Emotional multiagent reinforcement learning in spatial social dilemmas

Jasper Busschers, Chuhan Shao, Matthias Stübinger and Meihan Wang

Vrije Universiteit Brussel, Pleinlaan 2, Brussel
jasper.busschers@vub.be,
chuhan.shao@vub.be,
mstubing@vub.ac.be,
meihan.wang@vub.be

Abstract

The paper Emotional multiagent reinforcement learning proposes a new algorithm to solve multiagent reinforcement learning problems. Whereas other approaches such cooperative Q-learning introduce some form of communication to reach cooperative behaviour. This paper does not require communication but rather proposes 2 models for derivation of the agents emotions and makes decisions based on those emotions. It uses the weighted policy learned algorithm in order to learn which emotion derivation function is best to use (Abdallah and Lesser, 2006).

Other popular approaches to solving multiagent problems such as imitation learning and inverse reinforcement learning also exist. Here the goal is to copy desired behaviour in multiagent systems. These approaches consider cooperating or defecting, just as another part in copying complex behaviour Song et al. (2018).

Introduction

In terms of cooperation behaviour, it is counter-intuitive to observe that this behaviour takes charge of the populations, although it is more reasonable for the altruism to dominate the whole population. After doing related simulation experiments, cooperation behaviour can achieve more benefits than the altruism can. In this research area, it is crucial to understand how agents **achieve cooperation** in social dilemmas by *learning from local experience*. The paper (Yu et al., 2015) mainly discusses how the emotion derivation functions behave under three different types of networks. A variety of metaphors can be adopted to study social dilemmas among self-interested agents, in which Prisoners Dilemma (PD) is the most widely used one. Besides, the paper investigates the possibility of exploiting emotions to modify agent learning behaviours in MARL in order to facilitate cooperation in social dilemmas. In addition, the paper (Yu et al., 2015) provides a psychological explanation of cooperative behaviours in human societies from the theoretical perspective and a computationally sound model that incorporates emotions into MARL in order to solve social dilemmas from the technical perspective.

In this report, we first introduce several concepts related to the topic. Then, the implementation methods of the algorithm will be detailed explained, where we show our thoughts and comments on the algorithm. Finally, the results from our experiments and the discusses will be listed.

Materials and Methods

To implement the algorithm, some fundamental concepts should be introduced first. In this section, three types networks for representing *a spatial social dilemma* are defined first. Secondly, *MARL* and *the emotional MARL framework* are illustrated. At last, *emotional MARL in spatial social dilemmas* will be presented.

Spatial social dilemmas

In (Nowak and May, 1992), a spatial social dilemma can be represented as a graph $G = (V, E)$, in which the vertices V are the agents and the edges E are the connections between agents, where the social dilemma game occurs. In addition, we have to define the neighbours of an agent due to the concept used in (Yu et al., 2015). For a spatial social dilemma G, the neighbours of agent i, $N(i)$, are *a set of agents* such that $N(i) = \{v_j | (v_i, v_j) \in E\}$ and $N(i) \subset V$. As we define the spatial social dilemmas in abstract level, the implementation of that requires specific networks. Thus, we then introduce three different types of networks.

Grid networks. A grid network is a computer network consisting of a number of systems connected in a grid topology. In a regular grid topology, each node in the network is connected with two neighbours along one or more dimensions (Morris et al., 2000). In our case, we consider two-dimension grid networks. an important character of this network is that an agent has two neighbours in the corner, three neighbours at the edge of the network (not corner), and four neighbours in the inner of the network. In other words, the number of the neighbours is fixed, which depends on the position of the agent. In (Yu et al., 2015), GR_N is denoted for an N-node grid network.

Small-world networks. This network is derived from *the small-world phenomenon* in reality. An obvious feature of small-world networks is that each node has only a small number of neighbours and yet can reach any other node in a

small number of hops (Yu et al., 2015). Furthermore, it features *a high clustering coefficient* and *a short average path length*. They use $SW_N^{\kappa,\rho}$ to represent this network, in which κ is the average size of the neighbourhood, and $\rho \in [0, 1]$ shows the different orders of network randomness.

Scale-free networks. A scale-free network is a connected graph with the property that the number of links k originating from a given node exhibits a power law distribution $P(\kappa) \sim \kappa^{-\lambda}$ (Barabási and Bonabeau, 2003). The network exhibits the feature of *preferential attachment*, which means that the likelihood of connecting to a node depends on the connectivity degree of this node (Yu et al., 2015). $SF_N^{\kappa,\lambda}$ is the denotation of the scale-free network here, where κ and λ are coefficients in the probability $\kappa^{-\lambda}$.

MARL

As we know in RL, basic reinforcement is modelled as a *Markov decision process* (MDP). It is formally defined as the following components (Liu, 2008)

- a set of environment and agent states, S;
- a set of actions, A, of the agent;
- $P_a(s, s') = Pr(s_{t+1} = s'|s_t = s, a_t = a)$ is the probability of transition from state s to states' under action a.
- $R_a(s, s')$ is the immediate reward after transition from s to s' with action a;
- rules that describe what the agent observes.

In this process, we focus on the concept of the *policy* $\pi : S \times A \rightarrow [0, 1]$. It is a probability distribution that maps a state $s \in S$ to an action $a \in A$ of the agent. Consequently, the policy π should be well learnt to achieve the goal to maximize the expected discounted reward $Q^\pi(s, a)$ for each state $s \in S$ and $a \in A$. Hence, the largest expectation of policy π for each agent indicates the best expected *discounted* reward. Equation 1 shows the strategy above, where $\gamma \in [0, 1)$ is a *discount* factor.

$$Q^\pi(s, a) = E_\pi \left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) | s_0 = s, a_0 = a \right] \quad (1)$$

We can always find at least one optimal policy π^* such that $Q^{\pi^*}(s, a) \geq Q^\pi(s, a)$, which means that the expected discounted reward under π^* is greater than that under any other else. A widely used RL approach is Q-learning, where a set of Q values is the core for agents to make a decision. The update rule for Q values is shown in Equation 2. This equation is for updating in a single step and $\alpha \in (0, 1]$ is the learning rate and γ is the discount factor (Russell and Norvig, 2016).

$$Q_{t+1}(s, a) = Q_t(s, a) + \alpha_t [R(s, a) + \gamma \max_{a'} Q_t(s', a') - Q_t(s, a)] \quad (2)$$

However, each agent makes a decision on the environment, and the state transitions and rewards are the result of the joint action of all agents (Yu et al., 2015). There are two solutions but they do not perform well. It is impractical for agents to make decision on the comprehensive information from the whole agents, while the learning environment will be a non-stationary if each agent learn individually.

On the other hand, the paper (Yu et al., 2015) limits itself by only focusing on static learning problems such as the prisoners dilemma. Here the rewards of cooperating and defecting can be instantly computed. Whereas in many reinforcement learning problems with different states and delayed rewards, this is not so trivial.

Emotional MARL Framework

The paper proposes an emotional framework existing of 2 parts, the inner layer learning and outer layer learning. In the outer layer learning, the weighted policy learning algorithm is used to learn which emotion derivation function to use.

The emotion derivation functions make use of 2 computed variables; the social fairness and individual well-being. The next section will cover how these variables are being computed.

Then the inner layer learning learns a Q function to choose an action based on the derived emotion.

Appraisal of Emotions

The social fairness and individual well-being parameters were added to derivate an agents current emotions.

In order to compute the social fairness, first we must compute the cooperation level of the group. This is done by using the Equation 3.

$$C = 1/N \sum_{i=1}^{i=N} ((n_c^i - n_d^i)/M) \quad (3)$$

where n_c^i and n_d^i are the number of times neighbour i cooperated and defected over a number of M games. The environment is called cooperative, whenever $C > 0$ and else uncooperative. Using this value, it is possible to compute the social fairness of the agent by comparing its own actions to those of the group.

$$F = C * ((n_c - n_d)/M) \quad (4)$$

where n_c and n_d are the number of times cooperated or defected by the current agent over M games.

For the individual well-being three different approaches are presented to computing this. The first being the absolute value approach given by Equation 5:

$$W = \frac{2R_t - M \times (T - S)}{M \times (T - S)} \tag{5}$$

with R_t being the cumulative reward received by the agent at episode t.

The second approach is called the variance-based approach and can be computed by Equation 6:

$$W = \frac{R_{t+1} - R_t}{M * (T - S)} \tag{6}$$

And at last the aspiration-based approach is given by Equation 7:

$$W = tanh[h(\frac{R_t}{M} - A_t) \tag{7}$$

where $tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$ is the monotonically increasing hyperbolic tangent function and A_t is the aspiration level that can be computed as shown in Equation 8:

$$A_{t+1} = (1 - \beta)A_t + \beta\frac{R_t}{M} \tag{8}$$

One remark on the way these values are computed is how they rely on the predefined reward matrix. There is a limitation when trying to generalize this to MDP. Doing this requires some expected rewards for cooperating and defecting over longer sequences of actions.

Emotion derivation model

The paper presents two different derivation functions based on the same model. These functions take in as input the social fairness and well being and should assign 1 of 4 possible emotions to the agent.

The formal definition for the emotion derivation model is given by Equation 9.

$$E_x(c, s) = f(D_x) * g(I_x) \tag{9}$$

where C is the core appraisal variable and S is the secondary appraisal variable. D_x is the desirability of emotion x and I_x the intensity of that emotion. The values of which will be defined for two different derivation strategies. The functions $f(d_x)$ and $g(I_x)$ are defined as shown in Equation 10 and 11.

$$f(D_x) = D_x \tag{10}$$

$$g(I_x) = (1 + I_x)/2 \tag{11}$$

Using this formal definition, two different derivation functions can be defined, one taking the social well fair as core variable (FW) and one taking the individual well being as core variable (WF). Whereas the first should offer a more fair agent, the WF function should lead to a more greedy policy.

Figure 1 illustrates how emotions are determined in these different derivation functions.

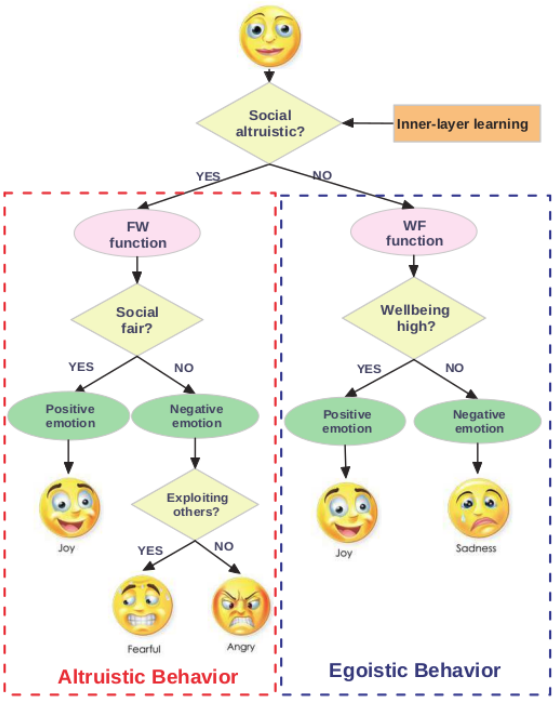


Figure 1: Emotion derivation tree. (Yu et al., 2015)

One limitation with such approach is how it only models a small part of emotions and also how it relies on predefined conditions in order to derivate emotions. It could be argued that implicitly learning these derivation functions would benefit the generalisation of such approach.

Results and Discuss

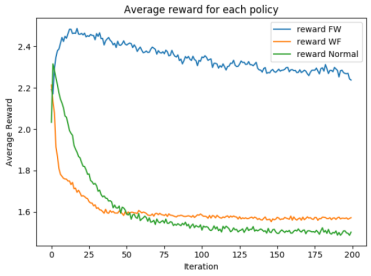
We use Table 1 to represent a social dilemma game. The Watts-Strogatz model (Watts and Strogatz, 1998) is used to generate a small-world network, and the Barabasi-Albert model (Barabási and Bonabeau, 2003) is used to generate a scale-free network. To use the Barabasi-Albert model, we start with 5 agents and add a new agent with 1 edge to the network at every time step. Learning rate $\alpha = 0.5$ and discount factor $\gamma = 0$. Scale-able parameter h in the aspiration-based approach is 10 and learning rate to update aspiration level β is 0.5. We choose linear functions $f(D_x) = D_x$ and $g(I_x) = (1 + I_x)/2$ to map the value of D_x and I_x to [0, 1]. All results are averaged over 100 independent runs.

Table 1: Payoff matrix of the Prisoners' Dilemma game

	Cooperate (C)	Defect (D)
Cooperate (C)	R(= 3),R(= 3)	S(= 0),T(= 5)
Defect (D)	T(= 5),S(= 0)	P(= 1),P(= 1)

Figure 2 shows the average rewards generating from different appraisal approaches and different networks. For the normal agents, we assume that they play games without

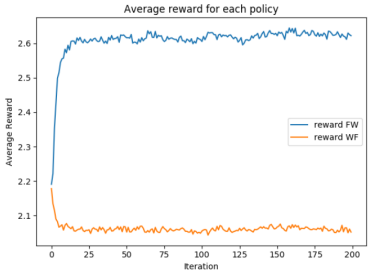
emotion parameter. From the results, we can see that different kinds of agents can learn distinct behaviour and FW agents show the best performance in all the different combination of variables. These figures are corresponding to the outcomes of the paper (Yu et al., 2013), which confirms the conclusion that both social fairness and individual wellbeing are fundamental factors in the appraisal of emotions, but social fairness must be considered to be the core appraisal variable to guarantee a high level of cooperation.



(a) Absolute-value & Scale Free

(b) Aspiration-based & Scale Free

(c) Absolute-value & Small World



(d) Aspiration-based & Small World

Figure 2: Impact for different kinds of agents using different appraisal approaches and networks

Figure 3 shows the impact of frequencies (M) to update emotions. It is obvious that the most appropriate value for M is 3 in such described condition. A higher updating frequency (a smaller number of interaction steps M) makes the social dilemma environment more complex and flexible. Eventually it slows down the emergence of cooperation, while a lower updating frequency (a larger number of interaction steps M) makes the agents torpid to the current learning situation and lead to an incomplete convergence of cooperation.

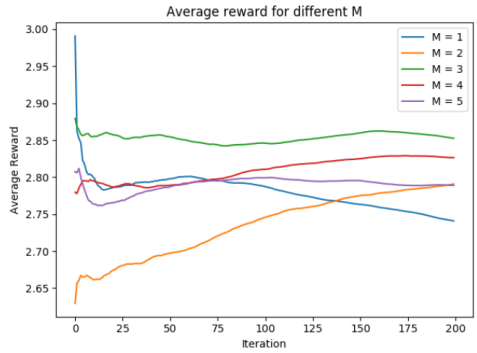


Figure 3: Impact for different values of M using Aspiration-based appraisal approach and small-world network.

From the results of Figure 4, FW Agents can learn to achieve cooperation using different approaches of wellbeing appraisal approaches and Aspiration-based algorithm outperforms other two methods.

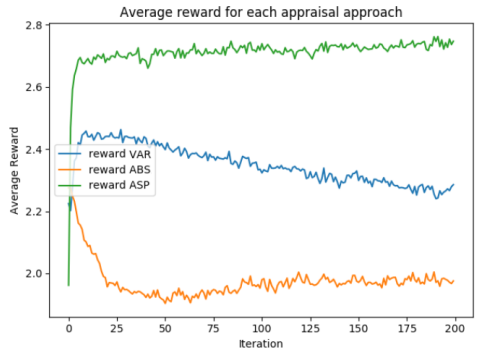


Figure 4: Impact for different kinds of wellbeing appraisal approaches using FW policy and scale-free network.

Figure 5 and Figure 6 show how the emotion appraisal variables change with learning episodes. Figure 5 gives the values of all FW Agents averaged in 100 runs while Figure 6 shows the learning dynamics in one FW agent in one particular run. F increasing means that more and more agents choose to cooperate with each other. It works with E in the same way. The value of M will finally converge to 0 because

the difference between average reward and cooperation reward is normalized by $T - S$.

Figure 5: Dynamics of appraisal variables for all agents using Aspiration-based appraisal approach and scale-free network.

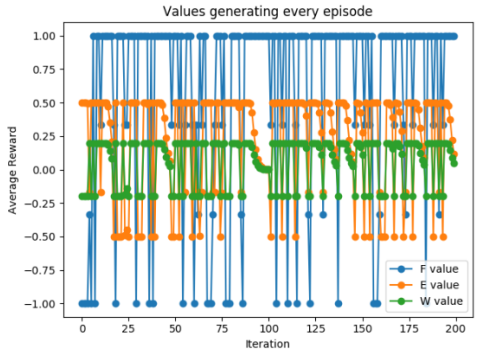


Figure 6: Dynamics of appraisal variables for one agent using Aspiration-based appraisal approach and scale-free network.

Algorithm

The algorithm used in this section is listed in **Algorithm 1**. It consists of an outer and inner learning loop. The inner learning loop is used to learn every agent what emotion derivation function is best to use, for this the weighted policy learning algorithm was used in the paper. However little clarification or reasoning behind this choice was given.
After the agent choose its derivation function it gets into the outer layer learning loop, where it will pick actions based on the chosen derivation function. This was implemented by giving each agent a Q-matrix containing values for 2 actions (cooperate or defect) for every emotion.

Algorithm 1 The Interaction protocol

Initialize network and learning parameters;
for each learning episode t $(t = 1, \ldots, T)$ **do**
  for each agent i $(i = 1, \ldots, N)$ **do**
    if $e = FW$ **then**
      Chooses action a_i based on $Q_i^{FW}(a)$;
    else
      Chooses action a_i based on $Q_i^{WF}(a)$;
  for each agent i $(i = 1, \ldots, N)$ **do**
    for each neighbour $j \in N(i)$ **do**
      Plays action a_i with agent j;
      Receives $r_{i,j}$ after interacting with agent j;
    Calculates its sense of social fairness;
    Calculates its individual wellbeing;
    Calculates R_{int} using selected function e;
    Updates $Q_i^e(a)$ using intrinsic reward R_{int}.

Conclusion

From many different perspectives, it would be beneficial to model the complex working of emotions on decision making. However the model that this paper provides is a very simplified approach to doing this, which relies a lot on predefined variables and conditions to determine emotions. It also only performs experiments using the prisoners dilemma, which leaves the question open how the approach generalises to different problems.
The paper shows that this simple representation reaches better results on the prisoners dilemma than in a rational agent, which may indicate that including emotions in decision making can benefit the overall performance. However it seems far off from modelling the complex influence of emotions on the decision making process. I would argue that in real emotional models, the emotions should be implicitly learned by correlating bad and good past experiences with each other. Something this paper neglects by using a predefined model.

References

Abdallah, S. and Lesser, V. (2006). Learning the task allocation game. ACM.

Barabási, A.-L. and Bonabeau, E. (2003). Scale-free networks. *Scientific american*, 288(5):60–69.

Liu, J. S. (2008). *Monte Carlo strategies in scientific computing*. Springer Science & Business Media.

Morris, R., Jannotti, J., Kaashoek, F., Li, J., and Decouto, D. (2000). Carnet: A scalable ad hoc wireless network system. In *Proceedings of the 9th workshop on ACM SIGOPS European workshop: beyond the PC: new challenges for the operating system*, pages 61–65. ACM.

Nowak, M. A. and May, R. M. (1992). Evolutionary games and spatial chaos. *Nature*, 359(6398):826.

Russell, S. J. and Norvig, P. (2016). *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,.

Song, J., Ren, H., Sadigh, D., and Ermon, S. (2018). Multi-agent generative adversarial imitation learning. *CoRR*.

Watts, D. and Strogatz, S. (1998). Collective dynamics of small-world networks. *Nature*, 393(6684):440–442.

Yu, C., Zhang, M., and Ren, F. (2013). Emotional multiagent reinforcement learning in social dilemmas. In *International Conference on Principles and Practice of Multi-Agent Systems*, pages 372–387. Springer.

Yu, C., Zhang, M., Ren, F., and Tan, G. (2015). Emotional multiagent reinforcement learning in spatial social dilemmas. *IEEE transactions on neural networks and learning systems*, 26(12):3083–3096.