Representation discovery for MDPs using bisimulation metrics

Sherry Shanshan Ruan
McGill University
shanshan.ruan@mail.mcgill.ca

Prakash Panangaden
McGill University
prakash@cs.mcgill.ca

Gheorghe Comanici
McGill University
gcoman@cs.mcgill.ca

Doina Precup
McGill University
dprecup@cs.mcgill.ca

Solving large sequential decision problems modelled as Markov Decision Processes (MDPs) requires the use of approximations to represent the state space. Popular approximation methods include state aggregation, linear function approximation and kernel-based methods. In this work we are mainly interested in state aggregation, in which the state space is partitioned into disjoint subsets and values are associated with each partition. The goal is to construct a partition incrementally, in such a way as to provide a good approximation to the true value function. One approach for this problem is to use bisimulation relations [6], also known as MDP homomorphisms [8], or their relaxation as *bisimulation metrics* [4]. Bisimulation metrics in particular are attractive because they allow quantifying the approximation error for *any* state space partitioning, or more generally, any linear function approximator [3]. However, bisimulation metric computation is very expensive (in the worst case, more expensive than performing dynamic programming in the original state space). Indeed, recent work [5] has shown that computing the metric amounts to solving an MDP resulting from a coupling of the state space with itself; such a coupling has size quadratic in the number of states.

In this work, we tackle this problem by proposing a significant improvement in how bisimulation metrics are computed. Our approach constructs an iteratively improving sequence of state space partitions, which still converges in the limit to the bisimulation relation. We prove that at each step the error of the value function computed over this partition (compared to the true optimal value function) is bounded. Since at each step, the value function approximation is computed over partitions rather than states, this approach can generate substantial space and computation time savings, as illustrated in our experiments.

The second contribution is an algorithm for asynchronous updates of the metric and the representation. We provide theoretical conditions which allow computational effort to be focused on parts of the state space where changes are happening rapidly, similar to successful asynchronous or distributed dynamic programming techniques such as [1, 2, 7]. Empirical results illustrate the use of heuristics that can substantially speed up the computation.

In conclusion, we presented two new ways of describing bisimulation metrics from a theoretical perspective, and we used these to design novel iterative refinement algorithms. These algorithms provide substantial improvement in terms of time and memory usage and more flexibility in terms of guiding the search for alternative state space representations for MDPs.

The approach presented opens the door to more specialized strategies to finding bisimulation-based MDP representations. We illustrated the advantage of using heuristic based search strategies, but the strategy we used (which attempts to keep the size of state partitions roughly the same) is very simple, and it is likely that more sophisticated approaches would work better. For example, one could try strategies similar to prioritized sweeping, which focus on areas of the state space where the metric is changing drastically. Investigating more sophisticated heuristics and applying them to larger problems is a worthwhile approach for future work.

REFERENCES

[1] D P Bertsekas and John N Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, Bellman, MA, 1996.

[2] Dimitri P. Bertsekas and David A. Castanon. Adaptive aggregation methods for infinite horizon dynamic programming. *IEEE transactions on automatic control*, 34, 1989.

[3] Gheorghe Comanici and Doina Precup. Basis Function Discovery using Spectral Clustering and Bisimulation Metrics. 2011.

[4] Norm Ferns, Prakash Panangaden, and Doina Precup. Metrics for finite Markov decision precesses. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pages 162–169, July 2004.

[5] Norma Ferns and Doina Precup. Bisimulation metrics are optimal value functions. In *UAI*, 2014.

[6] Robert Givan, Thomas Dean, and Matthew Greig. Equivalence notions and model minimization in Markov decision processes. *Artificial Intelligence*, 147(1-2):163–223, 2003.

[7] Andrew Moore and Chris Atkeson. Prioritized sweeping: Reinforcement learning with less data and less real time. *Machine Learning*, 13:103–130, 1993.

[8] Balaraman Ravindran and Andrew G. Barto. Model minimization in hierarchical reinforcement learning. In *SARA*, pages 196–211, 2002.