrease the climb rate.

In the Bucket Brigade, cash resulting from a given payoff propagates backwards along a tree of rules that I call the *reward tree*. Paths in that tree constitute a sample. More branching makes the sample larger and thus increases the sampling rate. Rules are reward-linked in the tree partly via message list messages, whereas in the Pitt approach, rules are reward-linked by being in the same Pitt individual. Such linkage in both cases is in a sense arbitrary and evolves by selection, but the unit of selection is the reward tree or the Pitt individual.

These two units of selection look similar, and I am trying to show formal equivalence in some sense. This would help connect human learning with genetic evolution. The argument may need to use rules each of which has more than one condition, hence more that one precursor. Arguments that use such rules seem plagued with performance instability. Even when performance is stable, it appears that the Bucket Brigade needs to operate strangely. Every time a rule passes some of its cash back, all that cash must go to one precursor and the other precursors must get matching sums of newly created cash. Surprisingly, this seems to work. (It was evidently Riolo's practice.)

It may be possible to stick to one condition per rule and proceed as in my article "Implicit Group Selection in a Michigan Classifier System", where groups of Michigan Classifiers resemble Pitt individuals.

At present, there are gaps and possibly errors in the plot line, but eventually we should be able to properly characterize the essential differences between genetic systems and subgoal reward systems, or better, show that there are none.