دوفصلنامه محاسبات و سامانه های توزیع شده
سال سوم، شماره دوم، شماره پیاپی ۶، صفحه ۱۲ تا ۲۱، سال ۱۳۹۹

دستیابی به همکاری از طریق یادگیری تقویتی چند عاملی در معمای زندانی تکرارشونده

سمیرا فرزانه[۱]، فرشته زندی[۲]، جواد سلیمی سرتختی*[۳]
[۱و۲]دانشجوی کارشناسی ارشد مهندسی کامپیوتر نرم افزار ، دانشکده مهندسی برق و کامپیوتر، دانشگاه کاشان، کاشان، ایران.
[۳]عضو هیئت علمی، استادیار، دانشکده مهندسی برق و کامپیوتر، دانشگاه کاشان، کاشان، ایران.

چکیده

امروزه معمای زندانی یکی از مسائل اولیه و مهم در نظریه بازی‌ها است. در این معما نقطه تعادل نشی وجود دارد و چنانچه عامل‌ها منطقی رفتار کنند در آن نقطه بازی می‌کنند؛ بدین منظور عامل‌ها برای دستیابی به سود بیشتر از بین دو عمل همکاری و عدم همکاری، عدم همکاری را انتخاب می‌کنند. در حالیکه برای عامل‌ها نقطه بهتری نسبت به نقطه نش وجود دارد و آن هم این است که هر دو عامل همکاری را انتخاب کنند. بنابراین، در جهت افزایش میزان همکاری عامل‌ها معمای زندانی به صورت معمای زندانی تکرارشونده با یک رویکرد یادگیری تقویتی در نظر گرفته‌شده است. نتایج مقاله نشان‌دهنده این است که رویکرد مورد نظر سبب افزایش میزان همکاری عامل‌ها شده است و اگر عاملی همکاری را پیشه کند عامل دیگر نیز همکاری را انتخاب می‌کند و بالعکس.

کلمات کلیدی: عدم همکاری متقابل، معمای زندانی تکرارشونده[۱]، یادگیری تقویتی، همکاری متقابل، LSTM[2]

Achieving Cooperation Through Multi agent Reinforcement Learning In Iterated Prisoner's Dilemma

Samira Farzaneh[1], Fereshteh Zandi[2], Javad Salimi Sartakhti*[3]

[1,2]Master Student of Computer Software Engineering, Faculty of Electrical and Computer Engineering, University of Kashan, Kashan, Iran.

[2]Faculty member, Assistant Professor, Faculty of Electrical and Computer Engineering, University of Kashan, Kashan, Iran.

Abstract

Nowadays, the prisoner’s dilemma is one of the primary and important issues in game theory. In this dilemma, there is a Nash Equilibrium, and if the agents behave rationally, they play at point; For this purpose, the agents choose defection between the two actions of cooperation and defection to achieve greater profit. However there is a better point for the agents than the Nash Equilibrium, it is that both agents choose the cooperation. However there is a better point for the agents than the Nash Equilibrium, it is that both agents choose the cooperation. Therefore, in order to increase the rate of cooperation of the agents, the prisoner's dilemma has been considered as iterated prisoner's dilemma with a reinforcement learning approach. The results of the article show that the desired approach let has increased the rate of cooperation of the agents, and if one agent choose the cooperation, the other agent also chooses cooperation and vice versa.

تاریخچه مقاله:

تاریخ ارسال: ۱۳۹۹/۱۰/۲۹
تاریخ اصلاحات: ۱۳۹۹/۱۱/۲۲
تاریخ پذیرش: ۱۳۹۹/۱۱/۱۹
تاریخ انتشار: ۱۳۹۹/۱۲/۲۹

Keywords:

Mutual Defection
Iterated Prisoner’s Dilemma
Reinforcement learning
Mutual Cooperation
LSTM(Long Short Term Memory)

*ایمیل نویسنده مسئول:
salimi@kashanu.ac.ir

۱- مقدمه

معضلات اجتماعی بیانگر وضعیتی هستند که در آن بازیکن‌ها با توجه به منافع شخصی خودشان عدم همکاری را انتخاب می‌کنند در حالیکه اگر همکاری انتخاب شود، بازیکن‌ها به سود بیشتری دست می‌یابند.

حال این مشکلات زمانی بوجود می‌آیند که بازیکن‌ها تصمیم بگیرند برای دستیابی به سود بیشتر رفتارهای خودخواهانه‌ای از خود نشان دهند و از منافع شخصی خودشان پیروی کنند و در جهت منافع بلند مدت گروهی هیچ اقدامی انجام ندهند. اینگونه معضلات اشکال مختلفی دارند و در اکثر رشته‌ها مانند روانشناسی، اقتصاد، علوم سیاسی و... مورد مطالعه قرار گرفته‌اند. نمونه‌هایی که می‌توان با استفاده از معضلات اجتماعی تبیین کرد شامل میزان مشارکت کم در انتخابات، جمعیت بیش از حد، تخلیه منابع و... است. چنین مشکلاتی را می‌توان از طریق تجزیه و تحلیل نظریه بازی‌ها درک و حل کرد. در این مقاله برای درک بهتر معضلات اجتماعی، بازی معمای زندانی معرفی شده است؛ زیرا در این بازی نشان داده می‌شود که پیامدهای منافع فردی با منافع گروهی در تضاد هستند.

مدل معمای زندانی متشکل از دو بازیکن A و B است. دو بازیکن قادر به برقراری ارتباط با یکدیگر نیستند. برای بازیکن‌ها دو عمل همکاری و عدم همکاری وجود دارد. اگر بازیکن A تصمیم به عدم همکاری و بازیکن B تصمیم به همکاری بگیرد، بازیکن A به سود بیشتری می‌رسد در حالیکه بازیکن B ضرر می‌کند و بالعکس. اما اگر هر دو بازیکن ترجیح دهند با یکدیگر همکاری کنند، هر دو سود خوبی دریافت می‌کنند. ولی اگر هر دو بازیکن عدم همکاری را انتخاب کنند، هر دو سود کمی دریافت می‌کنند. در این شرایط به نظر می‌رسد که هر بازیکن باید با یکدیگر همکاری کنند تا سود خوبی دریافت کنند؛ اما نتیجه بالعکسی بدست می‌آید. زیرا، در معمای زندانی نقطه تعادل نشی وجود دارد و چنانچه بازیکن‌ها منطقی رفتار کنند در نقطه تعادل نش بازی می‌کنند. بنابراین، هر دو عدم همکاری را انتخاب می‌کنند. در حالیکه برای عامل‌ها نقطه بهتری نسبت به نقطه نش وجود دارد و آن هم این است که هر دو عامل همکاری را انتخاب کنند.

طی دو دهه گذشته، مجموعه بزرگی از تکنیک‌های یادگیری چند عاملی مانند Nash Q-learning، Conditional-JAL و -min max Q-learning در جهت افزایش همکاری عامل‌ها ارائه شده است [۳-۱].

حال در این مقاله در راستای افزایش همکاری عامل‌ها بازی معمای زندانی، به معمای زندانی تکرارشونده تبدیل گشته است. معمای زندانی تکرارشونده تعمیمی از بازی معمای زندانی است. بدین صورت که بازی به صورت مکرر توسط همان عامل‌ها چندین بار انجام می‌شود و آنها استراتژی خود را بر اساس اقدامات قبلی رقیب خود انتخاب می‌کنند. سپس محیط بازی معمای زندانی تکرارشونده به صورت ماتریسی و برای هر حرکت در محیط یک هزینه در نظر گرفته می‌شود.

با الگوریتمی که در ادامه ارائه می‌شود، یک رویکرد یادگیری تقویتی پیشنهاد می‌گردد که منجر به افزایش همکاری عامل‌ها می‌شود. رویکرد یادگیری تقویتی گونه‌ای از روش‌های یادگیری ماشین و مبتنی بر پاداش است و عامل‌ها در اینگونه محیط‌ها در تلاش هستند تا پاداش خودشان را به حداکثر برسانند. این نوع یادگیری برای محیط‌هایی که عامل هیچ نوع شناختی نسبت به آنها ندارد، مورد استفاده قرار می‌گیرد.

رویکرد پیشنهادی شامل دو مرحله است: مراحل آفلاین و آنلاین.

مرحله آفلاین سیاست‌ها را با درجه‌های مختلف همکاری تلفیق می‌کند؛ سیاست تعیین‌کننده نحوه رفتار عامل با هر اقدام و تصمیم-گیری در شرایط مختلف است و عامل طبق سیاست بهترین اقدامش را انتخاب می‌کند. سپس شبکه شناسایی درجه همکاری را آموزش می‌دهد و درجه‌های همکاری‌های مختلفی به عنوان سیاست تولید می‌شوند [۴].

در جهت تولید سیاست‌ها و آموزش سیاست‌های اولیه با درجه‌های مختلف همکاری، پیشنهاد می‌شود از طرح بازیگر منتقد مشترک[3] (JAC) استفاده شود. JAC به عامل‌ها به صورت یکسان آموزش می‌دهد. بنابراین هر دو عامل در محیط مشترکی قرار می‌گیرند تا در همان محیط مشترک سیاست‌های اولیه برای مرحله آفلاین فرابگیرند. سیاست‌ها و درجه همکاری‌های تولیدشده برای بازی در مرحله آنلاین درنظر گرفته می‌شود.

در مرحله آنلاین عامل بر اساس میزان همکاری مشخص‌شده رقیب، استراتژی خود را انتخاب می‌کند. در این قسمت یک شبکه عصبی LSTM جهت پیش‌بینی درجه همکاری رقیب در نظر گرفته می‌شود.

نتایج بدست آمده حاکی از افزایش میزان همکاری در بین عامل‌ها است و نشان داده می‌شود که عامل با تغییر سیاست‌های ثابت در برابر رقیب عملکرد خوبی داشته باشد.

ادامه مقاله به صورت زیر سازماندهی شده است. در بخش دوم به تعریف مفاهیم اولیه، در بخش سوم بطور مختصرکارهای گذشته مرور

[1] Iterated Prisoner's Dilemma
[2] Long Short Term Memory
[3] Joint Actor Critic

می‌شود، در بخش چهارم به رویکرد پیشنهادی و در بخش پنجم به ارزیابی پرداخته می‌شود. در بخش ششم و پایانی نتیجه‌گیری ارائه می‌شود.

۲- تعریف مفاهیم پایه

با در نظر گرفتن بازی معمای زندانی ساده نمی‌توان میزان همکاری عامل‌ها را افزایش داد. بنابراین بازی معمای زندانی به صورت تکرار-شونده و در محیط ماتریسی در نظر گرفته می‌شود. در ادامه تمامی موارد فوق به صورت کامل توضیح داده شده است.

۲-۱- بازی معمای زندانی:

در حالت کلی بازی معمای زندانی را می‌توان به صورت مجموعه چندتایی $\langle N, \{A_i\}, \{R_i\}\rangle$ نشان داد؛ N تعداد عامل‌ها، A_i مجموعه اقداماتی است که برای هر عامل وجود دارد و R_i پاداش برای هر عامل است. ماتریس پرداخت[4] بازی مذکور در جدول ۱ نشان داده شده است. در این بازی عامل‌ها منطقی رفتار می‌کنند همچنین برای هر عامل دو عمل همکاری (C) و عدم همکاری (D) و پاداش‌های S, P, R و T در نظرگرفته شده است که مقادیر نتایج نهایی آنها در زیر آورده شده است:

- $R > P$ یعنی همکاری متقابل به عدم همکاری متقابل ترجیح داده می‌شود.
- $R > S$ یعنی همکاری متقابل یکی از عامل‌ها به عدم همکاری عامل دیگر ترجیح داده می‌شود.
- $2R > S+T$ یعنی همکاری متقابل با احتمال برابری به همکاری و عدم همکاری یک جانبه ترجیح داده می‌شود.
- $T > R$ یعنی عدم همکاری نسبت به همکاری متقابل ترجیح داده می‌شود.
- $P > S$ یعنی عدم همکاری متقابل به همکاری متقابل ترجیح داده می‌شود.

ماتریس پرداخت: در نظریه بازی‌ها، ماتریس پرداخت جداولی است که در آن استراتژی‌های یک عامل در ردیف‌ها و استراتژی‌های عامل دیگر در ستون‌ها ذکر شده و سلول‌ها نشان‌دهنده پرداخت برای هر عامل است به گونه‌ای که در هر سلول ابتدا نتیجه عامل اول ذکر می‌شود. بنابراین، ماتریس پرداخت یک بازی متشکل از سود، ضرر یا هزینه اضافی است و با اجرای استراتژی عامل با توجه به استراتژی عامل دیگر به عامل تعلق می‌گیرد. استفاده از این ماتریس در نظریه بازی‌ها به عامل در جهت کسب اطلاعات بیشتر و دستیابی به نقطه تعادل نش کمک می‌کند.

لازم به ذکر است، اگر عامل‌ها منطقی باشند هر دو عدم همکاری را انتخاب می‌کنند. انتخاب عدم همکاری همان نقطه تعادل نش است.

ساده‌ترین نوع بازی، بازی معمای زندانی تک مرحله‌ای است؛ که در مثال زیر توضیح داده شده است.

مثال) بازی معمای زندانی متشکل از دو بازیکن A و B است که به جرم متهم شده‌اند. این دو بازیکن منطقی رفتار می‌کنند و قادر به برقراری ارتباط با یکدیگر نیستند. برای بازیکن‌ها دو عمل همکاری (اعتراف نکردن به جرم) و عدم همکاری (اعتراف کردن به جرم) وجود دارد. اگر بازیکن A تصمیم به خیانت به بازیکن B بگیرد، بازیکن A هیچ وقت زندانی نخواهد شد در حالیکه بازیکن B حکم حبس قابل توجهی دریافت می‌کند و بالعکس. اما اگر هر دو بازیکن ترجیح دهند در مورد جنایت سکوت کنند و اعتراف نکنند، هر دو مجازات کمتری دریافت می‌کنند ولی اگر هر دو بازیکن به یکدیگر خیانت کنند و دیگری را تحویل دهند، هر دو مجازات‌های قابل توجه‌تری را دریافت می‌کنند. در این شرایط به نظر می‌رسد که هر بازیکن باید سکوت را انتخاب کند تا هر دو مجازات کمتری دریافت کنند اما به دلیل منطقی رفتار کردن آنها و وجود نقطه تعادل نش در بازی، هر دو در نقطه تعادل نش بازی می‌کنند، بنابراین، هر دو عدم همکاری را برای رسیدن به سود بیشتر انتخاب می‌کنند.

(جدول-۱): ماتریس پرداخت معمای زندانی

بازیکن ۱ \ بازیکن ۲	**C**	**D**
C	R,R	S,T
D	T,S	P,P

استراتژی: قاعده از پیش تعیین‌شده‌ای است و مشخص می‌کند هر عامل در مقابل حرکت عامل دیگر چه واکنشی باید داشته باشد.

سیاست: تعیین‌کننده نحوه رفتار عامل با هر اقدام و تصمیم-گیری در شرایط مختلف است و عامل طبق سیاست بهترین اقدامش را انتخاب می‌کند.

۲-۲- بازی معمای زندانی تکرارشونده

معمای زندانی تکرارشونده تعمیمی از معمای زندانی ساده است [۵]. بدین صورت که بازی به طور مکرر توسط همان عامل‌ها چندین بار

[4] Payoff matrix

انجام می‌شود و آنها استراتژی خود را بر اساس اقدامات قبلی رقیب خود انتخاب می‌کنند.

اگر بازی N بار تکرار شود و هر دو عامل از تعداد تکرار بازی اطلاع داشته باشند، هر دو در نقطه تعادل نش بازی می‌کنند و عدم همکاری را انتخاب می‌کنند. حال در این مقاله با ارائه الگوریتمی (بخش ۴) میزان همکاری عامل‌ها در معمای زندانی تکرار شونده افزایش می‌یابد.

۳- مرور کارهای گذشته

در دهه‌های گذشته محققان معمولا بازی معمای زندانی را برای دستیابی به هدف همکاری انتخاب و بررسی می‌کردند و پی بردند در اینگونه بازی‌ها اگر هر دو عامل منطقی بازی کنند، به نقطه تعادل نش می‌رسند و در آن نقطه بازی می‌کنند؛ عدم همکاری همان نقطه تعادل نش است. بنابراین، هیچ همکاری رخ نمی‌دهد. در همین راستا محققان برای انگیزه دادن عامل‌ها به سمت همکاری متقابل بازی‌های معمای زندانی تکرارشونده با تغییراتی در محیط بازی پیشنهاد دادند [۶-۸].

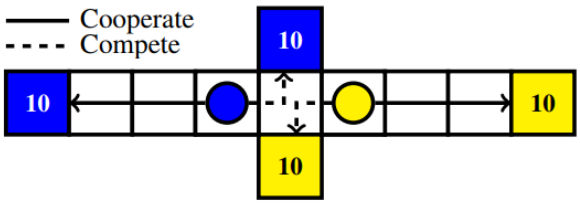
کراندال بازی معمای زندانی را به صورت تکرارشونده در نظر گرفته است و الگوریتمی را در راستای همکاری متقابل عامل‌ها ارائه داد [۹]. سپس برخی از محققان راهکار ارائه شده کراندال را در سناریو‌های مختلف گسترش دادند. به عنوان مثال، بازی با تغییر استراتژی در مقابل حریف‌ها [۱۰].

لیبو و همکاران در راستای بهتر نشان دادن ویژگی‌های معضل اجتماعی در دنیای واقعی با حفظ ویژگی‌های بازی‌های معمای زندانی تکرار شونده، یک بازی دو بعدی جدید را معرفی کرده‌اند [۱۱].

کلیمان و همکاران در جهت بررسی همکاری یا عدم همکاری عامل‌ها، جامعه تعاملات استراتژیک افراد را به عنوان بازی‌های ماتریسی در نظر گرفته‌اند [۱۲].

۴- رویکرد پیشنهادی

در بیشتر تحقیقات قبلی بازی معمای زندانی به صورت تکراری حل گشته‌اند و برخی از محققان با محدود کردن اطلاعات عامل‌ها سبب افزایش همکاری در بین آنها شده‌اند [۷]. همچنین برخی دیگر معمای زندانی تکرارشونده را با استفاده از راهکاری مبتنی بر یادگیری تقویتی عمیق حل کرده‌اند. بدین صورت، برای هر کدام از عامل‌ها بافر جداگانه‌ای در راستای آموزش استراتژی‌ها در نظر گرفته شده است و هر کدام ار آنها به تنهایی در محیط آموزش می‌بینند [۱۳]. همچنین وانگ و همکاران از یک رویکرد یادگیری تقویتی چند عاملی برای افزایش همکاری متقابل در محیط بازی‌های ویدئویی مانند دو بازی Fruit-Gathering و Apple-Pear استفاده کرده‌اند [۱۴]. حال در این مقاله در پیرو همان رویکرد، محیط بازی تغییر کرده و به محیط ماتریسی تبدیل شده است.



(شکل-۱): معمای زندانی تکرارشونده

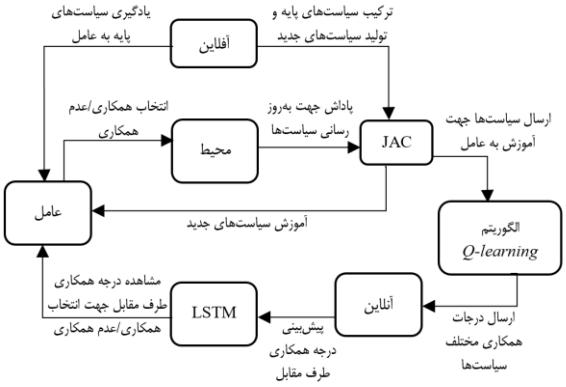
محیط ماتریسی بازی معمای زندانی تکرارشونده در شکل ۱ نشان داده شده است.

در این بازی اصل بر این است که اگر هردو عامل به خانه منقطع بروند یعنی عدم همکاری داشته‌اند؛ بنابراین، عامل‌ها تا زمانی که بر روی خطوط ممتد و به سمت خانه شماره ۱۰ حرکت کنند، همکاری کرده و زمانی که بر روی خطوط منقطع و به سمت خانه ۱۰ حرکت کنند، عدم همکاری داشته‌اند. همچنین اگر هر دو به سمت خانه منقطع بروند به صورت تصادفی یکی از عامل‌ها انتخاب می‌شود به آن خانه می‌رود و دیگری شکست می‌خورد و به خانه قبلی برمی‌گردد. این فرایند باعث دریافت پاداش بیشتری برای عاملی که در خانه دارای خط منقطع است، می‌گردد و عامل دیگری که به خانه قبلی بازگشته هزینه بیشتری پرداخت می‌کند. تلاش ما در این بازی به این صورت است که عامل‌ها به سمت خانه‌هایی که در طرفین قرار دارد رفته تا میزان همکاری عامل‌ها افزایش یابد. این فرایند دارای هزینه حرکت بر روی خطوط می‌باشد اما از هزینه بالا استفاده از خطوط منقطع جلوگیری می‌شود.

در این مقاله بازی به صورت زندانی تکرارشونده و دو عمل همکاری و عدم همکاری در محیط ماتریسی برای عامل‌ها درنظر گرفته شده است. حال برای افزایش میزان همکاری عامل‌ها از یک رویکرد یادگیری تقویتی چند عاملی استفاده می‌شود.

رویکرد پیشنهادی شامل دو مرحله آفلاین و آنلاین است. در مرحله آفلاین، سیاست‌ها با درجه‌های مختلف همکاری تلفیق می‌شوند و سپس شبکه شناسایی درجه همکاری، آموزش می‌بیند؛ و درجه‌های همکاری‌های مختلفی به عنوان سیاست‌های جدیدی تولید می‌شوند. در مرحله آنلاین از سیاست‌ها و درجه همکاری‌های تولیدشده بازی استفاده می‌شود. در این مرحله عامل بر اساس میزان همکاری

مشخص‌شده رقیب، استراتژی خود را انتخاب می‌کند. شکل ۲ نشان دهنده چارچوب کلی پیشنهادی است.



(شکل-۲): چارچوب کلی پیشنهادی

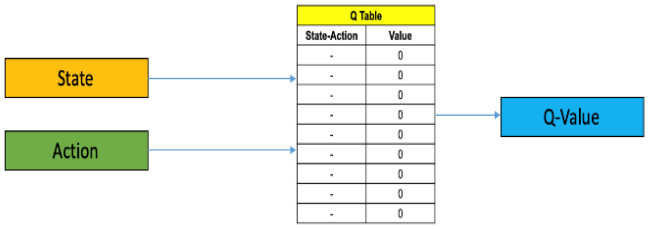
شاین ذکر است، در رویکرد پیشنهادی از رویکرد یادگیری تقویتی و الگوریتم Q-learning جهت یادگیری عامل‌ها استفاده شده است. از این‌رو، قبل از توضیح رویکرد پیشنهادی دو مفهوم رویکرد یادگیری تقویتی و الگوریتم Q-learning توضیح داده شده است.

رویکرد یادگیری تقویتی: یادگیری تقویتی مبتنی بر پاداش می‌باشد. این نوع یادگیری برای محیط‌هایی که عامل هیچ شناختی نسبت به آنها ندارد، مورد استفاده قرار می‌گیرد [۱۵].

الگوریتم Q-learning: یک الگوریتم یادگیری تقویتی مبتنی بر ارزش است که با یادگیری یک تابع Q-value سیاست مشخص را برای انجام حرکات مختلف در وضعیت‌های گوناگون دنبال می‌کند [۱۶]. این الگوریتم دارای یک جدول Q است که سطرهای نشان‌دهنده مکان عامل در محیط و ستون‌های آن نشان‌دهنده اقداماتی است که عامل در هر مکان می‌تواند انجام دهد. یکی از نقاط قوت این روش، توانایی یادگیری تابع مذکور بدون داشتن مدل معینی از محیط است. شمای کلی این الگوریتم در شکل ۳ نشان داده شده است [۱۷]. با این الگوریتم و طبق فرمول ۱ مقادیر Q-value (مقادیری که درون جدول هستند) بروزرسانی می‌گردد.

$$New\ Q(s,a) = Q(s,a) + \alpha[R(s,a) + \gamma \max Q'(s',a') - Q(s,a)] \quad (۱)$$

New Q(s, a) بیانگر Q-value جدید برای آن حالت و اقدام، Q(s, a) بیانگر Q-value کنونی، α بیانگر نرخ یادگیری؛ در هر گام به چه میزان در مسیر حداکثری تابع گام برداشته شود. R(s, a) بیانگر پاداش برای انجام آن اقدام در همان حالت، γ بیانگر نرخ تخفیف؛ چه میزان به بیشترین مقدار Q-value مکان جدید اهمیت می‌دهیم. Max Q'(s', a') بیانگر حداقل پاداش آینده پیش‌بینی شده با توجه به 's جدید و همه اعمال ممکن در حالت جدید است.



(شکل-۳): شمای جدول Q

به وسیله الگوریتم پیشنهادی عامل‌ها در محیط بازی آموزش داده می‌شوند.

آموزش عامل‌ها به این صورت است که در ابتدا برای هر کدام از آنها یک جدول Q درنظر گرفته می‌شود و سپس از نرخی به نام اپسیلون استفاده می‌گردد. نرخ اپسیلون به دلیل ناشناخته بودن محیط برای عامل استفاده می‌شود بنابراین، عامل در ابتدا به صورت تصادفی بازی می‌کند. سپس مقادیر جدول Q بروزرسانی می‌گردد و نرخ اپسیلون در طی هر بار تکرار بازی کاهش می‌یابد و هرگاه این مقدار در کمترین میزان خود بود عامل‌ها برای بازی از جدول Q استفاده می‌کنند. بنابراین، در مقاله از یک تابع تولید عدد تصادفی که عددی بین صفر و یک را انتخاب می‌کند، استفاده می‌شود تا خروجی حاصل از این تابع تصادفی در هر بار اجرا با نرخ اپسیلون مقایسه شود و اگر نرخ اپسیلون بیشتر از مقدار تصادفی بود حرکت بعدی نیز تصادفی انتخاب شود و اگر کمتر بود حرکت بعدی از روی جدول Q انتخاب شود.

لازم به ذکر است، در هر بار انتخاب حرکت بعدی نتیجه این حرکت یک پاداش یا هزینه می‌باشد. پاداش یا هزینه مکان فعلی برای محاسبه مقدار Q خانه قبلی طبق فرمول ۱ محاسبه و جدول Q بروز رسانی می‌گردد. در انتهای هر بار اجرای این بازی مجموع پاداش برای هر عامل محاسبه شده و به عنوان نتایج حاصل از این بازی که عامل، پاداشی جهت همکاری دریافت کرده یا خیر در ارزیابی استفاده می‌گردد.

از سوی دیگر، در الگوریتم پیشنهادی سعی می‌شود مانع بازی کردن عامل‌ها در نقطه تعادل نش شود و میزان همکاری عامل‌ها افزایش یابد. در خلال این فرایند تلاش می‌شود تا به عامل‌ها آموخته شود که اگر رقیب همکاری نمود، عامل هم همکاری کند و در غیر اینصورت عامل عدم همکاری را داشته باشد.

۴-۱- تولید سیاست‌های اولیه در محیط آفلاین:

در گذشته برخی از محققان، برای تولید سیاست‌ها با درجه‌های مختلف همکاری، از رویکردهای تغییر پارامترهای کلیدی محیط‌ها استفاده می‌کردند. به عنوان مثال لیبو و همکارانش تأثیر درجه فراوانی منابع محیط را در تمایل به افزایش همکاری در بازی‌های معضلات اجتماعی تکراری بررسی می‌کنند که در آن عامل‌ها برای منابع محدود با یکدیگر رقابت می‌کنند و مشخص گردید با استفاده از الگوریتم Q-learning می‌توان به عامل‌ها برای همکاری بیشتر انگیزه داد [۱۱].

در این مقاله، برای هر عامل یک جدول Q در نظر گرفته می‌شود و در ازای هر بار بازی که در انتهای آن به همکاری یا عدم همکاری منجر می‌گردد؛ عامل پاداش یا هزینه‌ای دریافت می‌کند. پاداش‌های حاصل از آن با استفاده از سیاست JAC در نظر گرفته شده است. در ادامه به توضیح JAC پرداخته شده است.

رفتار این سیاست به این گونه است که از یک تابع ارزش برای دریافت پاداش استفاده می‌گردد و JAC، پاداش هر عامل را بر اساس ضریبی از پاداش‌های مجموع عامل‌ها محاسبه می‌کند و به عنوان پاداش کلی در نظر گرفته و به همه عامل‌ها همان پاداش را می‌دهد؛ اگر عاملی همکاری و دیگری عدم همکاری را انتخاب کند، مجموع پاداش‌های آنها کاهش می‌یابد و اگر هر دو همکاری کنند، پاداش‌ها افزایش یافته و برای هر دو عامل محاسبه می‌شود. این فرایند باعث ایجاد رغبت به همکاری و دوری از عدم همکاری می‌گردد؛ زیرا اگر عاملی عدم همکاری را پیشه کند شاید پاداش خود را افزایش دهد اما سهمی نیز از هزینه‌های عامل دیگر در پاداش خود دارد.

فرمول ۲ نشان‌دهنده محاسبه پاداش در سیاست JAC است:

$$r_{total} = \sum_{i \in N} att_i \times r_i \qquad (۲)$$

att_i بیانگر این است که عامل چه میزان به پاداش‌های خود و دیگران اهمیت می‌دهد که مقدار آن بین صفر تا یک به صورت احتمالی می‌باشد؛ r_i بیانگر میزان پاداش هر عامل است.

شایان ذکر است، پاداش‌های حاصل از یک دور بازی کامل را به عنوان درجه همکاری در نظر گرفته شده است و اگر این مقدار زیاد باشد به معنای همکاری بالا و اگر این مقدار کم باشد به معنای همکاری پایین است. همچنین جدول Q به عنوان سیاست بازی ذخیره می‌گردد. این اقدام جهت ذخیره نحوه بازی با پاداش‌های مختلف انجام می‌شود.

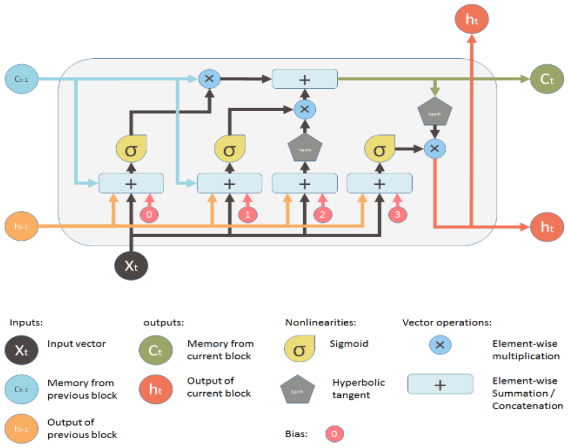
بازیگر منتقد مشترک(JAC) : در مرحله آفلاین برای تولید سیاست‌ها و آموزش سیاست‌های اولیه با درجه‌ها مختلف همکاری از بازیگر منتقد مشترک JAC استفاده می‌شود. JAC به عامل‌ها به صورت یکسان آموزش می‌دهد. بنابراین، هر دو عامل در محیط مشترکی قرار می‌گیرند تا در همان محیط مشترک سیاست‌های اولیه را فرابگیرند. در هنگام استفاده از JAC نیازی به کنترل سیاست-هایی که ممکن است رقیب در مرحله آفلاین استفاده کند، نیست. بدین ترتیب می‌توان از مساله عدم همکاری عامل دیگر جلوگیری کرد.

۴-۲- مرحله آنلاین: استفاده از سیاست‌های تولید شده در مرحله آفلاین با استفاده از LSTM:

در مرحله آفلاین نزدیک به ۱۰۰ جدول Q به همراه پاداش-های آنها که به عنوان درجه همکاری در نظر گرفته‌شده است، ذخیره می‌گردد.

در مرحله آنلاین عامل هیچ نوع آشنایی از رقیبش ندارد. در این مرحله از شبکه عصبی LSTM که از خانواده شبکه‌های عصبی بازگشتی است، استفاده می‌گردد. LSTM سلول ویژه‌ای را معرفی می‌کند که می‌تواند در هنگام شکاف زمانی (وقفه) پردازش داده‌ها را انجام دهد. در واقع LSTM نوعی مدل یا ساختار برای داده‌های ترتیبی است. این نوع مدل شامل بلاکی با نام بلاک LSTM است که در آن گیت‌های مختلفی وجود دارد؛ در این نوع شبکه عصبی به دلیل وجود حافظه، اطلاعات به بلاک‌های موازی دیگر ارسال می-گردد تا رابطه بین داده‌ها حفظ شود. شمای کلی این بلاک در شکل ۴ نمایش داده‌شده است [۱۸].

این نوع ساختار یکی از بهترین روش‌ها برای پیش‌بینی هر فرایندی که به صورت سری و دنباله‌ای رفتار کند، است. حال علت استفاده از LSTM در مقاله، پیش‌بینی درجه همکاری رقیب در تکرارهای بازی است. این رفتار نوعی سری زمانی است که با LSTM قابل پیش‌بینی می‌باشد.



(شکل-۴): شمای شبکه LSTM به همراه توضیح پارامترها

۵- ارزیابی

در این مقاله، از کتابخانه‌هایی همچون Numpy و Pandas استفاده شده است.

در مرحله آنلاین جدول Q که در بخش آفلاین به همراه درجه همکاری آنها ذخیره شده بود، فراخوانی می‌گردد. برای پیش‌بینی درجه همکاری عامل رقیب نیاز به آموزش مدل LSTM از داده-های گذشته آن است و در ارزیابیمان اندازه پنجره ساختار LSTM برابر با ۳، ۵ و ۷ در نظر گرفته‌شده است.

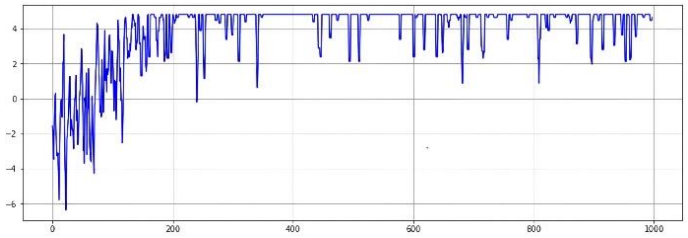
در شروع بازی به صورت آنلاین به علت عدم وجود درجه همکاری، عامل‌ها به صورت تصادفی تا ۱۰ قدم بازی می‌کنند. سپس درجه همکاری‌های هر عامل بعد از ۱۰ قدم به صورت ورودی به مدل LSTM داده می‌شود تا برای عامل قدم بعدی را پیش‌بینی کند. در اصل عامل‌ها در ابتدا برای بازی با رقیب نیاز به انتخاب جدول Q دارند و در این قسمت است که نیاز به LSTM احساس می‌شود، یعنی طبق درجه همکاری‌های گذشته، عامل درجه همکاری‌های چند گام گذشته رقیب را مشاهده می‌کند و نتیجه می‌گیرد که با آن همکاری کند یا خیر. از این‌رو، درجه همکاری‌هایی که از ۱۰ گام اول بدست آمده است، به صورت ورودی به مدل LSTM داده می‌شود و مدل LSTM برای عامل درجه همکاری کنونی رقیب را پیش‌بینی می‌کند و عامل نیز براساس درجه همکاری‌هایی که خود در مرحله آفلاین آموخته است، یکی از سیاست‌هایی که می‌تواند منجر به همکاری یا عدم همکاری باشد را برای مقابله با رقیب انتخاب می‌کند.

۶- نتایج تجربی

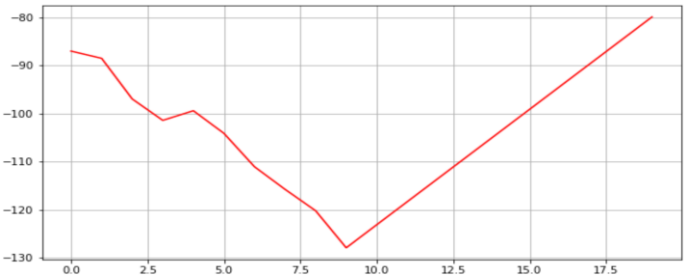
در آزمایش‌ها نشان داده می‌شود که چگونه هر دو عامل در آموزش آفلاین می‌توانند پاداش‌های خود را بهبود بخشند و به بیشترین پاداش دست یابند.

شکل ۵ نشان می‌دهد که عامل‌ها در مجموع، پاداش خود را در هر بار اجرای بازی افزایش داده و به سمت بهترین پاداش حرکت می‌کنند. در شکل محور عمودی بیانگر میزان پاداش و محور افقی بیانگر تعداد تکرار بازی است.

عامل‌ها در مرحله آنلاین بر اساس درجه همکاری که از رقیب پیش‌بینی می‌کنند، حرکت خود را انتخاب می‌کنند. در شکل ۶ روند میزان پاداش حاصل از پیش‌بینی درجه همکاری رقیب برای هر دو عامل نمایش داده‌شده است. در شکل محور افقی بیانگر تعداد دفعات تکرار بازی و محور عمودی بیانگر پاداش حاصل است. همانطور که مشاهده می‌شود در ۱۰ حرکت اول عمل‌های هر عامل به صورت تصادفی انتخاب می‌شود و سیر نزولی دارد، ولی هنگامی که عامل‌ها از LSTM برای پیش‌بینی استفاده می‌کنند روند دریافت پاداش افزایش می‌یابد. همچنین، این نمودار این موضوع را نشان می‌دهد که اگر عاملی عدم همکاری را انتخاب کند، عامل دیگر نیز عدم همکاری و اگر عامل همکاری را انتخاب کند، عامل دیگر نیز همکاری را انتخاب می‌کند.

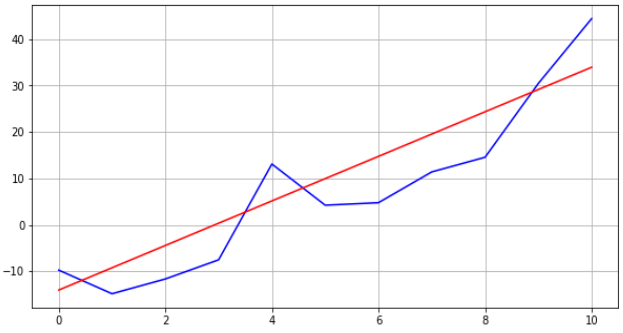


(شکل-۵): بیانگر روند افزایشی دریافت پاداش توسط عامل‌ها است. در محور عمودی میزان پاداش و در محور افقی تعداد تکرار بازی نشان داده شده است.

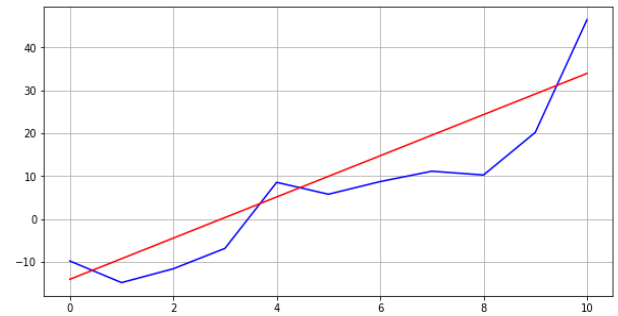


(شکل-۶): روند دریافت پاداش از پیش‌بینی درجه همکاری رقیب را نشان می‌دهد. محور افقی نشان‌دهنده تعداد دفعات تکرار بازی و محور عمودی نشان‌دهنده پاداش حاصل است.

در قسمت‌های(الف) و (ب) شکل ۷ دقت پیش‌بینی درجه همکاری عامل‌ها نشان داده شده است. همانطور که مشاهده می‌شود با افزایش اندازه پنجره در LSTM به مقدار ۷ دقت پیش‌بینی درجه همکاری عامل‌ها کاهش یافته است؛ اما طبق قسمت‌های (الف) و (ب) شکل ۸ با کاهش اندازه پنجره LSTM به ۵ دقت پیش‌بینی درجه همکاری عامل‌ها افزایش یافته است. همچنین طبق قسمت‌های (الف) و (ب) شکل ۹ اندازه پنجره LSTM به ۳ کاهش یافته است ولی دقت پیش-بینی درجه همکاری عامل‌ها افزایش یافته است.



(الف)

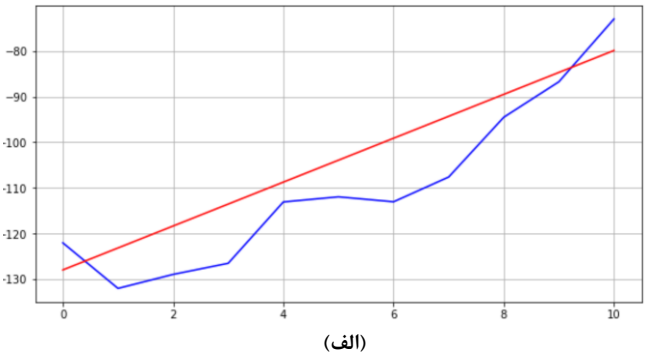


(ب)

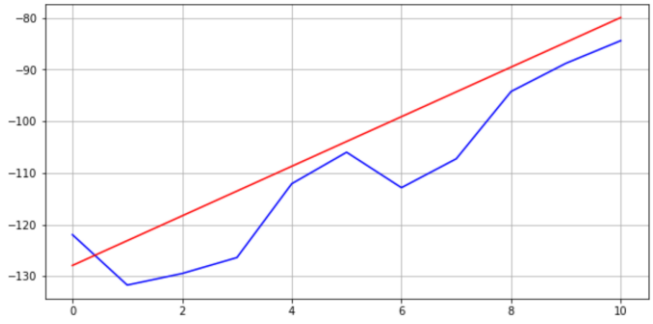
(شکل-۷):

(الف)پیش‌بینی درجه همکاری عامل اول با اندازه پنجره۷

(ب) پیش‌بینی درجه همکاری عامل دوم با اندازه پنجره۷



(الف)

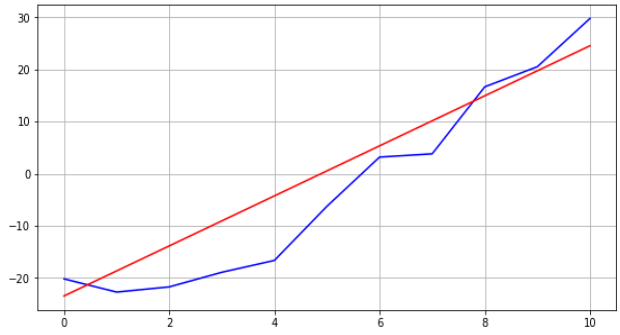


(ب)

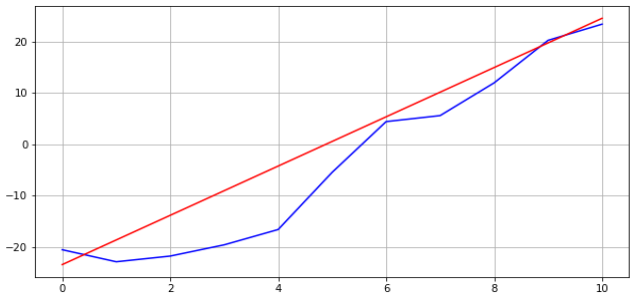
(شکل-۸):

(الف)پیش‌بینی درجه همکاری عامل اول با اندازه پنجره۵

(ب) پیش‌بینی درجه همکاری عامل دوم با اندازه پنجره۵



(الف)



(ب)

(شکل-۹):

(الف)پیش‌بینی درجه همکاری عامل اول با اندازه پنجره۳

(ب) پیش‌بینی درجه همکاری عامل دوم با اندازه پنجره۳

۷- نتیجه‌گیری

یکی از چالش برانگیزترین مسائل در نظریه بازیها، آموزش عامل‌ها در بازی‌هایی با معماری معمای زندانی در جهت همکاری عامل‌ها می‌باشد. ما در این مقاله رویکردی جهت بهبود و هوشمند سازی عامل‌ها برای بازی در مقابل یکدیگر و یا حتی بازی در مقابل چندین عامل متفاوت پیشنهاد کردیم. رویکرد دارای دو مرحله آفلاین و آنلاین می‌باشد. در مرحله آفلاین ما عامل‌ها را با رویکرد JAC آموزش دادیم

و سیاست‌های خاصی را برای مقابله با چندین عامل متفاوت در نظر گرفتیم. برای این سیاست‌ها درجه‌ای در نظر گرفته شد که آن را درجه همکاری می‌نامیم. این درجه همکاری میزان همکاری رقیب را در مقابله با عامل نشان می‌دهد. در این مرحله عامل‌ها می‌آموزند که چگونه همکاری کنند. در مرحله آنلاین عاملی که آموزش دیده است در مقابل رقیب خود به بازی می‌پردازد. مرحله آنلاین به اینگونه است که عامل درجه همکاری رقیب خود را پیش بینی کرده و یک استراتژی برای مقابله با آن انتخاب می‌کند. که برای این منظور از LSTM استفاده شده است. طبق نتایج بدست آمده ما نشان دادیم که رویکرد پیشنهادی ما می‌تواند میزان همکاری عامل‌ها را افزایش دهد.

مراجع

[1] *J. Hu, and M. P. Wellman, "Nash Q-learning for general-sum stochastic games," Journal of machine learning research, vol. 4, no. Nov, pp. 1039-1069, 2003.*

[2] *D. Banerjee, and S. Sen, "Reaching pareto-optimality in prisoner's dilemma using conditional joint action learning," Autonomous Agents and Multi-Agent Systems, vol. 15, no. 1, pp. 91-108, 2007.*

[3] *M. L. Littman, "Markov games as a framework for multi-agent reinforcement learning," Machine learning proceedings 1994, pp. 157-163: Elsevier, 1994.*

[4] *R. S. Sutton, D. A. McAllester, S. P. Singh et al., "Policy gradient methods for reinforcement learning with function approximation." pp. 1057-1063.*

[5] *D. B. Fogel, "Evolving behaviors in the iterated prisoner's dilemma," Evolutionary Computation, vol. 1, no. 1, pp. 77-97, 1993.*

[6] *L. Busoniu, R. Babuska, and B. De Schutter, "A comprehensive survey of multiagent reinforcement learning," IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 38, no. 2, pp. 156-172, 2008.*

[7] *J. W. Crandall, and M. A. Goodrich, "Learning to teach and follow in repeated games."*

[8] *S. Damer, and M. L. Gini, "Achieving Cooperation in a Minimally Constrained Environment." pp. 57-62.*

[9] *J. W. Crandall, "Just add Pepper: extending learning algorithms for repeated matrix games to repeated markov games." pp. 399-406.*

[10] *P. Hernandez-Leal, and M. Kaisers, "Towards a fast detection of opponents in repeated stochastic games." pp. 239-257.*

[11] *J. Z. Leibo, V. Zambaldi, M. Lanctot et al., "Multi-agent reinforcement learning in sequential social dilemmas," arXiv preprint arXiv:1702.03037, 2017.*

[12] *M. Kleiman-Weiner, M. K. Ho, J. L. Austerweil et al., "Coordinate to cooperate or compete: abstract goals and joint intentions in social interaction."*

[13] *J. Foerster, G. Farquhar, T. Afouras et al., "Counterfactual multi-agent policy gradients."*

[14] *W. Wang, J. Hao, Y. Wang et al., "Achieving cooperation through deep multiagent reinforcement learning in sequential prisoner's dilemmas." pp. 1-7.*

[15] *L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: A survey," Journal of artificial intelligence research, vol. 4, pp. 237-285, 1996.*

[16] *V. Mnih, K. Kavukcuoglu, D. Silver et al., "Human-level control through deep reinforcement learning," nature, vol. 518, no. 7540, pp. 529-533, 2015.*

[17] *C. J. Watkins, and P. Dayan, "Q-learning," Machine learning, vol. 8, no. 3-4, pp. 279-292, 1992.*

[18] *S. Hochreiter, and J. Schmidhuber, "Long short-term memory," Neural computation, vol. 9, no. 8, pp. 1735-1780, 1997.*

سمیرا فرزانه در حال حاضر دانشجوی کارشناسی ارشد مهندسی کامپیوتر گرایش نرم افزار نوبت روزانه دانشگاه کاشان می‌باشد.

samira.farzaneh@grad.kashanu.ac.ir

فرشته زندی در حال حاضر دانشجوی کارشناسی ارشد مهندسی کامپیوتر گرایش نرم افزار نوبت روزانه دانشگاه کاشان می‌باشد.

f.zandi@grad.kashanu.ac.ir

جواد سلیمی سرتختی در حال حاضر استادیار گروه مهندسی کامپیوتر و از سال ۱۳۹۹ مدیر تحصیلات تکمیلی دانشگاه کاشان می‌باشد. او در سال ۱۳۹۶ مدرک دکترای خود را با معدل ۱۹.۶۱ و رتبه نخست از دانشگاه صنعتی اصفهان اخذ نمود. زمینه‌های تحقیقاتی ایشان یادگیری عمیق، یادگیری تقویتی، تئوری بازی‌ها، طراحی مکانیزم و پردازش زبان طبیعی می‌باشد.

salimi@kashanu.ac.ir

روش ارجاع به مقاله : س. فرزانه، ف. زندی، ج. سرتختی. دستیابی به همکاری از طریق یادگیری تقویتی چند عاملی در معمای زندانی تکرارشونده ، دوفصلنامه محاسبات و سامانه های توزیع شده، سال سوم، شماره دوم، شماره پیاپی ۶، صفحه ۱۲ تا ۲۱،

How to cite: Samira Farzaneh, Fereshteh Zandi, Javad Salimi Sartakhti. Achieving Cooperation Through Multi agent Reinforcement Learning In Iterated Prisoner's Dilemma , Journal of Distributed Computing and Systems(JDCS), Vol 3, Issue 2, Page 12-21, 2021.