

استفاده از شبکه اتوماتاهای یادگیر در حل مسائل تصمیم‌گیری غیرمتتمرکز چند عامله

بهروز معصومی^۱، محمد رضا میبدی^۲

^۱ دانشکده مهندسی برق و رایانه، دانشگاه آزاد اسلامی قزوین و دانشگاه آزاد اسلامی واحد علوم و تحقیقات، تهران، bmasoumi@Qazviniau.ac.ir

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

چکیده- اتوماتاهای یادگیر در حال حاضر به عنوان ابزاری ارزشمند در طراحی الگوریتمهای یادگیری تقویتی بوده و حتی در سیستمهایی که از چندین اتوماتای یادگیر استفاده می نمایند نیز ویژگیهای خوبی را ارائه داده اند. اتوماتاهای یادگیر در مسائل تصمیم‌گیری های غیرمتتمرکز قادر به کنترل زنجیره های مارکوف محدود و حتی حل بازی های مارکوفی نیز هستند. این بازیها توسعه ای از فرآیندهای تصادفی مارکوف با چندین عامل و بازی های ماتریسی با چندین حالت بوده و هدف هر عامل پیدا کردن سیاست بهینه ای است که امید ریاضی مجموع کاهش یافته پاداشها را بیشینه نماید. در این مقاله روشی مبتنی بر شبکه اتوماتاهای یادگیر و مفهوم آنتروپی برای حل بازی های مارکوفی در شرایط ارگودیک به خصوص مسائل تصمیم‌گیری مارکوف چند عامله (*mmdp*) برای یافتن خطمشی بهینه پیشنهاد شده است. در روش پیشنهادی، در هر حالت محیط به ازای هر عامل یک اتوماتای یادگیر قرار گرفته شده است. اعمال انتخابی اتوماتای یادگیر با توجه به پاداش تجمعی دریافتی اتوماتاهای یادگیر و یا آنتروپی بردار احتمال اعمال اتوماتای یادگیر حالت جدید، پاداش یا جریمه دریافت می کنند. نتایج آزمایشها انجام گرفته نشان داده اند که الگوریتم ارائه شده از کارایی مناسبی در سرعت همگرایی (رسیدن به راه حل بهینه) برخوردار است.

کلید واژه-، آنتروپی، اتوماتاهای یادگیر، بازی های مارکوفی، یادگیری تقویتی چند عامله.

بازی مارکوفی با یک حالت بصورت یک بازی نرمال تکراری^۳ در تئوری بازی ها شناخته می شود و هربازی مارکوفی با یک عامل بصورت یک فرآیند تصمیم‌گیری مارکوفی^۴ (*MDP*) است^[۵].

در بازی های مارکوفی بر خلاف *MDP* ها در حالت وجود پاداش های متفاوت پیدا کردن راه حل بهینه ای که مستقل از عاملهای دیگر باشد و بتواند امید ریاضی مجموع کاهش یافته پاداشها را برای همه عاملها بیشینه نماید امکان پذیر نیست. لذا نقاط تعادل در بازی جستجو می شوند. در صورتیکه پاداش یکسان برای عاملها در نظر گرفته شود شود آنها را کاملا همکارانه^۵ گویند و به آن فرآیندهای تصادفی مارکوف چندعامله^۶ (*MMDP*) نیز می گویند. در *MMDP* ها عاملها بایستی یادگیرند که چگونه بر روی سیاست بهینه مشترک توافق نمایند. در بازی های مارکوفی کلی^۷ بایستی سیاست متعادل جستجو شود، وضعیتی که هیچ عاملی به تنها یی نمی تواند برای بهبود پاداشش سیاستش را تغییر دهد تا زمانیکه تمام عاملهای دیگر سیاستشان را ثابت نگه می دارند. در^[۴] نشان داده شده است که شبکه ای از اتوماتاهای یادگیر قادر به رسیدن به استراتژی های تعادل در بازی های مارکوفی با فرض ارگودیک بودن آنها خواهد بود. در^[۶] یک راه حل کلی برای بازی های مارکوفی با مجموع کلی با

۱- مقدمه

اتوماتاهای یادگیر، ابزارهای تصمیم‌گیری تطبیقی و یادگیرنده های مستقل هستند که به عنوان الگوریتمهای جدید یادگیری تقویتی چند عامله بسیار مناسب و مفیدند^[۱]. یکی از کاربردهای نظریه اتوماتاهای یادگیر این است که در مجموعه تصمیم‌گیری های غیرمتتمرکز قادر به کنترل زنجیره های مارکوف محدود با احتمالات گذار و پاداش های نامشخص هستند^[۲]. اخیرا این نتایج برای چارچوب بازی های مارکوفی^۸ [۳] به عنوان توسعه ای از مسائل تصمیم‌گیری تک عامله در مسائل تصمیم‌گیری چند عامله توزیع شده مورد استفاده قرار گرفته اند^[۴]. این بازی ها برای مدل سازی سیستمهای چند عامله بسیار مورد استفاده قرار گرفته و بصورت توسعه ای از فرآیندهای تصادفی مارکوف با چندین عامل و بازی های ماتریسی با چندین حالت هستند که در آن اعمال انتخابی در هر حالت نتیجه ترکیب اعمال مستقل انجام شده تمام عاملهای است و گذار از حالتی با حالت دیگر وابسته به این اعمال گروهی است. علاوه بر این هر عامل تابع پاداش خاص خودش را دارد. هر

ارگودیک باشد، می توان نشان داد که بهترین استراتژی در هر حالت استراتژی محض بوده و مستقل از زمانی است که هر حالت تصرف می شود [۲]. یک زنجیره مارکوف مانند $\{x_n, n \geq 0\}$ را ارگودیک گویند وقتیکه توزیع زنجیره به یک توزیع محدود $\pi(\alpha) = (\pi_1(\alpha), \dots, \pi_n(\alpha))$ برای $i, \forall i, \pi_i(\alpha) \geq 0, n \rightarrow \infty$ با توجه به این موضوع معادله ۱ را می توان بصورت زیر نوشت :

$$J(\alpha) = \sum_{i=1}^N \pi_i(\alpha) \sum_{j=1}^N T^{ij}(\alpha) R^{ij}(\alpha) \quad (2)$$

۲-۲- تعریف بازی مارکوف

بازی های مارکوف تعیین فرآیندهای تصادفی مارکوف به حالت چند عامله است و بصورت چند تابی $\langle n, S, A_{1..n}, T, R_{1..n} \rangle$ بیان می شود که در آن n تعداد عامل ها، S مجموعه حالات کل محیط، A_i مجموعه اعمال هر عامل i (در فضای اعمال مشترک $S \times A \times S \rightarrow [0,1]$)، $T(A_1 \times A_2 \times \dots \times A_n \rightarrow S \times A \times S)$ تابع انتقال باشد برای عامل i ام $S \times A \rightarrow \mathbb{R}$ است [۵]. با توجه به رابطه (۲) امید ریاضی مجموع کاهش یافته پاداشها برای هر عامل k $J_k(\alpha)$ با همان فرضیات ارگودیک بودن بصورت زیر تعریف می گردد :

$$J_k(\alpha) = \sum_{i=1}^N \pi_i(\alpha) \sum_{j=1}^N T_k^{ij}(\alpha) R_k^{ij}(\alpha) \quad (3)$$

در حالتی که هر عامل پاداش مختلفی را دارد و به آن بازی های مارکوف کلی گویند، پیدا کردن سیاست بهینه برای تمام عاملها کاری مشکل تا حد غیر ممکن خواهد بود لذا بجا ای آن نقاط تعادل در بازی جستجو می شوند، وضعیتی که هیچ عاملی به تنهایی نمی تواند تا زمانیکه تمام عاملهای دیگر سیاستشان را ثابت نگه می دارند برای بهبود پاداش سیاستش را تغییر دهد. در بازی های مارکوفی دارای ویژگی ارگودیک نشان داده شده است که حداقل یک نقطه موازن در سیاستهای ایستارا دارند [۷].

۳- اتماتاهای یادگیر

اتوماتاهای یادگیر یکی از مدلهای یادگیری تقویتی است که در آن یک اتماتا یک عمل بهینه را با توجه به اعمال گذشته و بازخورد محیط فرا می گیرد. هدف نهایی این است که اتماتا یاد بگیرد تا از بین اعمال خود، بهترین عمل را انتخاب کند. بهترین عمل، عملی است که احتمال دریافت پاداش از محیط را به حداقل برساند. کارکرد اتماتای یادگیر در تعامل با محیط، در شکل ۱ مشاهده می شود.

استفاده از اتماتای یادگیر بدون در نظر گرفتن مدل ارگودیک ارائه گردیده است.

در این مقاله با استفاده از اتماتاهای یادگیر روشنی بهبود یافته برای تصمیم گیری های غیر متصرف کردن بازی های مارکوفی بادر نظر گرفتن پاداش میانی که بر پایه آنتروپی بردار احتمالات اعمال اتماتاهای یادگیر تعیین می گردد ارائه می گردد و نشان داده شده است که شبکه اتماتاهای یادگیر که از پاداش میانی استفاده می کنند قادر به بهبود سرعت پیدا کردن تعادل در بازی های مارکوفی هستند. در ادامه سازماندهی این مقاله بصورت زیر است، در بخش ۲ به تعریف و بررسی MDP و بازی های مارکوفی و همچنین مفهوم راه حل در آنها پرداخته شده است. در بخش ۳ مفهوم اتماتای یادگیر و استفاده از آن در حل بازی های مارکوفی و ارائه راه حل پیشنهادی ارائه گردیده است. در بخش ۴ مثالی از بازی های هماهنگی مارکوفی با دو عامل و چهار حالت به عنوان بستر حل مساله ارائه شده و در بخش ۵ آزمایشها و نتایج دیده می شوند.

۲- بازی های مارکوفی

۲-۱- تعریف MDP

مساله کنترل کردن یک زنجیره مارکوفی محدود به نام مساله تصمیم گیری مارکوفی خوانده می شود که در آن احتمالات گذار حالت و پاداشها ناشناخته اند و به صورت زیر تعریف می شود.

تعریف ۱. فرآیند تصادفی مارکوف بصورت چند تابی $\langle S, A, R, T \rangle$ نشان داده می شود که در آن S مجموعه متناهی از وضعیت ها، A مجموعه عملیات قابل دسترس برای عامل، R ضریب کاهش و $T: S \times A \times S \rightarrow [0,1]$ احتمال انتقال از وضعیت جاری به وضعیت بعدی با انجام عمل a است و $R: S \times A \rightarrow \mathbb{R}$ تابع پاداش است که یک مقدار عددی را بر می گرداند.

هدف کلی در فرآیند های تصادفی مارکوف، پیدا کردن سیاستی مانند α است بطوریکه امید ریاضی مجموع کاهش یافته پاداشها $J(\alpha)$ را بیشینه نماید که در رابطه (۱) دیده می شود. سیاستهای در نظر گرفته شده بصورت ایستا بوده و غیر تصادفی هستند.

$$J(\alpha) = \lim_{l \rightarrow \infty} \frac{1}{l} E \left[\sum_{t=0}^{l-1} R^{x(t)x(t+1)}(\alpha) \right] \quad (1)$$

با فرض اینکه زنجیره مارکوفی متناظر با هر سیاست α

خیلی کوچکتر باشد، الگوریتم را L_{RP} و اگر b مساوی صفر باشد آن را L_{RI} می‌نامیم [۸]. شمای $S-L_{RP}$ برای مدل‌های Q و S براساس رابطه (۵) بیان می‌شود اگر عمل α_i در مرحله n انتخاب شود در این صورت طبق معادله (۶) داریم:

$$\begin{aligned} p_i(n+1) &= p_i(n) + a(1 - \beta_i(n))(1 - p_i(n)) \\ &\quad - a\beta_i(n)p_i(n) \\ p_j(n+1) &= p_j(n) - a(1 - \beta_i(n))p_j(n) + \\ &\quad a\beta_i(n)\left[\frac{1}{r-1} - p_j(n)\right] - a(1 - \beta_i(n))p_j(n) \end{aligned} \quad (6)$$

که در آن r تعداد اعمال ممکن، a پارامتر پاداش و b پارامتر جریمه می‌باشند. برای اطلاعات بیشتر در باره اتوماتاهای یادگیر می‌توان به [۹] مراجعه نمود.

۱-۳ استفاده از شبکه‌ای اتوماتای یادگیر در حل MDP‌ها

مساله کنترل زنجیره‌های مارکوف می‌تواند به صورت شبکه‌ای از اتوماتاهای بیان گردد که در آن، کنترل از اتوماتایی به اتوماتای دیگر منتقل می‌شود. در این مدل، هر حالت از زنجیره مارکف دارای یک اتوماتای یادگیر است که با استفاده از روابط (۴) و (۵) سعی می‌کند توزیع احتمال بهینه اعمال در آن وضعیت را یاد بگیرد. عامل با شروع از حالت اولیه تا رسیدن به هدف بر روی این شبکه از اتوماتاهای یادگیر حرکت می‌کند و در هر حالت، از اتوماتای یادگیر آن حالت، برای انتقال به یکی از حالت‌های مجاور کمک می‌گیرد که این کار با استفاده از بردار احتمال اعمال اتوماتای یادگیر انجام می‌گیرد. در هر لحظه فقط یک اتوماتای یادگیر فعال بوده و انتقال از یک وضعیت به وضعیت دیگر، زمانیکه بردار احتمالهای کلیه اتوماتاهای یادگیر به پایداری برسد و یا شرط خاصی برقرار گردد تکرار می‌شود.

فرض می‌شود که اتوماتای LA_i در حالت i از پاداش یک مرحله‌ای $R^{ij}(a_i)$ با توجه به اقدام a_i در حالت s_i که منجر به حالت s_j می‌شود آگاهی ندارد. وقتیکه حالت s_i مجدداً دیده می‌شود اتوماتای مربوطه دو نوع داده را دریافت می‌نماید: پاداش تجمعی بدست آمده تا به حال و زمان سراسری فعلی. با توجه به این دو داده پاسخ محیط یا ورودی اتوماتا در حالت i بر اساس رابطه (۷) محاسبه می‌گردد که در آن $\rho^i(t_i+1)$ پاداش تجمعی بدست آمده از عمل a_i در حالت s_i و $\eta^i(t_i+1)$ پارامتر گذشت زمان است.



شکل ۱- ارتباط بین اتوماتای یادگیر و محیط

محیط را می‌توان توسط سه‌تایی $E \equiv \{\alpha, \beta, c\}$ نشان داد که در آن $\alpha \equiv \{\alpha_1, \alpha_2, \dots, \alpha_r\}$ مجموعه ورودی‌ها، $c \equiv \{c_1, c_2, \dots, c_r\}$ مجموعه خروجی‌ها و $\beta \equiv \{\beta_1, \beta_2, \dots, \beta_m\}$ مجموعه احتمال‌های جریمه می‌باشد. هرگاه β مجموعه‌ای دو عضوی باشد، محیط از نوع P است. در چنین محیطی $\beta_1 = 1$ به عنوان جریمه و $\beta_2 = 0$ به عنوان پاداش در نظر گرفته می‌شود. در محیط از نوع Q ، $\beta(n)$ می‌تواند به طور گسسته یک مقدار از مقادیر محدود در فاصله $[0, 1]$ را اختیار کند و در محیط از نوع S ، $\beta(n)$ متغیر تصادفی در فاصله $[0, 1]$ است. c_i احتمال اینکه عمل i نتیجه نامطلوب داشته باشد می‌باشد. در محیط ایستا، مقادیر c_i بدون تغییر می‌مانند، حال آن‌که در محیط غیرایستا این مقادیر در طی زمان تغییر می‌کنند. اتوماتاهای یادگیر به دو دسته اتوماتای یادگیر با ساختار ثابت اتوماتای یادگیر با ساختار متغیر^۷ (VSLA) دسته بندی می‌شوند. اتوماتای یادگیر با ساختار متغیر را می‌توان توسط چهارتایی اتوماتا، $\{\alpha, \beta, p, T\}$ نشان داد که $\alpha \equiv \{\alpha_1, \alpha_2, \dots, \alpha_r\}$ مجموعه عمل‌های اتوماتا، $\beta \equiv \{\beta_1, \beta_2, \dots, \beta_r\}$ ورودی‌های اتوماتا، $p = \{p_1, p_2, \dots, p_r\}$ بردار احتمال انتخاب هر یک از عمل‌ها و $T[\alpha(n), \beta(n), p(n)]$ الگوریتم یادگیری می‌باشد. الگوریتم زیر براساس روابط (۴) و (۵) یک نمونه از الگوریتم‌های یادگیری خطی است. فرض می‌کنیم عمل i در مرحله n انتخاب شود.

- پاسخ مطلوب از محیط

$$\begin{aligned} p_i(n+1) &= p_i(n) + a[1 - p_i(n)] \\ p_j(n+1) &= (1 - a)p_j(n) \quad \forall j \neq i \end{aligned} \quad (4)$$

- پاسخ نامطلوب از محیط

$$\begin{aligned} p_i(n+1) &= (1 - b)p_i(n) \\ p_j(n+1) &= (b/r - 1) + (1 - b)p_j(n) \\ &\quad \forall j \neq i \end{aligned} \quad (5)$$

در روابط (۳) و (۴)، a ، پارامتر پاداش و b پارامتر جریمه می‌باشند. با توجه به مقادیر a و b سه حالت را می‌توان در نظر گرفت: اگر a و b با هم برابر باشند، الگوریتم را L_{RP} هنگامی که از

اعمال اتوماتاهای یادگیر هر حالت بعدی را تعیین می-کنند. در ابتدا، اتوماتاهای یادگیر تمام عملهای خود را با احتمالی یکسان انتخاب می-کنند. در صورتیکه عمل اتوماتاهای یادگیر منجر به حالتی گردد که قبلاً مشاهده شده است اتوماتای یادگیربرمبنای رابطه (۷) پاداش می-گیرد، در غیر اینصورت از آنتروپی بردار احتمال اتوماتای یادگیر حالت بعد برای تعیین پاداش یا جریمه عمل انتخابی استفاده می-شود.

آنتروپی بردار احتمال میزان عدم قطعیت اتوماتای یادگیر حالت بعد را در انتخاب عمل خود نشان می-دهد. هر چه آنتروپی بیشتر باشد میزان عدم قطعیت بیشتر است. عدم قطعیت بالا در بردار احتمال اتوماتای یادگیر به این معنی است که این اتوماتا دارای اطلاعات مفیدی برای رسیدن به هدف نیست و عملهای خود را به صورت تصادفی انتخاب می-کند (جستجو^۱). ولی چنانچه عدم قطعیت کم باشد به این معنی است که اتوماتا با احتمال بالایی یکی از اعمال خود را انتخاب می-کند و دارای اطلاعات مفیدی برای رسیدن به هدف می-باشد و از این اطلاعات بهره-برداری^۲ می-نماید. فرض کنید که $\{p_1, p_2, \dots, p_r\}$ بردار احتمال اعمال یک اتوماتای یادگیر باشد. آنتروپی این بردار احتمال به شکل زیر (رابطه ۸) تعیین می-شود [۱۰] و [۱۱]:

$$H(X) = - \sum_{i=1}^r p_i \log(p_i) \quad (8)$$

زمانی آنتروپی بیشترین مقدار را خواهد داشت که تمام اعمال احتمالی یکسان داشته باشند: $p_1 = p_2 = \dots = p_r = 1/r$ و زمانی کمترین مقدار یعنی (برابر با ۰) را خواهد داشت که یکی از اجزاء بردارهای احتمال به یک برسد ($p_j = 0 \forall j \neq i$). برای اینکه مقدار آنتروپی را به مقداری بین ۰ و ۱ تبدیل کنیم تا به عنوان بردار تقویتی در اتوماتای ساختار متغیر مدل S قابل استفاده باشد. فرض کنید که عامل k در حالت s_i باشد و اتوماتای یادگیر آن LA_k^i عامل را به حالت s هدایت کند. در اینصورت تعیین سیگنال تقویتی طبق معادله ۸ تعیین می-شود که به جای p_i ها بردار احتمال اعمال اتوماتای یادگیر k در حالت i یعنی $P(LA_k^i | s_i)$ قرار می-گیرد. رابطه ۹ میزان پاداش را تعیین می-کند. الگوریتم کلی به نام $ENLA$ در شکل ۲ دیده می-شود.

$$\beta_k^i(t_i+1) = \begin{cases} \frac{\rho_k^i(t_i+1)}{\eta_k^i(t_i+1)} & \text{if state } i \\ H(\text{prob}(LA_k^i)) & \text{visited again} \\ H(\text{prob}_{\text{equal}}) & \text{otherwise} \end{cases} \quad (9)$$

$$\beta^i(t_i+1) = \frac{\rho^i(t_i+1)}{\eta^i(t_i+1)} \quad (7)$$

این روش بروز رسانی تحت عنوان روش TI خوانده می-شود و نتایج زیر اثبات گردیده اند [۴]:

لم ۱. مساله کنترل زنجیره های مارکوفی می-تواند با یک اتوماتای بازی با پاداش یکسان متشکل از N اتوماتا تخمین زده شود.

لم ۲. فرض کنید هر حالت عمل گرا مانند s_i از N حالت زنجیره مارکوفی دارای یک اتوماتای LA_i با استفاده از روش TI داشتن r_i عمل باشد. فرض کنید هر زنجیره مارکوف به هر سیاست α ارجو دیک باشد. بنابراین تطبیق غیر متمرکز LA ها بطور سراسری بهینه ϵ نسبت به پاداش مورد انتظار بلند مدت در هر مرحله زمانی یعنی (α) است. (اتوماتای یادگیر را بهینه ϵ گویند بردار احتمال روش به کار گرفته شده نزدیک بهینه باشد)

۲-۳- استفاده از اتوماتای یادگیر در حل بازی های مارکوفی

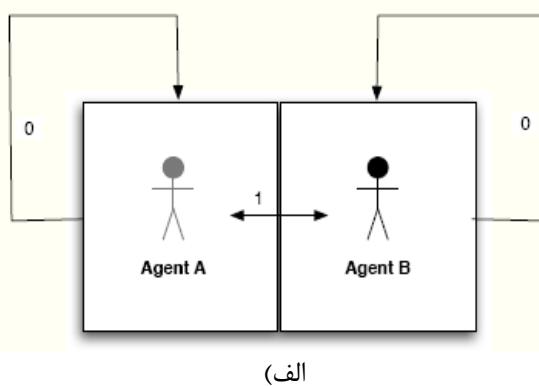
در یک بازی مارکوفی عمل انتخابی در هر حالت نتیجه ترکیب اعمال مستقل انجام شده عملهای موجود در سیستم است. شبکه اتوماتای یادگیر به کار رفته برای MDP ها می-تواند برای بازی های مارکوفی نیز با قرار دادن یک اتوماتا بازی هر عامل در هر حالت توسعه یابد [۴]. در این روش، در هر حالت s_i بازی هر عامل k ، یک اتوماتای یادگیر قرار می-گیرد. در هر لحظه فقط اتوماتاهای یک حالت فعال شده و عمل گروهی اتوماتاهای یادگیر، حالت بعدی را فعال می-سازد. با توجه به حالت قبل پاسخ سیستم نیز برای هر اتوماتای k مشابه رابطه (۶) محاسبه می-گردد. در [۴] قضیه زیر اثبات گردیده است.

قضیه ۳. مدل اتوماتای پیشنهادی در بازی های مارکوف ارجو دیک با قابلیت مشاهده کامل قادر به پیدا کردن یک نقطه تعادل در استراتژی های محضور است.

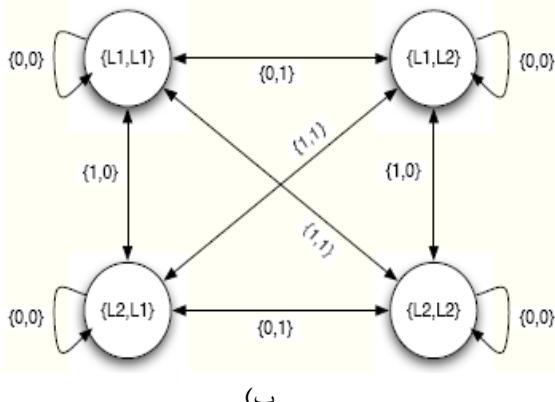
۳-۳- الگوریتم بهبود یافته پیشنهادی (ENLA)

در یک بازی مارکوفی عمل انتخاب شده در هر حالت با توجه به نتیجه اعمال گروهی عملهای مستقل در سیستم است. در این مدل در هر حالت s_i از محیط ($1 \leq i \leq m$ ، m تعداد حالات)، بازی هر عامل k یک اتوماتای یادگیر نظری LA_k^i با ساختار متغیر و مدل S قرار داده می-شود. با توجه به تعداد حالات های مجاور، تعداد اعمال اتوماتا در هر حالت تعیین می-شود. ترکیب

تابع پاداش متفاوت نشان داده شده است که در آن ستون ۱ مکان انتقال یافته را نشان می‌دهد و ستون وسط در آن پاداش در حالت $MMDP$ و ستون سمت راست بیانگر پاداش بازی مارکوف کلی است. احتمال رفتن از حالتی به حالت دیگر را با توجه به احتمالات می‌توان محاسبه کرد. به عنوان مثال در موقعیت $\{L_1, L_1\}$ با توجه به عمل $\{0, 0\}$ احتمال ماندن در $\{L_1, L_1\}$ برابر $81/81$ است و احتمال متناظر به رفتن به حالت های $\{L_2, L_1\}$ و $\{L_1, L_2\}$ برابر $0/0.9$ و $0/0.9$ است.



(الف)



(ب)

شکل ۳. (الف) بازی GridWorld با دو مکان و دو عامل [۴] (ب) بازی مارکوفی همان بازی

با توجه به مدل پیشنهادی بازی هر عامل در هر حالت یک اتوماتای یادگیر در نظر گرفته می‌شود در این صورت به چهار اتوماتای یادگیر برای هر عامل و در مجموع ۸ اتوماتای یادگیر نیاز داریم. فرض می‌شود که LA_k^i اتوماتای یادگیر برای عامل a_i در حالت i از پاداش R^{ij} تک مرحله‌ای ناشی از عمل مشترک k در حالت i که منجر به حالت j می‌شود آگاهی نمی‌یابد، در عوض به محض دیدن دوباره حالت i دو نوع اطلاعات مطرح شده را دریافت می‌دارد. همچنین فرض شده است که عاملها فقط می‌توانند موقعیت و اعمال خودشان را ببینند. هر اتوماتا فقط در همان موقعیت حالت خود را می‌بینند و دیدکلی و سراسری از

۴- بازی هماهنگی GRIDWORLD و نتایج آزمایشها

یکی از مثالهای بازی‌های مارکوفی، مساله هماهنگی چند عامل در چند حالت است که جزء بازی‌های *GridWorld* مطرح شده است. این بازی شامل دو موقعیت است که در آن دو عامل وجود دارند و سعی در هماهنگ نمودن رفتارشان به منظور رسیدن به پاداش بیشینه دارند. در هر لحظه هر دو عامل می‌توانند یکی از دو اعمال ۰ یا ۱ را انجام دهند. اگر عامل عمل ۰ را انتخاب کند در همان موقعیت قبلی باقی می‌ماند و اگر ۱ را انتخاب کند به مکان بعدی می‌رود. انتقال از حالتی به حالت بعدی احتمالاتی است بطوریکه با احتمال $9/10$ در موقعیت فعلی باقی می‌ماند و با احتمال $1/10$ به موقعیت بعدی می‌رود. عاملها متناسب با مکان جابجا شده پس از انتقال پاداش دریافت می‌کنند [۱۲].

ENLA (*MarkovGame,a,b,M*)

Inputs: a, b : reward and penalty parameter for LA , M : total training time

Initialize : s_0, a_1, \dots, a_n , initialization probablity of all of LA

$s_i = \text{Random (State)}$;

1. **for** TimeStep =1 to M **do**

2. **for** each agent k **do concurrently**

3. $s_i^k = \text{Active } LA_k^i$

4. Choose action a_k^i in state s_i^k

5. Observe Rewards r_k^i and Next State s'

6. Compute Average Reward , ρ_k^i, η_k^i

7. Compute β_k^i signal based on EQ (9)

8. Train LA_k^i residing in state S_i according visited stste and β_k^i

$s=s'$

end for

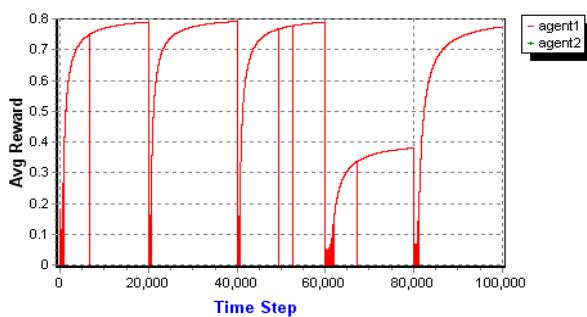
شکل ۲. الگوریتم پیشنهادی ENLA

در شکل ۳ بخش بالایی بازی هماهنگی و در بخش پایین آن، بازی مارکوفی معادل آن نشان داده شده است. همانطور که در شکل دیده می‌شود بازی دارای چهار حالت $S_1=\{L_1, L_1\}, S_2=\{L_1, L_2\}, S_3=\{L_2, L_1\}, S_4=\{L_2, L_2\}$ است که در آن S_1 و S_4 با توجه به عمل مشترک دو عامل $MMDP$ و بازی محيط تغییر می‌کند. این بازی در دو حالت بازی $MMDP$ و بازی مارکوف کلی مطرح شده است که در جدول ۱ دو نوع

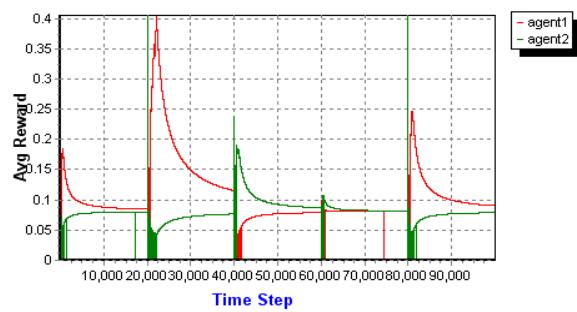
وضعیتهای محیط را ندارد.

جدول ۱. دو نوع تابع پاداش با توجه به وضعیت مکانی انتقال یافته

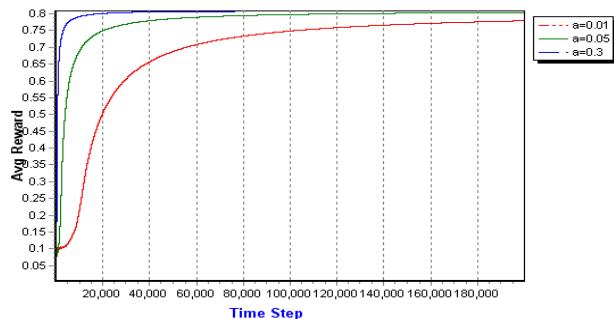
stste	R1	R2
{L ₁ ,L ₁ }	0.01,0.01	0.5,0.5
{L ₁ ,L ₂ }	1.0,1.0	0.0,1.0
{L ₂ ,L ₁ }	0.5,0.5	1.0,0.0
{L ₂ ,L ₂ }	0.01,0.01	0.1,0.1



شکل ۳. نمودار میانگین پاداش برای بازی مارکوف چند عامله MMDP (پاداش R1) با پارامترهای یادگیری a=0.1, b=0 در پنج اجرای مختلف



شکل ۴. نمودار میانگین پاداش برای بازی مارکوف کلی (پاداش R2) با پارامترهای یادگیری a=0.1, b=0 در پنج اجرای مختلف



شکل ۵. مقایسه پارامتر یادگیری a در میزان همگرایی برای R1 a={0.01,0.05,0.3}

۵- نتیجه گیری

در این مقاله رفتار عاملهای مستقل برای تصمیم گیری های غیر متراکر چند عامله مورد بررسی قرار گرفت و روشی بهبود یافته برای حل بازی های مارکوفی با استفاده شبکه ای از اتوماتاهای یادگیر با توجه به پاداش میانی پرایه آنتروپی ارائه گردید. با توجه به مقایسه بین الگوریتم اولیه والگوریتم بهبود یافته در بازی های مارکوفی دیده می شود که پاداش میانی می تواند سرعت همگرایی را بهبود بخشد. استفاده ازروش L_{RI} در اتوماتای یادگیر میتواند نقاط تعادل و نزدیک بهینه را ایجاد نموده و میزان پارامتر پاداش در اتوماتای یادگیر در سرعت همگرایی

برای مقایسه دو الگوریتم (الگوریتم پیشنهادی و الگوریتم پایه) از میانگین پاداش اخذ شده توسط عامل ۱ برای بازی در حالت MMDP و سپس بازی مارکوف کلی نشان داده شده است. برای اینکه همگرایی الگوریتم نشان داده شود به جای تکرار آزمایش و میانگین به دست آوردن، در هر مرحله (۲۰۰۰۰) اجرا اتوماتاهای یادگیر مجدداً با احتمالات تصادفی راه اندازی شده نقاط تعادل مختلف نشان داده شوند. شکل ۳ و ۴ نتایج بدست آمده از دو نوع بازی مارکوفی را با توجه به الگوریتم پایه نشان میدهد. در این آزمایش ضریب یادگیری برابر ۰/۱ و پارامتر جریمه را صفر یعنی (L_{RI}) در نظر گرفته شده است. همانطور که مشاهده می شود با توجه به پاداش R1 عاملها هم می توانند به نقطه تعادل بهینه و نزدیک بهینه دست یابند. این کار با توجه به مقدار تعادل اولیه داده شده برای احتمالات اعمال اتوماتای یادگیر است. در مورد R2 دیده می شود همواره به یک نقطه تعادل می رسد.

برای بررسی سرعت همگرایی ابتدا پارامتر یادگیری بازی پارامترهای مختلف بررسی گردید. انتخاب پارامتر $a=\{0.01,0.05,0.3\}$ نشان می دهد هرچه پارامتر یادگیری بیشتر باشد سرعت همگرایی نیز بیشتر خواهد بود. البته انتخاب پارامترهای بالای ۰/۵ باعث میگردد الگوریتم فقط به یکی از نقاط تعادل نزدیک بهینه و نه لزوماً بهینه همگرایی داشته باشد. شکل ۵ میزان همگرایی را برای مقادیر مختلف پارامتر a (پارامتر یادگیری اتوماتا) نشان می دهد.

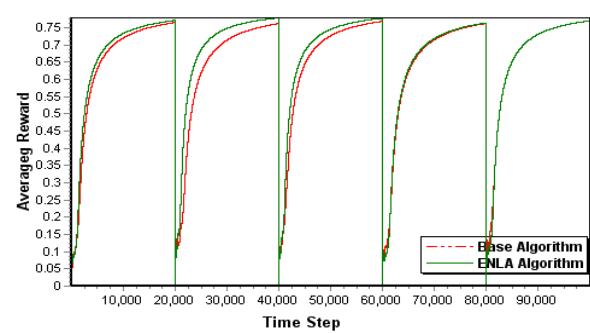
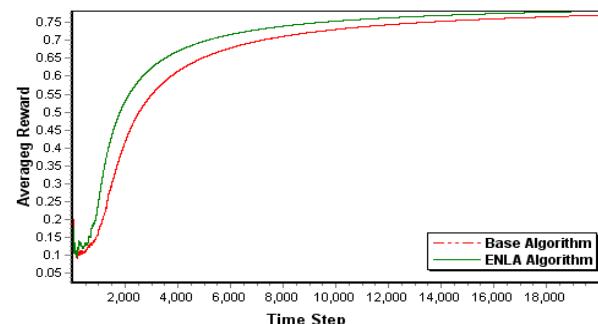
در آزمایش بعدی روش پایه و روش پیشنهادی که در آن از آنتروپی به عنوان پاداش میانی استفاده می شود، دردو حالت بازی های MDP و بازی مارکوف کلی مورد ارزیابی قرار گرفته اند. همانطور که در شکل ۶ دیده می شود سرعت همگرایی در الگوریتم پیشنهادی افزایش یافته است. برای بررسی نقش پارامتر یادگیری در الگوریتم پیشنهادی آزمایشی با چهار مقدار مختلف پارامتر $a=\{0.01,0.05,0.1,0.5\}$ صورت گرفت که نتایج در شکل ۷ دیده می شود. همانطور که در شکل دیده می شود تغییر پارامتر یادگیری می تواند در سرعت همگرایی و انتخاب نقاط تعادل نقش داشته باشد.

مراجع

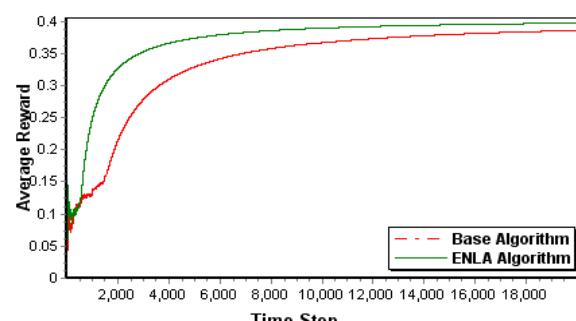
- [1] A. Now'e, K.Verbeeck, and M.Peeters. "Learning Automata as a Basis for Multi-agent Reinforcement Learning," *Lecture Notes in Computer Science*, pp. 71–85, 2006.
- [2] R. M. Wheeler and K. S. Narendra, "Decentralized Learning in Finite Markov Chains," *IEEE Transactions on Automatic Control, AC-31*:pp. 519 – 526, 1986.
- [3] M. Littman. "Markov Games as a Framework for Multi-agent Reinforcement Learning." In *Proceedings of the 11th International Conference on Machine Learning*, pp. 322 – 328, 1994.
- [4] P. Vrancx, K. Verbeeck, and A. Nowé, "Decentralized Learning in Markov Games," *IEEE Transactions on Systems, Man and Cybernetics, Part B: Special Issue on Approximate Dynamic Programming and Reinforcement Learning for Control*, 38(4), 2008.
- [5] J. Osborne and A. Rubinstein, *A Course in Game Theory*. MIT Press, Cambridge, MA, 1994.
- [6] B. Masoumi , M. R. Meybodi, and B. Jafarpour, "Solving General Sum Stochastic Games using Learning Automata," *Proceedings of the second Joint Congress on Fuzzy and Intelligent Systems*, Malek Ashtar University of Technology, Tehran, Iran, 28-30 October, 2008.
- [7] M. Sobel, "Noncooperative Stochastic Games," *The Annals of Mathematical Statistics*, vol. 42, no. 6, pp. 1930–1935, 1971.
- [8] K. S. Narendra and M. A. L Thathachar, *Learning automata: An introduction*. Prentice Hall, 1989.
- [9] M. A. L. Thathachar and Sastry, "Varieties of Learning Automata: An Overview," *IEEE Transaction on Systems, Man, and Cybernetics-Part B: Cybernetics*, Vol. 32, no. 6, pp. 711-722, 2002.
- [10] M. Costa, A. Goldberger and C. Peng, "Multi-scale Entropy Analysis of Complex Physiologic Time Series," *Physical Review Letters* 89, 068102 , 2002.
- [11] Z. Dianhu, F. Shaohui and D. Xiaojun, " Entropy – A Measure of Uncertainty of Random Variable," *Systems Engineering and Electronics*, no. 11, pp. 1-3, 1997.
- [12] P. Vrancx, K. Verbeeck, and Nowé, "Limiting Games of Multi-agent Multi-state Problems," in *Proc. Workshop on Adaptive and Learning Agents 2007, 6th International Conference on Autonomous Agents and Multi-Agents Systems (AAMAS 2007)*, Hawaii, USA, 2007.

-
- 1 Markov Game
 - 2 Repeated Normal Game
 - 3 Markov Decision Process
 - 4 Fully Cooperative
 - 5 Multi-Agent MDP
 - 6 General Markov Game
 - 7 Variable Sturcture Learning Automata
 - 8 Exploration
 - 9 Exploitation

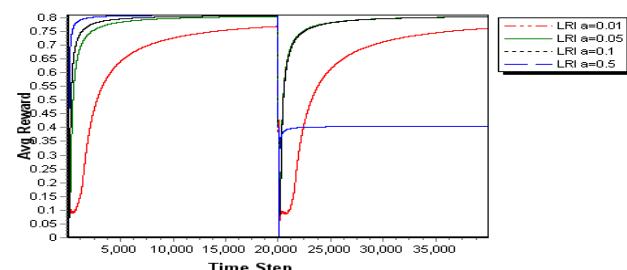
نقش مهمی را دارد. با توجه به آزمایش‌های انجام گرفته اتوماتاهای یادگیر مدل مناسبی برای تصمیم گیری های غیر متتمرکز چند عامله هستند.



شکل ۶. مقایسه الگوریتم پایه و الگوریتم پیشنهادی (ENLA) برای بازی (R1) با پارامترهای یادگیری $a=0.1, b=0$ در دو حالت یک مرحله اجرا و پنج مرحله اجرا



شکل ۷. مقایسه الگوریتم پایه و الگوریتم پیشنهادی (ENLA) برای بازی مارکوف کلی (پاداش R2) با پارامترهای یادگیری $a=0.1, b=0$



شکل ۸. مقایسه پارامتر یادگیری a در میزان همگرایی در الگوریتم پیشنهادی $a=\{0.01, 0.05, 0.1, 0.5\}$ MMDP برای بازی