تعیین آرایش بهینه یگان‌ها با استفاده از یادگیری تقویتی چندعاملی در بازی‌جنگ

علی اکرمی‌زاده[۱]

احمد افشار، محمد باقر منهاج، سمیرا جعفری[۲]

چکیده

در این مقاله، به مسئله یادگیری تقویتی چندعاملی با کاربرد در بازی‌جنگ پرداخته شده‌است. ساختارهای نظامی باعث ایجاد اولویت در اجرای تصمیمات بین عامل‌های درگیر در صحنه نبرد می‌شود. حالت‌های استاتیک تصمیم‌گیری بین عامل‌ها در این ساختارهای را می‌توان در قالب بازیهای بسیط بیان کرد. فرآیند مزبور در چارچوب بازیهای مارکوف بسیط مدل شده که عمل مشترک بهینه از طریق محاسبه نقطه تعادل نش کامل زیربازی به دست می‌آید. با استفاده از مفهوم ارزشهای انجمنی، امکان ایجاد مصالحه در انتخاب عمل بهینه نقطه تعادل نش و اکتشاف عمل‌های جدید فراهم شده‌است. شبیه‌سازی انجام شده بر روی نسخه ساده‌ای از یک بازی‌جنگ واقعی، علاوه بر تأیید همگرایی، کارآمدی این روش را در بررسی پدیده‌های مختلف جنگ نشان می‌دهد.

کلیدواژه

بازی‌جنگ، یادگیری تقویتی چندعاملی، مصالحه بین اکتشاف و استخراج، بازیهای مارکوف بسیط، آرایش‌بندی

۱. دانشگاه صنعتی امیرکبیر akramizadeh@aut.ac.ir
۲. دانشگاه صنعتی امیرکبیر.
تاریخ دریافت: ۸۹/۱۱/۲۰ تاریخ پذیرش: ۸۹/۱/۲۲

مقدمه

خطر جنگ حتی امروزه که شاهد پیشرفت‌های قابل توجهی در صنایع، فناوری و گسترش مبانی معنوی و حقوق بشر بوده‌ایم همچنان به صورت بالقوه وجود دارد. این مسئله باعث شده تا ملاحظات مربوط به امنیت دفاعی، سرلوحه امور بسیاری از کشورهای جهان قرار گیرد. رشد فنآوری و پیدایش تجهیزات و تسلیحات متنوع نظامی باعث ایجاد تکنیکها و تاکتیکهای جدیدی گشته است که قدرت مانور و عملکرد واحدهای نظامی را در حمله و دفاع با تأثیرپذیری کاملاً متفاوت از یکدیگر تحت تأثیر قرار داده‌است. بنابراین تعیین استراتژیها و آرایش‌بندهای مناسب در عملیات نظامی از اهمیت بسزائی برخوردار شده است. روشی که از دیرباز به این منظور مورد استفاده قرار می‌گرفته‌است، تحت عنوان بازی جنگ شناخته می‌شود. طی اعصار مختلف اشکال مختلفی برای آن مطرح شده که می‌توان به رزمایش‌های صحرائی، و مانور روی نقشه اشاره کرد. استفاده از الگوهای ریاضی و نظریه بازیها [۱] و نهایتاً مفهوم عامل‌ها و سیستمهای چندعاملی [۲] به مرور جایگزین روشهای قبلی شد. این مفاهیم در بسیاری از کاربردهای امروزه مورد استفاده قرار می‌گیرد [۳]. اشاره به این نکته ضروری است که ذات بسیاری از مسائل به گونه‌ای است که به کارگیری سامانه های چندعاملی نه‌تنها راه‌حل مناسبی است، بلکه در اصول ممکن است اجباری باشد. سئوالی که مطرح می‌شود این است که عامل‌ها چه عمل‌های‌ی باید اتخاذ کنند که آنها را به اهدفشان نزدیکتر کند [۴]؟ به ویژه اگر اطلاعات کاملی از محیط و نحوه تعامل با سایر عامل‌ها وجود نداشته باشد.این عدم شناخت مذکور باعث می‌شود که عامل درک مناسبی از بایدها و نبایدها نداشته باشد. این بایدها و نبایدها شالوده اصلی تدبیر[1] عامل در تعامل با محیط و سایر عامل‌ها است. نگاهی به انسان به عنوان یک عامل هوشمند متعالی، مؤید این نکته است که هوشمندی تنها راهنمای انسان در حل مسائل و شناخت بایدها و نبایدهایش نبوده، بلکه یادگیری از تجربیات گذشته به ویژه در بدست آوردن یک مهارت نقش مهمی دارد. بنابراین حجم گسترده‌ای از فعالیتهای تحقیقاتی طی سالیان متمادی به مسئله یادگیری اختصاص داشته‌است. لزوم استفاده از آن در حوزه بازی‌جنگ نیز اهمیت داشته [۵] که در یکی از اولین اقدامات صورت گرفته در این راستا سعی شد از MDP[2] و POMDP[3] به جای تحقیق در عملیات استفاده شود [۶]. البته ساختار بازی‌جنگ، که یک سامانه چندعاملی است [۲]، لزوم استفاده از روشهای یادگیری چندعاملی را مطرح می‌کند [۷]، [۸]. اکثر این روشها تعمیم روش‌های یادگیری تک عاملی به چندعاملی هستند که از آنجمله می‌توان به روشهای یادگیری تکاملی[4] و هم‌تکاملی[5] [۹,۱۰]،

1. Policy
2. Markov Decision Process
3.Partial Observable Markov Decision Process
4. Evolutionary learning
5. Coevolutionary techniques

نظریه بازیهای تکاملی[1] [۱۱,۱۲]، جستجوی هدایت‌شده[2] [۱۳]، تعمیم روشهای مطرح در یادگیری تقویتی [۱۴] و نظریه بازیها [۱۵] و ترکیب روش یادگیری تقویتی و نظریه بازی [۱۶] اشاره کرد. هر کدام از روشهای مطرح شده در این حوزه دارای مفروضاتی هستند که آنها را برای دسته خاصی از کاربردها مناسبتر می‌سازند. البته در این بین، ترکیب مبانی یادگیری تقویتی و نظریه بازی نسبت به سایر روشها توانایی بالاتری در حل مسائل مختلف به ویژه در حوزه مسائل دینامیک دارد. این رهیافت اولین بار در [۱۷] مطرح شد. این الگوریتم، که تحت عنوان MinMax-Q نامیده شد، برای حالت خاصی از سامانه های چندعاملی کاربرد دارد که در آن دو عامل کاملاً رقیب به صورت متوالی اقدام به انتخاب عمل می‌کنند. مدتی بعد، حالت کلی‌تری از سامانه های چندعاملی مدنظر قرار گرفت که در آن تعدادی عامل خودخواه[3] به صورت همزمان اقدام به اخذ تصمیم می‌کردند [۱۶]. بازه‌های تصمیم‌سازی توسط بازیهای با فرم نرمال مدل شدند که به منظور ارضاء شرایط همگرایی، لازم بود نقطه تعادل نش بازی یکتا باشد. این شرط که در کاربردهای عملی بسیار محدود کننده است، کمی تخفیف یافت ولی فرض بر این شد که نقش عامل‌ها در ابتدا مشخص باشد [۱۸]. البته تلاش‌های دیگری نیز در همین راستا انجام شد [۱۹,۲۰]، که مقایسه و ارزیابی آنها تا حدودی در [۲۱] ارائه شده است. از این ساختارها در بازی‌جنگ نیز استفاده شد [۲۲]. البته ساختار سلسله مراتبی در کاربردهای نظامی، مسئله مهمی است که باید در نظر گرفته شود. در این راستا دو راه‌حل قابل تصور است. راه‌حل اول استفاده از روشهای یادگیری تقویتی سلسه مراتبی است [۲۳] که کارآیی خوبی در حالت تک-عاملی از خود نشان داد [۲۴]. البته مسئله تعمیم آن به حالت چندعاملی دشواریهای زیادی را به دنبال دارد. راه‌حل دوم، استفاده از فرم بسیط بازیها به جای فرم نرمال است. به منظور تعمیم اقدامات اولیه در [۱۷] روش دیگری ارائه شد [۲۵] که متأسفانه همچنان محدودیت‌های قبل را تا حدود زیادی داشت [۲۶]. روش یادگیری برای بازیهای بسیط چندعاملی جمع- عمومی[4] و اثبات همگرایی آن بوسیله نویسندگان این مقاله ارائه شد [۲۷]. مصالحه بین اکتشاف و استخراج در روش یادگیری مطرح شده در مقاله بعدی مطرح گردید [۲۸].

1 Evolutionary game theory
2 Directed search
3 Self-interest
4 General-sum

از جمله نکات مهمی که در ارائه روشهای فوق مطرح است، نحوه پیاده‌سازی آنها در کاربردهای عملی مانند بازی جنگ است. در این مقاله، کاربرد روش‌های فوق در بازی‌جنگ مورد توجه قرار گرفته‌است. ادامه مقاله به این صورت است که در فصل ۲ به بررسی مبانی اولیه مورد نیاز با استفاده از اصطلاحات تخصصی حوزه یادگیری پرداخته شده است. سپس در فصل ۳ الگوریتم مناسب یادگیری تقویتی چندعاملی مطرح شده است. پس از آن در فصل ۴ مسئله بازی‌جنگ مطرح شده و امکان استفاده از یادگیری تقویتی در آن بررسی شده است. سپس درفصل ۵ نتایج شبیه‌سازی مورد بررسی قرار گرفته است. نهایتاً در فصل ۶ جمع‌بندی مطالب ارائه گردیده است.

مبانی اولیه

در ادامه به صورت خلاصه مبانی اولیه مربوط به یادگیری تقویتی چندعاملی مرور شده‌است. به منظور هماهنگی، اصطلاحات مورد استفاده بر مبنای ادبیات مطرح شده در حوزه یادگیری تقویتی است.

یادگیری تقویتی تک- عامله

اساس یادگیری تقویتی بر پایه سعی و خطا بنا شده‌است. عامل با مشاهده وضعیت محیط عملی را انجام می دهد. انجام این عمل باعث تغییر وضعیت محیط شده و متناسب با این تغییر، پاداشی از طرف محیط به عامل داده می‌شود. با تکرار سعی و خطا، عامل به دنبال پیدا کردن تدبیر بهینه رسیدن به اهدافش در محیط است. عموماً، مدل محیط توسط یک فرآیند تصمیم‌گیری مارکوف متناهی مدلسازی می‌شود. این مدل چارچوب ریاضی بیان مسئله انتخاب عمل در حالتی است که فقط قسمتی از نتایج خروجی‌ها تحت کنترل عامل است و قسمتی دیگر به صورت اتفاقی است.

تعریف ۱) فرآیند تصمیم‌گیری مارکوف یک چندگانه (S, A, R, P) است، که:

- S مجموعه حالت‌های محیط است،
- A مجموعه عمل‌های قابل‌انجام است،
- $R = \{R \mid R : S \times A \to \Re\}$ تابع پاداش است،
- $P : S \times A \to \Delta(S)$ تابع تغییر حالت است، که $\Delta(S)$ مجموعه توزیع احتمال‌ها بر روی مجموعه S است.

عامل به دنبال پیدا کردن تدبیر بهینه‌ای است که پاداش مورد انتظار تنزلی آینده را بیشینه سازد.

$$V^{\pi}(s) = E^{\pi}\left\{r^{k+1} + \gamma r^{k+2} + \gamma^2 r^{k+3} + \ldots \mid s^k = s\right\}$$
$$= E^{\pi}\left\{r^{k+1} + \gamma V^{\pi}(s^{k+1}) \mid s^k = s\right\}$$
$$= \sum_{a \in A} \pi(s,a)\left[r(s,a) + \gamma \sum_{s'} P_{ss}^{a} V^{\pi}(s')\right]$$

این تدبیر نگاشتی از حالت‌ها به عمل‌های قابل انجام بوده، $\pi : S \times A \to [0,1]$ که $\forall s, s' \in S, a \in A$، داریم:

r^k: پاداش لحظه‌ای،

$V^{\pi}(s)$: ارزش حالت s تحت تدبیر π که ارزش- حالت خوانده می‌شود،

$\pi(s,a)$: احتمال انتخاب عمل $a \in A$ در حالت s،

$\gamma \in [0,1]$: فاکتور تنزل

$P_{ss'}^{a}$: تابع تغییر حالت که $P_{ss'}^{a} = \Pr\left\{s^{k+1} = s' \mid s^k = s, a^k = a\right\}$

ارزش هر حالت با توجه به تدبیر بهینه توسط تابع بهینه ارزش- حالت بدست می‌آید:

$$V^{*}(s) = \max_{\pi} V^{\pi}(s)$$
$$= \max_{a \in A} E\left\{r^{k+1} + \gamma V^{*}(s^{k+1}) \mid s^k = s\right\}$$
$$= \max_{a \in A}\left[r(s,a) + \gamma \sum_{s'} P_{ss}^{a} V^{*}(s')\right]$$

ارزش انتخاب عمل a در حالت s تحت تدبیر π، که توسط $Q^{\pi}(s, a)$ نشان داده می‌شود، مجموعه پاداش‌های مورد انتظار تنزلی است که با شروع از حالت s و انجام عمل a بر اساس تدبیر π حاصل خواهد شد.

$$Q^{\pi}(s,a) = E\left\{r^{k+1} + \gamma r^{k+2} + \gamma^2 r^{k+3} + \ldots \mid s^k = s, \pi(s) = a\right\}$$
$$= r(s,a) + \gamma \sum_{s'} P_{ss'}^{a} \sum_{a'} \pi(s',a') Q^{\pi}(s',a')$$

تابع فوق را ارزش- عمل خوانده که مقدار بهینه آن به صورت زیر قابل محاسبه است:

$$Q^*(s,a) = \max_{\pi} Q^{\pi}(s,a)$$
$$= r(s,a) + \gamma \sum_{s'} P^{a}_{ss'} \max_{a'} Q^*(s',a')$$

بر اساس تحقیقات انجام شده، پیدا کردن تدبیر بهینه معادل با پیدا کردن تابع ارزش- عمل بهینه با کمک رابطه بازگشتی زیر است:

$$Q(s,a) := (1-\alpha)Q(s,a) + \alpha\left(r + \gamma \max_{a'} Q(s',a')\right)$$

نظریه بازی

بازی پدیده‌ای است که در آن دو یا چند عامل در تعامل با یکدیگر باید از بین مجموعه عمل‌های قابل انجام خود یکی را انتخاب و اجرا کرده و بر اساس این انتخاب و آنچه دیگر عامل‌ها انتخاب کرده‌اند، پاداش دریافت کنند [۲۹]. هر عامل ممکن است به دنبال افزایش بازده[۱۰] آنی، بازده دراز مدت و یا مسائلی پیچیده‌تری باشد که به سایر عامل‌ها نیز مرتبط است. در این مقاله فرض بر این است که هر عامل دارای اطلاعات کامل[۱۱] و کافی[۱۲] است. در این دسته از بازیها، هر عامل همه چیز را درباره عمل‌های قابل انجام سایر عامل‌ها و ارزش آنها را می‌شناسد. مجموعه این ارزشها که مقادیر Q خوانده می‌شوند تحت عنوان مقادیر Q بسط یافته[۱۳] نامیده شده‌است.

تعریف۲) یک بازی بسیط با اطلاعات کامل توسط چندگانه $g = (X, \Sigma, f, Q)$ بیان می‌شود، که:

- $X = \{x_1, x_2, \ldots, x_N\}$ مجموعه عامل‌ها است،
- $\Sigma = \{\sigma \mid \sigma \in \langle A_1, A_2, \ldots, A_N \rangle\}$ مجموعه عمل‌های مشترک است که A_i مجموعه عمل‌های مجاز عامل iام است،

10.Payoff
11.Complete
12.Perfect
13.Extended Q-values

- $f\left(\hat{\sigma}_i\right)$ تابع تعیین اولویت است که بعد از هر زیردنباله $\hat{\sigma}_i$ تعیین می‌کند که نوبت کدام عامل است که عملش را انتخاب و اجرا کند.[۱]
- $Q=\{Q_i \mid Q_i : A \to \Re\}$ ارزش هر عمل مشترک را برای هر عامل تعیین می‌کند.

از جمله انتخاب‌های بهینه در یادگیری تقویتی کلاسیک، انتخاب حریصانه عمل است که بیشترین ارزش ممکن را در هر حالت نتیجه می‌دهد. در حالت چندعاملی، عامل باید انتخاب سایر عامل‌ها را نیز در نظر بگیرد. انتخاب عمل بهینه در حالتی که تعدادی عامل دیگر نیز حضور داشته که هر یک به دنبال رسیدن به بیشترین ارزش هستند، تحت عنوان نقطه تعادل نش در نظریه بازیها مورد بررسی قرار گرفته‌است.

تعریف ۳) مجموعه عمل‌های $\sigma^*=\left(a_1^*,a_2^*,...,a_N^*\right)$ نقطه تعادل نش می‌باشد اگر هیچ عمل دیگری برای عامل i ارزش بیشتری را نتیجه ندهد، به شرط آنکه سایر عامل‌ها همان عمل‌های قبلی را همچنان انجام دهند [۳۰]. به عبارت دیگر:

$$Q_i\left(\sigma^*\right)\geq Q_i\left(a_i,\sigma_{-i}^*\right) \quad \forall i,a_i\in A_i$$

که $\sigma_{-i}^*=\left\langle a_1^*,...,a_{i-1}^*,a_{i+1}^*,...,a_N^*\right\rangle$.

به منظور تعیین نقطه تعادل نش در بازیهای بسیط باید مفهوم زیربازی نیز مدنظر قرار گیرد. یک زیربازی به هر قسمتی از یک بازی گفته می‌شود که خودش یک بازی بسیط است[۲]. نقطه تعادل نش در بازیهای بسیط که تحت عنوان نقطه تعادل کامل زیربازی[۳] نامیده می‌شود به صورت زیر تعریف می‌شود.

تعریف ٤) عمل مشترک $\sigma^*=\left\langle a_1^*,a_2^*,...,a_N^*\right\rangle$ نقطه تعادل کامل زیربازی است، اگر در هر زیربازی یک نقطه تعادل نش باشد. آنچه که تا کنون بیان شد، مربوط به بازیهای استاتیک است. با الهام از فرآیند تصمیم‌گیری مارکوف، حالت استاتیک بازی به صورت دینامیک بیان شده که در آن تعدادی عامل به صورت دینامیک با هم تعامل داشته و هر یک قصد دارند با انتخاب عمل مناسب، به بیشترین

۱. هر دنباله $\hat{\sigma}_i=\left\langle a_1,a_2,...,a_i\right\rangle$ را با توجه به مجموعه عمل‌های $\sigma=\left\langle a_1,a_2,...,a_N\right\rangle$ یک زیردنباله از عمل‌ها می‌نامند.
۲. در این مقاله فرض شده است که همه زیربازیها صحیح هستند.
3. Subgame perfect equilibrium point

پاداش ممکن برسند. این فرآیند تحت عنوان بازی مارکوف[1] نامیده می‌شود [۳۱]. در صورتی که این فرآیند بر پایه بازیهای بسیط نهاده شده باشد، تحت عنوان بازی مارکوف بسیط خوانده می‌شود [۳۲].

تعریف ٥) بازی مارکوف بسیط توسط چندگانه $\Psi = \langle G, X, \Sigma, P, R \rangle$ بیان می‌شود، که:

- G مجموعه بازیهای بسیط با اطلاعات کامل است،
- $X = \{x_1, x_2, \ldots, x_N\}$ مجموعه عامل‌ها است که اولویت آنها در انتخاب عمل در طول فرآیند ثابت است،
- $\Sigma = \{\sigma \mid \sigma \in \langle A_1, A_2, \ldots, A_N \rangle\}$ مجموعه عمل‌های مشترک قابل دسترس است که A_i مجموعه عمل‌های قابل انجام عامل i است،
- $P : G \times \Sigma \to \Delta(G)$ تابع تغییر حالت بوده که، $\Delta(G)$ بیانگر توزیع احتماتی بر روی G است،
- $R = \{R_i \mid R_i : G \times \Sigma \to \Re\}$ تابع پاداش است.

مفهوم نقطه تعادل نش را می‌توان در یک بازی مارکوف بسیط به صورت زیر بیان کرد.

تعریف ٦) پروفایل تدابیر π^* یک نقطه تعادل نش در یک بازی مارکوف بسیط G است، اگر هیچ عامل i قادر نباشد نتیجه بهتری در هر بازی $g \in G$ با تغییر تدبیر خود کسب کند، به شرط آنکه سایر عامل‌ها همان تدابیر قبلی را همچنان انجام دهند. به عبارت دیگر:

$$Q_i\left(g, \pi^*(g)\right) \geq Q_i\left(g, \pi_i(g), \pi^*_{-i}(g)\right) \quad \forall g, i, a_i \in A_i$$

که $Q_i\left(g, \pi(g)\right)$ بیانگر ارزش انتخاب عمل مشترک $\pi(g)$ در بازی g برای عامل i است و $\pi_{-i}(g)$ شامل تمام تدابیر عامل‌ها به جز عامل i است.

1.Markov game

یادگیری تقویتی چندعاملی

یادگیری

یادگیری در سامانه های چندعاملی فرآیندی است که در آن تعدادی عامل با درجه پایین‌تری از یک عامل کاملاً منطقی برای رسیدن به بهینگی تلاش می‌کنند [۳۳]. این فرآیند در شکل ۱ در قالب یک فرآیند مارکوف بسیط نمایش داده شده‌است. این فرآیند شامل تعداد بازی با فرم بسیط است، که $g^k \in G, k = 1,...,K$. هر بازی را می‌توان توسط یک درخت محدود نشان داد با مجموعه‌ای از گره‌ها به عنوان عامل، مجموعه‌ای از کمان‌ها به عنوان عمل‌ها، و مجموعه‌ای از ارزش‌ها که اولویت‌بندی مجموعه عمل‌های مشترک را نشان می‌دهد. در هر بازی g^k، عامل i بازی را به سمت زیربازی $\hat{g}_i^k \in G, i = 1,...,N$ هدایت می‌کند که که ارزش بیشتری را به دنبال دارد. عامل به دنبال پیدا کردن تدبیری است که در نهایت بیشترین مجموع پاداش‌های مورد انتظار تنزلی را نتیجه دهد.

$$V_i^{\pi}(g) = E^{\pi}\left[r_i^{k+1} + \gamma r_i^{k+2} + \gamma^2 r_i^{k+3} + ... \mid g^k = g, \pi\right]$$
$$= E^{\pi}\left[r_i^{k+1} + \gamma V_i^{\pi}\left(g^{k+1}\right) \mid g^k = g\right]$$
$$= \sum_{\sigma \in \Sigma} \pi(g,\sigma)\left[r_i(g,\sigma) + \gamma \sum_{g'} P_{gg}^{\sigma} V_i^{\pi}(g')\right]$$

که برای هر $g, g' \in G, \sigma \in \Sigma$ داریم:

r_i: پاداش لحظه‌ای عامل i

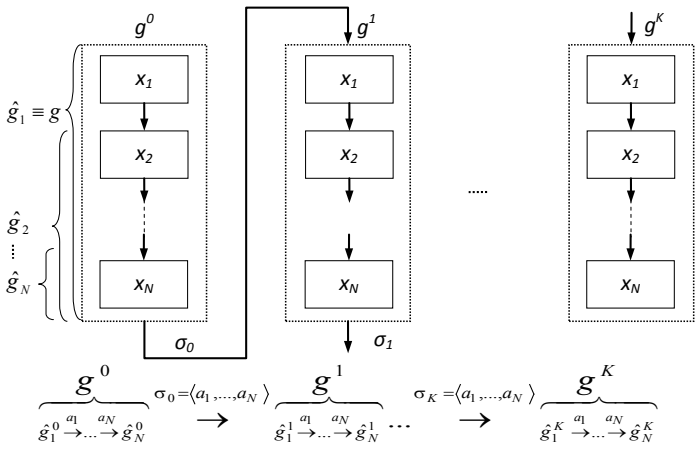
$V_i^{\pi}(g)$: ارزش بازی g تحت تدبیر $\pi = (\pi_1,...,\pi_N)$ که تحت عنوان تابع ارزش- بازی بیان می‌شود،

$\pi(g,\sigma)$: احتمال انتخاب عمل $\sigma = (a_1,...,a_N) \in \Sigma$ در بازی g که
$\pi(g,\sigma) = \pi_1(g,a_1) \times \pi_2(g,a_2) \times ... \times \pi_N(g,a_N)$

$\pi_i(g,a_i)$: تدبیر عامل i در انتخاب عمل a_i در بازی g،

$\gamma \in [0,1]$: ضریب تنزل

$P_{gg'}^{\sigma}$: تابع احتمال تغییر بازی که $P_{gg'}^{\sigma} = \Pr\left\{g^{k+1} = g' \mid g^k = g, \sigma^k = \sigma\right\}$



شکل ۱) فرآیند تصمیم‌گیری مارکوف بسیط

مقدار بهینه تابع ارزش- بازی ماکزیمم مقدار مجموع کل پاداش‌های تنزلی است که عامل i در صورتی بدست می‌آورد که از تدبیر بهینه پیروی کند. همانگونه که در تعریف ۶ بیان شد، نقطه تعادل نش بیانگر پروفایلی از تدبیرها است که یک عامل بیشترین ارزش ممکن را می‌تواند کسب کند.

$$\begin{aligned} V_i^*(g) &= \max_{\pi} V_i^{\pi}(g) \\ &= E^{\pi_N \mathrm{E}}\left[r_i^k + \gamma V_i^*\left(g^{k+1}\right) \mid g^k = g \right] \\ &= \mathrm{SPEV}_i\left[r_i(g,\sigma) + \gamma \sum_{\sigma'} P_{gg}^{\sigma} V_i^*(g') \right] \end{aligned}$$

که $\mathrm{SPEV}_i(.)$ ارزش حاصله برای عامل i در نقطه تعادل نش کامل زیربازی است.

تعبیر مشابهی می‌توان در رابطه با تابع ارزش- عمل ارائه کرد. ارزش حاصله از انتخاب عمل مشترک σ تحت تأثیر تدبیر π در بازی g برای عامل یادگیرنده i، که توسط $Q_i^{\pi}(g,\sigma)$ نمایش داده می‌شود، برابر است با:

$$\begin{aligned} Q_i^{\pi}(g,\sigma) &= r_i(g,\sigma) + \gamma \sum_{g'} P_{gg}^{\sigma} V_i^{\pi}(g') \\ &= r_i(g,\sigma) + \gamma \sum_{g'} P_{gg'}^{\sigma} \sum_{\sigma' \in \Sigma} \pi(g,\sigma') Q_i^{\pi}(g,\sigma') \end{aligned}$$

بنابر روش مشابهی، تابع بهینه ارزش- عمل به صورت زیر است:

$$Q_i^*(g,\sigma)=r_i(g,\sigma)+\gamma\sum_{g'}P_{gg}^{\sigma}V_i^*(g')$$

بنابراین می‌توان یادگیری تقویتی Q-Learning را توسط فرمول بازگشتی زیر در یک فرآیند تصمیم‌گیری مارکوف بسیط انجام داد [۳۲]:

$$Q_i^{k+1}(g,\sigma)=(1-\alpha^k)Q_i^k(g,\sigma)+\alpha^k\left[r_i(g)+\gamma\mathrm{SPEV}_i\left(\bar{Q}_i^k(g',\sigma')\right)\right]$$

که $\mathrm{SPEV}_i\left(\bar{Q}_i^k(g',\sigma')\right)$ ارزش انجام عمل نقطه تعادل نش کامل زیربازی برای عامل i در بازی بعدی g' با توجه به تمام عمل‌های مشترک امکانپذیر σ' است. عامل‌ها برای انتخاب عمل خود نیاز دارند که اولویت‌بندی سایر عامل‌ها را نیز داشته باشند. بنابراین عامل i عمل‌ها و پاداش‌های آنی سایر عامل‌ها را نیز مشاهده می‌کند و همزمان، باور خود از اولویت‌های سایر عامل‌ها را نیز به‌روز می‌رساند:

$$Q_j^{k+1}(g,\sigma)=(1-\alpha^k)Q_j^k(g,\sigma)+\alpha^k\left[r_j(g)+\gamma\mathrm{SPEV}_j\left(\bar{Q}_j^k(g',\sigma')\right)\right]\quad j\neq i, j=1,\ldots,N$$

فرآیند یادگیری فوق را می‌توان تحت الگوریتم زیر نشان داد.

1. Initialize:
   - 1.1. All extended Q-tables,
   - 1.2. k ← 0,
   - 1.3. Reset to g_0
2. Do until reaching goals:
   - 2.1. k ← k+1
   - 2.2. g ← current game
   - 2.3. For agent i, i = 1,…,N
     - 2.3.1. Select action a_i based on action selection strategy,
     - 2.3.2. Execute action,
   - 2.4. For agent i, i = 1,…,N
     - 2.4.1. Observe all actions, immediate rewards, and g',
     - 2.4.2. Update Q-values based on equation (5) and (6).
3. End.

انتخاب عمل

متناسب با ساختار بازی، نقطه تعادل نش [۳۴] و بعضی از مفاهیم توسعه یافته آن مانند نقطه تعادل نش کامل زیربازی[۳۲]، و نقطه تعادل استاکلبرگ[۳۵][1] به عنوان عمل بهینه در یادگیری تقویتی چندعاملی مورد استفاده قرار گرفت. از آنجا که اثبات همگرایی در الگوریتم‌های یادگیری تقویتی تدبیر- خارج[2] مستقل از روش اکتشاف[3] است و فقط به عنوان یک فرض اصلی مطرح می‌شود، در بسیاری از الگوریتم‌های مطرح شده یادگیری تقویتی چندعاملی به مسئله اکتشاف به صورت جدی پرداخته نشده است. در مسئله اکتشاف باید مصالحه مناسبی بین حالتی که عامل‌ها با درجه پایین‌تری از عقلانیت اقدام به انتخاب عمل‌هایی دورتر از نقطه تعادل نش کرده و حالتی که عامل‌ها اقدام به انتخاب نقطه تعادل نش می‌کنند، ایجاد کرد. به گونه‌ای که رفتارها به مرور از حالت اول به حالت دوم تغییر پیدا کنند. با توجه به الگوریتم یادگیری تقویتی چندعاملی ارائه شده [۲۷]، دو روش جدید به منظور ایجاد مصالحه بین اکتشاف و استخراج مطرح شد [۲۸]. مبنای این دو روش الگوریتم شناخته شده استقراء معکوس[4] است. این روش که به صورت حریصانه اقدام به انتخاب عمل می‌کند منجر به نقطه تعادل نش کامل زیربازی می‌شود.روش اجرای الگوریتم در یک بازی بسیط با اطلاعات کامل در شکل ۲ نشان داده شده‌است. این بازی بین سه بازیکن که هر یک قادر به انتخاب دو عمل هستند اجرا می‌شود.

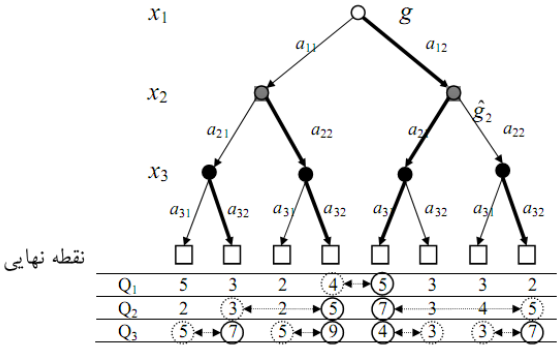
همانگونه که قابل مشاهده است، تعبیر ارزش عمل‌ها در زمان انتخاب مطابق با ارزش عمل‌ها برای سایر عامل‌ها تغییر می‌کند. این واقعیت، باعث بروز مفهوم جدید در انتخاب عمل‌ها می‌شود که آنرا ارزش- انجمنی[5] نامیده‌ایم.

تعریف ۷) تخمین ارزش یک عمل در یک بازی/ زیربازی با توجه به مجموعه عمل‌های مشترک سایر عامل‌ها ارزش- انجمنی نامیده می‌شود.

بر مبنای روشی که برای تخمین ارزش- انجمنی به کار گرفته می‌شود، می‌توان به طرق مختلف بین اکتشاف و استخراج مصالحه ایجاد کرد. در روش استفاده شده در این مقاله عامل فرض می‌کند عامل‌های بعدی یادگیری نداشته و بنابراین به همان روش استقراء معکوس اقدام به تعیین انتخاب بهینه عامل‌های بعدی کرده ولی خودش به جای انتخاب حریصانه با استفاده از توزیع بولتزمن اقدام

1. Stackelberg's equilibrium points
2.Off-policy
3. Exploration
4.Backward induction
5.Associative Q-values

به ایجاد مصالحه بین اکتشاف و استخراج می‌کند. این روش تحت عنوان استقراء معکوس محتملترین مسیر نامیده شد.



شکل ۲) الگوریتم استقراء معکوس در یک بازی بسیط با اطلاعات کامل. انتخاب حریصانه بین دو ارزش که با دایره نشان داده شده‌اند، اتفاق افتاده که ارزش برگزیده توسط دایره سیاه و عمل آن نیز با فلش سیاه نشان داده شده‌است.

در این روش مقادیر ارزش- انجمنی برای عامل یادگیرنده i که بعد از زیردنباله $\hat{\sigma}_{i-1} = (a_1,\ldots,a_{i-1})$ می‌خواهد عملش را انتخاب کند به صورت زیر محاسبه می‌شود:

که $\hat{g}_{i+1}^{\hat{\sigma}_i}$ بیانگر انتخاب بهینه عامل‌های بعدی در زیربازی $a_{i+1}^*,\ldots,a_N^* \in (A_{i+1}^* \times \ldots \times A_N^*)$ می‌باشد.

بازی جنگ

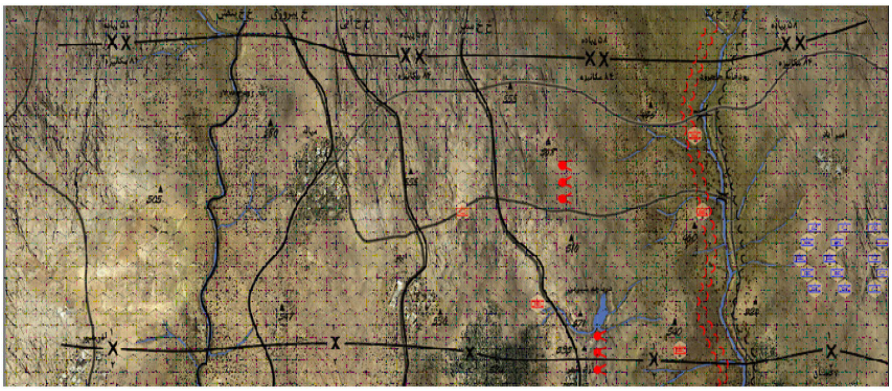
بازی جنگ یکی از مهمترین سامانه‌های دفاعی هر کشور ابزار کارآمدی در راستای ارزیابی راهبردها و فنون نظامی، آموزش فرماندهان و افسران نظامی، و تجزیه و تحلیل ادوات و تجهیزات نظامی می‌باشد. این سامانه در نسخه‌های امروزی در قالب نرم‌افزارهای کامپیوتری ارائه می‌شود، که از انواع نظامی آن می‌توان به JLASS[۳۶]، Army After Next [۳۷] و Joint Warrior[۳۸] اشاره کرد. واژه‌هایی دیگری همچون جبهه نبرد مجازی[1] وتمرین تاکتیکی بدون سرباز[2] نیز تحت همین مفهوم مورد استفاده قرار می‌گیرند. المان‌های سازنده این سامانه، یگان‌ها و تجهیزات نظامی هستند که هر یک به صورت یک عامل در سامانه های چندعاملی درنظر گرفته می‌شوند[۳۹].هر یک از این عامل‌ها قادر به انجام عملیات اصلی مربوط به خود هستند. معمولاً در میدان رزم، محدوده اختیارات و عملکرد

1 Virtual battlefield
2Tactical exercise without troop

عامل‌ها به صورت دقیق مشخص می‌شود. وظیفه اصلی کاربر این است که با توجه به قابلیت‌های هر یک از عامل‌ها، روش نبرد مناسب را تعیین کند. تعیین آرایش مناسب، مهمترین قسمت در طرح‌ریزی روش نبرد است. فاکتورهای زیادی در تعیین روش نبرد تأثیرگذار است که باعث پیچیدگی مسئله می‌شود. مواردی همچون تجربیات گذشته، اصول و قواعد اساسی رزم و خلاقیت مهمترین عواملی هستند که در صحنه نبرد بر روی آنها تکیه می‌شود. ولی در بازی جنگ، هدف این است که با استفاده از ابزارهای مناسب مبتنی بر شبیه‌سازی، بهترین روش نبرد تعیین شود. همانگونه که بیان شد، الگوریتم‌های یادگیری تقویتی در شرایطی که دانش کمی در رابطه با مسئله وجود دارد قادرند به منظور استخراج تدابیر بهینه به کار گرفته شوند. ساختار سلسله مراتبی در نیروهای نظامی باعث می‌شود که بازی مارکوف بسیط جهت مدلسازی فرآیند یادگیری مناسب باشد.

شکل۴ نمایشی از یک صحنه نبرد مجازی را نشان می‌دهد. در این مثال یگان‌های عمل کننده به دو دسته قرمز و آبی تقسیم می‌شوند. نیروهای خودی باید با عبور از خطوط دشمن که همراه با درگیری است، قوای خود را حفظ کرده و در پشت مواضع فعلی آنها استقرار یابند. نیروهای خودی به سه دسته

یگان فرماندهی، یگان مکانیزه و یگان پیاده تقسیم می‌شوند. از طرفی موقعیت‌های نظامی نیز به سه دسته ارزیابی، درگیری و استقرار تقسیم می‌شوند. در هر موقعیت نظامی، آرایش یگان‌ها با توجه به خط مقدم نبرد به صورت: یگان عمل‌کننده، یگان پشتیبان و یگان امن می‌باشد. در موقعیت ارزیابی، هدف انجام عملیات شناسایی با رزم، و شناخت توان نظامی طرف مقابل با کمترین تلفات نظامی است. یگان پیاده با کمک نیروهای مکانیزه اقدام به انجام عملیات شناسایی کرده و یگان فرماندهی در محل امن استقرار می‌یابد. در موقعیت درگیری، یگان مکانیزه نقش اصلی را ایفا کرده و یگان پیاده به آن کمک می‌کند. نهایتاً در موقعیت استقرار، یگان فرماندهی وارد صحنه شده و اختیار عمل در منطقه را در دست می‌گیرد. در این حالت یگان مکانیزه و یگان پیاده در پشت سر یگان فرماندهی قرار دارند.



شکل۱) شبیه‌سازی صحنه نبرد در سامانه بازی‌جنگ. نیروهای آبی باید ضمن حرکت از بین مواضع دشمن به موقعیت مناسب در پشت خطوط آنها رسیده و استقرار یابند.

اولویت در حفظ امنیت متعلق به یگان فرماندهی بوده و پس از آن یگان مکانیزه و یگان پیاده قرار دارند. بنابراین در هر موقعیت نظامی، یگان‌ها با توجه به اولویت مذکور اقدام به تعیین محل خود در آرایش نظامی می‌کنند. توان رزمی هر یگان در شروع برابر با ۱۰۰ است. در صورت حضور مؤثر در هر موقعیت نظامی، ۲۵ واحد از توان رزمی آن کاهش می‌یابد. ولی اگر هر یگان در محل مناسب خود قرار نگیرد، متحمل خسارت شدیدتری خواهد شد و ممکن است کل عملیات با شکست مواجه شود. روش محاسبه خسارت مازاد ناشی از جاگیری اشتباه در جدول ۲ نشان داده شده‌است.

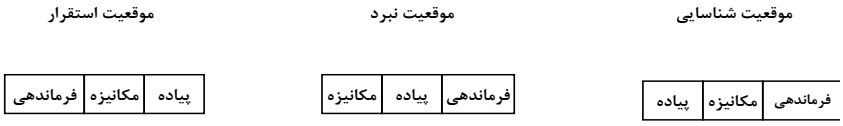
جدول ۲) محاسبه میزان اضافه خسارت یگان‌ها در صورت حضور در جایگاه اشتباه

	۲ سلول	۱ سلول
فرماندهی	۳۰	۲۰
مکانیزه	۲۰	۱۰
پیاده	۱۰	۰

جدول ۲ بیان می‌کند که به عنوان مثال اگر یگان فرماندهی به اندازه ۱ سلول در تعیین موقعیت مناسب خود اشتباه کند، ۲۰ واحد بیشتر توان رزمی خود را از دست خواهد داد؛ ولی اگر ۲ سلول در تعیین موقعیت خود اشتباه کند میزان مازاد خسارت وارده ۳۰ واحد خواهد بود. به عنوان مثال اگر در موقعیت نظامی ارزیابی یگان فرماندهی به جای حضور در جایگاه سوم در جایگاه اول قرار بگیرد، به

میزان ۳۰+۲۵ واحد از توان رزمی خود را از دست خواهد داد. این مسئله ناشی از آسیب‌پذیری بیشتر یگان فرماندهی نسبت به سایر یگان‌ها است. همانگونه که در جدول ۲ نشان داده شده‌است، یگان مکانیزه و یگان پیاده به ترتیب آسیب‌پذیری کمتری در برابر جاگیری اشتباه دارند.

در صورتی که توان رزمی یگان برای انجام یکی از مراحل کافی نباشد (مثلاً یگان فرماندهی در دو موقعیت قبلی دچار خسارت شدیدی شود)، یعنی کمتر از ۲۵ واحد از توان رزمی برای مرحله آخر باقی مانده باشد، کل عملیات با شکست مواجه شده و خسارت وارده به همه یگان‌ها برابر ۱۰۰ خواهد بود.



شکل۱) آرایش مناسب نظامی در موقعیت‌های مختلف

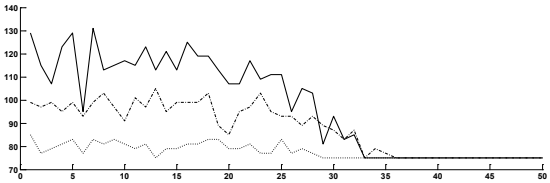
هدف یگان‌ها این است که با عبور از این سه موقعیت، کمترین میزان خسارت را متحمل شوند. به منظور استفاده از الگوریتم یادگیری تقویتی چندعاملی و پیدا کردن تدبیر بهینه برای هر عامل می‌توان این مسئله را در قالب بازیهای مارکوف بسیط مطرح کرد که چندگانه $\Psi = \langle G, X, \Sigma, P, R \rangle$ به صورت زیر است:

مجموعه بازیها G: موقعیت‌های نظامی شامل {ارزیابی، نبرد، استقرار}،

- مجموعه عامل‌ها X: یگان‌های عمل‌کننده به ترتیب اولویت در انتخاب عمل، شامل {فرماندهی، مکانیزه، پیاده}،
- مجموعه عمل‌های مشترک امکان‌پذیرS: کلیه حالت‌های ممکن آرایش‌بندی یگان سه‌گانه،
- تابع تغییر بازیها P: به سادگی قابل بیان است،
- تابع پاداش R: به سادگی قابل بیان است.

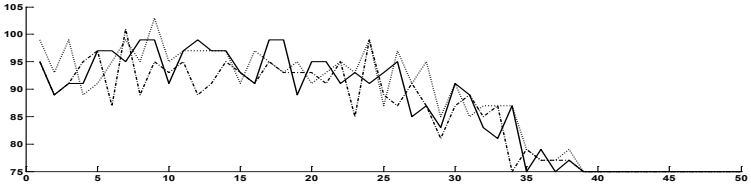
شبیه سازی

در اکثر مقالات، الگوریتم‌ها یادگیری تقویتی چندعاملی در بازیهایی با محیط مشبک[1] پیاده‌سازی شده‌اند[۴۰] و [۲۵]. این محیط‌ها قادرند بسیاری از مسائل مربوط به پیاده‌سازی را نشان دهند، ولی کارآیی این الگوریتم‌ها در مسائل دنیای واقعی همچنان به صورت مبهم قابل طرح است. با استفاده از مسئله بازی‌جنگ مطرح شده در بخش ۴، می‌توان کارآیی الگوریتم‌های ارائه شده را بررسی کرد. شکل ۵نتایج مربوط به شبیه‌سازی را نشان می‌دهد. پس از هر ۱۰ بار اجرا، متوسط کل خسارت وارده بر نیروها محاسبه شده‌است. میزان خسارتی که یگان‌ها به ترتیب اولویت در تصمیم‌گیری متحمل می‌شوند به مرور کاهش می‌یابد. میزان خسارت یگان فرماندهی بیشتری از دو یگان دیگر می‌باشد.



شکل ۵) میزان خسارات وارده به یگان‌ها در پایان عملیات به سمت میزان مطلوب همگرا می‌شود. میزان خسارت یگان فرمانده با خط مشکل، یگان مکانیزه با خط-نقطه و یگان پیاده با استفاده از خط‌چین نشان داده شده‌است.

همانگونه که درشکل ۶ نشان داده شده‌است، اگر در تعریف صورت مسئله، میزان خسارت یگان‌ها با هم مساوی باشد، مدت زمان لازم جهت همگرایی اندکی افزایش می‌یابد.



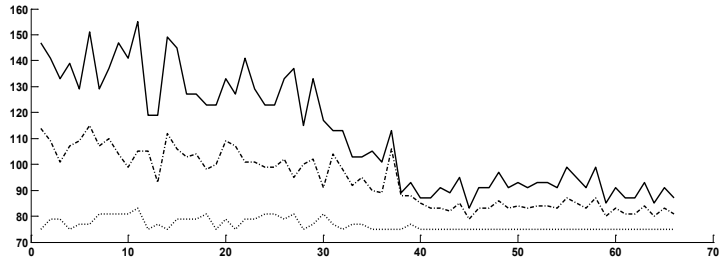
شکل ۶) در صورتی که میزان خسارت وارده به هر یگان در ازاء تصمیم اشتباه به صورت مساوی تعریف شود، یگان‌ها دیرتر همگرا می‌شوند.

مسئله‌ای که تاکنون از آن چشم‌پوشی شده‌است، قدرت تصمیم‌گیری یگان پیاده است. در صورت مسئله کنونی، یگان پیاده مجبور به رعایت آرایشی است که سایر یگان‌های بالاتر نسبت به انتخاب آن

1.Grid world games

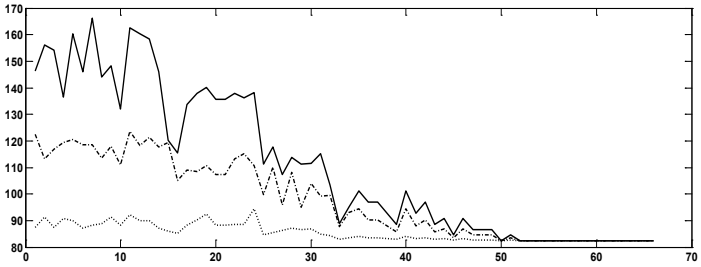
اقدام کرده‌اند. حال آنکه در صحنه نبرد ممکن است که یکی از یگان‌ها از شرکت در عملیات سرباز زند و از حضور در نبرد خودداری کند. در این صورت، خسارت بیشتری به سایر یگان‌ها وارد می شوند، ولی یگان فراری آسیب نخواهد دید.

در شبیه‌سازی شکل ۷ یگان فرماندهی و یگان مکانیزه، خسارت بیشتری را متحمل می‌شوند که ناشی از عدم همکاری یگان پیاده در آرایش‌بندی است. اگر چه یگان پیاده در سطح پایین‌تری از تصمیم‌گیری قرار دارد، ولی ممکن است با انتخاب نامناسب امکان رسیدن به سطح مطلوبی از خسارت را غیرمحتمل سازد. یگان پیاده انگیزه‌ای برای شرکت در آرایش‌بندی نداشته و لازم است به منظور غلبه بر این مشکل تمایل به همکاری بین یگان‌ها را ایجاد کرد.



شکل ۷) خسارت وارد به یگان‌های فرماندهی (خط پر) یگان مکانیزه (خط- نقطه) و یگان پیاده (خط‌چین)

به این منظور، در پایان هر موقعیت نظامی، ۵ درصد از میزان خسارتی که هر یگان متحمل شده‌است، به سایر یگان‌ها منتقل خواهد شد. بنابراین یگان پیاده که به صورت غیرمسئولانه فقط به فکر منفعت خود بوده و در شبیه‌سازی قبلی انگیزه‌ای برای شرکت در آرایش‌بندی نداشته است، ملزم به نقش‌پذیری در عملیات خواهد شد.



شکل ۸) تأثیر ایجاد همکاری در بین یگان‌ها در صورتی که امکان رفتارهای غیر مسئولانه نیز وجود داشته باشد.

همانگونه که در شکل ۸ مشخص شده‌است، یگان پیاده پس از مدتی ملزم به رعایت موقعیت مناسب در عملیات شده که باعث می‌شود سایر یگان‌ها نیز به حد مناسبی از خسارت برسند. در صورتی که سطح همکاری بین یگان‌ها کمتر از ۵ درصد لحاظ شود، یگان‌پیاده با تأخیر بیشتری در عملیات شرکت می‌کند که چندان مطلوب نیست. از طرفی افزایش آن نیز باعث می‌شود ریسک‌پذیری آنها در اکتشاف آرایش‌بندیهای جدید کاهش یابد.[۱] به گونه‌ای که افزایش آن تا ۲۰درصد اندکی باعث تسریع در همگرایی می‌شود، ولی پس از آن تأثیر چندانی در عملکرد یگان‌ها ندارد.

شبیه‌سازیهای فوق کاربری یادگیری تقویتی چندعاملی در بازی‌جنگ و بررسی پدیده‌های مختلف جنگی و تأثیر آن در بروز رفتارهای معقول را نشان می‌دهند. به این صورت می‌توان راهکارهای مقابله با رفتارهای نامناسب را نیز بررسی کرده و تدبیر مناسبی جهت برخورد با آنها طراحی کرد.

نتیجه‌گیری

بازی‌جنگ یکی از مهمترین سامانه‌های دفاعی هر کشور است که نقش مهمی درتعیین دکترین دفاعی آنها تعیین می‌کند. یکی از کاربردهای مهم این سامانه، تعیین آرایش بهینه در هنگام نبرد است. در این مقاله، با استفاده از یادگیری تقویتی چندعاملی روشی جهت تعیین آرایش بهینه نیروهای نظامی در بازی‌جنگ ارائه شده‌است. به این منظور، الگوریتم‌یادگیری تقویتی چندعاملی مبتنی بر بازیهای بسیط Extensive-Q و روشهای مناسب ایجاد مصالحه بین اکتشاف و استخراج مطرح شد. ساختار سلسله مراتبی بین یگان‌ها باعث می‌شود که بسیاری از الگوریتم‌های مطرح شده، که در قالب بازیهای نرمال هستند، قابل استفاده نباشند. به ویژه که محاسبه نقطه تعادل نش در بازیهای نرمال با بیشتر از ۲ عامل بسیار پیچیده است [۴۱]. نتایج شبیه سازی نشان می‌دهد که این الگوریتم به تدبیر بهینه همگرا شده و می‌توان آنرا در سامانه بازی‌جنگ به کار گرفت. به دلیل استفاده از الگوریتم استقراء معکوس، افزایش تعداد عامل‌ها تأثیر چندانی بر مسئله ندارد.

حل مسئله آرایش‌بندی در بازی‌جنگ در حالتی که اطلاعات قسمتی از عامل‌ها مخدوش شده با استفاده از الگوریتم‌های یادگیری مبتنی بر بازیهای بسیط با اطلاعات ناکامل[۲] در مقالات بعدی مورد توجه قرار خواهد گرفت.

۱. البته در صورت مسئله فعلی که تعداد بازیهای قابل تصور زیاد نیست، تأثیر خود را چندان نشان نمی‌دهد.

۲. Incomplete information extensive form games

مراجع

1.Ali Akramizadeh(3), Mohammad B. Menhaj, and Ahmad Afshar, "Extensive Q-learning, a new approach in multiagent reinforcement learning," in *Adaptive Dynamic Programming and Reinforcement Learning (ADPRL2009), Series Symposium in Computational Intelligenence (SSCI2009)*, Nashville, 2009.

2.G. M. Whittaker, «Asymmetric Wargaming: Toward A Game Theoretic Perspective,» 2000.

3.Alan Washburn and Moshe Kress, *Combat Modeling*. New York: Springer, 2009.

4.Michael Luck, Peter McBurney, and Chris Preist, *Agent Technology: Enabling Next Generation Computing*.: AgentLink community, 2003.

5.Michael Wooldridge, *An Introduction to MultiAgent Systems*. Chichester, England: John Wiley & Sons, 2002.

6.Gerhard Weiss, *Multiagent Systems: A Modern Approach to Distributed Modern Approach to Artificial Intelligence*. London: MIT Press, 1999..

7.M. Wooldridge, *Intelligent agents. In Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence*.: MIT Press, 2000.

8.R.W. Pew and A.S. Mavor, *Modeling Human and Organizational Behavior :Application to Military Simulations*. Washington, DC: National Academy Press, 1998.

9.Abdeslem Boukhtouta, «Planning Military Opearations under Uncertainty,» 2002.

10.Eduardo Alonso and Daniel Kudenko, "Logic-based Multi-Agent Systems for

Conflict Simulations: A preliminary report on architecture and implementation,"

in *Third Workshop of the UK Special Interest Group on Multi-Agetnt Systems*, 2000.

11.E. Alonso and D. Kudenko, «Machine learning techniques for Adaptive Logic-based 14.Multi-Agent Systems,» in *In Proceedings of The Second Workshop of the UK Special Interest Group on Multi-Agent Systems*, Bristol, 1999.

12.L. Panait and S. Luke, "Cooperative multi-agent learning: The state of the art," *Autonmous Agents Multi-Agent Systems, vol. 11, no. 3*, p. 387–434, 2005.

13.M. A. Potter and K. A. D. Jong, "A cooperative coevolutionary approach to function optimization," , Jerusalem, Israel, 1994.

14.Karl Tuyls, Ann Nowe, Tom Lenaerts, and Bernard Manderick, "An Evolutionary game theoretic perspective on learning in multi-agent systems," *Kluwer Academic Publishers*, 2004.

15.Karl Tuyls, Pieter Jan 't Hoen, and Bram Vanschoenwinkel, "An Evolutionary Dynamical Analysis of Multi-Agent Learning in Iterated Games," in *Autonomous Agents and Multi-Agent Systems*.: Springer, 2006.

16.F. Ho and M. Kamel, "Learning coordination strategies for cooperative multiagent systems," *Machine Learning,vol. 33, no.2–3*, p. 155–177, 1998.

17.C. Claus and C. Boutilier, "The dynamics of reinforcement learning in cooperative multiagent systems," in *Proceedings of the Fifteenth National Conference of Artificial Intelligence(AAAI-98)/Proceedings of the Tenth Conference of Innovative Applications of the Artificial Intelligence (IAAI-98)*, Madison, 1998.

18.R. Powers and Y. Shoham, "New criteria and a new algorithm for learning in multi-agent systems," in *Proceeding of Advance Neural Information Process*

System, Vancouver, 2004, p. 1089– 1096.

19.J. Hu and P. Wellman, "Multiagent reinforcement learning: Theoretical framework and an algorithm," in *In Proceedings of the Fifteenth International Conference on Machine Learning*, 1998, p. 242–250.

20.Michael Lederman Littman, "Algorithms for Sequential Decision Making," Providence, Rhode Island, 1996.

21.M. L. Littman, "Friend-or-foe Q-learning in general-sum games," , 2001

22.X. Wang and T. Sandholm, "Reinforcement learning to play an optimal Nash equilibrium in team Markov games," *Advances Neural Information Processing System (NIPS-02), Vancouver, Canada*, p. 1571–1578, 2002.

23.V. Conitzer and T. Sandholm, "AWESOME: A general multiagent learning algorithm that converges in self-play and learns a best response against stationary opponents," , 2003.

24.Lucian Busoniu, Robert Babuska, and Bart De Schutter, "A Comprehensive Survey of Multiagent Reinforcement Learning," *IEEE Transaction on Systems, Man, and Cybernetics,Part C: Applications and Reviews, Vol. 38, No. 2,* pp. 156-172, 2008.

25.Charles Madeira, Vincent Corruble, and Geber Ramalho, "Designing a Reinforcement Learning-based Adaptive AI for Large-Scale Strategy Games ," in *American Association for Artificial Intelligence*, 2006.

26.Amandeep Singh Sidhu, Narendra S. Chaudhari, and Ghee Ming Goh,

"Hierarchical Reinforcement Learning Model for Military Simulations," in *International Joint Conference on Neural Networks*, Vancouver, 2006 , pp. 2572-2576.

27.Folkert Huizinga, "Hierarchical Reinforcement Learning and Safe State Abstraction on Realtime Strategy Games," 2007.

28.Ville Kononen, "Asymmetric multiagent reinforcement learning," *Web Intelligence and Agent Systems: An international journal*, p. 105–121, 2004.

29.Ali Akramizadeh(2), Ahmad Afshar, and Mohammad –B. Menhaj, "Different Forms of the Games in Multiagent Reinforcement learning: Alternating vs. simultanous movements," in *17th Mediterranean Conference on Control and Automation*, Thessaloniki, Greece, 2009.

30.Ali Akramizadeh, Ahmad Afshar, and Mohammad -B Menhaj, "Exploration strategies in n-Person general-sum multiagent reinforcement learning with sequential action selection," *Accepted to be published in Intelligent Data Analysis*, 2010.

31.J. Von Neumann and O. Morgenstern, *Theory of Games and Economic Behavior*.: Princeton University Press, 1944

32.Martin J. Osborne, *An Introduction to Game Theory*.: Oxford University Press, 2000

33.Guillermo Owen, *Game Theory: Third edition*. Orlando, Florida: Academic Press, 1995.

34.Ali Akramizadeh, Ahmad Afhsr, and Mohammad Bagher Menhaj, "Multiagent 38.Reinforcement Learning in Extensive Form Games with

Perfect Information," *Journal of Applied Science 9(11)*, pp. 2056-2066, 2009.

35.D. Fundenberg and D. k. Levine, *The theory of learning in games*.

Cambridge, Massachusetts: MIT Press, 1998.

36.J. Hu and P. Wellman, "Multiagent reinforcement learning: Theoretical framework and an algorithm," in *In Proceedings of the Fifteenth International Conference on Machine Learning*, 1998, p. 242–250.

37.Ville Kononen, "Asymmetric multiagent reinforcement learning," *Web Intelligence and Agent Systems: An international journal*, vol. 2, no. 3, p. 105–121, 2004.

38.James C. Hyde and Michael W. Everett. (1996) JLASS: Educating Future Leaders in Strategic and Operational Art. [Online]. http://www.au.af.mil/au/awc/awcgate/jfq/0912.pdf

39.Walter L. Perry, Bruce R. Pirnie, and John Gordon, *Issues Raised During the Army After Next 43.Spring Wargame*, 4th ed. Washington, USA: RAND Publicaitons, 1999. [Online]. http://www.boozallen.com/consulting-services/services_article/981533

40.Bobby J. Wilkes and Barrett S. Elliott. (2003) COLLEGE OF AEROSPACE DOCTRINE, RESEARCH AND EDUCATION. [Online]. http://www.cadre.maxwell.af.mil/

41.(2004) A Multi-Agent System for Efficient Strategies and Tactics in Wargames. [Online]. http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=AE71A0314C173C25468E38E931792E34?doi=10.1.1.103.4059&rep=rep1&type=pdf

42.J. Hu and M. Wellman, “Nash Q-learning for general-sum stochastic games,” *Journal of Machine Learning Research*, vol. 4, p. 1039–1069, 2003.

43.C. Daskalakis, P.W. Goldberg, and C.H. Papadimitriou, “The Complexity of Computing a Nash Equilibrium,” *ECCC, TR05-115*, 2005.