Multi-Agent, Reward Shaping for RoboCup KeepAway (Extended Abstract)

Sam Devlin
University of York, UK

Marek Grześ
University of Waterloo, CA

Daniel Kudenko
University of York, UK

ABSTRACT

This paper investigates the impact of reward shaping in multi-agent reinforcement learning as a way to incorporate domain knowledge about good strategies. In theory [2], potential-based reward shaping does not alter the Nash Equilibria of a stochastic game, only the exploration of the shaped agent. We demonstrate empirically the performance of state-based and state-action-based reward shaping in RoboCup KeepAway. The results illustrate that reward shaping can alter both the learning time required to reach a stable joint policy and the final group performance for better or worse.

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning; I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence—*Multiagent Systems*

General Terms

Experimentation

Keywords

Reinforcement Learning, Reward Shaping,
Multiagent Learning, Reward Structures for Learning.

1. INTRODUCTION

Most multi-agent, reinforcement learning agents are implemented under the assumption that there is no prior knowledge available. This is, however, often not the case in many practical applications. In many domains, heuristic knowledge can be easily identified by the designer of the system.

In the area of single-agent reinforcement learning, incorporating heuristic knowledge by a potential-based reward shaping has been proven to be both sufficient and necessary to not modify the optimal policy of the agent [7]. However, in multi-agent the implications of the method are different [2].

To date, only relatively simple multi-agent scenarios have been studied with regard to potential-based reward shaping [1, 2, 5]. The contribution of this work is the first application of potential-based reward shaping [7] to a complex MAS, the first application of potential-based advice [3] to any MAS and the proposal of three, generally applicable, multi-agent specific categories of domain knowledge.

Cite as: Multi-Agent, Reward Shaping for RoboCup KeepAway (Extended Abstract), Sam Devlin, Marek Grześ and Daniel Kudenko, *Proc. of 10th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2011)*, Tumer, Yolum, Sonenberg and Stone (eds.), May, 2–6, 2011, Taipei, Taiwan, pp. 1227-1228.

Figure 1: A 3 vs. 2 KeepAway game [8].

2. KEEPAWAY

KeepAway [8] is a sub-problem of the complete game of soccer/football. In this task (see Figure 1), N players (keepers) learn how to keep the ball when attacked by $N-1$ takers within a small, fixed area of the football pitch.

Most published learning agents in KeepAway learn the behaviour of the Keeper in possession of the ball, but as at any one time only one agent has possession this research is more relevant to single-agent reinforcement learning. Instead, we focus on learning the behaviour of the Takers who must simultaneously decide to mark a specific keeper or tackle for the ball.

3. EMPIRICAL STUDY

Our baseline learner combines the approaches of two existing published learning takers [4, 6]. Specifically we use the reward function and state representation of Min et al. [6] and the SARSA algorithm with tile coding and ϵ-greedy action selection method as Iscen and Erogul [4] did. The resulting takers outperform both existing agents gaining possession on average in just 4.8 seconds in a game of 3v2 on a pitch of size 20x20.

To extend this baseline, we treat the agents as black boxes and simply provide an additional potential-based reward. To demonstrate both state-based [7] and state-action based [3] reward shaping, three heuristics were designed:

1. *Separation-Based:* Encourage takers to spread out.
2. *Role-Based:* Encourage one taker to tackle, the others to mark.
3. *Combined:* The combination of (1) and (2).

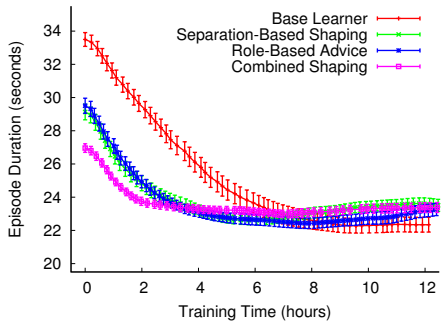

Figure 2: 3v2 at 50x50.

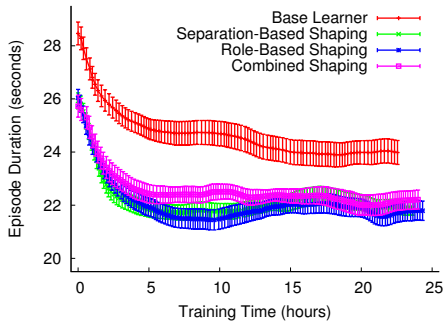

Figure 3: 4v3 at 50x50.

The separation-based heuristic is homogeneous, as all takers receive the same additional reward at all times, but the others are heterogeneous, rewarding different behaviours unique to each taker. The roles assigned are not hard-coded, only encouraged. Therefore, the taker receiving additional positive reinforcement to tackle can still learn to deviate from its assigned role when necessary.

3.1 Results

All graphs presented plot the mean of at least 25 repeats, with the standard error from the mean illustrated by error bars. As we are learning the behaviour of the takers trying to win possession, the episode time the better the agents are performing.

The results shown have been chosen to represent the benefits of shaping in MAS. Both graphs show shaped agents that require less time to reach a stable joint policy than the baseline learner. Furthermore, in Figure 3, we demonstrate an example of where the joint policy learnt has changed due to reward shaping. This time the altered exploration has improved the final performance of the agents but, if the heuristic had been misleading, the opposite can also occur.

Other results, not shown here due to limited space, show similar benefits in all combinations of games of 5v4 and 4v3 on pitches of 40x40 and 50x50.

4. CONCLUSION

In conclusion, providing domain knowledge by an additional potential-based reward to agents affects their exploration. In single-agent reinforcement learning this only affects the time to convergence, but in multi-agent both the time to convergence and final performance can be changed.

Although the potential functions implemented have used domain specific knowledge the types of domain knowledge represented are generally applicable. The knowledge that keepers and takers should try to stay separate is an example of knowledge regarding how agents should maintain states relative to each other. Maintaining a state relative to either team-mates or opponents is a common type of knowledge applicable in many MAS. Similarly, having one tackler and one marker is specific to takers in KeepAway but the knowledge that agents should specialise into roles is also common in MAS.

Furthermore, neither type of knowledge used for reward shaping in our experiments explicitly defines the solutions. Each agent's policy is still learnt by the agent, the knowledge only directs the path exploration takes. Therefore, agents are still free to explore and converge to any equilibrium via self-learning without being limited to a pre-defined solution.

To close, we have demonstrated the benefits of applying potential-based reward shaping functions (both state based [7] and state-action based [3]) when multiple individual learners are acting in a common environment and so, given our recent theoretical guarantees [2], encourage their use in knowledge-based, multi-agent, reinforcement learning when suitable heuristics are known.

5. REFERENCES

[1] M. Babes, E. de Cote, and M. Littman. Social reward shaping in the prisoner's dilemma. In *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems*, volume 3, pages 1389–1392, 2008.

[2] S. Devlin and D. Kudenko. Theoretical considerations of potential-based reward shaping for multi-agent systems. In *Proceedings of The Tenth Annual International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2011.

[3] G. C. Eric Wiewiora and C. Elkan. Principled methods for advising reinforcement learning agents. In *Proceedings of the Twentieth International Conference on Machine Learning*, 2003.

[4] A. Iscen and U. Erogul. A new perspective to the keepaway soccer: the takers. In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems*, volume 3, pages 1341–1344, 2008.

[5] B. Marthi. Automatic shaping and decomposition of reward functions. In *Proceedings of the 24th International Conference on Machine learning*, page 608. ACM, 2007.

[6] H. Min, J. Zeng, J. Chen, and J. Zhu. A Study of Reinforcement Learning in a New Multiagent Domain. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT'08*, volume 2, 2008.

[7] 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.

[8] P. Stone, R. S. Sutton, and G. Kuhlmann. Reinforcement learning for RoboCup-soccer keepaway. *Adaptive Behavior*, 13(3):165–188, 2005.