بهبود عملکرد یادگیری تقویتی به کمک دانش تصمیم گیری در کنترل رفتار محور

حمیده سروری کاریزکی[۱]، ناصر مزینی[۲]، سید محمد حسین میرهاشمی[۳]

[۱]دانشگاه علم و صنعت ایران، دانشکده کامپیوتر ، hamide_sarvari@comp.iust.ac.ir

[۲] دانشگاه علم و صنعت ایران، دانشکده کامپیوتر ، mozayani@iust.ac.ir

[۳]دانشگاه علم و صنعت ایران، دانشکده کامپیوتر ، mirhashemi@comp.iust.ac.ir

چکیده

کنترل ربات‌های متحرک به دلیل بزرگی فضای حالت معمولا وظیفه‌ای دشوار و بسیار پیچیده قلمداد می‌شود. کنترل رفتار محور با تکیه بر مفهوم رفتار این پیچیدگی را کاهش داده و کل مسئله کنترل را در قالب چند مفهوم مجرد بنام رفتار توصیف می‌نماید. البته مسئله انتخاب و چگونگی هماهنگی رفتارها و بدست آوردن یک رفتار برآیند از ترکیب یا انتخاب رفتارها خود چالشی جدید است. اکثر روش‌های موجود برای این مسئله، معمولاً در فضای حالت و عمل بزرگ، ناکارآمد می‌شوند و یا از دانش به طور کامل استفاده نمی‌کنند. درنتیجه کاهش کارایی ربات را در پی دارند. از طرفی نظریه‌های تصمیم‌گیری چند معیاره راه حل‌های مناسبی برای مسئله هماهنگ‌سازی رفتار و مدیریت دانش موجود ارائه می‌کنند. ما در این مقاله با استفاده از روش الکتره۳، سازوکار یادگیری تقویتی را که در کنترل رفتار محور ربات بکار می‌رود، بهبود داده و نتایج آزمایشات نشان می‌دهند، دانش موجود در یک سامانه رفتار محور بخوبی در اصلاح فرایند یادگیری موثر بوده است.

واژه‌های کلیدی

کنترل رفتار محور، یادگیری تقویتی، دانش تصمیم گیری چند معیاره.

۱- مقدمه

کنترل رفتار محور[۱] روشی برای کنترل ربات‌های متحرک است که با شکستن وظیفه اصلی به زیروظایف کنترل را بین اجزای ساده‌ای به نام رفتار تقسیم می‌کند[۱]. هرچند این روش کنترلی، موفقیت‌های زیادی در کنترل ربات‌های متحرک داشته است اما طراحی آن وظیفه‌ای دشوار را بر دوش طراح می‌گذارد.

به نظر می‌رسد با تجمیع یادگیری در چهارچوب یک سامانه رفتارمحور می توان گام موثری در جهت طراحی خودکار برداشت. از سوی دیگر وجود یک سازوکار یادگیری در ربات، عملکرد ربات را در محیط‌های غیرقطعی بهبود می‌دهد. یادگیری تقویتی[۲] گزینه‌ای مناسب برای این منظور است. زیرا مسئله‌ای که یک عامل خود مختار با آن مواجه است یک مسئله یادگیری تقویتی است. به همین خاطر محققان زیادی یادگیری تقویتی را در ربات‌هایشان به کار برده‌اند. اما یادگیری تقویتی از مشکل نفرین ابعاد رنج می‌برد به این معنی که وقتی فضای حالت بسیار بزرگ می‌شود، یادگیری تقویتی عملاً ناکارآمد می‌شود. از انجا که فضای حالت برای یک ربات در محیط‌های واقعی بسیار بزرگ است، بیشتر روش‌های مرسوم برای تجمیع یادگیری تقویتی و رباتیک با شکست مواجه شده اند.

کنترل رفتارمحور با استفاده از مفهوم رفتار، گامی موثر در جهت شکستن فضای حالت بر می‌دارد اما مسئله اصلی در چگونگی هماهنگی رفتارها و بدست آوردن یک رفتار برآیند از ترکیب یا انتخاب رفتارهاست. روش‌های زیادی برای هماهنگی رفتارها در مقالات ارائه شده است. اما اکثر این روش‌ها یا در فضای حالت و عمل بزرگ، ناکارآمد می‌شوند و یا از دانش موجود به طور کامل استفاده نمی‌کنند و لذا کاهش کارایی ربات را در پی دارند.

در این مقاله یک معماری برای عاملی رفتار محور ارائه می‌شود که در آن، رفتارها یادگیرهای Q[۳] هستند که زیرفضای کوچکی از فضای حالت کلی را یاد می‌گیرند. دراین معماری مسئله هماهنگی رفتارها با استفاده از اطلاعات جدول Q از رفتارهای پایین دستی و مفاهیم موجود در علم تصمیم گیری چند معیاره[۴] حل می‌شود.

[1] Behaviour-based control

[2] Reinforcement Learning

[3] Q-learning

[4] Multiple Criteria Decision Making

نظریه‌های تصمیم‌گیری چند معیاره راه حل مناسبی را برای مسئله هماهنگ‌سازی رفتار ارائه می‌کنند. یکی از روش‌های محبوب در میان این روش‌ها، الکتره ۳ است.

در مقابل روش‌های سنتی که تنها دو موضوع برتری و بی تفاوتی را برای دو گزینه در نظر می‌گرفتند، مفهوم ارزش آستانه بی تفاوتی q، ارزش آستانه برتری p، و ارزش آستانه عدم برتری v، را تعریف می‌کند. بنابراین این روش در مقابل داده‌های اشتباه و نامعین مقاوم است. [۲]

مطابق با نظر پیرچانیان در [۳]، روش‌های موجود در تصمیم گیری چند معیاره، بیان می‌کنند که وقتی معیارهای متفاوت و گاهاً متضاد وجود دارند، پیدا کردن جواب بهینه به گونه ای که بتواند همزمان تمام معیارها را بهینه نماید، ناممکن است. در چنین شرایطی هدف پیدا کردن جوابی است که تا حد ممکن رضایت بخش باشد.

در این مقاله، الکتره ۳ به گونه ای مناسب برای مسئله هماهنگ‌سازی رفتار به کار گرفته شده است. به طوری که از نتایج بدست آمده از یادگیری به عنوان ورودی‌های این روش استفاده شده است. قابلیت یادگیری در کنار استفاده مناسب از روش‌های نظریه تصمیم‌گیری چند معیاره، کارایی عامل را در محیط‌های پیچیده و غیرقطعی افزایش می‌دهد. روش پیشنهادی برای یک مسئله که دارای فضای حالتی بزرگ و پیچیده است، به کار رفته و نتایج شبیه‌سازی موفقیت این روش را در میان سایر روش‌های موجود نشان می‌دهد.

۲- کارهای انجام شده

یکی از نخستین کارهایی که از یادگیری در طراحی سامانه‌های رفتارمحور استفاده کرده است، [۴] است. در این کار، هرچند مستقیماً از یادگیری تقویتی نام برده نمی شود، از سازوکاری مشابه با یادگیری تقویتی برای هماهنگی رفتارها استفاده می‌شود. در این کار فرض می‌شود که رفتارها ثابت و از پیش داده شده‌اند. در [۵]، رفتارها یادگیرنده‌های Q ساده هستند و معماری از نوع مرتبه ای است. در این کار برای غلبه بر مشکل بزرگی حالت از خوشه‌بندی استفاده شده است. در [۶]، طراحی رفتار و ساختار همزمان و با استفاده از یادگیری Q انجام می‌شود. در این روش یک رفتار بالاسری که خود یک یادگیر Q می‌باشد، کار هماهنگی رفتارها را انجام می‌دهد. یادگیر بالاسری کل فضای حالت را مشاهده می‌نماید و به جای انتخاب عمل، انتخاب رفتار را انجام می‌دهد. مشکل این روش این است که رفتار بالاسری هنوز کل فضای حالت را می‌بینند. این روش یادگیری تقویتی Q سلسله مراتبی نامیده می‌شود.

مشکل اصلی تمام این روش‌ها این است که مسئله بزرگی فضای حالت حل نشده باقی مانده است و یا برای اینکه با این مشکل مواجه نشوند از یک رویکرد اولویت بندی مانند معماری مرتبه‌ای برای هماهنگ‌سازی رفتار استفاده شده است که مبتنی بر طراح می‌باشد.

در [۶]، روشی دیگر نیز تحت عنوان یادگیری W ارائه شده است. در این روش که رفتارها یادگیرهای تقویتی Q هستند، هر رفتار عملی را با بیشترین مقدار Q پیشنهاد می‌کند. در نهایت رفتاری برای تولید عمل نهایی انتخاب می شود که اگر برنده نشود، بیش از باقی رفتارها متحمل ضرر می‌شود. در این روش اگرچه با بزرگی فضای حالت به گونه‌ای مقابله شده است اما هنوز از اطلاعات موجود در رفتارهای پایینی به طور کامل استفاده نمی‌شود و لذا روش کارایی چندان زیادی نخواهد داشت.

[۷]، استدلال می‌کند که استفاده از یادگیری Q در رفتارهای پایین اشتباه است و لذا رفتارهای پایین دستی با یادگیر سارسا پیاده سازی می شوند. چنین کاری اگرچه در مواردی باعث بهبود می‌شود اما اثبات همگرایی برای آن وجود ندارد.

فرهمند در [۸]، با استفاده از یادگیری تقویتی، طراحی معماری سامانه رفتارمحور از نوع مرتبه‌ای[5] را خودکار می‌نماید و اولویت بهینه هر رفتار را بدست می‌دهد. در این کار رفتارها با استفاده از یادگیری تکاملی بدست می‌آیند. مشکل بزرگی فضای حالت در این روش نیز وجود دارد.

در [۱] برای غلبه بر مشکل بزرگی فضای حالت که در یادگیری تقویتی Q ساسله مراتبی وجود داشت، از مفهوم انتزاع استفاده می‌کند، بدین ترتیب که رفتار بالاسری نیاز ندارد، تمام فضای حالت را مشاهده کند و میتواند دنیا را از طریق رفتارهای پایین دستی مشاهده کند. لذا ورودی فضای حالت برای یادگیرنده بالاسری را اطلاعات موجود در جدول Q یادگیرنده‌های پایین دستی می‌داند. مشکل این کار آن است که با زیاد شدن وظایف، فضای حالت برای رفتار بالاسری بزرگ می‌شود.

در [۹]، روشی به نام بزرگ ترین جرم ارائه شده است که در هر حالت عملی انتخاب می شود که مجموع ارزش Q برای آن عمل بیشینه باشد. این روش تا حدی شبیه روش SAW[6] در تصمیم گیری چند معیاره است. اما برای رفتارها وزن یکسان و مساوی با یک در نظر گرفته شده است و لذا اهمیت و وزن هر رفتار نادیده گرفته شده است.

[۱۰]، نیز برای کاهش فضای حالتی که یک ربات با آن مواجه است و برای اینکه بتواند از یادگیری تقویتی در رباتیک استفاده نماید، به جای استفاده از ورودی‌های حسگری ربات از نقشه مکانی محیط استفاده می‌کند که ربات رفتارمحور با مشاهده محیط ساخته است. این نقشه مکانی با استفاده از روشی ساخته می‌شود که ماتاریک در[۱۱] ارائه کرده است.

در تمام روش‌های موجود یا مشکل بزرگی فضای حالت و عمل حل نشده باقی می ماند و یا کارایی عامل به علت عدم استفاده از اطلاعات موجود، کاهش می‌یابد.

۳- فرمول بندی مسئله

فرآیند تصمیم مارکف فرموله‌بندی اساسی برای مسائل یادگیری تقویتی می‌باشد. در این فرموله‌بندی، محیط با مجموعه ای از حالات و اعمال مدل می‌شود و هدف، کنترل سیستم از طریق بیشینه‌سازی پاداش می‌باشد. فرآیند تصمیم مارکف به صورت یک چندتایی M =< S, A, T,R>، تعریف

[5] subsumption

[6] Simple Additive Weighting

می‌شود که S مجموعه حالات، A مجموعه اعمال، T تابع گذر و R تابع پاداش می‌باشد. تابع گذر احتمال رفتن از یک حالت به حالتی دیگر را نشان می‌دهد و به صورت $T : S \times A \times S \rightarrow [0, 1]$ تعریف می‌شود و تابع پاداش، ارزش بودن در یک حالت و یا ارزش بودن در یک حالت و عمل است که به صورت $R : S \rightarrow R$ و یا $R : S, A \rightarrow R$ نشان داده می‌شود. یافتن یک راه حل برای یک مسئله MDP، می تواند شامل پیدا کردن سیاستی[7] باشد که مقدار مورد انتظار پاداش را در طول زمان حدکثر سازد. یک سیاست به صورت $\pi : S \rightarrow A$ تعریف می‌شود [۱۲].

برای عامل، n رفتار و برای هر رفتار یک سیگنال پاداش جداگانه در نظر گرفته می شود. لذا سیگنال های پاداش R1,R2, …, Rn به رفتارها داده می شود. هر رفتار متناسب با محیطی که در آن قرار دارد، یک فضای حالت نیز دارد و n زیر فضای S1,S2,…,Sn وجود دارد که یکی برای هر رفتار می باشد. بنابراین برای هر رفتار یک MDP در نظر گرفته می شود که به صورت (Si, A, Ti, Ri) نشان داده می شود. فضای اعمال، با فضای اعمال در MDP کلی برابر و تابع گذر نیز از روی MDP کلی مشخص می شود. مجموعه رفتارها نیز با B1, B2, … , Bn نشان داده می‌شوند.

۴- یادگیری تقویتی

یادگیری تقویتی یک راه حل برای مسئله تصمیم گیری مارکف است. این روش، حتی در غیاب مدل محیط که شامل تابع گذر و تابع پاداش است، نیز می‌تواند پاسخ بهینه را بیابد. یکی از مشهورترین الگوریتم‌های یادگیری تقویتی یادگیری Q می‌باشد. علت محبوبیت این روش وجود اثبات‌های همگرایی در این روش، خارج سیاست بودن، ساده بودن و سرعت همگرایی مناسب آن می باشد. شبه کد این روش در شکل ۱ ارائه شده است[۱۳].

```
Initialize Q(s,a) arbitrarily
Repeat (for each episode):
. Initialize s
. Repeat (for each step of episode)
. . choose a from s using policy derived from Q
. . . (e.g., ε greedy)
. . Take action a observe r, s'
. . Q(s,a) ← Q(s,a)+ α[r+γmax_a, Q(s′, a′) − Q(s, a)]
. . s←s';
until s is terminal.
```

شکل ۱. شبه کد یادگیری Q

۴. روش الکتره ۳

فرآیند حل هر مسئله تصمیم‌گیری شامل دو مرحله است. یک مرحله ارزیابی و دیگری مرحله انتخاب. در مرحله اول شاخص‌های کلیدی برای ارزیابی گزینه‌ها تعیین می‌شود که نتیجه حاصل از این مرحله تشکیل ماتریس تصمیم‌گیری است و در مرحله دوم با توجه به ماتریس تصمیم، گزینه‌ها رتبه‌بندی می‌شوند. شکل ۲، این ماتریس را نشان می‌دهد. در این ماتریس، gj(0) معیار j ام، a1…an، مجموعه گزینه‌ها و gj(ai) مقدار گزینه i ام در معیار j ام است [۲].

روش الکتره یکی از روش‌های تصمیم‌گیری است که در پاسخ به کاستی‌های روش‌های موجود ارائه شد. نسخه‌های متعددی از این روش ارائه شده است و دسته‌ای از این روش‌ها تحت عنوان خانواده الکتره معرفی شده اند. یکی از پرکاربردترین آن‌ها، الکتره ۳ است. این روش از آن لحاظ بر دیگر روش‌ها برتری دارد که در مقابل داده نادرست، نامعین و بد تعریف، مقاوم است و لذا نتایج قابل اعتمادتری را ارائه می‌کند. در مقابل روش‌های سنتی که تنها برتری و بی تفاوتی را برای دو گزینه در نظر می‌گرفتند، این روش مفهوم آستانه ارزش بی تفاوتی q، ارزش آستانه برتری p، و ارزش آستانه عدم برتری v، را تعریف می کند. در ادامه مراحل الکتره ۳ باتوجه به [۱۴]، ارائه می شود.

a	$g_1(\cdot)$	$g_2(\cdot)$	$\cdots$	$g_j(\cdot)$	$\cdots$	$g_k(\cdot)$
a_1	$g_1(a_1)$	$g_2(a_1)$	$\cdots$	$g_j(a_1)$	$\cdots$	$g_k(a_1)$
a_2	$g_1(a_2)$	$g_2(a_2)$	$\cdots$	$g_j(a_2)$	$\cdots$	$g_k(a_2)$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
a_i	$g_1(a_i)$	$g_2(a_i)$	$\cdots$	$g_j(a_i)$	$\cdots$	$g_k(a_i)$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
a_n	$g_1(a_n)$	$g_2(a_n)$	$\cdots$	$g_j(a_n)$	$\cdots$	$g_k(a_n)$

شکل۲. ماتریس تصمیم گیری [۱۵].

۴-۱- محاسبه هماهنگی[8]

در این مرحله با مقایسه دو دویی گزینه‌ها و براساس روابط ۱و ۲ ماتریس هماهنگی گزینه‌ها که ماتریسی مربعی با ابعاد تعداد گزینه‌هاست، بدست می‌آید. C(a,b)، مقدار هماهنگی برای گزینه a و b است. kj وزن معیار j ام و cj(a,b)، مقدار هماهنگی برای گزینه a و b در معیار j ام، q آستانه بی تفاوتی و p، آستانه برتری و r تعداد معیارها است.

$$c(a,b) = \frac{1}{k}\sum kj.\,cj(a,b) \qquad (1)$$

$$k = \sum_{j=1}^{r} kj$$

$$cj(a,b) = \begin{cases} 1, & gj(b) - gj(a) \le qj \\ 0, & gj(b) - gj(a) \ge pj \\ \dfrac{pj + gj(a) - gj(b)}{pj - qj}, & \text{other} \end{cases} \qquad (2)$$

j=1,2, ...,r

۴-۱- محاسبه عدم هماهنگی[9]

برای محاسبه عدم هماهنگی برای هر دو گزینه از رابطه ۳ استفاده کنیم. آستانه وتویی v می تواند میزان اعتبار برتری یک گزینه به دیگری را کاملاً رد کند. ماتریس عدم هماهنگی برای هر شاخص بدست می‌آید و برخلاف ماتریس هماهنگی نمی توان تجمیعی از این شاخص‌ها داشت. dj(a,b) مقدار عدم هماهنگی برای گزینه a و b برای معیار j ام است.

[7] policy

[8] Concordance

[9] Discordance

$$dj(a,b)=\begin{cases}0\ ,\ gj(b)-gj(a)\le pj\\ 1,\ gj(b)-gj(a)\ge vj\\ \frac{gj(b)-gj(a)-pj}{vj-pj},\ other\end{cases}\quad (3)$$

j=1,2, ...,r

۴-۳- تشکیل ماتریس اعتبار[10]

پس از بدست آوردن مقدار هماهنگی و عدم هماهنگی برای هر دو گزینه میزان اعتبار برتری یک گزینه بر دیگری بدست می‌آید و پس از این مرحله ماتریس اعتبار بدست می‌آید در این رابطه، J(a,b) بیان‌گر معیارهایی است که در آن $dj(a,b)\le C(a,b)$ است.

$$S(a,b)=\begin{cases}C(a,b), if\ dj(a,b)\le C(a,b)\\ C(a,b)*\prod_{j\in J(a,b)}\frac{1-dj(a,b)}{1-C(a,b)}\end{cases}\quad (4)$$

j=1,2, ...,r ,

۴-۴- رتبه بندی گزینه ها

در این مرحله با داشتن ماتریس اعتبار، گزینه‌ها رتبه بندی می‌شوند. اگرچه روش عمومی که در مقالات برای الکتره ۳ ارائه شده است، روشی است که از ترکیب دو رتبه بندی نزولی و صعودی، رتبه بندی نهایی را بدست می‌دهد، در این مقاله از رابطه ۵ برای رتبه بندی نهایی گزینه‌ها استفاده شده است. این روش ترکیبی از روش الکتره ۳ و روش پرومته[11] [۱۵] است. در این رابطه، $\varphi^{+}(a)$ متوسط برتری معیار a بر دیگر معیارها و $\varphi^{-}(a)$ متوسط برتری دیگر گزینه ها بر b است. از اختلاف این دو مقدار رتبه گزینه a مشخص می شود. گزینه ها بر حسب مقدار φ، به صورت نزولی مرتب می شوند. A، در این رابطه مجموعه تمام گزینه ها است.

$$\varphi^{+}(a)=\frac{1}{n-1}\sum_{x\in A}S(a,x)\quad (5)$$
$$\varphi^{-}(a)=\frac{1}{n-1}\sum_{x\in A}S(x,a)$$
$$\varphi(a)=\varphi^{+}(a)-\varphi^{-}(a)$$

۵- روش پیشنهادی

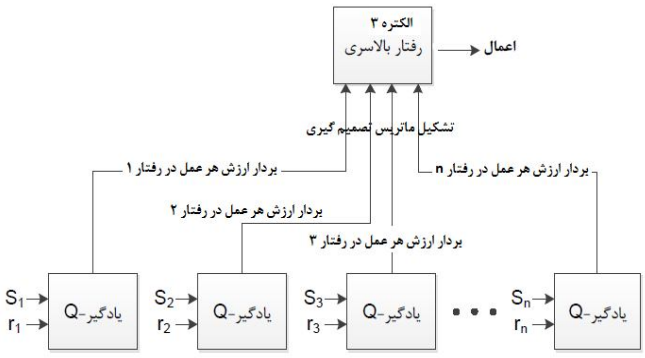
در روش ارائه شده در این مقاله هر رفتار یادگیر تقویتی Q می‌باشد که در زیر فضای حالت خود که فضایی به اندازه کافی کوچک است به یادگیری می‌پردازد.

پس از اتمام یادگیری مقادیر Q به رفتار بالاسری برده می‌شود. هر رفتار ازآنجا که یادگیر تقویتی Q است، جدولی از ارزش حالت و عمل در خود نگه داشته است. لذا اگر هر رفتار را به عنوان یک معیار ببینیم و هر عمل گزینه‌ای باشد که در ماتریس تصمیم‌گیری به آن اشاره کردیم. آنگاه مقدارهای Q مقدار هر عمل در هر معیار می‌باشند و برای مثال برای عملی با ۴ رفتار و ۲ عمل ماتریس تصمیم گیری به صورت شکل ۳، می‌شود. فقط باید توجه داشته باشیم از آن جا که هر رفتار یک زیرفضای حالت مخصوص به خود را دارد، زیرفضای حالت هر رفتار را می توان نماینده آن رفتار دانست.

جدول ۱. مثالی برای نمایش ماتریس تصمیم به عنوان نتایجی از یادگیری Q

	$B_1=s_1$	$B_2=s_2$	$B_3=s_3$	$B_4=s_4$
a_1	$Q(s1,a_1)$	$Q(s2,a_1)$	$Q(s3,a_1)$	$Q(s4,a_1)$
a_2	$Q(s1,a_2)$	$Q(s2,a_2)$	$Q(s3,a_2)$	$Q(s4,a_2)$

با داشتن چنین ماتریسی می توان در هر مرحله که عامل می‌خواهد عملی را انجام دهد با استفاده از روش الکتره ۳ بهترین عمل انتخاب شود. بنابراین معماری رفتار محور به صورت شکل ۳ می‌شود.



شکل ۳. معماری عامل رفتار محور

پارمترهای روش به صورت تعریف می شود که برای بدست آوردن ماتریس وزن مانند روشی که در [۶] با نام بیشینه کردن مجموع خوشبختی ارائه شده است وزن هر رفتار بدست می آید. بنابراین رفتاری که مجموع مقادیر Q برای اعمالش بیشینه است اهمیت بیشتری دارد و به صورت رابطه ۴ نوشته می‌شود.

برای هر یک از مقادیر p, q, v از آنجا که در مقالات این مقادیر از سوی خبره تعیین می شود و در روش ارائه شده در مقاله معیارها که همان رفتارها هستند، عینیت واقعی ندارند، مقادیر این سه آستانه از روی بیشینه اختلاف بین مقادیر گزینه‌ها برای هر معیار بدست آمده است. مقادیر p,q,v را بین این مقدارهای بیشینه و کمینه تغییر می‌دهیم و مقدار مطلوب با استفاده از سعی و خطا بدست آمده است. راهنما برای تنظیم این مقادیر متوسط پاداشی بوده که عامل از طریق عمل کردن بر حسب روش الکتره ۳ بدست آورده است. با تنظیم این مقادیر به مقدارهای جدول ۲، نتایج مطلوب بدست آمده است. از آنجا که با سعی و خطا مقادیر بهینه محلی بدست آمده است این امکان وجود دارد که پارامترهای بهتر کارایی بیشتری را نسبت به آنچه به دست آمده، نتیجه بدهد.

[10] Credibility matrix

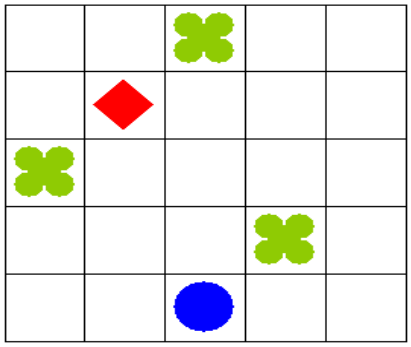
11 Promethee

جدول۲. مقادیر انتخاب شده برای پارامترهای روش الکتره ۳

	B1	B2	B3	B4
q	0.29	0.29	0.29	0.29
p	0.29	0.29	0.29	0.29
v	0.75	0.75	0.75	0.75

۶- پیاده سازی

برای پیاده سازی روش پیشنهادی مسئله جمع آوری غذا انتخاب شده است. در این مسئله که در یک محیط دو بعدی تعریف می شود، هدف عامل جمع آوری تکه‌های غذای ساکن و همزمان اجتناب از یک شکارچی می‌باشد. پیوسته سه تکه غذا در محیط وجود دارد. اگر عامل به یکی از این تکه‌های غذا برخورد کند، پاداشی به اندازه ۱ دریافت می‌کند و آن تکه غذا به مکان تصادفی دیگر منتقل می‌شود. عامل می‌تواند برای حرکت یکی از ۸ عمل انتخابی خودش را انجام دهد اما با احتمال ۰,۱ یک حرکت تصادفی انتخاب می‌کند. شکارچی در هر گام به صورت قطعی یک خانه به سمت عامل حرکت می‌کند. عامل هر گام زمانی که از شکارچی فرار می‌کند پاداشی به مقدار ۰,۵ دریافت می‌کند[۱]. در شکل ۴، این محیط برای ابعاد ۵*۵ نشان داده شده است. که عامل با لوزی قرمز، شکارچی با دایره آبی و تکه های غذا با رنگ سبز نشان داده شده اند.



شکل ۴. محیط مسئله جمع‌آوری غذا

فضای حالت در این محیط ۵*۵ با در نظر گرفتن مکان سه تکه غذا و شکارچی و عامل به اندازه 25^5 می‌باشد. این فضای حالت بزرگتر از آن است که یک یادگیر تقویتی ساده بتواند در آن موفق باشد. برای پیاده‌سازی روش پیشنهادی، فضای حالت به ۴ زیر فضا شکسته می‌شود. به طوری‌که در هر زیرفضا تنها عامل و یکی از ۴ عنصر شکارچی یا ۳ تکه غذا وجود دارد. هر رفتار بخشی از این زیر فضا را یاد می‌گیرد.

هر رفتار یک یادگیر Q ساده است که از سیاست اپسیلون- حریصانه[۱۲] استفاده می‌نماید. پارامتر اپسیلون ۰,۴ می‌باشد و به صورت خطی به صفر کاهش می‌یابد. در همه یادگیرها، نرخ یادگیری ثابت و ۰,۰۵ و نرخ کاهش ۰,۹ می‌باشد.

[12] e-greedy

پس از اتمام یادگیری مقادیر Q برای هر رفتار به رفتار بالاسری می‌رود و به صورت یک ماتریس به روش الکتره ۳ داده می‌شود و پارامترهای روش الکتره ۳ در جدول ۲ ارائه شده است.

برای این مسئله روش بزرگترین جرم، یادگیری W، روش ارائه شده در [۱] نیز پیاده سازی شده است و کارایی این روش ها با روش پیشنهادی مقایسه شده است.

۷- تحلیل نتایج

جدول ۴، نتایج بدست آمده از روش‌های متفاوت را نشان می‌دهد. مقادیر در این جدول متوسط پاداشی است که عامل با استفاده از روش‌های متفاوت بدست آورده است نمودار متناسب با این جدول در شکل ۵ نشان داده شده است.

باتوجه به جدول ۴، متوسط پاداش در روش ارائه شده در این مقاله در حدود یک می‌باشد که نتیجه‌ای بسیار خوب است. روش بزرگترین جرم، همین ماتریس تصمیم گیری را از رفتارهای پایین دستی می گیرد اما تنها عملی را انتخاب می کند که جمع مقادیر ارزش آن یعنی مجموع عناصر در هر سطر ماتریس تصمیم بیشینه باشد. در این روش اهمیت و وزن هر رفتار نادیده گرفته می شود و انتظار هم می رود که متوسط پاداشی که عامل با استفاده از این روش بدست می آورد مقدار کمی باشد. برای یادگیری w نیز چنین توجیهی وجود دارد. زیرا این اگرچه برای هر رفتار وزنی را انتخاب می کند در هر لحظه تنها عمل رفتاری را انتخاب می کند که وزن بیشتری دارد. مقدار وزن بستگی به مقادیر همین ماتریس و پاداشی که عامل بدست می آورد، دارد اما از اطلاعات این ماتریس به طور کامل استفاده نمی شود. این جدول نشان می‌دهد که روش پیشنهادی نتیجه‌ای نزدیک به بهترین نتایج بدست اورده است و از سویی با بزرگی فضای حالت و عمل ناکارآمد نمی‌شود.

جدول۳. متوسط پاداش بدست آمده توسط عامل با استفاده از روش‌های متفاوت

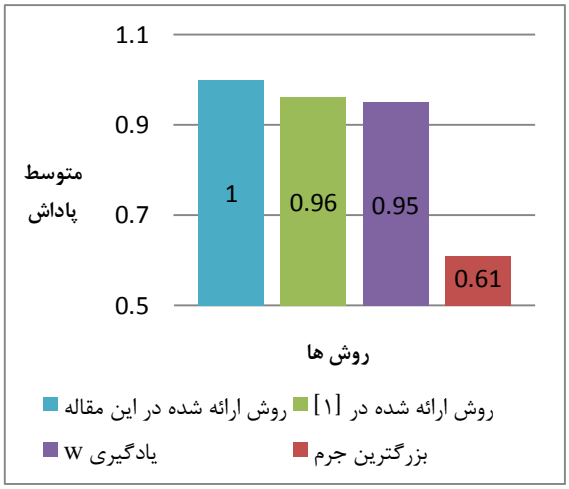
روش	متوسط پاداش (مقدار حدودی)
بزرگترین جرم	۰,۶۱
یادگیری w	۰,۹۵
روش ارائه شده در [۱]	۰,۹۶
روش ارائه شده در این مقاله	۱

۹- نتیجه گیری

در این مقاله روشی برای هماهنگی رفتار با استفاده از مفاهیم موجود در نظریه‌های تصمیم‌گیری چند معیاره ارائه شده است. بدین ترتیب یک معماری برای عامل رفتارمحور ارائه شده است که در آن رفتارهای پایین دستی یادگیرهای Q و رفتار بالاسری با استفاده از روش الکتره ۳ که یکی از روش‌های تصمیم‌گیری چند معیاره است، عمل برآیند را تعیین می‌کند. از مزایای روش ارائه شده می توان به این موارد اشاره کرد که این روش

مشکلی در فضای حالات و اعمال بزرگ ندارد و کارایی عامل همچنان حفظ می شود و از طرفی دیگر نتایج تولید شده نسبت به سایر روش های موجود بسیار قابل قبول است.

در کنار مزایایی که این روش دارد، معایبی هم می توان برای آن عنوان کرد و آن این است که این روش بسیار وابسته به پارامتر می‌باشد و تنظیم پارامتر در آن فرآیندی دشوار است. از سوی دیگر چون این پارامترها با سعی و خطا بدست می آیند، ممکن است مقدار بهینه برای آن ها حاصل نشود. بنابراین یک پیشنهاد می تواند این باشد که این مقادیر با استفاده از منطقی خاص به صورت خودکار بدست آیند.



شکل ۵. نمودار نتایج

مراجع

[۱] س. م. میرهاشمی، "به کارگیری یادگیری در طراحی خودکار سامانه های رفتارمحور،" پایان نامه برای دریافت درجه کارشناسی ارشد در رشته مهندسی کامپیوتر گرایش هوش مصنوعی، علم و صنعت ایران، دانشکده کامپیوتر، صفحه ۱-۹۰، ۱۳۹۱.

[۲] ا. کزازی، " ارزیابی و اولویت بندی استراتژ ی ها با استفاده از تکنیک الکتره ۳ در محیط فازی (مطالعه موردی: شرکت تماد)،" فصلنامه علمی پژوهشی مطالعات مدیریت صنعتی سال هشتم، شماره ۲۰، صفحه ۴۹-۷۹، ۱۳۹۰.

[3] P. Pirjanian, "Behavior coordination mechanisms-state-of-the-art," *Institute for Robotics and Intelligent Systems, School of Engineering, University of Southern California,Tech. Rep,* pp. 1-49, 1999.

[4] S. Mahadevan and J. Connell, "Automatic programming of behavior-based robots using reinforcement learning," *Artificial intelligence and mobile robots,* vol. 55, pp. 311-365, ۱۹۹۲.

[5] P. Maes and R. A. Brooks, "Learning to coordinate behaviors," *in Proceedings of the eighth National conference on Artificial intelligence (AAAI'90),* pp. 796-802., 1990.

[6] M. Humphrys, "Action Selection method using Reinforcement Learning," *Proceedings of Fourth International Conference on Simulation of Adaptive Behavior,* 1996.

[7] N. Sprague and D. H. Ballard, "“Multipe-Goal Reinforcement Leraning with Modular Sarsa(0)," *Proc. Int. Joint Conf. on Artificial Intelligence,* pp. 1445-1447, 2003.

[8] A. m. Farahmand, *et al.*, "Hybrid Behavior Co-evolution and Structure Learning in Behavior-based Systems," *Evolutionary Computation , IEEE Congress on Sheraton Vancouver Wall Centre Hotel, Vancouver, BC, Canada,* pp. 275-282, 2006.

[9] S. Whitehead, *et al* " ,.Learning Multiple Goal Behavior via Task Decompostion and Dynamic Policy Merging," *Kiluwer Academic Press,* pp. 45-78, 1993.

[10] G. D. Konidaris and G. M. Hayes, "An Architecture for Behavior-Based Reinforcement Learning," *International Society of Adaptive Behavior,* pp. 5-32, 2005.

[11] M. J. Matarić, "Learning in behavior-based multi-robot systems: Policies, models, and other agents," *Cognitive Systems Research,* vol. 2, pp. 81-93, 2001.

[12] M. v. Otterlo, "A Survey of Reinforcement Learning in Relational Domains," pp. 1-66, 2005.

[13] R. S. Sutton and A. G. Barto, "Reinforcement Learning: an Introduction," *The MIT Press,Cambridge,* 1998.

[14] J. BUCHANAN, *et al.*, "Project Ranking Using the ELECTRE Method," pp. 1-21, 1999.

[15] J. FIGUEIRA, *et al.*, Eds., *Multip ple criteria decisiom analysis*. Springer Science,pp. 1-1085, 2005.