Enhanced Learning from Multiple Demonstrations with a Flexible Two-level Structure Approach

Doctoral Consortium

Su Zhang
Washington State University
Pullman, Washington
su.zhang2@wsu.edu

ABSTRACT

Learning from demonstration (LfD) has been emerged as a successful transfer learning technique to speed up reinforcement learning (RL). However, the effectiveness of the LfD heavily depends on the quality of the demonstrations. This work investigates how to enable efficient human-agent (or agent-agent) knowledge transfer and allow the RL agent to extract useful information from multiple demonstrations of different quality. In particular, we aim to avoid the effect of noise or bad examples from the collected demonstration data. Inspired by the multi-armed contextual bandit problem and Human Agent Transfer algorithm, we developed a Flexible Two-level Structured Approach to address the above challenges. Evaluated with Mario, Cart Pole and RC Car domains, the experimental results show that this approach holds the promising capacity to successfully leveraging the demonstrations of different quality.

KEYWORDS

Learning from Demonstration; Multi-Armed Bandit; Reinforcement Learning; Transfer Learning

ACM Reference Format:
Su Zhang. 2019. Enhanced Learning from Multiple Demonstrations with a Flexible Two-level Structure Approach. In *Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019,* IFAAMAS, 3 pages.

1 INTRODUCTION

Reinforcement learning [9] (RL) has had many successes in sequential decision tasks, where an agent learns to maximize a real-valued reward. However, learning tabula rasa can be slow, particularly in difficult domains. Learning from demonstration [1] (LfD) is an alternative formulation where an agent typically learns to mimic a human demonstrator. Following this general idea, a rich set of LfD techniques have been developed. For example, one of those techniques is the Human Agent Transfer [11] (HAT) algorithm, which allows the agent to summarize the demonstrated policies with a rule-based learner, the learned rules are transferred to the agent [10] and used in a probabilistic policy reuse [3] (PPR) manner to improve the agent's performance. However, the majority of such techniques are still limited by the demonstrator's performance: an LfD agent's goal is typically to mimic the human, while an RL agent's goal is to maximize total reward.

Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), N. Agmon, M. E. Taylor, E. Elkind, M. Veloso (eds.), May 13–17, 2019, Montreal, Canada.

In this thesis, we are concerned with how to best combine LfD with RL so that we can reap the benefits of fast learning while not being limited by the demonstrator's ability. In particular, we focus on cases of heterogeneous demonstrations. One could consider combining demonstrations from sources with different average performances, or one demonstrator with a high variance in performance. In both cases, we would like to maximize how much we can learn from the demonstrators while minimizing how much poor demonstrations hurt the learner.

In dealing with this problem, several straightforward solutions have been developed, including removing unnecessary or inefficient examples from the demonstration [4], requesting additional clarifications [2], etc. However, those ideas try to eliminate the effects of heterogeneous and noisy demonstrations by removing them instead of trying to extract useful information and learn from them. The challenge of how to effectively learn from heterogeneous demonstrations, making use of the beneficial ones while avoiding the influence of bad ones, still exists.

To address this challenge, we draw inspiration from the multi-armed contextual bandit problem and propose a flexible two-level structure based on the probabilistic policy reuse scheme of HAT: level 0 uses multiple classifiers to summarize demonstrations and level 1 takes advice from the low-level classifiers and combines their opinions with a decision algorithm (e.g., majority voting).

Our experimental results show that the two-level structure can improve the overall performance (total reward) and initial performance (jumpstart), while minimizing the effects of bad demonstrations (relative to the existing HAT algorithm).

2 FLEXIBLE TWO-LEVEL STRUCTURED APPROACH

To address the problem of learning from multiple demonstrations, we propose a Flexible Two-level Structured Approach. This approach integrates with various decision algorithms and the PPR scheme of HAT, allowing the agent to leverage multiple demonstrations of different quality.

2.1 Problem Statements

We currently formalize the problem of learning from multiple demonstrations as a contextual multi-armed bandit problem: With N experts (or classifiers where each is trained on a demonstration), at each time step t, each expert will give out advice ξ based on the context vector x_t and the agent will select upon K actions according to a certain strategy. Such strategy will be adjusted according to the

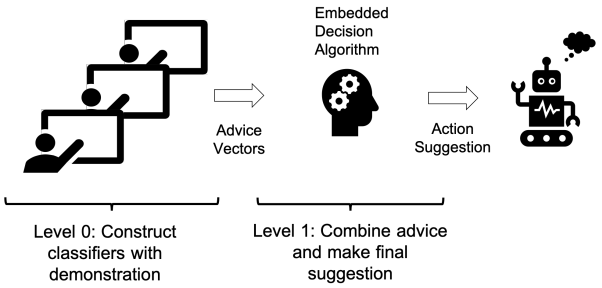

Figure 1: Two-level structure of bandit selection

reward r of the selected action. The goal of the agent is to maximize its reward.

2.2 Methodology

To capture the multi-armed bandit selection, we propose a Flexible Two-level Structured Approach (FTSA) based on the framework of HAT and adopt the Probabilistic Policy Reuse (PPR) [3] scheme to determine whether the agent will accept the suggested actions. With a certain probability ϕ, the agent will follow the recommended actions, otherwise, continue following its value estimation or explore around. The probability ϕ usually less than 1 and decays exponentially over time. The two levels of the FTSA approach will be responsible for extracting policies from the demonstration with classifiers, then work as a decision component to select/combine multiple advice and make the suggestion to the agent. Level 0 will take the provided demonstrations as inputs, use the action transactions as labels and build "expert" classifiers with any supervised learning methods. Each classifier would be trained with only one demonstration, and it will summarize the policies according to the transactions of the demonstration. Those "experts" will then generate advice according to the context while the agent performing its task and interacting with the environment; Level-1 will take the advice from the level-0 classifiers, and the embedded decision component will combine the advice and make a final suggestion to the agent. For example, majority voting, Exponential Weighted Algorithm, or meta-classifier, etc.

2.3 Evaluation

To evaluate the proposed approach, we choose three domains for experiments. First, the **Mario** simulator released by Karakovskiy [5], is based on the Nintendo's game Super Mario Bros. The state is represented as a 27-tuple vector; and the agent can choose from 12 different actions; and the detailed reward settings could refer to [8]. Second, we consider **Cartpole**. Third, we use an **RC Car** simulator.

To investigate whether our approach could learn reasonable performance with multiple demonstrations, we conduct experiments and provide the agent with: 1) good demonstrations, 2) bad demonstrations, and 3) a mixture of demonstrations with different quality. The demonstrations are collected with pre-trained agent — most of the good demonstrations have nearly optimal performance.

For the settings of the approach, we use J48 [7] as the base classifiers in level-0, and test with embedding 1) Majority Vote; 2) Exponential Weighted Algorithm; 3) Meta-classifier (using J48 as level-1 classifier) into level-1.

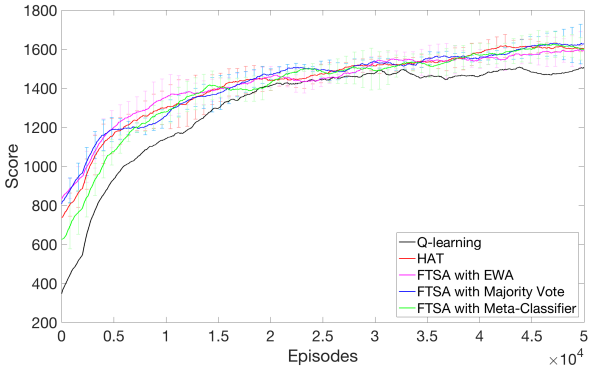

Figure 2: Mario with 10 good demonstrations and 20 bad demonstrations

We adopt the Q-learning agent without any prior knowledge as the baseline and HAT with J48 as benchmark method. And for the evaluation metrics, we consider the *Jumpstart* and the *Total Reward*. Figure 2 shows the performance of using the blending of 10 good and 20 bad demonstrations in the Mario domain.

According to the experimental results, we conclude that this 2-level structured approach could efficiently capture the knowledge from multiple demonstrations, eliminate the effects of bad demonstrations better than the classical HAT approach, and enable the agent to make better usage from the provided demonstrations of different quality, without the assumption of using good or expert level demonstrations only.

3 FUTURE WORK

In future work, we may want to extend this work in three directions: First, we are going to investigate the possibility of combining the confidence-based decision scheme into this approach (e.g., CHAT [12]) to access the reliability of action advice given by the demonstrations. Second, instead of using the PPR scheme of HAT, we want to explore the possibility of introducing "advising reward" [6] to enable the agent to develop a decision policy from sketch. Third, we want to further validate the effectiveness of this approach in additional complex application scenarios, such as online settings, multi-agent systems, etc.

REFERENCES

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

[2] Sonia Chernova and Manuela Veloso. 2007. Confidence-based policy learning from demonstration using gaussian mixture models. In *Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems*. ACM, 233.

[3] Fernando Fernández and Manuela Veloso. 2006. Probabilistic policy reuse in a reinforcement learning agent. In *Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems*. ACM, 720–727.

[4] H Friedrich and R Dillmann. 1995. Obtaining good performance from a bad teacher. In *Workshop: Programming by Demonstration vs Learning from Examples; International Conference on Machine Learning*.

[5] Sergey Karakovskiy and Julian Togelius. 2012. The Mario AI benchmark and competitions. *IEEE Transactions on Computational Intelligence and AI in Games* 4, 1 (2012), 55–67.

[6] Shayegan Omidshafiei, Dong-Ki Kim, Miao Liu, Gerald Tesauro, Matthew Riemer, Christopher Amato, Murray Campbell, and Jonathan P How. 2018. Learning to Teach in Cooperative Multiagent Reinforcement Learning. *arXiv preprint arXiv:1805.07830* (2018).

[7] J Ross Quinlan. 2014. *C4. 5: programs for machine learning*. Elsevier.

[8] Halit Bener Suay, Tim Brys, Matthew E Taylor, and Sonia Chernova. 2016. Learning from demonstration for shaping through inverse reinforcement learning. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 429–437.

[9] Richard S Sutton and Andrew G Barto. 1998. *Reinforcement learning: An introduction*. Vol. 1. MIT press Cambridge.

[10] Matthew E. Taylor and Peter Stone. 2007. Cross-Domain Transfer for Reinforcement Learning. In *Proceedings of the Twenty-Fourth International Conference on Machine Learning (ICML)*.

[11] Matthew E Taylor, Halit Bener Suay, and Sonia Chernova. 2011. Integrating reinforcement learning with human demonstrations of varying ability. In *The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*. International Foundation for Autonomous Agents and Multiagent Systems, 617–624.

[12] Zhaodong Wang and Matthew E Taylor. 2017. Improving reinforcement learning with confidence-based demonstrations. In *Proceedings of the 26th International Conference on Artificial Intelligence (IJCAI)*.