Engineering Applications of Artificial Intelligence 58 (2017) 111–122

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

The dynamics of reinforcement social learning in networked cooperative multiagent systems

Jianye Hao[a], Dongping Huang[b], Yi Cai[b,*], Ho-fung Leung[c]

[a] School of Software, Tianjin University, China
[b] South China University of Technology, China
[c] The Chinese University of Hong Kong, Hong Kong

ARTICLE INFO

Keywords:
Multiagent social learning
Multiagent coordination
Cooperative games

ABSTRACT

Multiagent coordination in cooperative multiagent systems, as one of the fundamental problems in multiagent systems, and has been widely studied in the literature. In real environments, the interactions among agents are usually sparse and regulated by their underlying network structure, which, however, has received relatively few attentions in previous work. To this end, we firstly systematically investigate the multiagent coordination problems in cooperative environments under the *networked social learning* framework under four representative topologies. A networked social learning framework consists of a population of agents where each agent interacts with another agent randomly selected from its neighborhood in each round. Each agent updates its learning policy through repeated interactions with its neighbors via both individual learning and social learning. It is not clear a priori whether all agents are able to learn towards a consistent optimal coordination policy. Two types of learners are proposed: *individual action learner* and *joint action learner*. We evaluate the learning performances of both learners extensively in different cooperative (both single-stage and Markov) games. Besides, the influence of different factors (network topologies, different types of games, different topology parameters) is investigated and analyzed and new insights are obtained.

1. Introduction

One fundamental property of an agent in Multiagent Systems (MASs) is its ability of adaptively adjusting its behaviors in response to other agents in order to achieve effective coordination on desirable outcomes. Recent years have witnessed significant efforts in researching on the coordination problem within cooperative MASs (Claus and Boutilier, 1998; Matignon et al., 2012). In cooperative MASs, the agents share common interests (e.g., the same reward function), thus the increase in individual's benefit also leads to the increase of the benefits of the whole group.

A number of challenges exist for them to overcome when the agents learn in cooperative multiagent environments. First, one major difficulty is the *equilibrium selection problem* (Fulda and Ventura, 2007), i.e., multiple optimal joint actions exist under which the agents needs coordinated behaviors to reach a consistent optimal joint action among multiple optimal ones. Second, another further challenging issue is the *Pareto selection problem* (Fulda and Ventura, 2007), in which there exist multiple Nash equilibria and also some of them are Pareto-dominated by the rest. The challenging question is how to make sure that the agents would effectively coordinate on one of the Pareto-optimal equilibria. Third, we also need to tackle the *stochasticity problem* (Matignon et al., 2012), i.e., the game itself can be non-deterministic. In this case, the difficulty is how the agents can distinguish whether the different payoffs received by selecting the same action come from the explorative behaviors of other agents or the stochasticity of the game (environment) itself.

A number of different multiagent reinforcement learning algorithms (Claus and Boutilier, 1998; Lauer and Riedmiller, 2000; Kapetanakis and Kudenko, 2002; Wang and Sandholm, 2002; Brafman and Tennenholtz, 2004; Matignon et al., 2012; Modeling and Thomas, 2015) have been proposed in the literature to handle the coordination issue in cooperative MASs. The most commonly adopted learning framework is the fixed players repeated interaction framework, in which two (or more) agents learn their optimal coordination policies through repeated interactions with the same opponent(s) (Matignon et al., 2012). Their pairwise interactions can usually be modeled as a repeated normal-form (or Markov) game. However, in real distributed environments, the chance that an agent always interacts with the same agent is quite small, and it is very likely that the interacting partners of an agent vary frequently. Due to diversity of different interaction partners, agent's optimal coordination policy

* Corresponding author.
E-mail addresses: jianye.hao@tju.edu.cn (J. Hao), huang.dp@scut.edu.cn (D. Huang), ycai@scut.edu.cn (Y. Cai), lhf@cuhk.edu.hk (H.-f. Leung).

http://dx.doi.org/10.1016/j.engappai.2016.11.008
Received 1 November 2015; Received in revised form 17 September 2016; Accepted 22 November 2016

towards one partner may become suboptimal to another partner. Therefore, the feature of non-fixed partner interaction adds additional dimension of difficulty to achieve effective coordinations in cooperative MASs.

Hao and Leung proposed a multiagent social learning framework to investigate multiagent coordination problem in cooperative games assuming that the agents' interactions are random. However, the interaction pattern among agents varies depending on the specific applications. For example, considering the ad-hoc networks (Jhaveri and Patel, 2015) or distributed multi-robot coordination problem, it might be reasonable to model the interaction among agents as random interaction. In contrast, for other systems such as distributed sensor network (Aldosari and Moura, 2006), the interaction among sensors is not random, and is determined by the system's underlying topology, which might have significant influence on the system's overall performance. It is not clear a priori if and how the agents are able to eventually coordinate on optimal solutions under such a networked social learning framework and whether different topologies can have predominant impact on the coordination performance among agents. Another related question is what kind of impact that different topology parameters could have on the learning performance of agents in different cooperative environments.

To this end, in this work, we propose a generic networked social learning[1] framework to investigate the multiagent coordination problem in cooperative MASs by explicitly modeling different network topologies. In this framework, each agent learns its policy through repeated interactions with its neighboring agents in the system. We consider a number of representative network topologies: ring network, small-world network and scale-free network. During each round each agent interacts with one of its neighbors randomly, and the interactions between each pair of agents are modeled as two-player cooperative Markov games. If no underlying topology exists, then one agent is randomly selected as its partner from the population.[2] Each agent learns its policy concurrently over repeated interactions with randomly selected partners from its neighborhood. Besides, apart from learning from its own experience, each agent may also learn from the experience of its neighbors.

We distinguish two different types of learning environments within the *networked social learning framework* depending on the amount of information available to the agents, and propose two types of learners accordingly: individual action learners (IALs) and joint action learners (JALs). IALs learn the values of each individual action directly by viewing their neighbors as part of the environment, while JALs learn the values of each action indirectly based on the learned values of the joint actions together with their partners. Both IALs and JALs employ the optimistic assumption and the FMQ heuristic to utilize the learning experience of their own and their neighbors obtained based on the observation mechanism. We extensively evaluate the learning performances of both types of learners in different types of (both single-stage and Markov) cooperative games with different topologies. Through the experimental results and analysis, new insights are also obtained regarding the impact of different factors (i.e., different topology parameters, different topology, different learners (IAL/JAL) and different game structures) on the learning performance of agents.

The structure of the paper is presented as follows. In Section 2, we introduce previous work related with coordination in cooperative MASs. In Section 3, we define the networked social learning framework and describe both IALs and JALs. The evaluation results of both types of learners in different cooperative games and the influence of different factors are presented in Section 4. Lastly we conclude with some remarks on future research directions in Section 5.

2. Related work

Coordination in cooperative MASs, as a fundamental issue in MASs, has received wide attention from the multiagent learning community. Usually a cooperative multiagent environment is modeled as a two-player cooperative repeated (or stochastic) game. In the work of Claus and Boutilier (1998), two different types of learners (without optimistic exploration) are distinguished based on the traditional Q-learning algorithm: independent learner (IL) and joint-action learner (JAL), and investigate their performance in the context of two-agent repeated cooperative games. Empirical results show that both types of learners can successfully coordinate on the optimal joint actions in simple cooperative games without significant performance difference. However, both types of learners fail when the domain becomes more complex. Specifically the authors consider two more complicated types of games: the climbing game (see Fig. 1a) and the penalty game (see Fig. 1b) (Claus and Boutilier, 1998). The former game models the coordination problems of a single optimal joint action with high mis-coordination penalty, while the latter game represents the cases of multiple optimal joint actions with high mis-coordination penalty. For the climbing game, the high penalty induced by achieving either (b, a) or (a, b) can make the agents find action a very unattractive, which thus usually results in convergence to the suboptimal outcome (b, b). This is also known as the Pareto-selection problem (Fulda and Ventura, 2007). For the penalty game, apart from the effect of high penalty induced when achieving (c, a) or (a, c) if the value of k is very small (e.g., k=−30), the coexistence of two optimal joint actions ((a, a) and (c, c)) makes the task of coordination on one of them more challenging, which is also known as the *shadowed equilibrium selection problem* (Matignon et al., 2012).

A number of improved learning algorithms have been proposed afterwards. In general, the mis-coordination problems in the above two types of games can be handled from two different perspectives: altering the Q-function update strategy (Kapetanakis and Kudenko, 2002; Lauer and Riedmiller, 2000) and altering the policy selection strategy (Lauer and Riedmiller, 2000). Lauer and Riedmiller (2000) propose a coordination algorithm based on the optimistic assumption under which each agent only takes into consideration the maximum payoff of each action when updating its strategy. The optimistic assumption ensures that the agents can eventually update their Q-values of their individual actions to their corresponding maximum payoffs and thus can eliminate the effect of penalty in the previously two types of games. Besides, to ensure that the agents can coordinate on the same optimal joint action, the authors also propose a modified policy selection strategy under which the agents always choose the corresponding action of the first mutually encountered optimal joint action as their policies. It is proved that the agents can be guaranteed to converge to optimal joint actions in two-player repeated cooperative games with deterministic payoffs; however, it fails when dealing with stochastic environments.

Kapetanakis and Kudenko (2002) propose the FMQ heuristic to alter the Q-value estimation function to handle the stochasticity of the games. Under FMQ heuristic, the original Q-value for each individual action is modified by incorporating the additional information of how frequent the action receives its corresponding maximum payoff. Experimental results show that FMQ agents can successfully coordinate on an optimal joint action in partially stochastic climbing games, which is a significant improvement compared with the original Q-learning approach based on the optimal assumption (it fails on partially stochastic climbing games). However the FMQ heuristic based Q-learning fails when it comes to fully stochastic climbing games (e.g., Fig. 3b). An improved version of FMQ (recursive FMQ) is proposed in Matignon et al. (2008). The major difference with the original FMQ is

[1] It's worth mentioning that there also exists a huge body of literature from the area of economics and social science (e.g., Leonard et al., 2012; Jadbabaie et al., 2012), in which the definition of *social learning* is quite different from what we use here in multiagent systems area. This work extends our previous paper (Hao et al., 2014).

[2] Note that in this case, the framework itself reduces to be equivalent with the one used in Hao et al. (2014).