The exploration/exploitation dilemma in Markov decision problems

Peter Auer

Montanuniversität Leoben

The classical model of the exploration/exploitation dilemma is the bandit problem with stationary reward distributions for the possible actions: a good exploration strategy finds the optimal action quickly without trying suboptimal actions too often. In an MDP the exploration problem is more complicated since an action – besides yielding some reward – also causes a state transition according to an unknown distribution.

After reviewing some results for the bandit problem, we analyse an effective exploration strategy for MDPs. Although doing exploration, this strategy yields rewards which are very close to the rewards of an optimal policy (which always chooses optimal actions). We also show that the gap between the sum of optimal rewards and the rewards of our strategy cannot be improved significantly.