بهبود یادگیری Q با استفاده از هم‌زمانی به روز رسانی و رویه تطبیقی بر پایه عمل متضاد

مریم پویان، شهرام گلزاری، امین موسوی و احمد حاتم

چکیده: روش یادگیری Q یکی از مشهورترین و پرکاربردترین روش‌های یادگیری تقویتی مستقل از مدل است. از جمله مزایای این روش عدم وابستگی به آگاهی از دانش پیشین و تضمین در رسیدن به پاسخ بهینه است. یکی از محدودیت‌های این روش کاهش سرعت همگرایی آن با افزایش بعد است. بنابراین افزایش سرعت همگرایی به عنوان یک چالش مطرح است. استفاده از مفاهیم عمل متضاد در یادگیری Q، منجر به بهبود سرعت همگرایی می‌شود زیرا در هر گام یادگیری، دو مقدار Q به طور هم‌زمان به روز می‌شوند. در این مقاله روشی ترکیبی با استفاده از رویه تطبیقی در کنار مفاهیم عمل متضاد برای افزایش سرعت همگرایی مطرح شده است. روش‌ها برای مسئله Grid world شبیه‌سازی شده است. روش‌های ارائه‌شده بهبود در میانگین درصد نرخ موفقیت، میانگین درصد حالت‌های بهینه، متوسط تعداد گام‌های عامل برای رسیدن به هدف و میانگین پاداش دریافتی را نشان می‌دهند.

کلیدواژه: **رویه تطبیقی، سرعت همگرایی، عمل متضاد، هم‌زمانی به روز رسانی، یادگیری Q.**

۱- مقدمه

یادگیری تقویتی یک روش هوشمند هدف‌محور است که به عنوان نتیجه‌ای از تلاش بشر برای آنالیز رفتار سیستم‌های مصنوعی توسعه یافته است. این روش برای حل مسایل در محیط‌های ناشناخته در نظر گرفته می‌شود. در این محیط‌ها، عامل باید بتواند با استفاده از تجربیات خودش و به شیوه آزمون و خطا یادگیری را انجام دهد.

حجم و تنوع تحقیقاتی که در زمینه حل مسایل مختلف از جمله کنترل هوشمند و علوم شناختی با استفاده از این روش گزارش شده‌اند، اشاره بر توانایی الگوریتم‌های یادگیری تقویتی در حل مسایل دارد. یکی از مزایای استفاده از یادگیری تقویتی، عدم وابستگی به آگاهی از دانش پیشین است. با توجه به این که روش‌های یادگیری تقویتی با افزایش بعد بسیار کند و غیر کاربردی می‌شوند، در نتیجه توسعه روش‌هایی که به افزایش سرعت یادگیری تقویتی انجامد دارای اهمیت بسیار هستند.

روش یادگیری Q یکی از مشهورترین روش‌های یادگیری تقویتی است [۱]. یادگیری Q کاربردهای وسیعی در زمینه‌های ره‌یابی ربات [۲] تا [۴]، مسایل اقتصادی [۵] و [۶] و تشخیص الگو [۷] دارد. هدف از یادگیری، یافتن تخمینی از تابع ارزش- عمل بهینه است که مقادیر Q نامیده می‌شود. مقادیر Q سودمندی مورد انتظار از انجام یک عمل در یک وضعیت را نشان می‌دهد.

چالش پیش روی یادگیری Q، زمانبر بودن و کندبودن سرعت همگرایی آن است زیرا برای تضمین همگرایی یادگیری، تمامی زوج‌های حالت- عمل باید بی‌نهایت بار بازدید شوند. تا کنون روش‌های گوناگونی برای بهبود یادگیری Q ارائه شده‌اند که عمده این روش‌ها به پنج دسته اصلی تقسیم می‌شوند که عبارتند از استراتژی به روز رسانی مقادیر Q [۸] و [۹]، کاهش فضای حالت [۱۰]، استفاده از دانش پیشین و مقداردهی اولیه Q [۱۱] و [۱۲]، شکل‌دهی تابع پاداش [۱۳] و [۱۴]، و استراتژی انتخاب عمل [۱۵] تا [۱۷]. تمامی این روش‌ها تلاش بر بهبود یادگیری دارند تا بتوانند در کمترین زمان به پاسخ بهینه یا شبه‌بهینه برسند.

ایده تضاد در یادگیری تقویتی اولین بار توسط تیزهوش مطرح و در [۸] سه روش یادگیری Q مبتنی بر تضاد معرفی شده است. نتایج آزمایش‌ها نشان می‌دهد که افزایش تعداد به روز رسانی مقادیر Q با استفاده از عمل متضاد باعث افزایش سرعت همگرایی شده است. در [۱۶] و [۱۷] اکتشاف بر پایه تفاضل ارزش برای بهبود رویه انتخاب عمل بیان شده و به صورت کلی معایب و مزایای پژوهش‌های پیشین در [۱۸] شرح داده شده است.

اگرچه تا کنون کارهای زیادی برای افزایش سرعت روش یادگیری Q انجام شده است اما هنوز افزایش سرعت همگرایی این الگوریتم به عنوان چالشی مطرح است. در این نوشتار از الگوریتم یادگیری Q مبتنی بر تضاد الهام گرفته شده است. با تغییر در تابع پاداش و قسمت به روز رسانی مقادیر Q، روشی جدید مطرح می‌شود. در روش ارائه‌شده، مقادیر Q اشتباه که در [۸] وجود داشته است و عامل سردرگمی بوده است، برطرف شده است. علاوه بر این در این مقاله استراتژی‌های به روز رسانی مقادیر Q و انتخاب عمل در جهت بهره‌گیری از مزایای هر دو روش ترکیب شده است. ایده ترکیب این روش‌ها به منظور تسریع در فرایند یادگیری و بهبود سرعت همگرایی ارائه شده است.

در بخش ۲ ابتدا یادگیری Q به صورت مختصر شرح داده شده و سپس مروری بر پژوهش پیشین در زمینه یادگیری Q مبتنی بر تضاد و انتخاب عمل شده است. در بخش ۳ روش‌های پیشنهادی برای افزایش سرعت همگرایی در یادگیری Q بیان شده است. ارزیابی و نتایج آزمایش‌ها در بخش ۴ و در نهایت نتیجه‌گیری در بخش ۵ آورده شده است.

۲- مفاهیم مورد نیاز

۲-۱ یادگیری Q

یادگیری Q روش کنترلی تفاضل زمانی off-policy[1] است که به طور

این مقاله در تاریخ ۷ اسفند ماه ۱۳۹۳ دریافت و در تاریخ ۱۶ مهر ماه ۱۳۹۴ بازنگری شد.

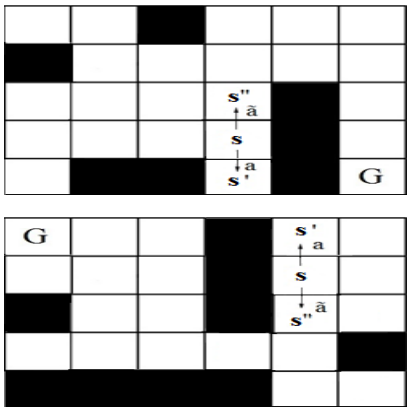
مریم پویان، گروه برق و کامپیوتر، دانشگاه هرمزگان، بندر عباس، (email: pouyan.student@hormozgan.ac.ir).

شهرام گلزاری، گروه برق و کامپیوتر، دانشگاه هرمزگان، بندر عباس، (email: golzari@hormozgan.ac.ir).

امین موسوی، گروه برق و کامپیوتر، دانشگاه هرمزگان، بندر عباس، (email: mousavi@hormozgan.ac.ir).

احمد حاتم، گروه برق و کامپیوتر، دانشگاه هرمزگان، بندر عباس، (email: a.hatam@hormozgan.ac.ir).

1. Temporal Difference



شکل ۱: مثال‌هایی از مقادیر Q اشتباه. در محیط‌ها، موقعیت هدف با G مشخص شده است. عامل در موقعیت s قرار دارد، عمل a را انجام می‌دهد، به حالت s' منتقل می‌شود، پاداش دریافت می‌کند و هم‌زمان برای عمل متضاد $\tilde{a}$ مجازات می‌شود. بنابراین، مقادیر Q این حالت‌ها اشتباه است.

هم‌زمان در محیط کاوش انجام می‌دهد و رویه بهینه[1] را می‌آموزد. این روش مستقل از محیط، اولین بار توسط واتکینز [۱۹] در سال ۱۹۸۹ بیان شد. روش یادگیری Q را می‌توان به کمک یک جدول پیاده‌سازی نمود. هر خانه جدول به یک جفت حالت- عمل تعلق دارد که هر سطر جدول یک حالت از مجموعه حالت‌ها و هر ستون آن یک عمل از مجموعه عمل‌ها را مشخص می‌کند. با هر بار فراخوانی رابطه به روز رسانی که در (۱) آمده است یک خانه از جدول تغییر می‌کند

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma \max_{a'} Q(s',a') - Q(s,a)] \quad (۱)$$

در (۱) s حالت فعلی، a عمل، s' حالت بعدی، a' عمل بعدی، r پاداش فوری، α پارامتر نرخ یادگیری، γ فاکتور تخفیف یا ضریب تنزیل و $Q(s,a)$ ارزش حالت- عمل برای حالت s و عمل a است.

یادگیری در الگوریتم Q بدین صورت است که در هر دوره عامل در یک حالت تصادفی یادگیری را شروع می‌نماید، عملی را طبق سیاست مشتق‌شده از مقادیر Q انتخاب نموده و پس از انجام عمل، پاداش آن را از محیط دریافت می‌کند. حالت بعدی محیط را مشاهده می‌کند و تا رسیدن به حالت پایانی این روند را تکرار می‌کند.

۲-۲ یادگیری بر مبنای تضاد

روش‌های متفاوتی برای تخمین شبکه‌های عصبی، بهینه‌سازی الگوریتم‌های تکاملی و جستجوی رویه بهینه وجود دارد. در بسیاری موارد مانند وزن‌ها در شبکه عصبی، کروموزوم‌ها در الگوریتم ژنتیک و رویه انتخاب عمل در یادگیری تقویتی، فرایند جستجو و یادگیری با مقداردهی اولیه تصادفی شروع می‌شود. حدس تصادفی اگر به راه حل بهینه نزدیک باشد می‌تواند باعث افزایش سرعت همگرایی شود اما اگر حدس اولیه در وضعیت متضاد راه حل بهینه قرار گرفته باشد، زمان یافتن راه حل بهینه افزایش پیدا می‌کند. بنابراین در صورت عدم دانش اولیه، باید در همه جهات به طور هم‌زمان به دنبال راه حل بود. واضح است که اگر جستجو در جهت مخالف معنادار باشد از همان ابتدا به طور هم‌زمان استفاده از مقدار متضاد می‌تواند مفید باشد [۸].

الگوریتم‌های مبتنی بر تضاد با افزایش تعداد به روز رسانی مقادیر Q، باعث افزایش سرعت همگرایی می‌شوند زیرا اگر عامل ارزش عمل مخالف را نیز بداند، به جای یک مقدار می‌تواند به طور هم‌زمان دو مقدار از تابع ارزش را به روز رسانی کند که به طور کلی فرایند یادگیری را سرعت می‌بخشند و به طور خاص موجب کوتاه‌شدن زمان اکتشاف می‌شوند [۸].

سه الگوریتم مبتنی بر تضاد مشتق‌شده از یادگیری Q در [۸] معرفی شده است. ایده اصلی این الگوریتم‌ها این است که در هر گام یادگیری، اگر عامل به ازای یک عمل پاداش دریافت کند برای عمل متضاد متناظر با آن مجازات می‌شود.

در اولین نسخه الگوریتم (OQL۱)[2] به روز رسانی برای مقادیر حالت- عمل و حالت- عمل متضاد متناظر، مطابق (۲) انجام می‌گیرد

$$\begin{aligned} Q(s,a) &\leftarrow Q(s,a) + \alpha[r + \gamma \max_{a'} Q(s',a') - Q(s,a)] \\ Q(s,\tilde{a}) &\leftarrow Q(s,\tilde{a}) + \alpha[\tilde{r} + \gamma \max_{a'} Q(s',a'') - Q(s,\tilde{a})] \end{aligned} \quad (۲)$$

در این الگوریتم فرض شده است در هر گام یادگیری عامل، با انجام عمل a و دریافت پاداش r، برای عمل متضاد متناظر $\tilde{a}$، پاداش متضاد $\tilde{r}$ دریافت می‌کند.

در دومین نسخه الگوریتم (OQL۲) برای مقدار حالت- عمل به روز رسانی انجام می‌گیرد و نرخ یادگیری به صورت تابع کاهشی محاسبه می‌شود. سپس مقدار تابع ارزش برای جفت حالت- عمل متضاد نیز به روز می‌شود. به عبارت دیگر مانند (۳) داریم

$$\begin{aligned} &Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma \max_{a'} Q(s',a') - Q(s,a)] \\ &\tilde{\alpha} = \sqrt{1 - \frac{i}{n_E}} \\ &Q(s,\tilde{a}) \leftarrow Q(s,\tilde{a}) + \tilde{\alpha}[\tilde{r} + \gamma \max_{a'} Q(s',a'') - Q(s,\tilde{a})] \end{aligned} \quad (۳)$$

که در آن i نشان‌دهنده تکرار و n_E تعداد اپیزود است [۸]. در سومین نسخه الگوریتم (OQL۳) به روز رسانی اضافی برای تعداد محدودی از دوره‌ها در آغاز یادگیری، مثلاً ۱/۴ تعداد کل دوره‌ها انجام می‌گیرد.

در پژوهش [۸] با فرض این که موقعیت هدف شناخته شده است، فاصله اقلیدسی بین هدف و عامل برای تعریف تابع پاداش در نظر گرفته شده است. تابع پاداش مطابق (۴) تعریف شده است

$$r = \begin{cases} +1 & \text{if} \quad d_i \le d_{i-1} \\ -1 & \text{if} \quad d_i > d_{i-1} \end{cases} \quad (۴)$$

که در آن d فاصله عامل از هدف را نشان می‌دهد که به صورت (۵) است

$$\sqrt{(x_G - x_A)^2 + (y_G - y_A)^2} \quad (۵)$$

که (x_G, y_G) موقعیت هدف و (x_A, y_A) موقعیت عامل را نشان می‌دهد [۸].

به کارگیری این تابع پاداش در محیط‌هایی که مانع وجود دارد، باعث گمراه‌نمودن عامل می‌شود زیرا وجود مانع بین عامل و هدف در نظر گرفته نشده است. دو محیط نشان داده شده در شکل ۱ نشان‌دهنده این امر هستند.

در شکل ۱ نمایی از دو محیط نشان داده شده که به دلیل انتساب پاداش‌های نادرست، باعث ایجاد مقادیر Q اشتباه می‌شوند. فرض شده که عامل در موقعیت s قرار دارد و عمل a را انجام می‌دهد و به حالت s' می‌رود، چون فاصله تا هدف کاهش یافته، بنابراین پاداش ۱+ دریافت می‌کند و هم‌زمان برای عمل مخالف $\tilde{a}$ مجازات ۱- دریافت می‌کند زیرا

1. Optimal Policy

2. Opposite Q-Learning

فاصله تا هدف برای انتساب پاداش و جریمه برای عامل استفاده شده است. تولید مقادیر Q اشتباه باعث گمراهی عامل در انتخاب عمل می‌باشد و شکست عامل در رسیدن به هدف را در پی دارد.

۲-۳ انتخاب عمل

در یادگیری تقویتی، رفتار عامل در هر زمان توسط رویه عمل تعریف می‌شود. نگاشت بین حالت و عمل را رویه می‌نامند [۱]. استراتژی انتخاب عمل بر فرایند یادگیری تأثیر می‌گذارد و بنابراین سؤال مطرح‌شده این است که کدام استراتژی در انتخاب عمل آموزش بهینه‌تری را فراهم می‌کند. تعادل بین اکتشاف و اکتساب یکی از مسایل کلیدی انتخاب عمل در یادگیری تقویتی است.

اکتساب در استراتژی اگر به گونه‌ای باشد که صرفاً بر اساس مقادیر ارزش حالت- عمل فعلی صورت گیرد باعث رخداد این مشکل می‌شود که عامل به سرعت به یک رویه بهینه محلی همگرا شود. به عنوان مثال انتخاب حریصانه، دارای مشکل همگرایی به یک رویه بهینه محلی است. در مقابل، استراتژی اکتشاف مبتنی بر این فرض است که عامل برای رسیدن به یک رویه بهینه سراسری با توجه به شرایط فعلی یک عمل تصادفی را انتخاب کند. کاوش بیش از حد هر چند که ممکن است عملکرد یادگیری را بهبود بخشد و از رویه بهینه محلی فرار کند ولی در بسیاری موارد عملکرد الگوریتم را کاهش می‌دهد.

در حوزه یادگیری تقویتی $\varepsilon-\text{greedy}$ و Softmax دو رویه معروف هستند. رویه $\varepsilon-\text{greedy}$ مانند (۶) تعریف می‌شود

$$\pi(s)=\begin{cases}\text{random action from } A(s) \text{ if } \text{rand}(0,1)<\varepsilon \\ \arg\max_{a\in A(s)} Q(s,a) \quad \text{otherwise}\end{cases} \qquad (۶)$$

که $A(s)$ نشان‌دهنده مجموعه اعمال در حالت s است و $rand$ یک عدد تصادفی با توزیع یکنواخت در بازه $[0,1]$ تولید می‌کند.

در این رویه در هر گام زمانی، عمل تصادفی با احتمال ε و عمل حریصانه که دارای بالاترین ارزش است با احتمال $1-\varepsilon$ انتخاب می‌شود. یکی از معایب این رویه این است که عمل‌های اکتشافی با توزیع یکنواخت انتخاب می‌شوند و به عبارت دیگر در یک حالت، برای عمل‌های غیر بهینه نیز احتمال مساوی با دیگر عمل‌ها در نظر گرفته می‌شود. انتخاب عمل در رویه Softmax مانند (۷) محاسبه می‌شود

$$\pi(a|s)=\Pr\{a_t=a|s_t=s\}=\frac{e^{\frac{Q(s,a)}{\tau}}}{\sum_{a_j\in A(s)} e^{\frac{Q(s,a_j)}{\tau}}} \qquad (۷)$$

که در آن τ ضریب دما است [۱].

در [۱۶] به منظور کنترل رویه انتخاب عمل عامل، ایده اکتشاف بر پایه تفاضل ارزش مطرح شده و VDBE[1] نام‌گذاری شده است. در این روش، پارامتر احتمال اکتشاف وابسته به حالت[2] $\varepsilon(s)$ معرفی شده است. در آغاز فرایند یادگیری انتظار می‌رود عامل بیشتر اکتشاف انجام دهد. عمل‌های اکتشافی در موقعیت‌هایی که مقادیر ارزش در فرایند یادگیری دارای نوسان است و دانش در مورد محیط به قطعیت نرسیده است انتخاب می‌شود. زمانی که عامل به شناختی از محیط برسد مطلوب است که مقدار اکتشاف کاهش یابد. چنین رفتار انطباقی با استفاده از محاسبه احتمال اکتشاف وابسته به حالت بعد از هر گام یادگیری از (۸) به دست می‌آید. این احتمال بر طبق تفاوت توزیع بولتزمن مقادیر قبل و بعد از یادگیری محاسبه می‌شود

$$f(s,a,\sigma)=\left|\frac{e^{\frac{Q_t(s,a)}{\sigma}}}{e^{\frac{Q_t(s,a)}{\sigma}}+e^{\frac{Q_{t+1}(s,a)}{\sigma}}}-\frac{e^{\frac{Q_{t+1}(s,a)}{\sigma}}}{e^{\frac{Q_t(s,a)}{\sigma}}+e^{\frac{Q_{t+1}(s,a)}{\sigma}}}\right|=\frac{1-e^{\frac{-|\alpha.\Delta|}{\sigma}}}{1+e^{\frac{-|\alpha.\Delta|}{\sigma}}} \qquad (۸)$$

$$\varepsilon_{t+1}(s)=\delta.f(s_t,a_t,\sigma)+(1-\delta).\varepsilon_t(s)$$

σ ثابت مثبت است. $\delta(s)=1/|A(s)|$ است که در آن $|A(s)|$ نشان‌دهنده تعداد عمل‌ها است. $\varepsilon(s)$ برای تمامی حالت‌ها در آغاز با ۱ مقداردهی شده است [۱۶].

اگرچه روش اکتشاف بر پایه تفاضل ارزش برای حل مسایلی مانند Bandit کاربرد دارد اما یکی از نقاط ضعف آن، انتخاب عمل‌های اکتشافی با توزیع یکنواخت است [۱۷]. برای رفع این مشکل در [۱۷] ایده ترکیب روش اکتشاف بر پایه تفاضل ارزش با رویه Softmax مطرح شده و VDBE-Softmax نام‌گذاری شده است. در این روش یک عدد تصادفی تولید می‌شود، اگر مقدار عدد تصادفی از احتمال اکتشاف وابسته به حالت کمتر باشد، عمل‌های تصادفی بر اساس رویه Softmax انتخاب می‌شود، در غیر این صورت عمل با بالاترین ارزش انتخاب می‌شود [۱۷].

۳- روش‌های پیشنهادی برای افزایش سرعت همگرایی در یادگیری Q

۳-۱ به روز رسانی مقادیر Q

در روش پیشنهادی مقادیر Q برای هر جفت حالت- عمل مطابق (۹) به روز رسانی می‌شود [۲۰]

$$Q(s,a)\leftarrow Q(s,a)+\alpha[r+\gamma\max_{a'}Q(s',a')+(1-\gamma)\min_{a''}Q(s',a'')-Q(s,a)] \qquad (۹)$$

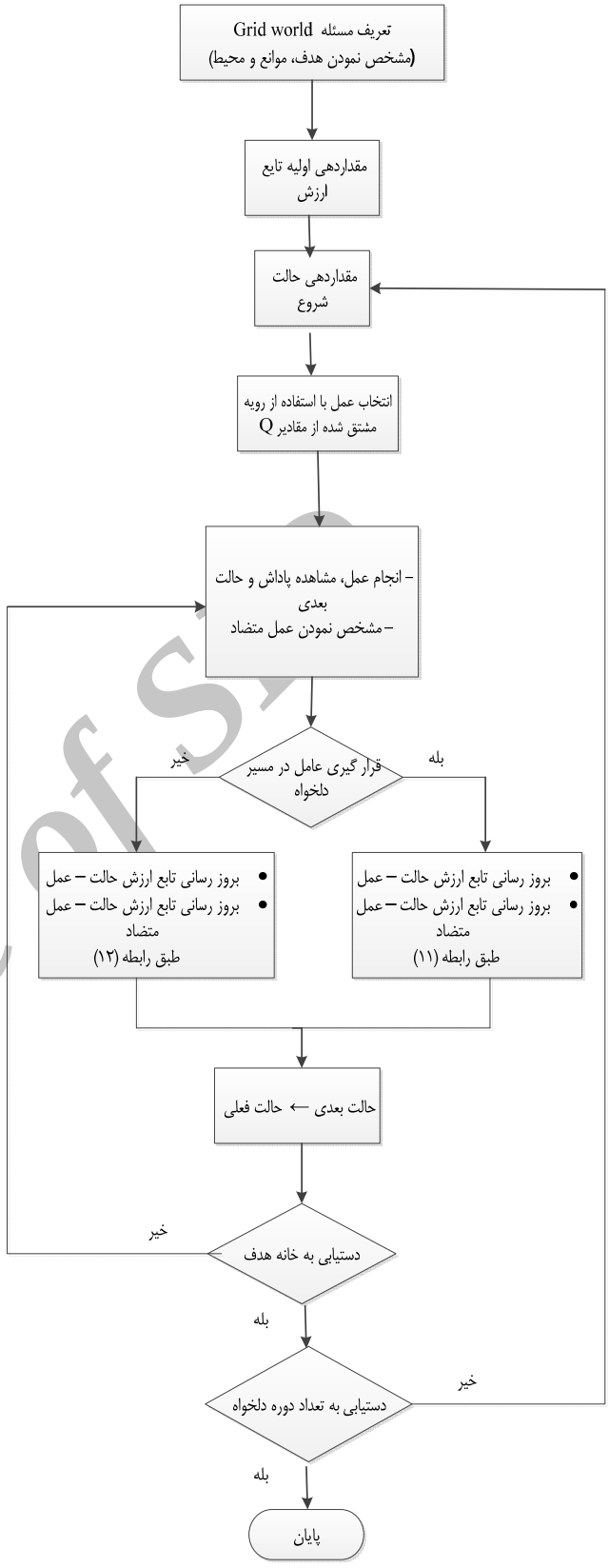
که α پارامتر نرخ یادگیری است و مقدار آن $0<\alpha<1$ است. این پارامتر سرعت تغییر مقدار Q در هر به روز رسانی را مشخص می‌کند. r تابع پاداش و γ فاکتور تخفیف است.

$\max_{a'}Q(s',a')$ ارزش بهینه‌ترین عمل و $\min_{a''}Q(s',a'')$ ارزش بدترین عمل در حالت جدید یعنی s' است. به عبارتی این مقادیر به ترتیب بیشینه و کمینه ارزش در حالت s' به شمار می‌آیند. با فرض این که عامل در جهت دلخواه باشد ارزش حالت- عمل متضاد، با دریافت پاداش متناظر با آن، برای عمل با کمترین ارزش وزن بیشتری نسبت به عملی که بیشترین ارزش در حالت بعدی را دارد در نظر می‌گیرد.

عامل یادگیرنده، متناسب با عملی که انجام می‌دهد پاداشی مثبت یا منفی دریافت می‌کند و تابع ارزش را به روز رسانی می‌نماید. اگر عامل در محیط واقعی در هر گام به یک عمل بسنده نماید به دلیل افزونگی تعداد حالت‌های محیط، سرعت همگرایی به شدت کاهش می‌یابد.

در روش پیشنهادی نیز عامل مقادیر Q را برای هر عمل و عمل متضاد متناظر با آن به صورت هم‌زمان به روز رسانی می‌کند. در این روش عمل متضاد دارای جهتی مخالف با جهت عمل اصلی است [۲۱]. به عنوان مثال هنگامی که عمل اصلی دارای جهت رو به بالا است، عمل متضاد متناظر با آن جهت رو به پایین را داراست.

1. Value Difference Based Exploration
2. State-Dependent Exploration Probability



شکل ۲: روش یادگیری Q مبتنی بر تضاد پیشنهادی.

تابع پاداش به صورت ماتریسی از حالت- عمل در نظر گرفته شده است. زمانی که عمل متضاد مشخص شد، عامل می‌تواند $r(s,\tilde{a})$ را با استفاده از تابع پاداش محاسبه کند. در این روش دیگر نیازی به دانستن موقعیت هدف نیست.

در روش پیشنهادی، ابتدا مقادیر تابع ارزش به ازای تمامی حالت‌ها و عمل‌ها برابر صفر در نظر گرفته شده است. در هر گام یادگیری، عامل عملی را بر اساس مقادیر Q انتخاب می‌نماید. در ابتدا عمل‌ها تصادفی هستند زیرا تخمینی برای مقادیر Q به دست نیامده است اما با تکرار الگوریتم و افزایش دوره‌ها، شانس انتخاب عمل‌های تصادفی به مرور زمان کاهش می‌یابد. عامل با انجام عمل به حالت جدید انتقال می‌یابد. عمل متضاد که دارای جهتی مخالف عمل اصلی است [۲۱]، تعیین می‌گردد و نحوه انتخاب عمل متضاد به صورت دانش از پیش تعریف شده در مسأله وجود دارد. سپس بهترین و بدترین عمل در حالت جدید مشخص می‌شوند و قرارگیری عامل در جهت دلخواه نیز بررسی می‌شود.

به عنوان مثال اگر عامل در وضعیتی با مقدار ارزش بالاتری باشد (رابطه (۱۰)) به طور هم‌زمان دو مقدار Q برای جفت‌های حالت- عمل و حالت- عمل متضاد همانند (۱۱) به روز می‌شود. برای به روز رسانی جفت‌های حالت- عمل و حالت- عمل متضاد، برای عمل با بیشترین ارزش نسبت به عملی که کمترین ارزش را در حالت جدید دارد، به ترتیب وزن بیشتری و کمتری در نظر گرفته می‌شود

$$Q(s,a) < Q(s',a^*) \quad a^* \leftarrow \arg\max_{i\in A(s)} Q(s',i) \tag{۱۰}$$

$$\begin{aligned} Q(s,a) &\leftarrow Q(s,a)+\alpha[r(s,a)+\gamma \max_{a'} Q(s',a')+ \\ &(1-\gamma)\min_{a''} Q(s',a'')-Q(s,a)] \\ Q(s,\tilde{a}) &\leftarrow Q(s,\tilde{a})+\alpha[r(s,\tilde{a})+\gamma \min_{a'} Q(s',a')+ \\ &(1-\gamma)\max_{a''} Q(s',a'')-Q(s,\tilde{a})] \end{aligned} \tag{۱۱}$$

اگر پاسخ منفی بود مقادیر Q مانند (۱۲) به روز می‌شوند

$$\begin{aligned} Q(s,a) &\leftarrow Q(s,a)+\alpha[r(s,a)+\gamma \min_{a'} Q(s',a')+ \\ &(1-\gamma)\max_{a''} Q(s',a'')-Q(s,a)] \\ Q(s,\tilde{a}) &\leftarrow Q(s,\tilde{a})+\alpha[r(s,\tilde{a})+\gamma \max_{a'} Q(s',a')+ \\ &(1-\gamma)\min_{a''} Q(s',a'')-Q(s,\tilde{a})] \end{aligned} \tag{۱۲}$$

اگر عامل به هدف خود دست یافته باشد یک دوره یادگیری انجام گرفته است. روندنمای روش پیشنهادی در شکل ۲ ارائه شده است.

۳-۲ رویه تطبیقی

در این قسمت، روش یادگیری Q مبتنی بر تضاد با رویه تطبیقی شرح داده می‌شود. از ایده اکتشاف بر پایه تفاضل ارزش ترکیب‌شده با رویه Softmax [۱۷] برای روش یادگیری بر مبنای تضاد بهره برده شده است. در این روش برای برقراری تعادل بین اکتشاف و اکتساب، پارامتر احتمال اکتشاف وابسته به حالت لحاظ شده است. در آغاز فرایند یادگیری مقدار این پارامتر برای تمامی حالت‌ها برابر یک قرار داده می‌شود. در هر گام یادگیری، یک عدد تصادفی بین صفر و یک تولید می‌شود، اگر مقدار عدد ایجادشده کمتر از احتمال اکتشاف وابسته به حالت باشد طبق رویه Softmax یک عمل انتخاب می‌شود. در حین فرایند یادگیری مقدار این پارامتر به تدریج کاهش می‌یابد و عمل با بالاترین مقدار ارزش (رویه حریصانه) انتخاب می‌شود. نتایج به کارگیری این روش راندمان بالاتری نسبت به روش‌های $\varepsilon-$greedy و Softmax دارد چون در این روش از مزیت‌های هر دو روش استفاده شده است.

در روش‌های مبتنی بر تضاد، در هر گام یادگیری، دو مقدار Q به روز می‌شوند و در نتیجه برای محاسبه اکتشاف وابسته به حالت $\varepsilon(s)$، میانگینی از $f(s,a,\sigma)$ و $f(s,\tilde{a},\sigma)$ در نظر گرفته شده است. تفاوت توزیع بولتزمن مقادیر قبل و بعد از یادگیری [۱۷] برای جفت‌های حالت- عمل و حالت-عمل متضاد محاسبه می‌شود. سپس بعد از هر گام یادگیری احتمال اکتشاف وابسته به حالت از (۱۳) محاسبه می‌شود [۱۸]

$$f(s,a,\sigma)=\frac{1-e^{\frac{-|\alpha.\Delta_1|}{\sigma}}}{1+e^{\frac{-|\alpha.\Delta_1|}{\sigma}}}$$
$$\Delta_1 = r(s,a)+\gamma \max_{a'} Q(s',a')+(1-\gamma)\min_{a''} Q(s',a'')-Q(s,a)$$
$$f(s,\tilde{a},\sigma)=\frac{1-e^{\frac{-|\alpha.\Delta_2|}{\sigma}}}{1-e^{\frac{-|\alpha.\Delta_2|}{\sigma}}} \qquad (۱۳)$$
$$\Delta_2 = r(s,\tilde{a})+\gamma \min_{a'} Q(s',a')+(1-\gamma)\max_{a''} Q(s',a'')-Q(s,\tilde{a})$$
$$\varepsilon_{t+1}(s)=\frac{1}{2}\delta(f(s,a,\sigma)+f(s,\tilde{a},\sigma))+(1-\delta)\varepsilon_t(s)$$

در روش پیشنهادی، عامل عمل انتخاب‌شده را انجام می‌دهد و سپس پاداش حاصله را دریافت می‌کند و حالت جدید ناشی از انجام عمل را نیز مشاهده می‌نماید. عمل متضاد تعیین می‌گردد. بهترین عمل a^* و بدترین عمل $\tilde{a}^*$ در حالت جدید مشخص می‌شوند. خطای تفاضل زمانی وابسته به عمل انجام‌شده همانند (۱۴) محاسبه می‌شود

$$\begin{aligned}\Delta_1 &= r(s,a)+\gamma Q(s',a^*)+(1-\gamma)Q(s',\tilde{a}^*)-Q(s,a)\\ \Delta_2 &= r(s,\tilde{a})+\gamma Q(s',\tilde{a}^*)+(1-\gamma)Q(s',a^*)-Q(s,\tilde{a})\\ \Delta_3 &= r(s,a)+\gamma Q(s',\tilde{a}^*)+(1-\gamma)Q(s',a^*)-Q(s,a)\\ \Delta_4 &= r(s,\tilde{a})+\gamma Q(s',a^*)+(1-\gamma)Q(s',\tilde{a}^*)-Q(s,\tilde{a})\end{aligned} \qquad (۱۴)$$

اگر عامل با انجام عمل a به وضعیتی با مقدار ارزش بالاتر انتقال یافته باشد، مقادیر ارزش برای زوج‌های حالت- عمل و حالت- عمل متضاد همانند (۱۵) به روز رسانی می‌شوند

$$\begin{aligned}Q(s,a) &= Q(s,a)+\alpha.\Delta_1\\ Q(s,\tilde{a}) &= Q(s,\tilde{a})+\alpha.\Delta_2\end{aligned} \qquad (۱۵)$$

احتمال اکتشاف وابسته به حالت نیز مطابق (۱۶) محاسبه می‌شود

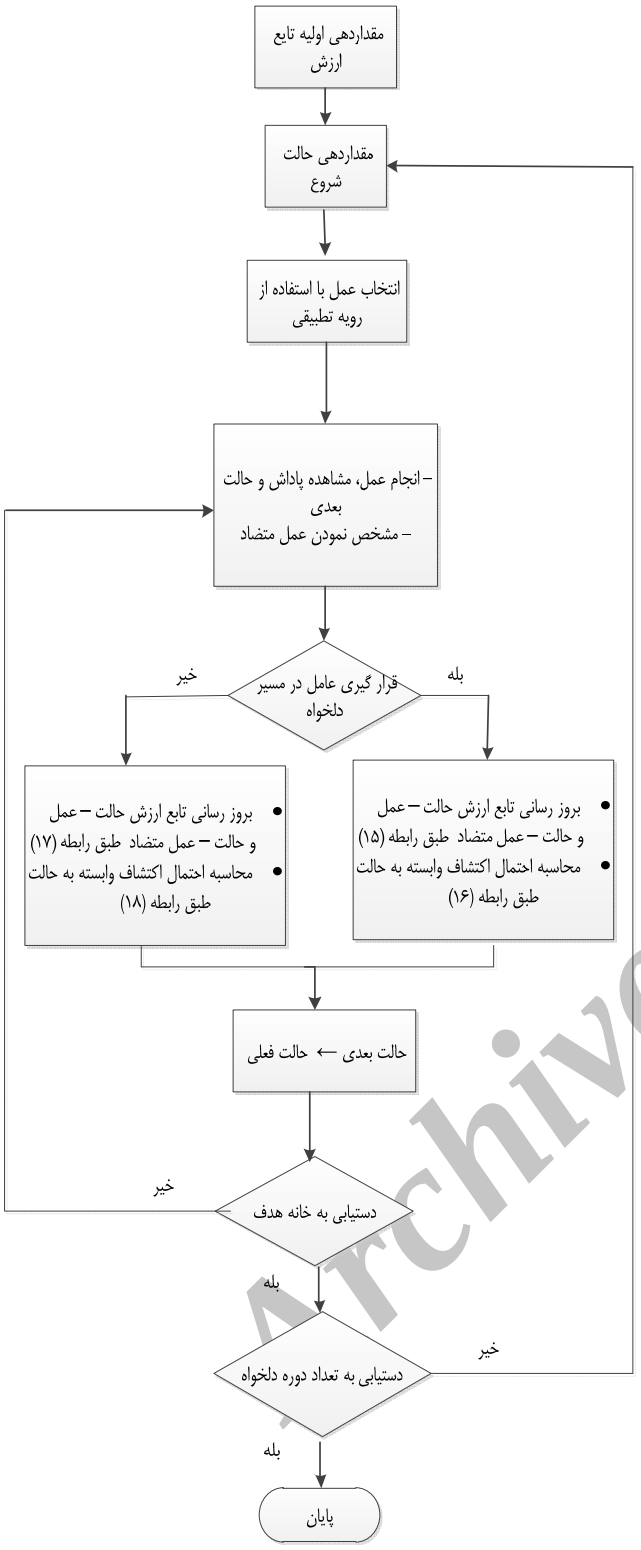
$$\varepsilon(s)=\frac{1}{2}\delta\left(\frac{1-e^{\frac{-|\alpha.\Delta_1|}{\sigma}}}{1+e^{\frac{-|\alpha.\Delta_1|}{\sigma}}}+\frac{1-e^{\frac{-|\alpha.\Delta_2|}{\sigma}}}{1+e^{\frac{-|\alpha.\Delta_2|}{\sigma}}}\right)+(1-\delta)\varepsilon(s) \qquad (۱۶)$$

در آغاز فرایند یادگیری، دانش عامل در مورد محیط به قطعیت نرسیده است و انتخاب عمل بیشتر به صورت اکتشافی انجام می‌پذیرد. بنابراین امکان دارد عامل با انجام عمل به وضعیتی با مقدار ارزش بالاتر منتقل نشود و به همین دلیل (۱۷) و (۱۸) لحاظ شده است

$$\begin{aligned}Q(s,a) &= Q(s,a)+\alpha.\Delta_3\\ Q(s,\tilde{a}) &= Q(s,\tilde{a})+\alpha.\Delta_4\end{aligned} \qquad (۱۷)$$

$$\varepsilon(s)=\frac{1}{2}\delta\left(\frac{1-e^{\frac{-|\alpha.\Delta_2|}{\sigma}}}{1+e^{\frac{-|\alpha.\Delta_2|}{\sigma}}}+\frac{1-e^{\frac{-|\alpha.\Delta_2|}{\sigma}}}{1+e^{\frac{-|\alpha.\Delta_2|}{\sigma}}}\right)+(1-\delta)\varepsilon(s) \qquad (۱۸)$$

روش پیشنهادی OQL-VDBE نام‌گذاری شده است. در شکل ۳، روندنمای آن رسم شده است.

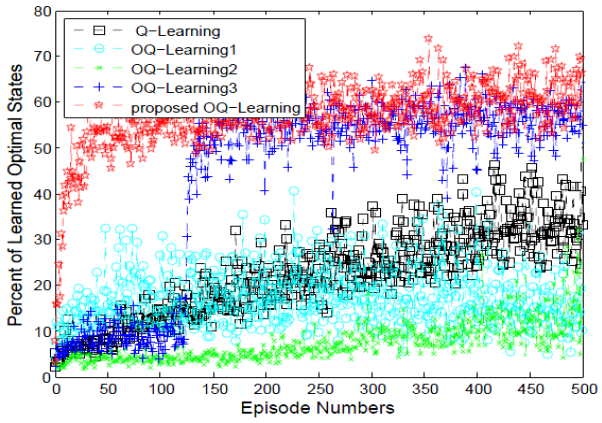


شکل ۳: روش OQL-VDBE پیشنهادی.

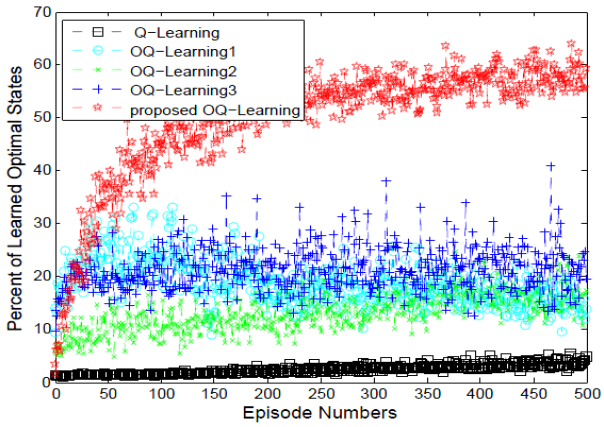
آزمایش ۱: ارزیابی روش‌های یادگیری مبتنی بر تضاد در محیط‌های بی‌مانع. هدف از انجام این آزمایش‌ها این است که آیا الگوریتم یادگیری Q مبتنی بر تضاد پیشنهادی، عملکرد بهتری نسبت به الگوریتم‌های یادگیری Q مبتنی بر تضاد پیشین [۸] و الگوریتم یادگیری Q دارد؟ به همین منظور شبیه‌سازی‌ها در محیط‌های بدون مانع [۸] با سایزهای متفاوت ۱۰×۱۰، ۵۰×۵۰ و ۱۰۰×۱۰۰ انجام گرفته است. درصد حالت‌های بهینه به عنوان معیار ارزیابی فرایند یادگیری در نظر گرفته شده است.

٤- آزمایش‌ها و ارزیابی

برای بررسی رفتار الگوریتم‌های پیشنهادی و ارزیابی آنها دو سری آزمایش انجام گرفته است که در ادامه شرح داده می‌شوند. الگوریتم‌ها برای مسئله Grid world با سایزهای متفاوت در نرم‌افزار Matlab شبیه‌سازی شده‌اند.



شکل ۴: میانگین درصد بهینه به تعداد دوره‌ها در محیط ۱۰×۱۰ با پارامترهای $\alpha=۰/۱$ و $\gamma=۰/۷$.



شکل ۵: میانگین درصد بهینه به تعداد دوره‌ها در محیط ۵۰×۵۰ با پارامترهای $\alpha=۰/۲$ و $\gamma=۰/۸$.

درصد حالت‌های بهینه با در نظر گرفتن نسبت بین تعداد گام‌های مسیر بهینه به تعداد گام‌های موجود در هر مسیر محاسبه می‌شود. در هر دوره برای یافتن مسیر بهینه مسئله از روش برنامه‌نویسی پویا استفاده می‌شود [۱۵]. تعداد گام‌های مسیر بهینه با پیاده‌سازی روش تکرار ارزش، پس از یافتن سیاست بهینه محاسبه شده است.

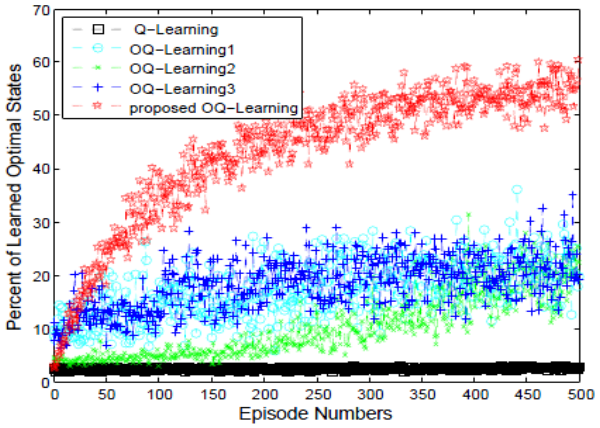
آزمایش‌ها به خاطر مستحکم‌نمودن نتایج حاصل‌شده ۲۰ مرتبه تکرار شده و در هر مرتبه ۵۰۰ دوره آزمایش شده است. در هر دوره مسیرهای یافته‌شده توسط الگوریتم‌های ذکرشده متفاوت است. هرچه میانگین درصد حالت‌های بهینه بیشتر باشد، الگوریتم کاراتر است.

عامل در هر دوره در یک محل تصادفی یادگیری را آغاز می‌کند. در هر قدم عامل می‌تواند در یکی از هشت جهت که شامل شمال، شمال شرق، شرق، جنوب شرق، جنوب، جنوب غرب، غرب و شمال غرب است، باشد. انتخاب عمل به وسیله رویه $\varepsilon-\text{greedy}$ انجام شده و عامل مسیر را برای رسیدن به خانه هدف پیمایش می‌کند. موقعیت خانه هدف ثابت و دارای مختصات $(X_{\max}, Y_{\max})$ است.

پارامترهای استفاده‌شده برای پیاده‌سازی‌ها در جدول ۱ آورده شده است. تابع پاداش تعریف‌شده بدین صورت است که هر حرکت پاداشی به اندازه ۱- دارد. زمانی که عامل به خانه هدف برسد پاداش ۱+ دریافت می‌کند.

نتایج حاصل از شبیه‌سازی در شکل‌های ۴ تا ۶ برای پنج روش ذکرشده در سه محیط نشان داده شده است.

شکل ۴ میانگین درصد حالت بهینه را در محیط ۱۰×۱۰ با پارامترهای $\alpha=۰/۱$ و $\gamma=۰/۷$ نشان می‌دهد. میانگین درصد حالت بهینه در روش



شکل ۶: میانگین درصد بهینه به تعداد دوره‌ها در محیط ۱۰۰×۱۰۰ با پارامترهای $\alpha=۰/۳$ و $\gamma=۰/۹$.

جدول ۱: مقداردهی پارامترها برای آزمایش ۱.

پارامتر	اندازه
نرخ یادگیری (α)	۰/۱، ۰/۲، ۰/۳
فاکتور تخفیف (γ)	۰/۷، ۰/۸، ۰/۹
تعداد دوره	۵۰۰
تعداد تکرار	۲۰
اپسیلون (ε)	۰/۱

پیشنهادشده دارای مقادیر بالاتری نسبت به دیگر روش‌ها می‌باشد. روش OQL۳ بعد از دوره ۱۲۵ بهبود چشمگیری داشته و دلیل این بهبود می‌تواند این باشد که تا دوره ۱۲۵ به روز رسانی اضافی با استفاده از مفهوم عمل متضاد صورت گرفته است. بنابراین استفاده از مفهوم تضاد در تعداد محدودی از دوره‌ها خصوصاً اوایل یادگیری می‌تواند مفید باشد.

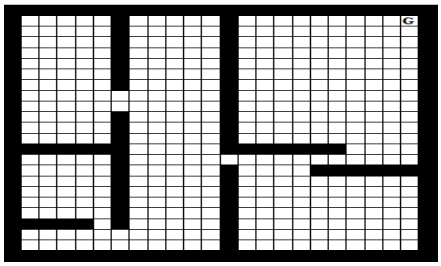
شکل ۵ میانگین درصد حالت بهینه را در محیط ۵۰×۵۰ با پارامترهای $\alpha=۰/۲$ و $\gamma=۰/۸$ نشان می‌دهد. میانگین درصد حالت بهینه در شکل ۶ برای محیط ۱۰۰×۱۰۰ با پارامترهای $\alpha=۰/۳$ و $\gamma=۰/۹$ آورده شده است. همان گونه که در شکل‌های ذکرشده مشاهده می‌گردد با افزایش بعد نیز روش پیشنهادی دارای کارایی بالاتر نسبت به سایر روش‌ها است.

میانگین درصد حالت بهینه در روش یادگیری Q با افزایش بعد به شدت کاهش یافته است. همان طور که قبلاً ذکر شد روش یادگیری Q با افزایش بعد از کارایی لازم برخوردار نیست.

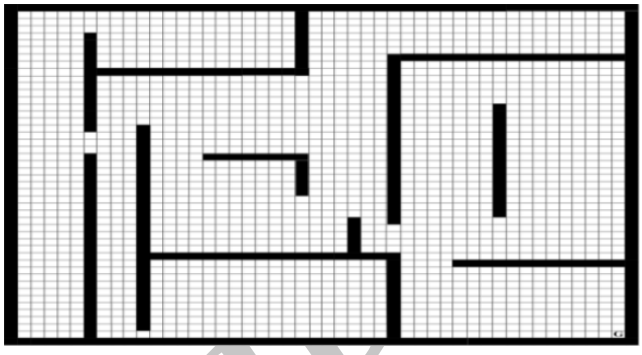
روش‌های یادگیری Q مبتنی بر تضاد پیشین به دلیل افزایش تعداد به روز رسانی و تابع پاداش تعریف‌شده باعث بهبود عملکرد یادگیری نسبت به روش یادگیری Q استاندارد می‌شوند.

آزمایش ۲: برای ارزیابی روش پیشنهادی (OQL_VDBE) و این که پژوهش‌های پیشین [۸] در محیط‌های با مانع موفق نبوده‌اند، از مسئله Grid world با سایزهای ۲۴×۲۴ و ۴۸×۴۸ [۱۸] استفاده شده است. شکل ۷ نمایی از دو محیط را نشان می‌دهد. هدف از این آزمایش‌ها این است که آیا الگوریتم OQL_VDBE که ترکیبی از روش یادگیری Q مبتنی بر تضاد و روش یادگیری Q با رویه تطبیقی است، عملکرد بهتری نسبت به روش‌های OQL، QL_VDBE و QL دارد؟

درصد نرخ موفقیت، میانگین درصد حالت‌های بهینه، میانگین تعداد گام‌های عامل برای رسیدن به هدف و میانگین پاداش دریافتی به عنوان معیارهای ارزیابی در نظر گرفته شده است. آزمایش‌ها ۵۰ مرتبه تکرار شده و در هر مرتبه ۴۰۰ دوره آزمایش شده است.



(الف)



(ب)

شکل ۷: نمایی از دو محیط شبیه‌سازی. موقعیت شروع، یکی از خانه‌های سفیدرنگ می‌باشد و موقعیت هدف با G مشخص شده است، (الف) محیط ۲۴×۲۴ و (ب) محیط ۴۸×۴۸ [۱۸].

جدول ۲: مقداردهی پارامترها برای آزمایش ۲.

پارامتر	اندازه
نرخ یادگیری (α)	۰/۱
فاکتور تخفیف (γ)	۰/۷
حداکثر تعداد گام در محیط (آ)	۳۰۰۰
حداکثر تعداد گام در محیط (ب)	۷۰۰۰
تعداد دوره	۴۰۰
تعداد تکرار	۵۰
دما (τ)	۰/۱
δ	۰/۱۲۵
σ	۱۰

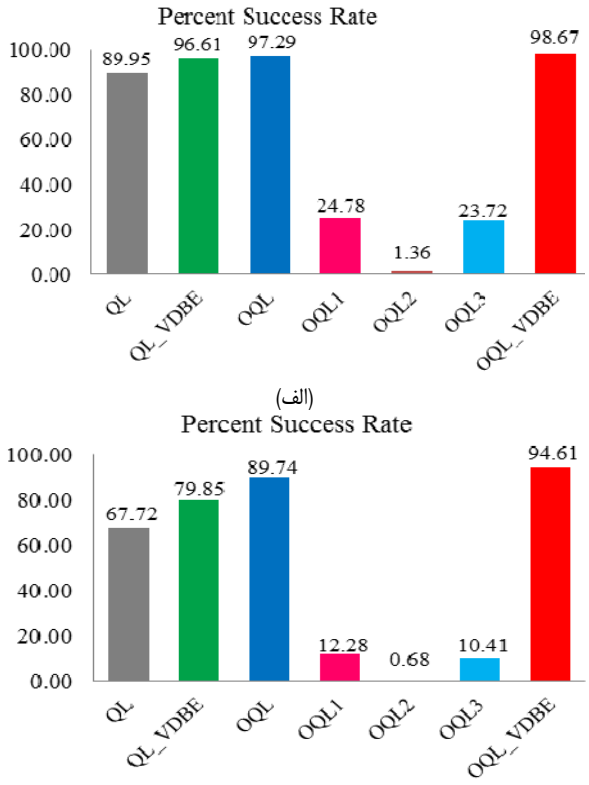
انتخاب عمل نیز به وسیله رویه Softmax انجام شده است. پارامترهای استفاده‌شده برای آزمایش ۲ در جدول ۲ آورده شده است.

هدف از یادگیری این است که عامل بتواند با پرداخت کمترین هزینه به خانه هدف برسد. تابع پاداش به صورت زیر تعریف شده است:

- فرض شده که هر حرکت پاداشی به اندازه ۱− دارد.
- حرکت‌هایی که باعث برخورد عامل به مانع یا دیوار می‌شود پاداش ۱۰− را در پی دارد.
- زمانی که عامل به خانه هدف برسد پاداش ۱+ دریافت می‌کند.

لازم به ذکر است که حرکت‌هایی که باعث برخورد عامل به مانع یا دیوار می‌شود محل عامل را تغییر نمی‌دهد. برای پیاده‌سازی پژوهش پیشین مبتنی بر تضاد از تابع پاداش ذکرشده در [۸] استفاده شده است.

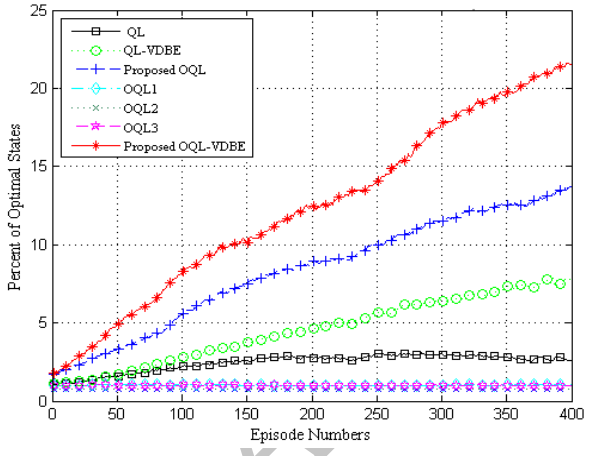
پایان هر دوره یادگیری، زمانی اتفاق می‌افتد که عامل به خانه هدف یا به حداکثر تعداد حرکات در نظر گرفته شده برای هر محیط برسد. نتایج حاصل از پیاده‌سازی‌ها در شکل‌های ۸ تا ۱۱ نشان داده شده است. در شکل‌های ذکرشده زیرنویس (الف) نتایج به دست آمده برای محیط (الف) را نشان می‌دهد و زیرنویس (ب) متناظر با نتایج به دست آمده در محیط (ب) است.



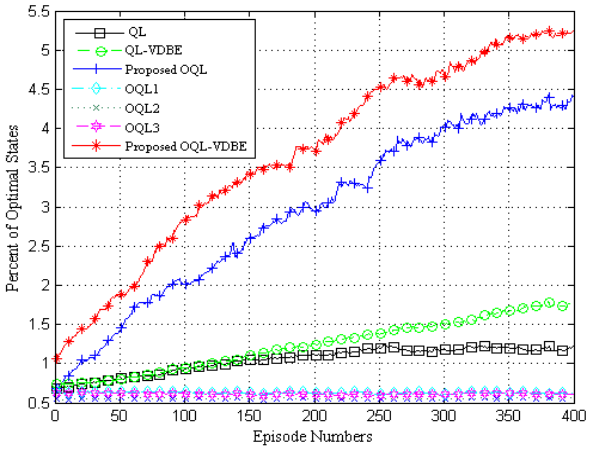
(الف)

(ب)

شکل ۸: درصد نرخ موفقیت با پارامترهای $\alpha=۰/۱$ و $\gamma=۰/۷$، (الف) محیط ۲۴×۲۴ و (ب) محیط ۴۸×۴۸.

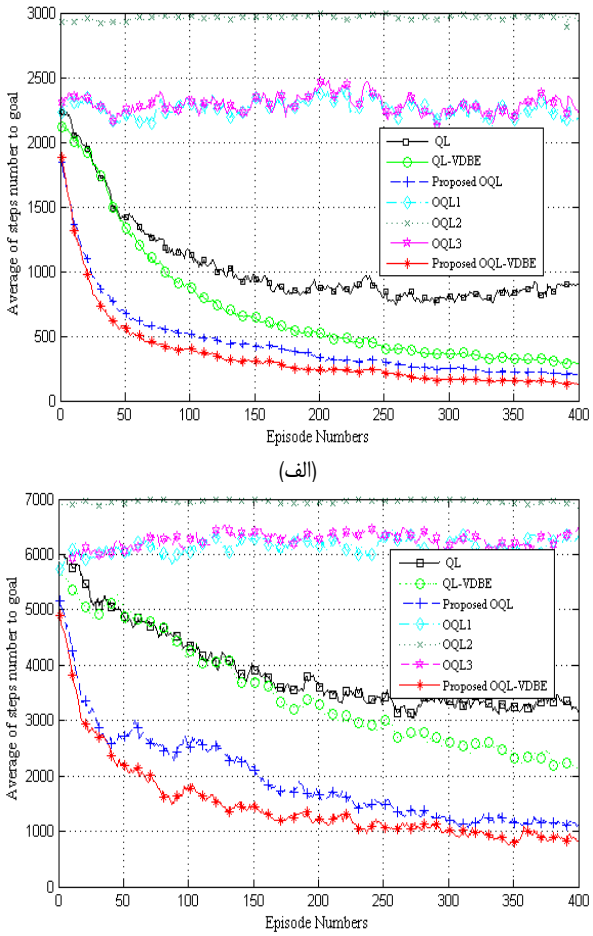


(الف)

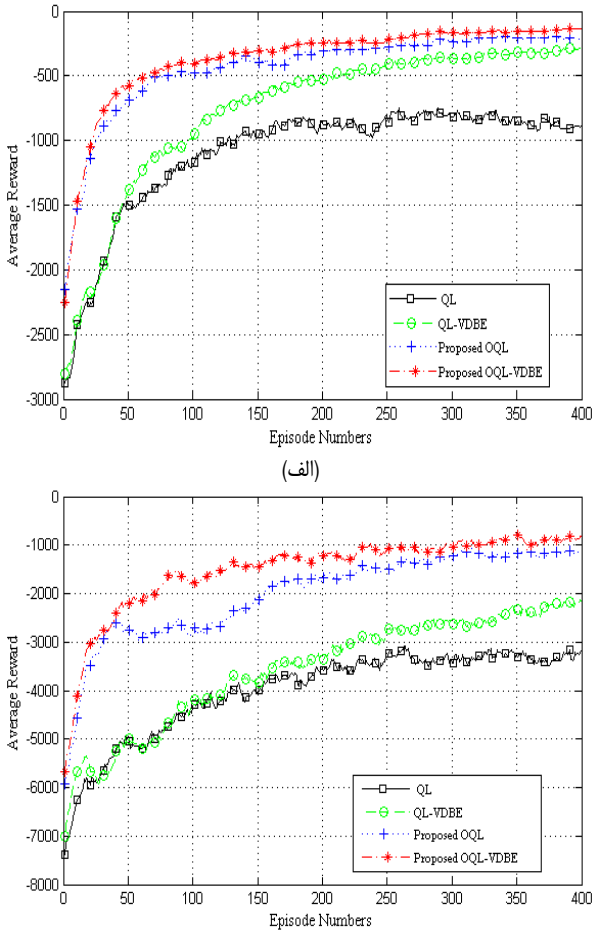


(ب)

شکل ۹: میانگین درصد بهینه به تعداد دوره‌ها با $\alpha=۰/۱$ و $\gamma=۰/۷$، (الف) محیط ۲۴×۲۴ و (ب) محیط ۴۸×۴۸.



شکل ۱۰: میانگین تعداد گام‌ها به تعداد دوره‌ها با $\alpha=۰/۱$ و $\gamma=۰/۷$، (الف) محیط ۲۴×۲۴ و (ب) محیط ۴۸×۴۸.



شکل ۱۱: میانگین پاداش به تعداد دوره‌ها با $\alpha=۰/۱$ و $\gamma=۰/۷$، (الف) محیط ۲۴×۲۴ و (ب) محیط ۴۸×۴۸.

شکل ۸ نمودار درصد نرخ موفقیت را نشان می‌دهد. برای محاسبه نرخ موفقیت، نسبت بین تعداد دفعاتی که عامل توانسته به خانه هدف برسد به تعداد دفعاتی که الگوریتم در هر محیط اجرا شده، در نظر گرفته شده است. کارهای مبتنی بر تضاد قبلی به دلیل مقادیر Q اشتباه، منجر به شکست در رسیدن به هدف می‌شوند که همین امر باعث می‌شود درصد نرخ موفقیت به دست آمده برای این روش‌ها بسیار پایین باشد. این نرخ برای روش پیشنهادی (OQL_VDBE) در مقایسه با سایر الگوریتم‌ها دارای مقادیر بیشتری است و نشان می‌دهد تعداد شکست این روش نسبت به سایر روش‌ها کمتر بوده است. روش یادگیری Q مبتنی بر تضاد پیشنهادشده، روش QL_VDBE و روش یادگیری Q استاندارد به ترتیب در رده‌های بعدی از نظر درصد نرخ موفقیت قرار گرفته‌اند.

شکل ۸-ب، نمودار درصد نرخ موفقیت را برای محیط (ب) نشان می‌دهد. برای تمامی روش‌ها نرخ موفقیت به دست آمده نسبت به شکل ۸- الف دارای درصد کمتری است زیرا تعداد حالت‌های محیط (ب) نسبت به محیط (الف) افزایش یافته است. همان طور که در شکل دیده می‌شود نسبت کاهش درصد نرخ موفقیت روش پیشنهادی نسبت به سایر روش‌ها کمتر بوده است که نشان‌دهنده کارایی بهتر است.

شکل ۹ میانگین درصد حالت‌های بهینه را با پارامترهای $\alpha=۰/۱$ و $\gamma=۰/۷$ برای هفت روش نشان می‌دهد. در شکل ۹ دیده می‌شود که این میانگین در روش پیشنهادی برای دو محیط دارای مقادیر بالاتری نسبت به دیگر روش‌ها است. روش OQL به دلیل استفاده از انتخاب عمل Softmax و عدم استفاده از رویه تطبیقی دارای مقادیر کمتری نسبت به روش OQL_VDBE است و چون به طور هم‌زمان دو مقدار از مقادیر Q را به روز می‌کند نسبت به روش QL_VDBE دارای مقدار بیشتری است. روش QL_VDBE نیز به دلیل بهبود رویه انتخاب عمل از روش QL بهتر عمل می‌کند.

روش‌های مبتنی بر تضاد (OQL۱,OQL۲,OQL۳) به دلیل انتساب پاداش‌های اشتباه، باعث ورود مقادیر اشتباه به جدول Q می‌شوند، بنابراین عامل در بسیاری از دوره‌ها نمی‌تواند به خانه هدف برسد و برای به پایان‌رساندن فرایند یادگیری، حداکثر تعداد گام‌های در نظر گرفته شده را طی می‌کند. انتخاب عمل نادرست موجب عدم همگرایی الگوریتم‌های یادگیری می‌شود و در نتیجه میانگین درصد حالت بهینه برای این روش‌ها بسیار پایین است.

شکل ۱۰، هفت الگوریتم را از لحاظ میانگین تعداد گام‌های عامل تا هدف مقایسه می‌کند. هرچه تعداد گام‌های لازم برای رسیدن به هدف کمتر باشد الگوریتم کاراتر خواهد بود.

با توجه به شکل ۱۰- الف مشاهده می‌شود که روش‌های مبتنی بر تضاد قبلی، بیشترین تعداد گام را دارند. در میان چهار الگوریتم دیگر، روش یادگیری Q تعداد گام‌های بیشتری برای رسیدن به هدف نیاز دارد. تعداد گام‌های لازم برای همگرایی در روش QL_VDBE نسبت به روش QL کاهش یافته که برتری استفاده از رویه تطبیقی نسبت به رویه Softmax در یادگیری Q را نشان می‌دهد و نتایج [۱۷] را تأیید می‌کند. روش یادگیری Q مبتنی بر تضاد پیشنهادشده (OQL) به دلیل کاهش زمان اکتشاف با افزایش تعداد به روز رسانی نسبت به روش QL_VDBE

جدول ۳: مقایسه روش‌ها از نظر میانگین درصد حالت‌های بهینه.

روش‌ها		p-value محیط ۲۴×۲۴	p-value محیط ۴۸×۴۸
OQL_VDBE روش پیشنهادی	QL	$۹٫۷۲\times۱۰^{-۱۲۴}$	$۸٫۳۲\times۱۰^{-۱۵۳}$
	OQL۱	$۱٫۲۴\times۱۰^{-۱۴۱}$	$۵٫۲۶\times۱۰^{-۱۷۳}$
	OQL۲	$۱٫۱۹\times۱۰^{-۱۴۴}$	$۱٫۲۱\times۱۰^{-۱۷۶}$
	OQL۳	$۸٫۰\times۱۰^{-۱۴۲}$	$۱٫۱۷\times۱۰^{-۱۷۳}$
	OQL	$۱٫۶\times۱۰^{-۳۰}$	$۱٫۳۴\times۱۰^{-۱۸}$
	QL_VDBE	$۱٫۴۲\times۱۰^{-۹۴}$	$۷٫۴۰\times۱۰^{-۱۴۵}$

جدول ۴: مقایسه روش‌ها از نظر میانگین تعداد گام‌های عامل تا هدف.

روش‌ها		p-value محیط ۲۴×۲۴	p-value محیط ۴۸×۴۸
OQL_VDBE روش پیشنهادی	QL	$۳٫۳۶\times۱۰^{-۱۳۶}$	$۳٫۷۱\times۱۰^{-۲۲۹}$
	OQL۱	۰	۰
	OQL۲	۰	۰
	OQL۳	۰	۰
	OQL	$۴٫۰۷\times۱۰^{-۶}$	$۵٫۲۶\times۱۰^{-۱۶}$
	QL_VDBE	$۷٫۴۵\times۱۰^{-۳۴}$	$۱٫۶۶\times۱۰^{-۱۴۲}$

جدول ۵: مقایسه روش‌ها از نظر میانگین پاداش.

روش‌ها		p-value محیط ۲۴×۲۴	p-value محیط ۴۸×۴۸
OQL_VDBE روش پیشنهادی	QL	$۴٫۲۲\times۱۰^{-۲۹}$	$۲٫۲۴\times۱۰^{-۱۹۹}$
	OQL	۰٫۰۰۳	$۳٫۴۴\times۱۰^{-۱۵}$
	QL_VDBE	$۱۰^{-۲۷}$	$۱٫۵۹\times۱۰^{-۱۲۵}$

با تعداد گام‌های کمتری به خانه هدف دست یافته است. تعداد گام‌های لازم برای همگرایی در روش OQL_VDBE کمترین مقدار را دارد، دلیل این امر به کار بردن مفهوم عمل متضاد در یادگیری و استفاده از رویه تطبیقی برای انتخاب عمل در هر حالت است.

در شکل ۱۰- ب مشاهده می‌شود که موارد ذکرشده برای محیط (ب) هم صادق است با این تفاوت که چون تعداد حالت‌ها در محیط (ب) افزایش یافته است میانگین تعداد گام‌های عامل تا هدف نیز برای روش‌ها افزایش یافته است.

شکل ۱۱ چهار الگوریتم را از لحاظ میانگین پاداش تجمعی به دست آمده در هر دوره مقایسه می‌کند. رفتار عامل یادگیری باید به گونه‌ای باشد که مجموع پاداش دریافتی ماکسیمم شود. بنابراین هرچه میانگین پاداش تجمعی به دست آمده برای الگوریتم بیشتر باشد، نشان‌دهنده کارایی بهتر الگوریتم است.

در شکل ۱۱ مشاهده می‌شود که الگوریتم یادگیری Q کمترین پاداش را به دست آورده است که دلیل آن استفاده از انتخاب عمل Softmax و به روز کردن تنها یک مقدار Q در هر گام یادگیری است. الگوریتم OQL_VDBE پیشنهادی به دلیل استفاده از رویه تطبیقی و به روز کردن هم‌زمان دو مقدار Q در هر گام یادگیری، موجب بهبود فرایند یادگیری و افزایش پاداش تجمعی شده است.

با توجه به نتایج به دست آمده از شبیه‌سازی‌ها مشاهده می‌شود که روش پیشنهادی باعث افزایش سرعت همگرایی در یادگیری Q شده است. برای اثبات معنی‌دار بودن آماری، برتری روش پیشنهاد شده در این مقاله با روش‌های پیشین از آزمون t-test [۲۲] و [۲۳]، دوطرفه با $\alpha = ۰٫۰۵$ استفاده شد. نتایج حاصل از این آزمون در جدول‌های ۳ تا ۵ گزارش شده است.

بر اساس مقادیر p-value جدول ۳ که در آن روش پیشنهادی (OQL_VDBE) با سایر روش‌ها از لحاظ میانگین درصد حالت‌های بهینه مقایسه شده است، مشاهده می‌شود که نتایج به دست آمده از ۰٫۰۵ کمتر است. روش پیشنهادی با سایر روش‌ها از لحاظ میانگین تعداد گام‌های عامل تا هدف و میانگین پاداش نیز مقایسه شد و مقادیر p-value آنها به ترتیب در جدول‌های ۴ و ۵ آورده شده است. مقادیر p-value کمتر از ۰٫۰۵ نشان‌دهنده این است که بهبود به دست آمده توسط روش پیشنهادی از نظر آماری معنی‌دار است.

۵- نتیجه‌گیری

در این مقاله، روش‌هایی برای افزایش سرعت همگرایی یادگیری Q مطرح شده و روش بهبودیافته یادگیری Q مبتنی بر تضاد ارائه شده است. روش دیگر ارائه‌شده برای برقراری تعادل بین اکتساب و اکتشاف در انتخاب عمل می‌باشد. اکتشاف بر پایه تفاضل ارزش ترکیب شده با رویه Softmax به عنوان یک رویه تطبیقی در روش یادگیری Q مبتنی بر تضاد به کار برده شده است. استفاده از این روش‌ها باعث بهبود فرایند یادگیری و افزایش سرعت همگرایی شده است. روش‌های پیشنهادی برای مسئله Grid world شبیه‌سازی شده است. نتایج نشان می‌دهد استفاده از این روش‌ها توانمندی عامل را در کشف محیط و رسیدن سریع‌تر به هدف مشخص‌شده افزایش می‌دهد. با توجه به این که در این مقاله از عمل متضاد استفاده شده است به کار بردن تضاد برای دیگر اجزای یادگیری تقویتی می‌تواند به عنوان کار آتی مطرح شود. به عنوان مثال، تضاد را می‌توان برای حالت و عمل به کار گرفت و چهار مقدار Q را به طور هم‌زمان به روز رسانی کرد. همچنین مفاهیم ارائه‌شده می‌توانند در دیگر الگوریتم‌های یادگیری تقویتی مانند $Q(\lambda)$، Sarsa و Sarsa(λ) بیان شوند.

مراجع

[1] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, MIT Press, Cambridge, MA, 1998.

[2] J. Qiao, R. Fan, H. Han, and X. Ruan, "Q-learning based on dynamical structure neural network for robot navigation in unknown environment," in *Proc. of the 6th Int. Symp. on Neural Networks: Advances in Neural Networks - Part III*, *ISNN'09,* pp. 188-196, 2009.

[3] W. Y. Kwon, I. H. Suh, and S. Lee, "SSPQL: stochastic shortest path-based Q-learning," *International J. of Control, Automation, and Systems*, vol. 9, no. 2, pp. 328-338, 2011.

[4] P. K. Das, S. C. Mandhata, H. S. Behera, and S. N. Patro, "An improved Q-learning algorithm for path-planning of a mobile robot," *International J. of Computer Applications*, vol. 51, no. 9, pp. 40-46, 2012.

[5] M. B. Naghibi-Sistani, M. R. Akbarzadeh-Tootoonchi, M. H. Javidi-Dashte Bayaz, and H. Rajabi-Mashhadi, "Application of Q-learning with temperature variation for bidding strategies in market based power systems," *Energy Conversion and Management*, vol. 47, no. 11, pp. 1529-1538, 2006.

[6] Y. Ozbek, A. Zeid, and S. Kamarthi, "A Q-learning-based adaptive grouping policy for condition-based maintenance of a flow line manufacturing system," *International J. of Collaborative Enterprise*, vol. 2, no. 4, pp. 302-321, 2011.

[7] R. A. Bianchi, A. Ramisa, and R. L. De Mantaras, "Automatic selection of object recognition methods using reinforcement learning," in *Advances in Machine Learning I,* Springer Berlin Heidelberg, pp. 421-439, 2010.

[8] H. R. Tizhoosh, "Opposition-based reinforcement learning," *J. of Advanced Computational Intelligence and Intelligent Informatics*, vol. 10, no. 4, pp. 578-585, 2006.

[9] X. Ma, Y. Xu, G. Q. Sun, L. X. Deng, and Y. B. Li, "State-chain sequential feedback reinforcement learning for path planning of autonomous mobile robots," *J. of Zhejiang University Science C*, vol. 14, no. 3, pp. 167-178, Mar. 2013.

[10] A. Lampton and J. Valasek, "Multiresolution state-space discretization method for Q-learning," in *Proc. American Control Conf.*, pp. 1646-1651, 2009.

[11] D. Vincze and S. Kovacs, "Incremental rule base creation with fuzzy rule interpolation-based Q-learning," in *Proc. Computational Intelligence in Engineering,* pp. 191-203, 2010.

[12] K. Terashima and J. Murata, "A study on use of prior information for acceleration of reinforcement learning," in *Proc. SICE Annual Conf.*, pp. 537-543, 2011.

[13] B. Marthi, "Automatic shaping and decomposition of reward functions," in *Proc. of the 24th Int. Conf. on Machine Learning*, pp. 601-608, 2007.

[14] S. Manju and M. Punithavalli, "An analysis of Q-learning algorithms with strategies of reward function," *IJCSE*, vol. 3, no. 2, pp. 814-820, Feb. 2011.

[15] M. Guo, Y. Liu, and J. Malec, "A new Q-learning algorithm based on the metropolis criterion," *IEEE Trans. Syst. Man Cybern. B*, vol. 34, no. 5, pp. 2140-2143, Oct. 2004.

[16] M. Tokic, "Adaptive ε-greedy exploration in reinforcement learning based on value differences," in *Proc. of the 33rd annual German Conf. on Advances in Artificial Intelligence*, *KI'10,* pp. 203-210, 2010.

[17] M. Tokic and G. Palm, "Value-difference based exploration: adaptive exploration between epsilon-greedy and softmax," in *Proc. of the 34rd annual German Conf. on Advances in Artificial Intelligence*, *KI'11*, pp. 335-346, 2011.

[۱۸] م. پویان، ا. موسوی، ش. گلزاری و ا. حاتم، "روشی نوین برای بهبود عملکرد یادگیری Q با افزایش تعداد به روز رسانی مقادیر Q بر پایه عمل متضاد،" *مجموعه مقالات بیستمین کنفرانس سالانه کامپیوتر ایران، دانشگاه فردوسی مشهد*، صص. ۲۳۳-۲۲۶، ۱۴-۱۲ اسفند ۹۳.

[19] C. J. C. H. Watkins, *Learning from Delayed Rewards*, Ph. D Thesis, Cambridge University, Cambridge, England, 1989.

[20] M. Pouyan, A. Mousavi, S. Golzari, and A. Hatam, "Improving the performance of Q-learning using simultanous Q-values updating," in *Proc. 2014 Int. Congress on Technology, Communication and Knowledge, ICTCK'14* , 6 pp., 26-27 Nov. 2014.

[21] M. Shokri, "Knowledge of opposite actions for reinforcement learning," *Applied Soft Computing*, vol. 11, no. 6, pp. 4097-4109, 2011.

[22] U. Nehmzow, Scientific Methods in Mobile Robotics: Quantitative Analysis of Agent Behavior, London: Springer-Verlag London Limited, 2006.

[23] L. A. Celiberto, J. P. Matsuura, D. Mantaras, R. Lopez, and R. A. Bianchi, "Using transfer learning to speed-up reinforcement learning: a cased-based approach," in *Proc. 2010 Latin American Robotics Symp. and Intelligent Robotic Meeting, LARS'10,* pp. 55-60, Sao Bernardo do Campo, Brazil, 23-28 Oct. 2010.

مریم پویان تحصیلات خود را در مقطع کارشناسی ارشد مهندسی کامپیوتر گرایش هوش مصنوعی در سال ۱۳۹۳ از دانشگاه هرمزگان به پایان رسانده است. زمینه‌های تحقیقاتی مورد علاقه ایشان عبارتند از: یادگیری تقویتی، الگوریتم‌های تکاملی و داده‌کاوی.

شهرام گلزاری تحصیلات خود را در مقاطع کارشناسی و کارشناسی ارشد مهندسی کامپیوتر- نرم‌افزار به ترتیب در سال‌های ۱۳۷۷ و ۱۳۸۰ از دانشگاه‌های صنعتی اصفهان و صنعتی امیرکبیر و در مقطع دکتری علوم کامپیوتر- هوش مصنوعی در سال ۱۳۹۰ از دانشگاه پوترای مالزی به پایان رسانده است و هم اکنون استادیار گروه مهندسی برق و کامپیوتر دانشکده فنی و مهندسی دانشگاه هرمزگان می‌باشد. زمینه‌های تحقیقاتی مورد علاقه ایشان عبارتند از: محاسبات زیست‌ملهم، داده‌کاوی، یادگیری ماشین کاربردی و یادگیری عمیق.

سیدامین موسوی تحصیلات خود را در مقطع کارشناسی مهندسی برق- کنترل در سال ۱۳۷۶ از دانشگاه صنعتی شریف و کارشناسی ارشد و دکتری را در رشته مهندسی برق- کنترل به‌ترتیب در سال‌های ۱۳۷۹ و ۱۳۹۵ از دانشگاه تهران به پایان رسانده است و هم‌اکنون عضو هیأت علمی گروه مهندسی برق و کامپیوتر دانشگاه هرمزگان می‌باشد. زمینه‌های تحقیقاتی مورد علاقه ایشان عبارتند از: یادگیری ماشین، هوش مصنوعی، یادگیری تقویتی و رباتیک.

احمد حاتم تحصیلات خود را در مقطع کارشناسی مهندسی برق- الکترونیک در سال ۱۳۶۸ از دانشگاه صنعتی اصفهان و کارشناسی ارشد مهندسی برق- الکترونیک (دیجیتال) در سال ۱۳۷۲ از دانشگاه شریف تهران و در مقطع دکتری مهندسی برق- مخابرات در سال ۱۳۸۹ از دانشگاه صنعتی اصفهان به پایان رسانده است و هم‌اکنون عضو هیأت علمی گروه برق و کامپیوتر دانشگاه هرمزگان می‌باشد. در ضمن نام‌برده در خلال تحصیل در مقطع دکتری در سال ۲۰۰۵ م. در دانشگاه کارلتون کانادا به عنوان فرصت مطالعاتی بوده است. زمینه‌های تحقیقاتی مورد علاقه ایشان عبارتند از: پردازش تصویر، تشخیص الگو، کدگذاری کانال و مخابرات بی‌سیم.