Overcoming Erroneous Domain Knowledge in Plan-Based Reward Shaping

(Extended Abstract)

Kyriakos Efthymiadis
University of York, UK

Sam Devlin
University of York, UK

Daniel Kudenko
University of York, UK

ABSTRACT

Reward shaping has been shown to significantly improve an agent's performance in reinforcement learning. Plan-based reward shaping is a successful approach in which a STRIPS plan is used in order to guide the agent to the optimal behaviour. However, if the provided domain knowledge is wrong, it has been shown the agent will take longer to learn the optimal policy. Previously, in some cases, it was better to ignore all prior knowledge despite it only being partially erroneous.

This paper introduces a novel use of knowledge revision to overcome erroneous domain knowledge when provided to an agent receiving plan-based reward shaping. Empirical results show that an agent using this method can outperform the previous agent receiving plan-based reward shaping without knowledge revision.

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning

General Terms

Experimentation

Keywords

Reinforcement Learning, Reward Shaping, Knowledge Revision

1. INTRODUCTION

Reinforcement learning (RL) has proven to be a successful technique when an agent needs to act and improve in a given environment. The agent receives feedback about its behaviour in terms of rewards through constant interaction with the environment. Traditional reinforcement learning assumes the agent has no prior knowledge about the environment it is acting on. Nevertheless, in many cases (potentially abstract and heuristic) domain knowledge of the RL tasks is available, and can be used to improve the learning performance through potential-based reward shaping [3].

Plan-based reward shaping [2] is an instance of reward shaping, where the agent is provided with a high level STRIPS plan which is used in order to guide the agent to the desired behaviour.

However, problems arise when the provided knowledge is partially incorrect or incomplete, which can happen frequently given that expert domain knowledge is often of a heuristic nature. It has been shown in [2] that if the provided plan is flawed then the agent's learning performance drops and in some cases is worse than not using domain knowledge at all.

This paper presents, for the first time, an approach in which agents use their experience to revise erroneous domain knowledge whilst learning and continue to use the then corrected knowledge to guide the RL process.

We demonstrate, in this paper, that adding knowledge revision to plan-based reward shaping can improve an agent's performance (compared to a plan-based agent without knowledge revision) when both agents are provided with erroneous domain knowledge.

Appears in: *Proceedings of the 12th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2013)*, Ito, Jonker, Gini, and Shehory (eds.), May, 6–10, 2013, Saint Paul, Minnesota, USA.

2. EVALUATION DOMAIN

We evaluate our method using the flag-collection domain, an extended version of the navigation maze problem which is a popular evaluation domain in RL. An agent is modelled at a starting position from where it must move to the goal position. In between, the agent needs to collect flags which are spread throughout the maze.

During an episode, at each time step, the agent is given its current location and the flags it has already collected. From this it must decide to move up, down, left or right and will deterministically complete their move provided they do not collide with a wall. Regardless of the number of flags it has collected, the scenario ends when the agent reaches the goal position. At this time the agent receives a reward equal to one hundred times the number of flags which were collected.

3. OVERCOMING INCORRECT KNOWLEDGE

While the agent is performing low level actions, it can gather information about the environment and in this specific case, information about the flags it was able to pick up. This information allows the agent to discover potential errors in the provided plan which in the incorrect case are goals that are present in the plan, but not in the simulation. Errors in the case of plan-based reward shaping come in the form of plan operator preconditions that cannot be satisfied.

When an error is found, the agent switches to verification mode trying to satisfy the precondition which is failing by

performing DFS. If the precondition cannot be satisfied then the knowledge base is contracted[1] and that particular precondition is removed from the plan's initial conditions. The plan is then recomputed.

4. OVERCOMING INCOMPLETE KNOWLEDGE

In the case of incomplete knowledge, while the agent performs low-level actions, it can satisfy important goals in the environment that are not present in the plan. If a new goal is discovered, the knowledge base is expanded[2] in order to include the new information.

The new information is then added to the initial conditions of the plan and a new plan is computed.

5. EVALUATION

In our experiments all agents implemented SARSA with $\epsilon-$greedy action selection and eligibility traces. For all experiments, the agents' parameters were set such that $\alpha = 0.1$, $\gamma = 0.99$, $\epsilon = 0.1$ and $\lambda = 0.4$. Each experiment lasted for 50000 episodes and was repeated 10 times for each instance of the erroneous knowledge.

5.1 Incorrect knowledge

In the incorrect knowledge case, the agents are provided with a plan which contains extra goals which cannot be achieved in simulation. The results are shown in Figure 1.

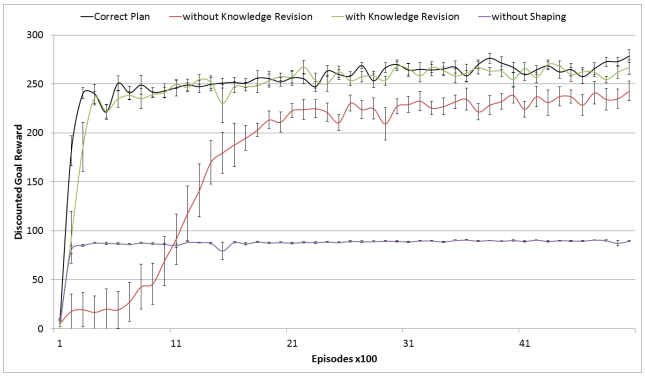

Figure 1: Incorrect knowledge.

It is apparent that the plan-based RL agent without knowledge revision is not able to overcome the incorrect knowledge and performs sub-optimally throughout the duration of the experiments. However, the agent with knowledge revision manages to identify the flaws in the plan and quickly rectify its knowledge. As a result after only a few hundred episodes of performing sub-optimally it manages to reach the same performance as the agent which is provided with correct knowledge.

5.2 Incomplete knowledge

In the incomplete case, the agents are provided with a plan which does not contain all the goals the agent should achieve in simulation. The results are shown in Figure 2.

[1]A rule ϕ, along with its consequences is retracted from a set of beliefs K.[1]

[2]A new information ϕ is added to the current belief base K. [1]

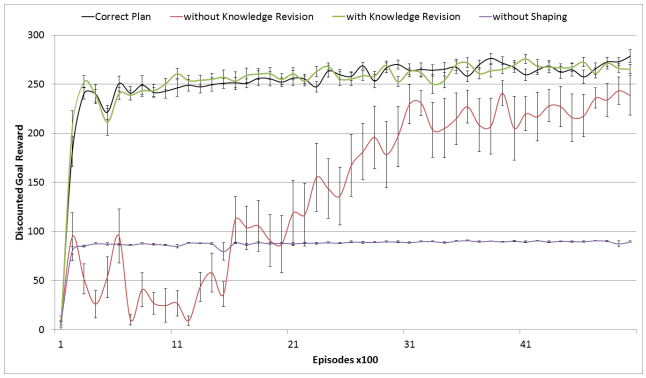

Figure 2: Incomplete knowledge.

Again, it is clear that the original plan-based RL agent without knowledge revision struggles to overcome the incomplete plan and performs sub-optimally throughout the course of the experiments. Our agent using knowledge revision manages very early on in the experiment to identify the flags which are missing from the plan and update its knowledge base. As a result it reaches a performance similar to the agent receiving the correct plan within a few hundred episodes.

6. CLOSING REMARKS

When an agent receiving plan-based reward shaping is guided by erroneous knowledge it can be led to undesired behaviour in terms of convergence time and overall performance in terms of total accumulated reward.

Our contribution is a novel generic method for overcoming erroneous knowledge in terms of incomplete and incorrect plans when provided to a plan-based RL agent.

Our experiments show that using knowledge revision in order to incorporate an agent's experiences to the provided high level knowledge can improve its performance and help the agent reach its optimal policy. The agent manages to revise the provided knowledge early on in the experiments and thus benefit from more accurate plans.

In future work we intend to investigate the approach of automatically revising erroneous knowledge in stochastic and dynamic domains, multi-agent environment and real life complex applications.

7. ACKNOWLEDGEMENTS

This study was partially sponsored by QinetiQ.

8. REFERENCES

[1] P. Gärdenfors. Belief revision: An introduction. *Belief revision*, 29:1–28, 1992.

[2] M. Grześ and D. Kudenko. Plan-based reward shaping for reinforcement learning. In *Proceedings of the 4th IEEE International Conference on Intelligent Systems (IS'08)*, pages 22–29. IEEE, 2008.

[3] A. Y. Ng, D. Harada, and S. J. Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the 16th International Conference on Machine Learning*, pages 278–287, 1999.