Encoding and Combining Knowledge to Speed up Reinforcement Learning

Tim Brys
Vrije Universiteit Brussel
Brussels, Belgium
timbrys@vub.ac.be

Abstract

Reinforcement learning algorithms typically require too many 'trial-and-error' experiences before reaching a desirable behaviour. A considerable amount of ongoing research is focused on speeding up this learning process by using external knowledge. We contribute in several ways, proposing novel approaches to transfer learning and learning from demonstration, as well as an ensemble approach to combine knowledge from various sources.

1 Introduction

Reinforcement learning [Sutton and Barto, 1998] is a paradigm that allows an agent to learn how to control a system in order to achieve specific goals. The agent is guided by reward/punishment received for the behaviour it exhibits, adjusting its behaviour in order to maximize the cumulative reward. In complex tasks, or tasks with sparse rewards, learning can be excruciatingly slow (as many learning algorithms take the *tabula rasa* approach), and the agent cannot do better than behaving randomly until feedback is received.

A lot of research in this domain is therefore dedicated to speeding up the learning process, relying on the incorporation of various pieces of external knowledge. Reward shaping is a popular technique with strong theoretical guarantees that allows for the incorporation of such knowledge [Ng *et al.*, 1999]. It modifies the typically sparser reward signal by adding extra intermediate rewards based on the prior knowledge encoded as a potential function. How to define knowledge as a potential function is not always trivial, and when several pieces of knowledge are available, how to combine these in the best way is also an open question. Our current contributions are three-fold: we formulate

- a reward shaping approach to policy transfer
- a reward shaping approach to learning from demonstration
- an ensemble approach to combining various pieces of heuristic knowledge encoded as reward shapings

We elaborate on these three topics in the following sections.

2 Transfer Learning

Transfer learning involves using knowledge learned in a previous, similar, task, to improve learning in the current task. Transfer learning has become quite popular in reinforcement learning in the past decade thanks to its many empirical successes [Taylor and Stone, 2009]. Yet, many of the existing techniques involve the use of low level information obtained in the source task, which may not be transferrable to or incompatible with the agent learning in the new task. In the most basic case, one can only assume access to a representation of the behaviour from the source task, as learned by a reinforcement learning or learning from demonstration algorithm, or learned or defined in some other way.

We have developed a novel approach to policy transfer, encoding the transferred policy as a dynamic potential-based reward shaping function [Brys *et al.*, 2015b]. This firmly grounds our approach in an actively developing body of theoretical research around reward shaping, resulting in a technique with important policy invariance guarantees. An experimental comparison with the state-of-the-art has shown how this technique outperforms and is more robust against the suboptimality of transferred knowledge than the state-of-the-art in several domains.

3 Learning from Demonstration

Learning from demonstration [Argall *et al.*, 2009] is an approach to behaviour learning where an agent is provided with demonstrations by a supposed expert, from which it should derive suitable behaviour. Yet, one of the challenges of learning from demonstration is that no guarantees can be provided for the quality of the demonstrations, and thus the learned behavior. We have investigated the intersection of learning from demonstration and reinforcement learning, where we leverage the theoretical guarantees provided by reinforcement learning, and use expert demonstrations to speed up the learning process by biasing exploration through reward shaping [Brys *et al.*, 2015a]. This approach allows us to leverage (human) demonstrations without making an erroneous assumption regarding demonstration optimality. We show experimentally that this approach requires significantly fewer demonstrations, is more robust against suboptimality of demonstrations, and achieves much faster learning than the state-of-the-art, while providing several theoretical guarantees on convergence and policy invariance.

4 Ensembles of Shapings

It is quite possible that an agent designer has several pieces of knowledge available to aid the agent in its learning process. Not only transferrable knowledge from previous tasks or human demonstrations, but also (and more often) heuristic rules devised by domain experts, or abstract knowledge obtained during learning could be available. Prior art dealt with the problem of combining these various pieces of knowledge by simply summing all of them in a single complex reward shaping function [Devlin *et al.*, 2011].

In contrast, we propose to use ensembles of shapings to handle the combination of these various pieces of knowledge [Brys *et al.*, 2014a]. Learning each heuristic in parallel allows an agent to see the bias introduced by each piece of knowledge, and avoids problems such as shapings of high magnitudes dominating others of lower magnitude. Assuming that the majority of the shapings is useful in any given situation, simple voting mechanisms allow the agent to combine the suggestions made by each shapings, creating a learner that makes the most out of each piece of knowledge [Brys *et al.*, 2014c].

Besides using simple voting mechanisms, we developed a more elaborate technique that estimates in each situation which of the shapings can be trusted most, and uses this confidence measure to make decisions [Brys *et al.*, 2014b]. We have shown how this ensemble technique makes very intuitive decisions about when to use which of the pieces of knowledge, and can even improve performance when each of the shapings in itself is detrimental to performance, i.e. it is robust against low quality heuristics.

5 Conclusions and Future Work

The aim of our work is to facilitate the inclusion of prior knowledge in reinforcement learning with as little tuning as possible. Reward shaping is an appealing framework to this end, as it provides several theoretical guarantees towards convergence and optimality of policies. Whereas in most work, the knowledge included comes in the form of heuristics, rules of thumb, defined by a domain expert, we have shown how knowledge transferred from a different previous task, and demonstrations by humans or other agents, can be leveraged within this framework. Furthermore, we have developed a robust approach to the combination of several such pieces of knowledge, using ensemble techniques. This aims to provide an off-the-shelf solution to reward shaping, removing the need for behind the scenes tuning.

Currently, we have only validated our ensembles approach to reward shaping with heuristics devised by domain experts. But, ensembles could also be used to achieve multi-task transfer for example, assuming different source tasks will contribute different information to the target task, or to incorporate demonstrations given by different experts, assuming different experts' demonstrations are significantly different. The ultimate goal is to demonstrate how an ensemble provided with a number of these types of information (transferred knowledge, demonstrations, heuristics) can allow a reinforcement learning agent to solve a complex practical application, such as a robotics manipulation task.

Acknowledgments

Tim Brys is funded by a Ph.D grant of the Research Foundation Flanders (FWO).

References

[Argall *et al.*, 2009] Brenna D. Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. A survey of robot learning from demonstration. *Robotics and autonomous systems*, 57(5):469–483, 2009.

[Brys *et al.*, 2014a] Tim Brys, Anna Harutyunyan, Peter Vrancx, Matthew E Taylor, Daniel Kudenko, and Ann Nowé. Multi-objectivization of reinforcement learning problems by reward shaping. In *International Joint Conference on Neural Networks (IJCNN)*, pages 2315–2322. IEEE, 2014.

[Brys *et al.*, 2014b] Tim Brys, Ann Nowé, Daniel Kudenko, and Matthew E. Taylor. Combining multiple correlated reward and shaping signals by measuring confidence. In *Twenty-Eighth AAAI Conference on Artificial Intelligence (AAAI)*, pages 1687–1693, 2014.

[Brys *et al.*, 2014c] Tim Brys, Matthew E Taylor, and Ann Nowé. Using ensemble techniques and multi-objectivization to solve reinforcement learning problems. In *European Conference on Artificial Intelligence (ECAI)*, 2014.

[Brys *et al.*, 2015a] Tim Brys, Anna Harutyunyan, Halit Bener Suay, Sonia Chernova, Matthew E. Taylor, and Ann Nowé. Reinforcement learning from demonstration through reward shaping. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2015.

[Brys *et al.*, 2015b] Tim Brys, Anna Harutyunyan, Matthew E. Taylor, and Ann Nowé. Policy transfer using reward shaping. In *International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, 2015.

[Devlin *et al.*, 2011] Sam Devlin, Daniel Kudenko, and Marek Grześ. An empirical study of potential-based reward shaping and advice in complex, multi-agent systems. *Advances in Complex Systems*, 14(02):251–278, 2011.

[Ng *et al.*, 1999] Andrew Y. Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the Sixteenth International Conference on Machine Learning*, volume 99, pages 278–287, 1999.

[Sutton and Barto, 1998] R.S. Sutton and A.G. Barto. *Reinforcement learning: An introduction*, volume 1. Cambridge Univ Press, 1998.

[Taylor and Stone, 2009] Matthew E. Taylor and Peter Stone. Transfer learning for reinforcement learning domains: A survey. *The Journal of Machine Learning Research*, 10:1633–1685, 2009.