Multi-Agent Path Finding via Reinforcement Learning with Hybrid Reward

Extended Abstract

Cheng Zhao  
Univ. of Sci. & Tech. of China  
Hefei, China  
zc16@mail.ustc.edu.cn

Liansheng Zhuang  
Univ. of Sci. & Tech. of China  
Hefei, China  
lszhuang@ustc.edu.cn

Haonan Liu  
Univ. of Sci. & Tech. of China  
Hefei, China  
phoenix_@mail.ustc.edu.cn

Yihong Huang  
Univ. of Sci. & Tech. of China  
Hefei, China  
hyh1109@mail.ustc.edu.cn

Jian Yang  
Beijing Institute of Technology  
Beijing, China  
yuhengzi_8205@163.com

ABSTRACT

Multi-agent path finding (MAPF) aims at finding a set of conflict-free paths for multiple agents so that each agent can reach its destination while optimizing a global cost. Recently, learning-based methods gain much attention due to their better real-time performance and scalability. However, most existing learning-based methods suffer from poor cooperation among agents since only local observations are used to make decisions. Meanwhile, methods that are bent on team benefits perform poorly due to a lack of individual exploration. To address this problem, this paper proposes a novel Hybrid Reward Path Finding (HRPF), which employs the global information to learn a cooperation mechanism for agents during the training, and embeds it in distributed networks to generate strategies during the execution. HRPF enforces agents to learn strategies from a new type of reward function that decomposes a complex MAPF task into a team task and individual tasks. Experiments on random obstacle grid worlds show that, HRPF performs significantly better in success rate and collision rate than state-of-the-art learning-based methods.

KEYWORDS

Multi-agent path finding; Multi-agent reinforcement learning; Hybrid reward function

ACM Reference Format:


1 INTRODUCTION

The task of multi-agent path finding (MAPF) aims at planning a group of conflict-free and shortest paths for multiple agents [13]. MAPF arises in many real world applications of multi-agent systems such as warehouse robots [4], office robots [15], aircraft-towing vehicles [8]. A key challenge of MAPF is to avoid frequent collisions and blockages, as multiple agents interact with each other. Many search-based methods [2, 3, 11, 12] perform well in finding collision-free solutions using global information in simple environments, but perform poorly in terms of real-time performance and scalability.

Recently, a great amount of work focus on learning-based methods [9] for better real-time performance and scalability. Learning-based methods typically generate one-step policies online for agents based on localised observations. In order to implement distributed planning with effective collaboration, existing learning-based methods usually use multi-agent reinforcement learning (MARL) algorithms in a centralised training with decentralised execution (CTDE) paradigm [1, 6, 10, 16]. The popular MARL methods use team rewards to encourage agents to complete team tasks. However, a MAPF task is complex, with different sub-tasks to be completed by each agent, and the team reward does not guide each agent through the sub-task in detail. As a result, typical MARL methods that do not make full use of individual rewards are less efficient in training and perform poorly in practice.

This paper proposes a novel Hybrid Reward Path Finding (HRPF) for MAPF, which considers both the global team task and individual sub-tasks. HRPF builds on the CTDE paradigm, which inherits the benefits of distributed execution, while enabling agents to use global information to acquire collaborative skills. During the training, HRPF uses a novel mechanism of reward function, hybrid reward, to guide the agents’ behaviour. Agents learn blue-print strategies for completing team tasks from team rewards and learn more refined strategies from individual rewards. During the execution, each agent independently makes decisions, only according to its individual observations, which guarantees the real-time performance and scalability of the method.

2 METHODOLOGY

HRPF defines hybrid reward for MAPF tasks, consisting of the team reward and the individual reward. The individual reward $R^{\text{in}}$ is only visible to each agent itself and is only related to that agent’s observation $o$ and action $a$. It is defined as follows:

$$R^{\text{in}}(o, a) = e^{\text{in}}_g(o, a)R^{\text{in}}_g + e^{\text{in}}_c(o, a)R^{\text{in}}_c + e^{\text{in}}_p(o, a)R^{\text{in}}_p.$$  \hspace{1cm} (1)

$e^{\text{in}}_g$, $e^{\text{in}}_c$, and $e^{\text{in}}_p$ are indicator functions for whether the agent first reaches the target, collides with other agents, and move or stay off the goal, respectively. $R^{\text{in}}_g$, $R^{\text{in}}_c$, $R^{\text{in}}_p$ are reward values for the agent performing these actions. The team reward $R^\text{te}$ is shared by the
whole team and is related to the joint observation $o$ and the joint action $a$. It is defined as follows:

$$R^i_n(o,a) = \epsilon^i_n(o,a) R^{i,n} + n^i_n(o,a) R^{i,e} + n^i_n(e) R^{i,e}.$$  \hfill (2)

The first term of RHS indicates the final reward when the task is completed, where $\epsilon^i_n$ is the indicator function for whether the task is completed. $n^i_n$ denote the number of agents that first reach the destination and the number of agents that collide at the current time step, respectively. $R^{i,n}$, $R^{i,e}$ are reward values.

HRPF trains two different Q-value networks for each agent, individual Q-network and team Q-network, based on DQN [7]. The input embedding and network structure of the Q-network is similar to that in PRIMAL [10], as shown in Fig. 1.

$$L^{n} = [(R^{i,n} + \gamma \max_{a_i^{t+1}} \sum_{i} \hat{Q}^i_{t+1}(o_i^{t+1}, a_i^{t+1}) - \sum_{i} \hat{Q}^i_{t}(o_i^t, a_i^t)]]^2. \hfill (4)$$

During the execution, each agent combines the Q-values of the two networks to get a new type of value, denoted $Q^{i,comb}$. We balance the weight of the individual Q-value and the team Q-value by a parameter $\beta$. The formula for $Q^{i,comb}$ is:

$$Q^{i,comb}_{t}(o_i^t, a_i^t) = \frac{1}{1+\beta} \hat{Q}^i_{t}(o_i^t, a_i^t) + \frac{\beta}{1+\beta} \hat{Q}^{i,e}_{t}(o_i^t, a_i^t). \hfill (5)$$

Typically, the individual Q-value and the team Q-value should be weighted approximately equally, so in practice $\beta$ is usually on the same scale as the number of agents $k$, as the total team Q-value is decomposed into $k$ components for each agent. Next, each agent generates an $\epsilon$-greedy policy based on $Q^{i,comb}$.

### 3 EXPERIMENT

We conduct experiments on random obstacle grid worlds with reference to the setup in PICO [5]. To test the performance under different settings, the number of agents varies in three cases: 8, 16, 32, and the density of obstacles varies in four cases: 0%, 10%, 20%, and 30%. The performance is evaluated on 6 different measurements compared with two baselines, PRIMAL [10] and PICO [5].

To further analyze the effectiveness of hybrid reward, we conduct ablation studies on the reward function. We propose two methods, IRPF and TRPF, with the team reward and the individual reward are removed respectively. Fig. 2 shows the result on grid worlds with an obstacle density of 20% and a number of agents of 8.

**Figure 1: Network structure of HRPF.**

**Figure 2: Ablation experiment result.**
REFERENCES


