ارزیابی‌اتوماتای یادگیربه‌عنوان‌راه‌حلی‌برای بازی‌های هماهنگی مشکل

بهروز معصومی[۱]؛ محمد رضا میبدی[۲]

چکیده

ایجاد هماهنگی یکی از مسائل مهم و اساسی در سیستمهای چندعامله است که توسط محققین بسیاری مورد مطالعه قرار گرفته است. یکی از بسترهای مناسب برای مطالعه مساله هماهنگی در سیستمهای چند عامله، بازی های هماهنگی است. در این مقاله کارایی اتوماتاهای یادگیر برای یادگیری رفتار هماهنگ در سیستمهای چند عامله با اهداف مشترک به خصوص سیستمهایی که هماهنگی با هزینه های بالایی مواجه هستند و به صورت بازی‌های هماهنگی مشکل مدل می شوند مورد بررسی قرار میگیرد. نتایج آزمایشات برروی چند بازی ماتریسی نمونه نشان داده است که اتوماتاهای یادگیر برای ایجاد هماهنگی حتی در سیستمهایی با هزینه های هماهنگی بالا، از عملکرد بهتری از نظر نرخ همگرایی و سرعت یادگیری در مقایسه با یادگیری Q دارا است.

کلمات کلیدی

اتوماتاهای یادگیر، بازی های هماهنگی، سیستمهای چند عامله.

Evaluation of Learning Automata in Solving Mis-coordination Games

Behrooz Masoumi[۱]; Mohammad Reza meybodi[۲]

[۱] Islamic Azad University -Qazvin Branch and Islamic Azad University science & Research Branch

[۲]Soft Computing Laboratory, Department of Computer Engineering and Information Technology Amirkabir University of Technology, Tehran, Iran

ABSTRACT

Coordination is an important issue in multi-agent systems and has been studied by many researchers. A common testbed for studying multi-agent coordination is the repeated cooperative single-stage games. In this paper, the efficiency of learning automata as coordination mechanisms in multi-agent systems with common goal specially, in scenarios with high miscoordination cost will be studied. The results of experimentations conducted on several matrix games have shown the performance of learning automata as a tool for coordination even when mis-coordination cost is high. It is shown that the learning automata approach has superiority over Q-learning in terms of rate of convergence and speed of learning.

KEYWORDS

Multiagent Systems, Coordination games, Learning Automata.

[۱]دانشگاه آزاد اسلامی قزوین، دانشکده مهندسی کامپیوتر دانشگاه آزاد اسلامی علوم و تحقیقات تهران، bmasoumi@Qazviniau.ac.ir

[۲]دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه صنعتی امیرکبیر،تهران، ایران، mmeybodi@aut.ac.ir

۱. مقدمه

یک سیستم چندعامله[۱]، شامل جامعه‌ای از عامل‌هاست که در یک محیط درکنار یکدیگر درحال کار بوده و سعی در انجام کاری خاص و رسیدن به هدفی مشخص دارند[۱]. این عامل‌ها هوشمند و خودمختار بوده، لذا بدون وجود روشی برای ایجاد هماهنگی میان آن‌ها، ممکن است سیستم دچار هرج و مرج شده و از رسیدن به هدف نهایی بازبماند. بسیاری از سیستم‌های چندعامله در قالب تئوری بازی و اکثراً به‌صورت بازی‌های استراتژیک، شامل ماتریس‌های پاداش[۲] برای هر عامل براساس اعمال گروهی[۳] آن‌ها مطرح می‌شوند[۲]. علاوه بر تحلیل سیستم‌های چندعامله با استفاده از تئوری بازی، برخی مسائل رایج از دامنه تئوری بازی نیز به‌عنوان مدل‌هایی از سیستم‌های چندعامله مورد استفاده قرار می‌گیرند.

بازی‌های هماهنگی[۴] یک‌مرحله‌ای تکراری، زمینه مناسبی برای مطالعه درباره مسئله هماهنگی چندعامله هستند. در این بازی‌ها، عامل‌ها دارای اهداف یکسان بوده و با توجه به اعمال جمعی، پاداش دریافت می‌کنند. در هر دور از بازی، هر عامل یک عمل را انتخاب می‌نماید. این اعمال به‌طور هم‌زمان اجرا شده و پاداش متناظر با عمل جمعی انجام‌شده، به عامل‌ها داده می شود. مطالعه بر روی بازی های هماهنگی کمک می کند تا بفهمیم چگونه می توان به نتایج برنده-برنده[۵] (وضعیتی که در آن هر کلیه طرفهای درگیر در آن وضعیت، به سود دست پیدا کنند) برسیم و از درجا زدن در نقاط متعادل غیر دلخواه پرهیز نماییم.

هدف از به کارگیری الگوریتم های یادگیری در بازی‌های هماهنگی آن است که بازیکن‌ها طی چندین دور بازی تکراری، هماهنگی بهینه را یاد بگیرند. برای رسیدن به این هدف می‌توان از یادگیری مستقل[۶] و یا یادگیری عمل جمعی[۷] استفاده نمود[۳]. تفاوت بین این دو روش از تفاوت میان اندازه اطلاعاتی که عامل‌ها می‌توانند درباره بازی به‌دست آورند، ناشی می‌شود. با این‌که در هر دو روش، عامل‌ها می‌توانند پاداش دریافتی را مشاهده نمایند، در روش یادگیری مستقل عامل‌ها اطلاعی از وجود عامل‌های دیگر ندارند. در مقابل، در یادگیری عمل جمعی عامل‌ها اعمال سایرین را نیز مشاهده می‌کنند، بنابراین می‌توانند مدلی از خط‌مشی عامل‌های دیگر را نگهداری کرده و اعمال خود را براساس مدلی که از خط‌مشی عامل‌های دیگر دارند انتخاب نمایند.

اخیرا استفاده از یادگیری تقویتی به عنوان ابزاری برای دستیابی به رفتار هماهنگ به خاطر عمومیت و استحکامی[۸] که دارد مورد توجه قرار گرفته است[۵] [۴]. اتوماتاهای یادگیرنیز در حال حاضر به عنوان ابزاری ارزشمند در طراحی الگوریتمهای یادگیری تقویتی بوده و ویژگیهای خوبی را ارائه داده اند[۷] و[۸]. در[۹] استفاده از اتوماتاهای یادگیر برای بازی های کاملا همکار (نوع خاصی از بازی های هماهنگی که در آنها تمام عامل ها، پاداش یکسانی دریافت می کنند) مورد استفاده قرار گرفت. در [۱۰] استفاده از یادگیری Q برای حل بازی های هماهنگی مشکل مورد استفاده قرار گرفته است.

هدف اصلی این مقاله، بررسی کارایی اتوماتاهای یادگیر در سیستمهایی است که‌درآنها هماهنگی مشکل و با هزینه است. این بررسی در دوحالت: عاملهای یادگیر مستقل و عاملهای یادگیرنده با اعمال جمعی انجام میگیرد و هدف آن است که بررسی شود آیا عاملهای یادگیرنده با کمک اتوماتای یادگیر در سیستمهای چندعامله‌ای که هماهنگی در آنها با هزینه ومشکل است به راه حل بهینه می رسند؟ وآیا تفاوتهایی بین عاملهای یادگیرنده مستقل و یادگیرندهٔ مشترک در این سیستمها وجود دارد؟ نتایج آزمایشات برروی چند بازی نمونه ازجمله مساله هماهنگی مشکل (بازی جرائم و بازی Climbing) مورد بررسی قرار می گیرند. همچنین عملکرد اتوماتای یادگیر در بازی های هماهنگی مشکل که پاداش آنها احتمالاتی است نیز مورد بررسی قرار گرفته است. ادامه مقاله بدین صورت سازماندهی شده است. در بخش ۲ بازی های هماهنگی و راه حل در آنها ارائه شده است. در بخش ۳ اتوماتاهای یادگیر در بخش ۴ به چگونگی بکارگیری اتوماتاهای یادگیر به منظور ایجاد هماهنگی در سیستمهای چند عامله پرداخته میشود و در بخش ۵ نتایج آزمایشها ارائه میگردد. بخش نهایی مقاله نتیجه گیری میباشد.

۲. بازی های هماهنگی وراه حل در آنها

بازی های هماهنگی نوعی خاص از بازی‌های ماتریسی[۹] یک مرحله ای هستند[۲]. بازی‌های ماتریسی یکی از ابتدایی‌ترین انواع بازی‌ها با هر تعداد بازیکن، و به‌طور خاص شامل دو بازیکن هستند که در آن، بازیکن‌ها اعمال خود را از فضای اعمال موجود انتخاب کرده و پاداشی متناسب با اعمال تمام عامل‌ها دریافت می‌نمایند. یک بازی ماتریسی به‌صورت $< n, A_1,..., A_n, R_1,..., R_n >$ تعریف می‌شود که در آن، n تعداد بازیکن‌ها، و A_i و R_i $(i = 1,..., n)$ به‌ترتیب مجموعه متناهی اعمال و تابع پاداش بازیکن i می‌باشندبه عنوان مثال، یکی از بازیهای هماهنگی که بین دو بازیکن انجام میگیرد و توسط بوتیلیر [۳] مطرح شده در شکل (۲) دیده می شود. در این بازی هر بازیکن دارای دو عمل میباشد. بازیکن اول دارای اعمال .a و a_1 و بازیکن دوم دارای اعمال .b و b_1 میباشد. به طور مثال اگر بازیکن اول عمل .a و بازیکن دوم عمل.b انتخاب کند هر یک از آنها پاداش ۱۰ را دریافت میکنند.

	a_0	a_1
b_0	۱۰	۰
b_1	۰	۱۰

شکل ۲: بازی هماهنگی

یکی از مباحثی که در بازی ها مطرح است این است که حالت پایدار بازی ها و راه حل بهینه چیست. یک نتیجه (راه حل) به صورت تعادل نش۱۰ در نظر گرفته می شود اگر استراتژی یک بازیکن، پاسخی بهینه به استراتژی های دیگر بازیکنان باشد[۱۱]. نظریهٔ تعادل‌های نش می‌گوید که در هربازی به فرض آنکه بازی کنان معقولانه استراتژی های خود را انتخاب کرده و به دنبال بدست آوردن حداکثر بهره(سود) از بازی باشند حداقل یک استراتژی برای بدست آوردن بهترین نتیجه برای هر بازیکن قابل انتخاب است که اگر بازیکن راهکار دیگری به غیر از آن را انتخاب کند، نتیجهٔ بهتری بدست نخواهد آورد. بازی های با تعادل نش در سیستمهای چند عامله بسیار مورد توجه هستند. زیرا عاملها نیازی ندارند استراتژی خود را مخفی نگه داشته و منابع را در پیدا کردن استراتژی های دیگر هدر دهند. یک مساله هماهنگی می تواند به عنوان مساله انتخاب عمل بهینه پارتو تعادل نش در یک بازی استراتژیکی بیان شود[۱۲]. (یک عمل جمعی، بهینه پارتو نامیده می‌شود اگر افزایش سود هریک از عامل‌ها به‌خاطر انجام هر عملی بجز آن‌چه که عامل در عمل جمعی بهینه پارتو انجام می‌دهد، کاهش سود یک یا چند عامل دیگر را درپی خواهد داشت).

مثال دیگری از انواع بازی های هماهنگی بازی های با هماهنگی مشکل است که‌توسط بوتیلیر تحت عنوان بازی جرائم(Penalty) و بازیClimbing مطرح گردید[۳]. هدف از طرح بازی Penalty این بود که نشان داده شود که درعامل های مستقل، رسیدن به هماهنگی با وجود جریمه سنگین به چه میزان مشکل است. شکل (۳) ساختار این بازی را نشان می دهد. میزان جریمه توسط پارامتر k مشخص می شود که می تواند عددی منفی بزرگ مثلا ۱۰۰- باشد. این بازی دارای سه نقطه تعادل نش است که عبارتند از $(a_0,b_0),(a_1,b_1),(a_2,b_2)$که با توجه به منفی بودن k عمل (a_2,b_2) جذب شونده می باشد.

	b_0	b_1	b_2
a_0	۱۰	۰	K
a_1	۰	۲	۰
a_2	K	۰	۱۰

شکل۳: بازی Penalty

ساختار بازی Climbing درشکل(۴) نشان داده شده است. هدف از طرح این بازی آن بود که نشان داده شود یادگیرنده های مستقل برای رسیدن به راه حل بهینه در بازی هایی که با جرائم سنگین محافظت شده اند ممکن است دچار مشکل‌گردند. این بازی نیز دارای سه نقطه تعادل نش است که عبارتند از $(a_0,b_0),(a_1,b_1),(a_2,b_2)$. در هردو بازی اخیر راه حل بهینه توسط جرائم سنگین محافظت شده اند که در بازی جریمه دو راه حل بهینه وجود دارد که عامل ها ممکن است در باره یکی از آنها توافق نمایند.

	b_0	b_1	b_2
a_0	۱۱	-۳۰	۰
a_1	-۳۰	۷	۶
a_2	۰	۰	۵

شکل ۴: بازی Climbing

۳. اتوماتاهای یادگیر

اتوماتای یادگیر، ماشینی است که می‌تواند تعدادی متناهی عمل را انجام دهد. هر عمل انتخاب شده توسط یک محیط احتمالی ارزیابی می‌شود و نتیجه ارزیابی در قالب سیگنالی مثبت یا منفی به اتوماتا داده می‌شود و اتوماتا از این پاسخ در انتخاب عمل بعدی تأثیر می‌گیرد. هدف نهایی این است

که اتوماتا یاد بگیرد تا از بین اعمال خود، بهترین عمل را انتخاب کند. بهترین عمل، عملی است که احتمال دریافت پاداش از محیط را به حداکثر برساند[۱۳].

اتوماتای یادگیر با ساختار متغیر را می‌توان توسط چهارتایی $\{\alpha, \beta, p, T\}$ نشان داد که α مجموعه عمل‌های اتوماتا، β مجموعه‌ورودی-های‌اتوماتا، $p = \{p_1, ..., p_r\}$ بردار احتمال‌انتخاب‌هریک‌ازعمل‌هاو $p(n+1)=T[\alpha(n),\beta(n),p(n)]$ الگوریتم یادگیری می‌باشد. الگوریتم زیربر اساس روابط (۱)و(۲) یک نمونه از الگوریتم‌های یادگیری است. فرض می‌کنیم عمل α_i در مرحله nام انتخاب شود.

- پاسخ مطلوب از محیط

$$\begin{aligned} p_i(n+1) &= p_i(n) + a[1 - p_i(n)] \\ p_j(n+1) &= (1-a)\, p_j(n) \qquad \forall j \quad j \neq i \end{aligned} \tag{۱}$$

- پاسخ نامطلوب از محیط

$$\begin{aligned} p_i(n+1) &= (1-b)\, p_i(n) \\ p_j(n+1) &= (b/r-1) + (1-b)\, p_j(n) \quad \forall j \quad j \neq i \end{aligned} \tag{۲}$$

در روابط (۱) و (۲)، a پارامتر پاداش و b پارامتر جریمه می‌باشند. با توجه به مقادیر a و b سه حالت را می‌توان درنظر گرفت: اگر a و b با هم برابر باشند، الگوریتم را L_{RP} ، هنگامی‌که b از a خیلی کوچکتر باشد، الگوریتم را $L_{R\varepsilon P}$ و اگر b مساوی صفر باشد آن را L_{RI} می‌نامیم.

در مدل Q متناظر با عمل α_i، خروجی محیط ممکن است تعداد متناهی از مقادیر اختیار کند. با نرمال سازی مقادیر خروجی، هر مدل Q با مقادیر متناهی از خروجیهای محیط در فاصلهٔ واحد [۰,۱] مشخص می گردد. تعداد این مقادیر خروجی از عملی به عمل دیگر متفاوت است و m_i برای عمل α_i $(i=1,2,.....,\gamma)$ بیان می شود. در مدل S، پاسخها می توانند مقادیری پیوسته در یک فاصله مشخص را اختیار کنند. با نرمال سازی پاسخها، می توان آنها را در فاصله [۰,۱] درنظرگرفت. اگر پاسخ محیط در مدل Q برای عمل α_i با $\beta^i_1, \beta^i_2, ..., \beta^i_{mi}$ مشخص شود که در آن $\beta^i_1 < \beta^i_2 < .... < \beta^i_{mi}$ ، مجموعه نرمال شدهٔ پاسخها $\{\beta^i_j\}$ به شکل $\beta^i_j = \beta^i_j - a / b-a \quad j=1,2,....,m_i$ تعریف می گردد که در آن: $a = \min_i \{\beta^i_j\}$ و $b = \max_i \{\beta^i_{mi}\}$ است. با توجه به ساختار بازی های مطرح شده در بازی هماهنگی اول به راحتی ساختار بازی را می توان به مدل P و دو بازی دیگر را به مدل Q تبدیل نمود. شمای S-LRP برای مدلهای Q و S براساس رابطه (۳) بیان می شود:

اگر عمل α_i در مرحله nام انتخاب شود در این صورت

$$\begin{aligned} p_i(n+1) &= p_i(n) + a.(1-\beta_i(n)).(1-p_i(n)) - a.\beta_i(n).p_i(n) \\ p_j(n+1) &= p_j(n) - a(1-\beta_i(n)).p_j(n) + a.\beta_i(n).\left[\frac{1}{r-1} - p_j(n)\right] - a.(1-\beta_i(n)).p_j(n) \quad \forall j \quad j \neq i \end{aligned} \tag{۳}$$

برای اطلاعات بیشتر در باره اتوماتاهای یادگیر می توان به[۱۴] مراجعه نمود.

۴. استفاده از اتوماتاهای یادگیر برای حل بازی های هماهنگی

برای بکارگیری اتوماتاهای یادگیر به منظور ایجاد هماهنگی بین بازیکنان برای حداکثر کردن دریافتی خود در بازیهای یک مرحله ای دو روش کلی وجود دارد که در ادامه این به دو روش پرداخته میشود.

- روش اتوماتای یادگیر مستقل

در این روش، هر یک از بازیکنان به یک اتوماتای یادگیر با مجموعه اعمال برابر با مجموعه اعمال آن بازیکن مجهز میشود و بردار احتمال انتخاب اعمال آن مقدار دهی اولیه میشود. در هر مرحله از بازی هر یک از بازیکنان یکی از اعمال خود را با استفاده از اتوماتای یادگیرش انتخاب میکند. با توجه به پاداشی که به بازیکنان تعلق میگیرد بردار احتمال انتخاب اعمال هر یک از اتوماتاهای یادگیر بر طبق الگوریتم یادگیری بروز میشود. فرایند انتخاب اعمال توسط بازیکنان و بروز در آوردن بردار احتمالات تا همگرایی اتوماتاهای یادگیر (اگاهی هر بازیکن به عملی که بیشترین پاداش را نصیب او مینماید) ادامه پیدا میکند. برای هر عامل، ماتریس پاداش و عامل دیگر محیط تصادفی برای اتوماتای یادگیر آن عامل است. اگر عاملها از وجود عاملهای

دیگر آگاه نباشند و یا نتوانند اعمال یکدیگر را تشخیص دهند آنگاه این روش برای هماهنگی روشی مناسب است. شکل ۵ شبه کد الگوریتم را نشان می دهد.

- روش اتوماتای یادگیر اعمال جمعی

در این روش یک اتوماتای یادگیر با $|A_1|\times|A_2|$ تعداد عمل مورد استفاده قرار میگیرد. مجموعه اعمال این اتوماتای یادگیر مجموعه جفتهای $(a_1,a_2)\in A_1\times A_2$ میباشد. بطور مثال برای بازی فوق اعمال اتوماتای یادگیر <a.,b.>، <a.,b۱>، <a۱,b.> و<a۱,b۱> میباشند. اتوماتای یادگیر که آنرا یادگیرندهٔ عمل مشترک مینامیم، احتمالات اعمال مشترک را بهنگام سازی می کند. در هر مرحله اتوماتای یادگیر یکی از اعمال خود را انتخاب میکند و سپس با توجه به پاداش دریافتی الگوریتم یادگیری بردار احتمال خود را بروز میکند. اینکار تا همگرایی اتوماتای یادگیر به یکی از اعمالش ادامه پیدا میکند. در این روش ماتریس پاداش، محیط اتوماتای یادگیر میباشد.

Initialize LA۱ ,LA۲;
Normalize *Payoff۱ ,Payoff۲* **into** ***Feedback۱,FeedBack۲***
for iteration = ۱ to M **do**
JointAction =∅
Action۱=LA۱.SelectAction ();
Action۲=LA۲.SelectAction ();
JointAction=Action۱ ∪Action۲;
LA۱.Update (Action۱, *Feedback۱*(JointAction);
LA۲.Update (Action۱, *Feedback۲*(JointAction);
end for;

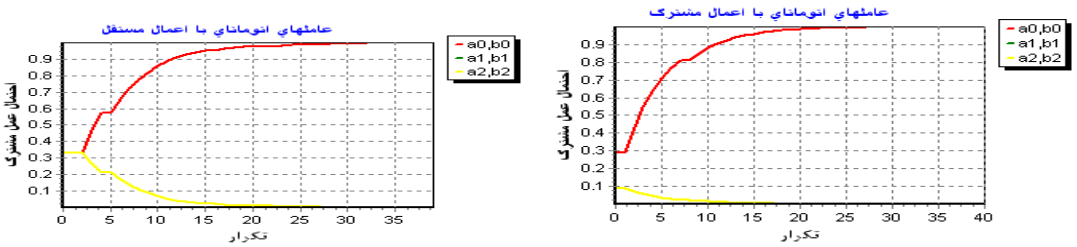
شکل ۵. شبه کد اتوماتای یادگیر

۵. ارزیابی رویکرد اتوماتاهای یادگیر درهماهنگی سیستمهای چند عامله

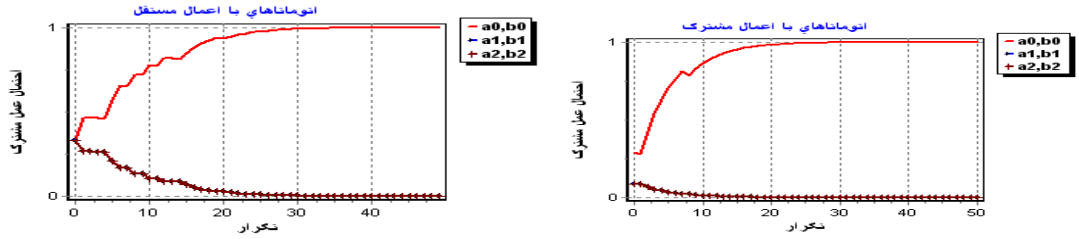
در این بخش‌کارایی اتوماتاهای یادگیر به منظور ایجاد هماهنگی در سیستمهای چند عامله مورد بررسی قرار میگیرد. درارزیابی عملکرد اتوماتاهای یادگیر بایستی بتوان به سوالات زیر پاسخ داد. آیا روش پیشنهادی در بازیهای هماهنگی با هزینه بالا به یک تعادل نش همگرا میشود؟ آیا اتوماتاهای یادگیر در بازی های هماهنگی که با جریمه زیاد محافظت شده اند همگرایی را تضمین می کند؟ آیا در یادگیرنده های مستقل و مشترک تفاوتی وجود دارد؟ نقش پارامتر جریمه در اتوماتاها چگونه است و نیز آیا در بازی های همکاری احتمالاتی نیز اتوماتاها همگرایی را تضمین می کنند؟ برای این منظور آزمایش های مختلفی با توجه به سوالات مطرح شده انجام و نتایج ارائه می شوند.

۱.۵ آزمایش سری اول: بررسی رفتار اتوماتای یادگیر در پیدا کردن راه حل تعادل نش

هدف اصلی این آزمایش بررسی رفتار اتوماتای یادگیر برای پیدا کردن راه حل بهینه تعادل نش در بازی های هماهنگی است. لذا دو عامل یادگیرنده اتوماتای یادگیر با هم بازی کرده و با توجه به پاسخ محیط که می تواند از نوع Q یا S باشد هریک احتمالات انجام اعمال خود را به روز می‌رسانند. در بازی های جرائم و Climbing پس از نرمال سازی از مدل Q برای بروز رسانی احتمالات انجام عمل اتوماتا استفاده می شود. در ابتدا تمامی احتمالات بطور یکسان در نظر گرفته شده است. در صورت تغییر احتمالات اولیه اتوماتاها در انتخاب نقطه موازنه به وِیژه در بازِی Penalty تاثیر گذار است. در بازی Climbing که دارای سه نقطه تعادل است که تنها (a_0,b_0) بهینه پارتو است که دیده می شود در این نقطه همگرایی حاصل می شود. شکل (۶) نتایج بدست آمده را نشان می دهد.



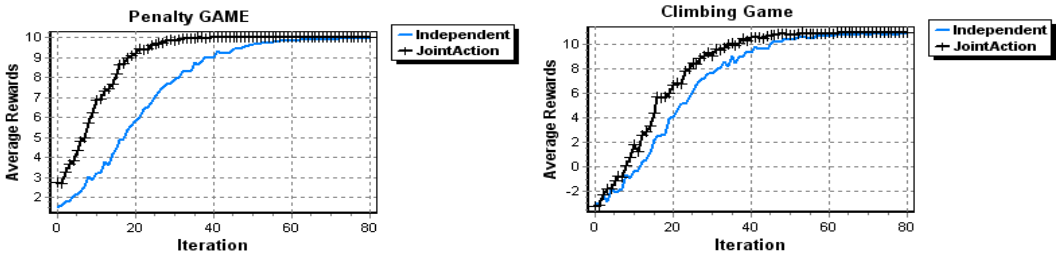
الف) بازی penalty



ب) بازی climbing

شکل ۶. بررسی رفتار اتوماتاهای یادگیری‌ردر بازی های هماهنگی پارامتر یادگیری a=۰.۲ و پارامتر b=۰.۰۰۳

برای بررسی دقیقتر، آزمایش دیگری که صورت گرفت مقایسه ای از نظر میانگین پاداش بدست آمده در ۱۰۰۰ آزمایش و برای هر آزمایش ۸۰ تکرار بود که برای عاملهای مستقل و عاملهای اعمال مشترک در بازی های مطرح انجام گردید. همانطور که در شکل(۷) دیده می شود روش اعمال مشترک نسبت به روش اعمال مستقل از رفتار و پایداری بهتری برخوردار است.



a=۰.۲ و پارامتر b=۰.۰۰۳

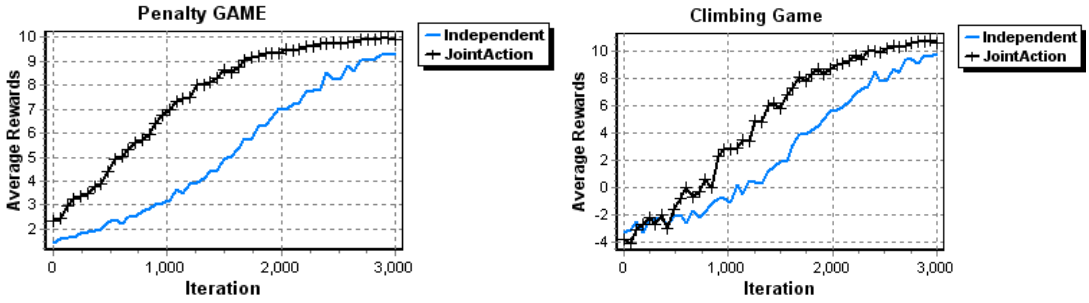
شکل ۷. مقایسه میانگین پاداش اخذ شده در دو حالت انوماتاهای مستقل و با اعمال مشترک در ۱۰۰۰آزمایش در بازیClimbing

۲.۵. آزمایشهای سری دوم، بررسی نقش پارامترهای یادگیری و پارامتر k

در آزمایش‌هایسری‌دوم نقش پارامترهای مختلف a,b و تاثیر آنها در سرعت همگرایی، نقش پارامتر k به عنوان پارامتر جریمه در دو بازی penalty و climbing و نیز رفتار اتوماتاهای مستقل و اعمال مشترک در این بازی های هماهنگی بررسی شده است. دربازی penaltyدرحالت عاملهای مستقل با کم کردن مقدار a به نزدیک ۰.۰۲ و ثابت نگهداشتن b همگرایی به شدت کاهش می یابد بطوریکه فقط با صفر نمودن مقدار b، همگرایی در بیش از ۳۰۰۰ تکرار حاصل می شود. شکل ۸ نتایج را برای هر دو بازی نشان می دهد.

جدول (۱) نقش پارامتریادگیری را در حالت های مختلف نشان می دهد.در حالت اول تعداد تکرار مورد نیاز برای همگرایی، حالت دوم در صد همگرایی به نقطه تعادل بهینه در حالت K= -۱۰۰ برای بازی جرائم و حالت سوم برای بازی Climbing در صد همگرایی به تعادل بهینه را نشان

می دهد. در آزمایش بعدی که به جهت نقش k جریمه مطرح است با توجه به نرمال سازی انجام گرفته می بینیم میزان پاداش در ابتدا کم و پس از ۲۰ تکرار به مقدار واقعی نزدیک می شود و از این لحظه به بعد میزان جریمه تاثیر قابل توجهی در نتایج نداشته و از رو نقش اتوماتا در اینجا بهتر خودش را نشان می دهد. شکل (۹) نمودار های مربوطه را برای بازی penalty در سه حالت بازای مقادیر مختلف پارامتر جریمه k=-۱۰۰ و k=-۵۰ و k=۰ نشان می دهد.همانطور که ازشکل پیداست با کاهش پارامترk میزان میانگین پاداش دریافتی تغییر می کند.



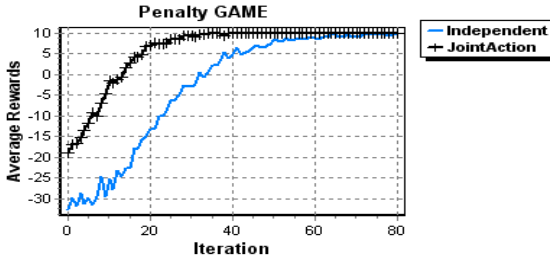
شکل۸. بررسی رفتار اتوماتاها در بازی های هماهنگی با تغییر پارامترهای a,b ۲ a=۰.۰۰ و پارامتر b=۰ و k=۰

جدول ۱. مقایسه پارامتر یادگیری در همگرایی بازی های جرائم و Climbing

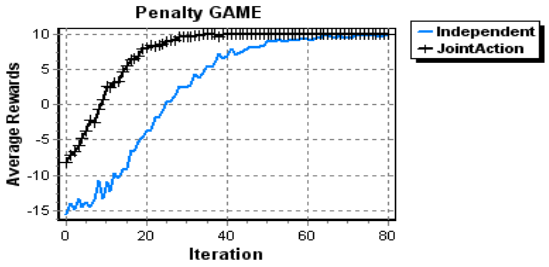
بازی جرائم k=۰ و b=۰	
تعداد تکرارمورد نیازبرای همگرایی به یکی از نقاط تعادل نش(نه لزوما بهینه)	a
۲۲۰۰	۰.۰۰۱
۴۲۰	۰.۰۰۵
۲۲۰	۰.۰۱
۲۲	۰.۱
۸	۰.۴

بازی جرائم K=-۱۰۰ در ۵۰۰۰ تکرار		
درصد همگرایی a۲,b۲	درصد همگرایی به a۰,b۰	a
۱۹	۱۹	۰.۰۰۱
۳۷	۳۷	۰.۰۰۵
۴۳	۴۳	۰.۰۱
۲۲	۲۲	۰.۱
۲۰	۲۰	۰.۴

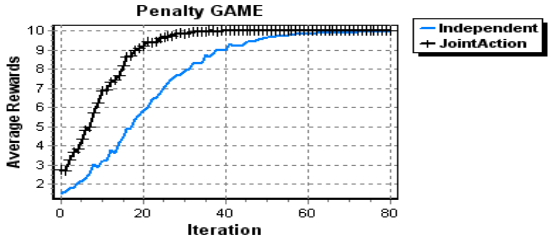
بازی Climbing		
درصد همگرایی تعادل نش	درصد همگرایی به نقطه تعادل بهینه	a
۱۷	۹	۰.۰۱
۳۲	۲۱	۰.۰۵
۳۹	۲۵	۰. ۱
۳۳	۲۲	۰.۲۵
۳۲	۱۶	۰.۵



K=-۱۰۰



K=-۵۰



K=۰

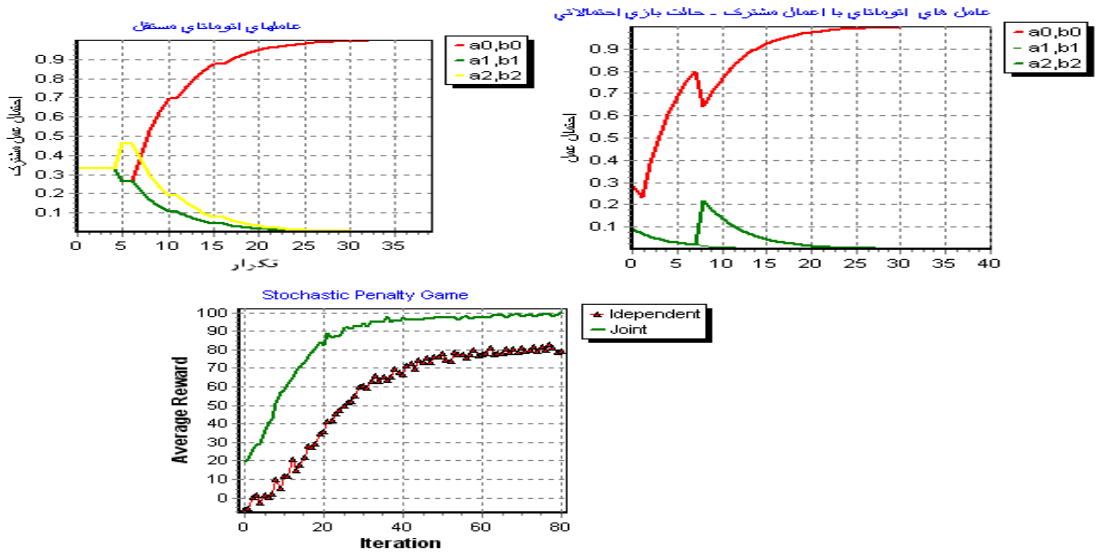
شکل۹. بررسی‌رفتار اتوماتا ازنظرنقش پارامترجری‌مه در بازی Penalty

۳.۵. آزمایش های سری‌سوم ،بررسی عملکرد اتوماتای یادگیر دربازیهای احتمالاتی

در این سری آزمایش، بررسی رفتار اتوماتاهادر بازیهای هماهنگی احتمالاتی انجام شده است. احتمالاتی بودن به این معناست که ماتریس پاداش در هر مرحله می تواند با احتمالاتی تغییر نماید. یکی از این بازی ها بازی احتمالاتی climbing است که در شکل(۱۰) دیده می شود[۱۰]. همانطور که در شکل دیده می شود ماتریس دارای دو مقدار در هر درایه است که با احتمال برابر ۰/۵ میتواند یکی از مقادیر انتخاب شود. در آزمایش انجام گرفته در دوحالت مختلف اعمال مستقل و مشترک با همان پارامترهای قبلی دیده شد که اتوماتای یادگیرنده در حالت مستقل مشابه اتوماتاهای مشترک همانند قبل در مراحل محدود به (a.,b.)همگرا گردید. نتایج بررسی ها در شکل (۱۱) دیده می‌شود.

	b_0	b_1	b_2
a_0	۸۰/۱۲۰	۳۰/۳۰	-۸۰/-۱۲۰
a_1	-۳۰/۳۰	۰/۴۰	-۳۰/۳۰
a_2	-۸۰/۱۲۰	۳۰/۳۰	۸۰/۱۲۰

شکل۱۰- بازی Stochastic Penalty Game



شکل۱۱- بررسی رفتار اتوماتاها در بازی های هماهنگی احتمالاتی *Climbing* با پارامتر a=۰.۲,b=۰.۰۰۲در دوحالت مستقل و عمل مشترک

۶. نتیجه‌گیری

در این مقاله استفاده ازاتوماتای یادگیر برای یادگیری رفتار هماهنگ در بازی های هماهنگی مشکل و احتمالاتی مورد بررسی قرار گرفت. با توجه به آزمایش های صورت گرفته و بررسی آنها دیده شد که اتوماتاهای یادگیر با توجه به سادگی محاسباتی همان نتایج الگوریتم های یادگیری تقویتی را می توانند تولید کنند و حتی در بازی های با هماهنگی مشکل ر فتار بهتری را از خود بروز می دهند. مهمترین پارامتر در سرعت همگرایی همان پارامتر های پاداش و جریمه a,b است که تاثیر به سزایی در مستقیم بودن مسیر همگرایی دارند.افزایش مقدار a باعث روان تر شدن حرکت به سمت نقطه موازنه می شود. با توجه به نتایج، روش LRI همگرایی را برای مساله تضمین می نماید. میزان جریمه با توجه به نرمال سازی ماتریس پاداش تاثیر زیادی در همگرایی اتوماتاها ایجاد نکرد. با توجه به آزمایش های صورت گرفته می بینیم اتوماتاهای یادگیر با وجود سادگی محاسباتی مدل مناسب یادگیری برای ایجاد هماهنگی در سیستمهای چند عامله محسوب می شوند.

۷. مراجع

[۱] G. Weiss; *Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence*, Cambridge, MA: MIT Press, ۱۹۹۹.

[۲] Y. Shoham; *Multiagent Systems: Algorithmic Game Theoretic and Logical Foundations*, Cambridge University Press, ۲۰۰۹.

[۳] C. Claus and C. Boutilier; "*The Dynamics of Reinforcement Learning in Cooperative Multiagent Systems* ", In Proceedings of the Fifteenth National Conference on Artificial Intelligence, pp. ۷۴۶–۷۵۲, ۱۹۹۸.

[۴] M. Lauer and M. Riedmiller; "*An Algorithm for Distributed Reinforcement Learning in Cooperative Multiagent Systems*", Proceedings of the Seventeenth International Conference in Machine Learning, ۲۰۰۰.

[۵] J. R. Kok; M. T. Spaan and N. Vlassis; "*Multi-Robot Decision Making using Coordination Graphs*", In Proc. ۱۱th Int. Conf. on Advanced Robotics, Coimbra, Portugal, ۲۰۰۳.

[۶] Kapetanakis and Kudenko; "*Reinforcement Learning of Coordination in Heterogeneous Cooperative Multi-agent Systems*", Proceedings of the Fourth AISB Symposium on Adaptive Agents and Multi-agent Systems (AISB/AAMAS-۲۰۰۴),Leeds, UK, ۲۰۰۴.

[۷] M. R. Khojasteh, and M. R. Meybodi; "*Evaluating Learning Automata as a Model for Cooperation in Complex Multi-Agent Domains*", Lecture Notes in Artificial Intelligence, Springer Verlag, LNAI ۴۴۳۴, pp. ۴۰۹-۴۱۶, ۲۰۰۷.

[۸] A. Now´e, K. Verbeeck, and M. Peeters; "*Learning Automata as a basis for Multi-agent Reinforcement Learning*", *Lecture Notes in Computer Science*", vol. ۳۸۹۸, pp. ۷۱–۸۵, ۲۰۰۶.

[۹] M. R. Shirazi and M.R. Meybodi; "Application of Learning Automata to Cooperation in Multi-Agent Systems", Proceedings of First International Conference on Information and Knowledge Technology (IKT۲۰۰۳), pp. ۳۳۸-۳۴۹, ۲۰۰۳.

[۱۰] M. Carpenter and D. Kudenko; "*Baselines for Joint-Action Reinforcement Learning of Coordination in Cooperative Multi-agent Systems*", Adaptive Agents and Multi-Agent Systems , pp. ۵۵-۷۲, ۲۰۰۵.

[۱۱] J. Nash; *Non-cooperative Games*, Annals of Mathematics, ۵۴, pp. ۲۸۶–۲۹۵, ۱۹۵۱.

[۱۲] N. Vlassis; *A Concise Introduction to Multiagent Systems and Distributed AI*, Informatics Institute, University of Amsterdam, http://www.science.uva.nl/~vlassis/cimasdai, ۲۰۰۳.

[۱۳] K. Narendra and M.A.L. Thathachar; *Learning Automata: An Introduction*, Prentice Hall, ۱۹۸۹.

[۱۴] M. A. L. Thathachar and Sastry; "*Varieties of Learning Automata: An Overview*", IEEE Transaction on Systems, Man, and Cybernetics-Part B: Cybernetics, vol. ۳۲, no. ۶. pp. ۷۱۱-۷۲۲, ۲۰۰۲.

زیر نویس ها

[۱] Multi-Agent System

[۲] Payoff Matrices

[۳] Joint Action

[۴] Coordination Games

[۵] Win-Win

[۶] Independent Lerners

[۷] Joint Action

[۸] Robustness

[۹] Matrix Games

[۱۰] Nash Equilibrium