SEQUENCING WITH NEURAL NETWORKS IN FLEXIBLE ASSEMBLING SYSTEMS

Automatic Formation of Optimal Control Strategies with TD-Method

Elgar Fleisch, Institut für Höhere Studien, Wien
Harald Hruschka, Universität Regensburg

A flexible assembling system (FAS) consisting of four partly different flexible assembling cells is to be controlled. The objective of the intelligent control system (agent) is minimization of mean flowtime, mean tardiness or any arbitrary combination of these competitive performance measures. The control problem involves sequencing and routing tasks.

The agent performs a heuristic dynamic programming approach in an adaptive heuristic critic framework to learn an optimal decision policy (strategy). The TD-Errors generated by application of the TD-method provide reinforcement signals used to adjust the parameters of the agent. The agent learns a strategic utility function and an optimal decision policy using a neural network approach. The applied design of adaptive heuristic critics works without built-in knowledge and without a model describing the controlled world.

Sequencing in FAS is a Markov decision problem and defined in terms of a discrete-time stochastic dynamical system with finite state set. At each time-step the controller observes the world's state i and selects an action k , which moves the world to state j with probability p_{ij}^k . The transition from state i to j produces a payoff r_{ij}^k , where $|r_{ij}^k| < \infty$ for each i, j and k . For any policy Π and state i one can define the expected infinite-horizon discounted return

$$E_{\Pi} \left[\sum_{t=0}^{\infty} \gamma^t r(t) \mid i(0) = i \right], \quad (1)$$

where $r(t)$ denotes the payoff at time t , $i(0)$ is the system's initial state and γ , $0 \leq \gamma \leq 1$ is a factor used to discount the future payoffs. The objective of the agent is to form a policy that maximizes the expected return defined by Equation 1 for each state i .

To provide suitable payoffs one has to measure the control-targets incrementally. The agent receives the states and the payoffs from the FAS online and performs direct learning. Applied to single-machine problems the agent converges to the well known optimal rules SPT and EDD. When applied to complex FAS problems the new control system converges to policies superior to advanced priority rules. The intricacy of tardiness-optimization strongly depends on the placed due dates. The automatically generated policies are priority rules and may be used for knowledge extraction.

References:

- Barto90** Barto, Andrew G.; Sutton, Richard S.; Watkins, Christopher J.C.H; Learning and Sequential Decision Making, in: Gabriel, M. and Moore, J. (editors); Learning and Computational Neuroscience: Foundations of Adaptive Networks; The MIT Press; Cambridge, MA, 1990
- Biendl84** Biendl, Peer; Ablaufsteuerung von Montagefertigungen; Haupt; Bern-Stuttgart, 1984
- Tempelmeier92** Tempelmeier, Horst; Kuhn, Heinrich; Flexible Fertigungssysteme; Entscheidungsunterstützung für Konfiguration und Betrieb; Springer-Verlag; Berlin, 1992