Chapter

Improvement of Cooperative Action for Multi-Agent System by Rewards Distribution

Mengchun Xie

Abstract

The frequency of natural disasters is increasing everywhere in the world, which is a major impediment to sustainable development. One important issue for the international community is to reduce vulnerability to and damage from disasters. In addition, a large number of injuries occur simultaneously in a large-scale disaster, and the condition of the injured will change over time. Efficient rescue activities are carried out using triage to determine the priority of injury treatment based on the severity of the persons' conditions. In this chapter, we discuss acquiring cooperative behavior of rescuing the injured and clearing obstacles according to triage of the injured in a multi-agent system. We propose three methods of reward distribution: (1) reward distribution responding to the condition of the injured, (2) reward distribution based on the contribution degree, and (3) reward distribution by the contribution degree responding to the condition of the injured. We investigated the effectiveness of the three proposed methods for a disaster relief problem by an experiment. The results of the experiment showed that agents gained high rewards by rescuing those in most urgent need under the method having the reward distributed according to the contribution degree responding to the condition of the injured.

Keywords: multi-agent system, reinforcement learning, reward distribution, triage, disaster relief problem

1. Introduction

The frequency of natural disasters is increasing everywhere in the world, which is a major impediment to sustainable development. In order to minimize the damage of disasters, the United Nations Office for Disaster Risk Reduction (UNISDR) calls for the promotion of disaster prevention and mitigation by local governments in each country. This is an important issue for the international community in order to reduce vulnerability to and damage from disasters.

In the case of a large-scale disaster, a large number of injuries occur simultaneously, and the condition of the injured changes with the lapse of time. This implies that, to conduct efficient treatment when resources are insufficient to immediately treat all the people who are injured, it is necessary to use triage, which is the process of determining the priority of treatment based on the severity of the injured person's condition [1].

To date, many different remote-controlled disaster relief robots have been developed. A further complication, besides the need for triage, is that these robots must work in environments in which communication is not always secure. For these reasons, there is a need for autonomous disaster relief robots, that is, robots which can learn from the conditions that they encounter and then take independent action [2]. Thus, efficient rescue needs to consider the condition of the injured, which changes with the lapse of time, even with the use of disaster rescue robots.

Reinforcement learning is one way that robots can acquire information about appropriate behavior in new environments. Under this learning system, robots can observe the environment, select and perform actions, and obtain rewards [3–6]. Each robot must learn what the best policy (i.e., the policy that obtains the largest amount of reward over time) is by itself.

Recent research on disaster relief robots has included consideration of multi-agent systems, that is, systems that include two or more disaster relief robots. A multi-agent system in which multiple agents explore sections of damaged building with the goal of updating a topological map of the building with safe routes is discussed [7, 8]. John et al. constructed a multi-agent systems approach to disaster situation management, which is a complex multidimensional process involving a large number of mobile interoperating agents [9]. However, to successfully interact in the real world, agents must be able to reason about their interactions with heterogeneous agents of widely varying properties and capabilities. It is necessary that agents are able to learn from the environment and implement independent actions by using perceptual and reasoning in order to carry out their task in the best possible way [10, 11].

Numerous studies regarding learning in multi-agent systems have been conducted. Spychalski and Arendt proposed a methodology for implementing machine learning capability in multi-agent systems for aided design of selected control systems allowed to improve their performance by reducing the time spent processing requests that were previously acknowledged and stored in the learning module [12]. In [13], a new kind of multi-agent reinforcement learning algorithm, called TM_Qlearning, which combines traditional Q-learning with observation-based teammate modeling techniques, was proposed. Two multi-agent reinforcement learning methods, both consisting of promoting the selection of actions so that the chosen action not only relies on the present experience but also on an estimation of possible future ones, have been proposed to better solve the coordination problem and the exploration/exploitation dilemma in the case of nonstationary environments [14]. In [15], the construction of a multi-agent evacuation guidance simulation that consists of evacuee agents and instruction agents was reported, and the optimum evacuation guidance method was discussed through numerical simulations by using the multi-agent system for post-earthquake tsunami events. A simulation of a disaster relief problem that included multiple autonomous robots working as a multi-agent system has been reported [16].

In disaster relief problems, it is important to rescue the injured and remove obstacles according to conditions that are changing with the passage of time. However, conventional research on multiple agents targeted for disaster relief has not taken into consideration the condition of the injured, so it is insufficient for efficient rescue.

In this chapter, we discuss acquiring cooperative behavior of rescuing the injured and clearing obstacles according to triage of the injured in a multi-agent system. We propose three methods of reward distribution: (1) reward distribution responding to the condition of the injured, (2) reward distribution based on the contribution degree, and (3) reward distribution by the contribution degree responding to the condition of the injured. We investigated the effectiveness of

these proposed methods for a disaster relief problem by an experiment. The results of the experiment showed that agents gained high rewards by rescuing those in most urgent need under the method having the reward distributed according to the contribution degree responding to the condition of the injured.

2. Learning of multi-agent systems and representation of disaster relief problem

2.1 Learning of multi-agent systems

Agents are a computational mechanism that exist in some complex environment, sense and perform actions in its environment, and by doing so realize a set of tasks for which it is assigned. A multi-agent system consists of agents that interact with each other, situated in a common environment, which they perceive with sensors and upon which they act with actuators (**Figure 1**). Agent and environment are relationships of the interaction. In the meantime, when the environments are inaccessible, the information which can be perceived from the environment is limited, and it is inaccurate, and it entails delay [2, 17, 18].

In [19], the following major characteristics of multi-agent systems were identified:

- Each agent has incomplete information and is restricted in its capabilities.
- The control of the system is distributed.
- The data are decentralized.
- The computation is asynchronous.

In multi-agent systems, individual agents are forced to engage with other agents that have varying goals, abilities, and composition. Reinforcement learning have been used to learn about other agents and adapt local behavior for the purpose of achieving coordination in multi-agent situations in which the individual agents are not aware of each another [20].

Various reinforcement learning strategies have been proposed that can be used by agents to develop a policy to maximizing rewards accumulated over time. A prominent algorithm in reinforcement learning is the Q-learning algorithm.

In Q-learning, the decision policy is represented by the Q-factors, which estimates long-term discounted rewards for each state-action pair. Let $Q(s, a)$ denote the

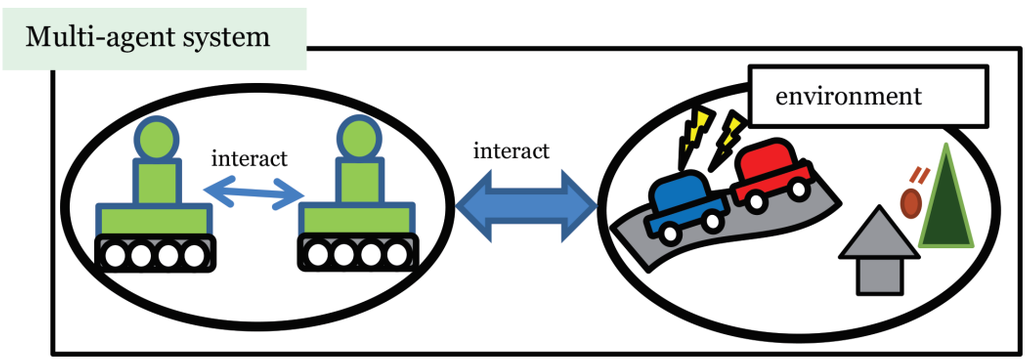

Figure 1.
Multi-agent systems.

Q-factor for state s and action a. If an action a in state s produces a reinforcement of r and a transition to state s_{t+1}, then the corresponding Q-factor is modified as follows:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r + \gamma max Q(s_{t+1}, a) - Q(S_t, a_t)] \quad (1)$$

where α is a small constant called learning rate, which denotes a step-size parameter in the iteration, and γ denotes the discounting factor. Theoretically, the algorithm is allowed to run for infinitely many iterations until the Q-factors converge.

2.2 The target problem

In this chapter, we focus on a disaster relief as a target problem similar to previous research [16]. In the disaster relief problem, agents must rescue the injured as quickly as possible, and the injured with different severity and urgency of the condition are placed on a field of fixed size. Because there are multiple injured and obstacles, the disaster relief problem can be considered to be a multi-agent system. Each agent focuses on achieving its own target, and the task of system is to efficiently rescue all of the injured and remove obstacles (**Figure 2**).

Efficient rescue is performed at a disaster site using triage to assign priority of transport or treatment based on the severity and urgency of the condition of the injured. In the disaster rescue problem, it is thus necessary to reflect triage based on the condition of the injured. For this purpose, in this chapter, we designate the condition of the injured as red (requiring emergency treatment), yellow (requiring urgent treatment), green (light injury), or black (lifesaving is difficult) in descending order of urgency.

The disaster relief problem is represented as shown in **Figure 3**. The field is divided into an N × N lattice. Agents are indicated by circles, ◎; the injured are indicated by R, Y, G, and B; removable obstacles are indicated by white triangles, △; and nonremovable obstacles are indicated by black triangles, ▲. The destination of injures is indicated by a white square, □; and the collection site of movable obstacles is indicated by a black square, ■. A single step is defined such that each of the agents on the field completes a single action, and the field is re-initialized once all of the injured have been moved.

The agents considered in this chapter are depicted in **Figure 4** and included two types, to obtain cooperative actions. The rescue agents have the primary function of rescuing the injured, and the removal agents have the primary function of removing obstacles, although either type can perform both functions. The agents

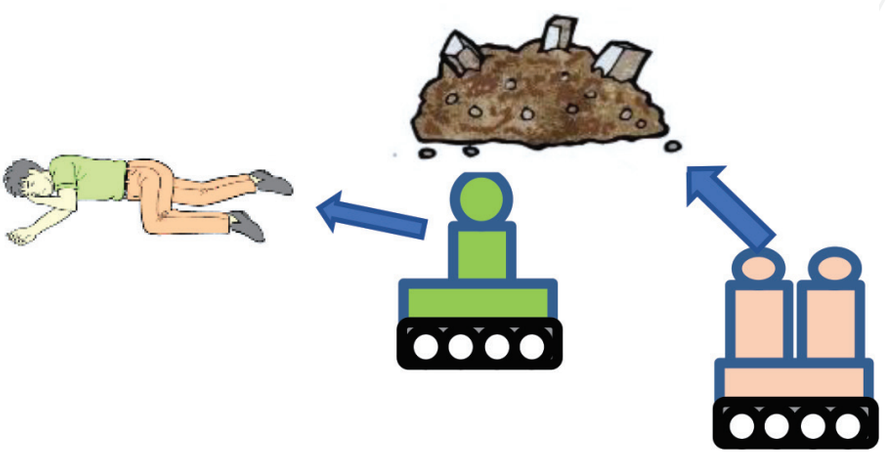

Figure 2.
Disaster relief problem.

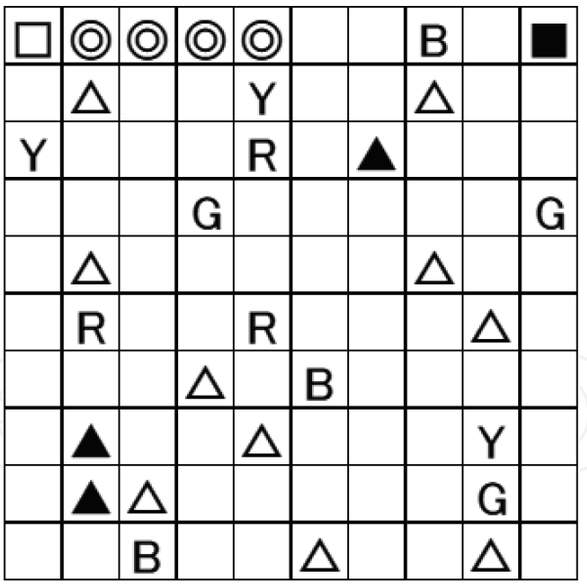

Figure 3.
An example of representation for a disaster relief problem.

Figure 4.
Agents of different functions: (a) removing agent and (b) relief agent.

recognize the colors of the injured on the field and identify the condition of the injured in correspondence with those colors.

Each agent considers the overall tasks in the given environment, carries out the tasks in accordance with the assigned roles, and learns appropriate actions that bring high rewards.

An agent can recognize its circumstance within a prescribed field of vision and move one cell vertically or horizontally, but will stay in place without moving if a nonremovable obstacle, injured transport destination, obstacle transport destination, or other agent occupies the movement destination or if the movement destination is outside the field. Each agent has a constant but limited view of the field, and it can assess the surrounding environment.

The available actions of agents are (1) moving up, down, right, or left to an adjacent cell; (2) remaining in the present cell when adjacent cell is occupied by an obstacle that cannot be removed or by another agent; and (3) finding an injured person or a movable obstacle and taking it to the appropriate location.

If an agent is processing an injured person and next action is moving it to the appropriate destination, then the task of the agent is completed. The agent can begin a new task for rescuing or removing. When all of the injured on the field have been rescued, the overall task is completed.

3. Improvement of Cooperative Action by Rewards Distribution

3.1 Cooperative agents

In multi-agent system, the environment changes from static to dynamic because multiple autonomous agents exist. An agent engaged in cooperative action decides its actions by referring to not only its own information and purpose but to those of other agents as well [16]. The cooperative agents are acquired by sharing sensation, sharing episodes, and sharing learned policies [17, 19–21]. Cooperative actions are important not only in situations where multiple agents have to work together to accomplish a common goal but also in situations where each agents has its own goal [22].

In this chapter, the multi-agent systems are composed of agents' different behaviors. One is to perform relief (*relief agents*), and the other is to remove obstacles (*removing agents*). Cooperation was achieved by giving different rewards to different behaviors.

3.2 Reward distribution with consideration of condition of the injured

It is necessary to have the multi-agent systems learn efficient rescue with the condition of the injured taken into consideration.

Prior studies used reward distribution with the reward value differing in accordance with the agent action but gave no consideration to the condition of the injured.

In this chapter, we propose three types of reward distribution as methods for obtaining cooperative action of injured rescue and obstacle removal in accordance with the urgency of the condition of the injured.

3.2.1 Method 1: reward distribution responding to the condition of the injured

A conferred reward is high in value when an injured person in a condition of high urgency is rescued and decreases in value for those in less urgent conditions. Thus, $R_r > R_y > R_g > R_b$, where R_r, R_y, R_g, and R_b are the reward values for the rescue of injured persons in the red, yellow, green, and black condition categories, respectively.

3.2.2 Method 2: reward distribution based on the contribution degree

In Method 2, the reward value reflects the time spent by the rescue agent as the contribution degree.

With R as the basic reward value when the rescue agent completes the injured rescue, C as the contribution degree, and λ as a weighting factor, the reward r earned by the rescue agent in learning is as given by Eq. (2). A large λ results in a reward that is greatly augmented relative to the basic reward, according to contribution degree.

Assessed contribution degree C increases with decreasing time spent in rescuing the injured, as shown in Eq. (3), in which T_e is the time of completion of rescue of all the injured by the rescue agents and T_i is the time spent by an agent to rescue an injured person.

$$r = (1 + \lambda C)R \tag{2}$$

$$C = T_i / T_e \tag{3}$$

3.2.3 Method 3: reward distribution by the contribution degree responding to the condition of the injured

In Eq. (2), basic reward value R at the time of completion of each task takes on one of the values $R_r > R_y > R_g > R_b$ according to the condition of the injured person.

4. Experimental results and discussion

4.1 Experimental conditions

In the study presented in this chapter, we experimented on obtaining cooperative action by agents for efficient rescue in accordance with the condition of the injured and obstacle removal, using the three proposed reward distributions. We assigned the injured and obstacle transport destinations to one cell each on the field shown in **Figure 3** and numbers of agents, injured persons, and obstacles as listed in **Table 1**. The mean of five simulation trials was taken as the result.

4.2 Effects of reward distribution timing

We investigated the effects of the three patterns of reward distribution timing on task completion on the efficiency of learning injured rescue. The reward is given when an injured person or obstacle is discovered in Pattern 1. The reward is given when an injured person or object is taken for removal to the appropriate location in Pattern 2. Rewards are given twice in Pattern 3: at the stage of discovering an injured person or obstacle and at the stage where transportation is completed.

The results of an experiment to compare the three reward distribution timing patterns are shown in **Figure 5**. The horizontal axis represents the episodes, and the vertical axis represents the number of steps for task completion by all agents. These results indicate that Pattern 3 allowed completion of the tasks in a smaller number

The setting of field		
The field size		10 × 10
The number of rescue agents		2
The number of clearing agents		2
The number of injured individuals	Red	3
	Yellow	3
	Green	3
	Black	3
The number of removal of possible obstacles		10
The number of removal of impossible obstacles		3
The setting of agent		
Learning rate α		0.1
Discount rate γ		0.9
Greedy policy ε		0.1

Table 1.
Experimental conditions.

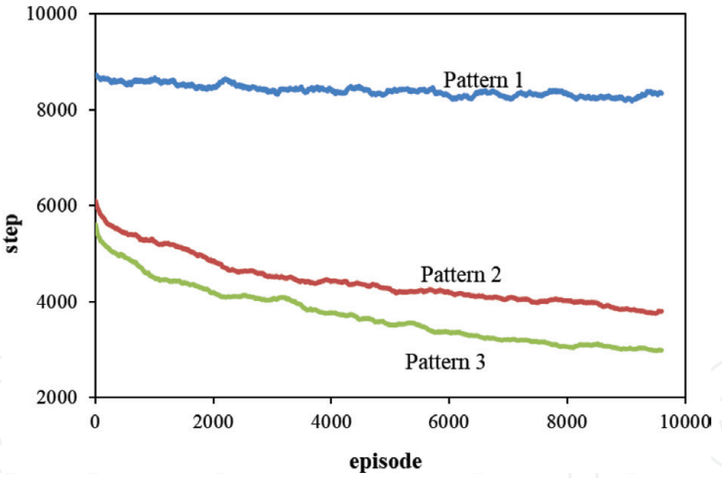

Figure 5.
Results of experiment to compare three reward distribution timing patterns.

of steps than did Patterns 1 and 2, which in turn indicates that efficient rescue and removal was learned by conferring rewards in two stages and thus led the agent to regard the course from discovery to transport as one task. We therefore applied Pattern 3 in the subsequent experiments.

4.3 Obtaining cooperative action that considers the condition of the injured and the effects of reward distribution timing

We applied the three types of reward distributions in experiments for efficient rescue in accordance with the urgency of the condition of the injured. In the following descriptions of experimental results, the horizontal axis represents episodes, and the vertical axis represents the number of steps for task completion by all agents.

Figure 6 shows the results of an experiment to investigate the effectiveness of a reward distribution in accordance with condition (Method 1) compared to the conventional method [16]. As shown, the number of steps is higher throughout with Method 1 than with the standard method, thus indicating learning to postpone rescue of the low urgency injured and prioritize rescue of the high urgency injured.

Finally, we performed an experiment to investigate the effectiveness of reward distribution based on contribution degree (Method 2) in comparison to Method 1 and another experiment to investigate the effectiveness of reward distribution by contribution degree in accordance with injured condition (Method 3) in

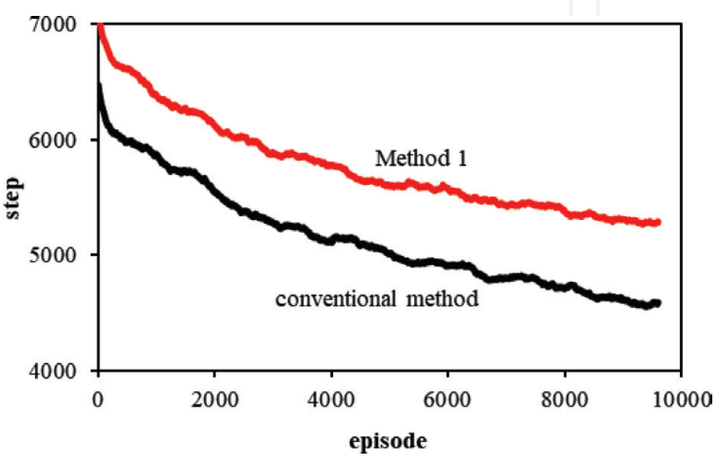

Figure 6.
Results of the conventional method and proposed Method 1.

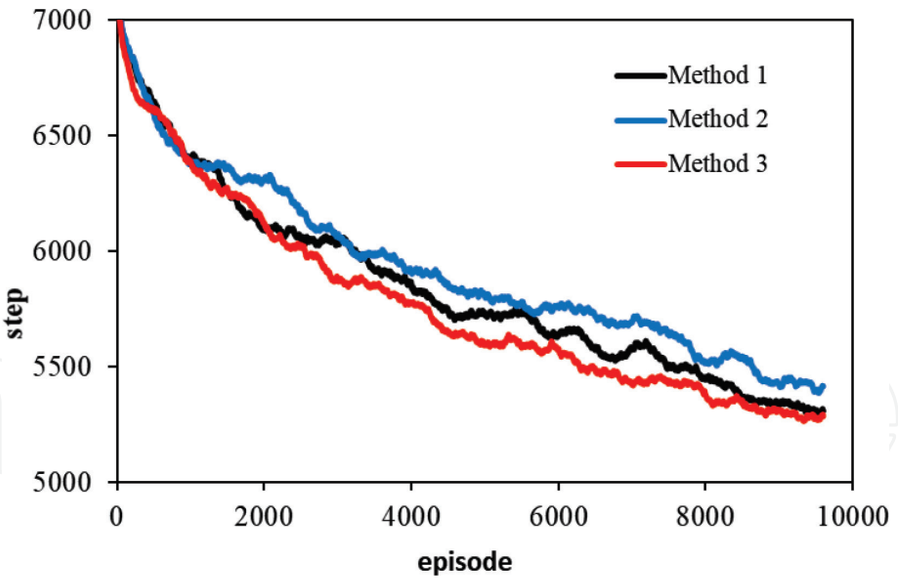

Figure 7.
Results of the different proposed methods.

Triage of injured	Method 1	Method 2	Method 3
Red	1126.73	1140.93	1102.93
Yellow	1518.20	1614.40	1269.07
Green	2404.53	2499.33	1685.47
Black	3284.27	2999.47	2447.33

Table 2.
Mean step numbers by different reward distribution.

comparison with the other proposed methods (Methods 1 and 2). The results are shown in **Figure 7** and **Table 2**.

As shown, Method 2 tends to yield a higher number of steps than Method 1 in and after around episode 6000. This indicates that, for efficient injured rescue with consideration for contribution degree, the rescue order learned was to first rescue those injured who were nearby and thus shorten the rescue time, leaving for later the rescue of those who were farther away.

Method 3 is approximately 2.2 and 3.4% superior to Methods 1 and 2, respectively. The agents were apparently able to learn rescue of the injured in accordance with urgency because a reward differing in accordance with injured condition was conferred on the agents. These results also show that the agents were able to learn efficient rescue action because the reward distribution reflected contribution degree.

5. Conclusion

In this chapter, we considered rescue robots as a multi-agent system and proposed three reward distributions for the agents to learn cooperative action with consideration given to the condition of the injured and obstacle removal in responding to our disaster rescue problem, as well as investigating the timing of reward conferral on the agents.

Comparative experiments showed that the timing of reward distribution enabling the agents to obtain the most efficient cooperative actions consisted of

reward conferral at the stages in which the agent discovered the injured person or obstacle and at completion of their transport. The results also showed that the capability of cooperative actions for the most efficient injured rescue and obstacle removal could be acquired through reward distribution by contribution degree in accordance with condition.

In this chapter, the multi-agent system corresponding to the disaster rescue problem, rescue simulations were performed with the condition of the injured determined in advance. In future studies, we plan to conduct simulations with dynamic changes over time in both the condition of the injured and removable versus nonremovable states of the obstacles.

Acknowledgements

This research was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Number JP16K01303.

Author details

Mengchun Xie
Department of Electrical and Computer Engineering, National Institute of Technology, Wakayama College, Wakayama-ken, Japan

*Address all correspondence to: xie@wakayama-nct.ac.jp

References

[1] Gautschi O, Cadosch D, Rajan G, Zellweger R. Earthquakes and trauma: Review of triage and injury-specific, immediate care. Prehospital and Disaster Medicine: The Official Journal of the National Association of EMS Physicians and the World Association for Emergency and Disaster Medicine in Association with the Acute Care Foundation. 2008;**23**(2):195-201. DOI: 10.1017/S1049023X00005847

[2] Xie M. Cooperative behavior rule acquisition for multi-agent systems by machine learning. In: Advances in Reinforcement Learning. Rijeka, Croatia: Intech; 2011. pp. 81-98

[3] Geron A. Hands-on Machine Learning with Scikit-Learn & TensorFlow. Sebastopol CA: O'Reilly Media Inc; 2018

[4] Sutton RS, Barto AG. Reinforcement Learning: An Introduction. Cambridge, Massachusetts, London, England: The MIT Press; 1998

[5] Conradie AVE, Aldrich C. Development of neurocontrollers with evolutionary reinforcement learning. Computers & Chemical Engineering. 2005;**30**:1-17

[6] Isbell CL, Shelton CR, Kearns M, Singh S, Stone P. A social reinforcement learning agent. In: Proceedings of the Fifth International Conference on Autonomous Agents. 2001. pp. 377-384

[7] Tatomir B, Rothkrantz L, Popa M. Intelligent system for exploring dynamic crisis environments. In: Proceeding of the Third International Conference on Information Systems for Crisis Response and Management. 2006

[8] Stone P, Veloso M. Using machine learning in the soccer server. In: Proc. of IROS-96 Workshop on Robocup. 1996

[9] Buford JF, Jakobson G, Lewis L. Multi-agent situation management for supporting large-scale disaster relief operations. The International Journal of Intelligent Control and Systems. 2006;**11**(4):284-295

[10] Matsubara H, Frank I, Tanaka K, et al. Automatic soccer commentary and RoboCup. In: The 2nd Proceedings of RoboCap Workshop. 1998

[11] Pereira R d P, Engel PM. A Framework for Constrained and Adaptive Behavior-Based Agents, arXiv preprint arXiv:1506.02312v12015. pp. 1-16

[12] Spychalski P, Arendt R. Machine learning in multi-agent systems using associative arrays. Parallel Computing. 2018;**75**:88-99

[13] Zhou P, Shen H. Multi-agent cooperation by reinforcement learning with teammate modeling and reward allotment. In: 8th International Conference on Fuzzy Systems and Knowledge Discovery; 4 February, 2011; FSKD. 2011. pp. 1316-1319

[14] Zemzemb W, Tagina M. Cooperative multi-agent systems using distributed reinforcement learning techniques. Procedia Computer Science. 2018;**126**:517-526

[15] Xie M, Murata M, Muraki Y. Tsunami evacuation guidance simulation using multi-agent systems based on OpenStreetMap. International Journal of Environmental Sciences. 2017;**2**:231-237. ISSN: 2367-8941

[16] Xie M, Murata M, Sato S. Acquisition of cooperative action by rescue agents with distributed roles. In: Leu G et al., editors. Intelligent and Evolutionary Systems, Proceedings in Adaptation, Learning and Optimization. Vol. 8. AG: Springer

International Publishing; 2017. pp. 483-493

[17] Jennings N, Sycara K, Wooldridge M. A roadmap of agent research and development. Autonomous Agents and Multi-Agent Systems. 1998;**1**:7-38

[18] Xie M. Representation of the perceived environment and acquisition of behavior rule for multi-agent systems by Q-learning. In: Proceedings of the 4th International Conference on Autonomous Robots and Agents. 2009. pp. 453-457

[19] Xie M, Tachibana A. Cooperative behavior acquisition for multi-agent systems by Q-learning. In: Proceedings IEEE Symposium on Foundations of Computational Intelligence. 2007. pp. 424-428

[20] Weiss G. Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence. Cambridge, Massachusetts, London, England: The MIT Press; 2000

[21] Panait L, Luke S. Cooperative multi-agent learning: The state of the art. Autonomous Agents and Multi-Agent Systems. 2005;**11**(3):387-434

[22] Son TC, Pontelli E, Nguyen N. Planning for Multiagent Using ASP-Prolog, Computational Logic in Multi-Agent Systems. New York, USA: Springer; 2010. pp. 1-19

[23] Xie M, Okazaki K. Application of multi-agent systems to disaster relief using Q-learning. In: Proceedings of the IASTED International Conference on Software Engineering and Applications. Calgary, Alberta, Canada: ACAT Press; 2008. pp. 143-147