مدل سازی تصمیم گیری اجتماعی با استفاده از رویکرد یادگیری تقویتی

سحر افتاده بالانی، علیرضا مرادی[1]*

۱- دانشجوی دکتری مدلسازی شناختی، موسسه آموزش عالی علوم شناختی،تهران،ایران

۲- استاد، عضو هیات علمی موسسه آموزش عالی علوم شناختی،عضو هیات علمی دانشگاه خوارزمی، تهران،ایران(نویسنده مسئول)

چکیده

در سال های اخیر شاهد تکامل چشمگیری در استفاده از مدل های یادگیری تقویتی (RL)[2] در علوم شناختی بوده ایم. با این حال، رشد استفاده از رویکردهای محاسباتی نسبتاً پیچیده منجر به تصورات نادرست بالقوه و تفسیرهای نادرست شده است. در این مقاله، یک چارچوب جامع برای بررسی تصمیم‌گیری اجتماعی با استفاده از رویکرد یادگیری تقویتی ارائه می‌کنیم. همچنین دانش و اطلاعات دریافتی از گروه را مورد پایش قرار داده و پیشنهادات کاربردی ارائه می دهیم. مرور مطالعات انجام شده در خصوص این مباحث نشان می دهد تصمیم گیری فرآیندی زمان بر است و انسان در آن واحد بیش از یک تصمیم نمی تواند اخذ کند. هیجانات مولفه ای اساسی در تنظیم فعل و انفعالات بین شرایط محیطی و فرآیند تصمیم گیری انسان هستند. سیستم های عاطفی، دانش ضمنی و صریح ارزشمندی را برای تصمیم گیری های سریع و عقلانی فراهم می کند. درنهایت اینکه تصمیمات ماهیت شناختی دارند و از این رو یافته های علوم شناختی می تواند به تقویت نظریه های تصمیم گیری برای درک کامل تر فرآیند انتخاب افراد کمک کند و دیدگاه های کامل تر و واقع گرایانه تری از تصمیم گیری ارائه کند. در این تحقیق، ابتدا نقش عملکردی یادگیری تقویتی در تصمیم گیری اجتماعی را بررسی نموده، سپس مدل پیشنهادی ارائه خواهد شد. هدف ما در این تحقیق ارائه توضیحات ساده ، مقیاس‌پذیر و دستورالعمل‌های عملی و استخراج زیر هدف ها و کسب اطلاعات و دانش جمعی جهت کمک به تصمیم گیری اجتماعی مبتنی بر رویکرد یادگیری تقویتی است. نتایج این تحقیق نشان می دهد استفاده از چند پارامتر یادگیری اجتماعی برای اخذ دانش و اطلاعات برتری های قابل ملاحظه‌ای برای روش پیشنهادی در این مقاله ایجاد می‌کند.

واژه‌های کلیدی

تصمیم گیری اجتماعی، مدل سازی محاسباتی، یادگیری تقویتی، یادگیری اجتماعی

مقدمه

انسان در طول زندگی با مسائل و مشکلات مختلفی مواجه است که ناگزیر به تصمیم گیری می انجامد. تصمیم گیری نقش زیادی در زندگی یک فرد اجتماعی دارد و همه چیز را از تصمیمات کوچک در مورد مسائل کوچک گرفته تا تصمیمات بسیار بزرگ و مهم را شامل می شود. در دنیای پیچیده امروزی که به عنوان عصر تغییرات سریع و عدم اطمینان شناخته می شود، موضوع تصمیم گیری از اهمیت بسزایی برخوردار است [۴].

با توجه به پیچیدگی ماهیت تصمیم گیری و دشواری درک آن از جنبه های مختلف، این موضوع در این مقاله از منظر یادگیری تقویتی مورد بررسی قرار گرفته است و به دنبال پاسخ به سوال هایی نظیر اینکه چگونه با استفاده از روش های یادگیری اجتماعی می توان در گروه های اجتماعی تصمیم بهینه را اتخاذ نمود، می باشیم.

یکی از مسائل مهمی که در تصمیم گیری نسبت به یک موضوع اثرگذار است، دانستن عقیده دیگران درباره آن موضوع است. از مدت ها پیش از ظهور وب، بسیاری از مردم از دیگران می خواستند تا درباره وسیله ای که قبلا خریده اند یا نماینده ای که می خواهند به آن رأی دهند نظرشان را بگویند. باگسترش اینترنت و حجم انبوه کاربران فرصت تازه ای برای تحلیل عقاید عموم فراهم شد. در این راستا مسئله جدیدی که توسط پژوهشگران مطرح شد این بود که سیستمی طراحی شود که بتواند احساس نویسنده را از داده متنی استخراج کند.

تصمیم گیری مناسب و بهینه برای افراد یک جامعه جهت امور مختلف از جمله سرمایه گذاری و مواردی از این دست نقش مهمی در زندگی افراد یک جامعه ایفا می کند. با این حال، دستیابی به تصمیم مناسب جهت اخذ تصمیم بهینه پیچیده، پویا و چالش برانگیز است. در این تحقیق پتانسیل یادگیری تقویتی را برای بهینه سازی تصمیم افراد و در نتیجه به حداکثر رساندن شانس یک تصمیم گیری خوب بررسی خواهیم کرد[۹].

---

[1] Moradi@khu.ac.ir

[2] Reinforcement Learning

تاریخ دریافت: ۱۴۰۱/۱۲/۲۲ تاریخ پذیرش: ۱۴۰۲/۰۲/۱۹

یادگیری تقویتی

یادگیری تقویتی شاخه ای از یادگیری ماشینی است که با الهام از رفتارگرایی و تمرکز بر رفتارهایی که ابزار برای به حداکثر رساندن پاداش ها باید انجام دهد[۱۲]. یادگیری تقویتی به یادگیری آنچه انجام شود، به منظور به حداکثر رساندن پاداش ها و ترسیم موقعیت ها به اقدامات گفته می شود. یادگیرنده یا عامل نمی داند چه اقداماتی را باید انجام دهد، در عوض اقداماتی را کشف می کند که امتحان کردن آنها بیشترین پاداش را به همراه دارد. اقدامات انجام شده ممکن است پاداش فوری را دریافت کنند یا نکنند، اما ممکن است در مدت زمان طولانی تری پاداش داشته باشند. ویژگی های مهم یادگیری تقویتی جستجوی آزمون و خطا و پاداش تاخیری است [۶].

در یادگیری تقویتی، اقدام عامل از قبل مشخص نیست در این روش عامل با جستجوی مبتنی بر آزمون و خطا یاد می گیرد رفتاری را انجام دهد که بیشترین پاداش را در بر داشته و منفعت کوتاه مدت فدای منفعت بلند مدت می شود. در جستجوی کسب پاداش بیشتر همواره دو رویکرد بهره مندانه و اکتشافی وجود دارد که می بایست بین آن ها تعادل ایجاد شود[۱۶].

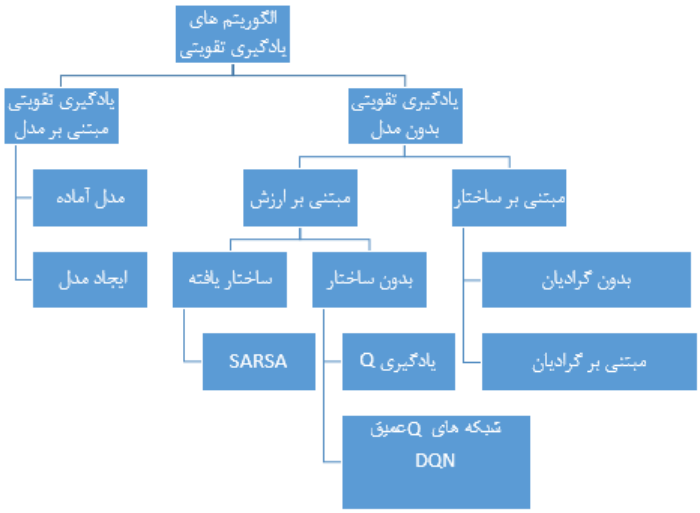
اساس یادگیری تقویتی، فرآیند تصمیم گیری مارکوف است. بر اساس فرآیند مارکوف، در هر مرحله از زمان، فرآیند در یک حالت از فضای حالت S قرار دارد و تصمیم گیرنده ممکن است یک عمل مثل A را که در حالت S در دسترس است انتخاب کند. این فرایند در مرحله بعدی با حرکت تصادفی به حالت جدید مثل S' پاسخ داده و به تصمیم گیرنده، پاداش مرتبط با آن تصمیم را می دهد که به صورت $R_a(s.s')$ است. ما نیز در این مقاله از فرآیند تصمیم گیری مارکوف[۳] استفاده خواهیم کرد[۴]. احتمال اینکه این فرآیند در مرحله بعدی در حالت S' قرار گیرد، بطور خاص توسط تابع تغییر حالت $R_a(s.s')$ ارایه می گردد. با دانستن وضعیت S و عمل a، وضعیت بعدی ازوضعیت ها و عمل های قبلی مستقل خواهد بود. به این معنا که انتقال حالت، دارای خاصیت مارکوف است[۲۱].

اجزای اصلی این روش عبارت است از:

۱- محیط: شامل مجموعه حالت هاست.
۲- عمل ها: تصمیم است که تصمیم گیرنده اتخاذ می کند.
۳- پاداش: پاداش یا جریمه ای است که تصمیم گیرنده بابت تصمیم اتخاذ شده دریافت می کند.

دو گروه عمده الگوریتم های یادگیری تقویتی عبارتند از :یادگیری تقویتی بدون مدل و یادگیری تقویتی بر مبنای مدل. در یادگیری تقویتی بدون مدل توابع ارزش(تخمین پاداش های آینده) به صورت سراسری و بر اساس تفاضل میان پاداش مورد انتظار از توابع ارزش کنونی و پاداشی که عامل در حال حاضر دریافت کرده بدست می آید[۴۳]. در این روش به دلیل به روزرسانی پیوسته توابع ارزش فرآیند انتخاب عمل در یک حالت داده شده ساده است. این سبک از یادگیری تقویتی دارای انعطاف می باشد و اغلب با عنوان یادگیری عادتی به آن ارجاع می شود.کاربرد این روش ها بیشتر در زمینه های اجتماعی تکراری و بازی های تکراری در محیط های آزمایشگاهی می باشد. یکی از معروف ترین الگوریتم ها در این زمینه الگوریتم Q-Learning است[۴۵].

در یادگیری تقویتی، نوع اقدامی که نماینده انجام خواهد داد از قبل تعیین نشده است، اما عامل رفتاری را می آموزد که بیشترین پاداش را به دست آورد و با آزمون و خطا، سود کوتاه مدت را فدای سود بلندمدت می کند. در استراتژی جستجو برای رسیدن به پاداش بیشتر، همواره دو رویکرد اصلی بهره مندانه و اکتشافی وجود دارد. چالش اصلی ایجاد یک تعادل با ترکیب دو رویکرد فوق است. یادگیری اساساً به سه آموزش با الگو، یادگیری بدون الگو و یادگیری تقویتی تقسیم می شود. در یادگیری الگو، داده‌های برچسب‌گذاری‌شده به شکل جفت‌های مشترک در فرآیند یادگیری استفاده می‌شوند و معمولاً هدف یافتن یک الگوی ساختار یافته در میان داده‌ها است. در یادگیری بدون مدل، داده ها بدون برچسب و اغلب بر اساس شباهت با یکدیگر تجزیه و تحلیل می شوند. در یادگیری تقویتی، داده ها بر اساس پاداش و تنبیه تجزیه و تحلیل می شوند [۲۹].



[۳] MDF

شکل ۱- طبقه بندی الگوریتم های غالب در یادگیری تقویتی مدرن که بصورت گسترده در مطالعات مختلف مورد استفاده قرار می گیرند[۴۵].

یادگیری اجتماعی

یادگیری اجتماعی فرآیند کسب دانش، مهارت و رفتار از طریق مشاهده، تعامل و ارتباط با دیگران است. این شامل یادگیری از تجربیات دیگران و انطباق رفتار خود بر اساس آن است. یادگیری اجتماعی می تواند در محیط های مختلفی از جمله آموزش رسمی، آموزش در محل کار و تعاملات اجتماعی غیررسمی رخ دهد. یادگیری اجتماعی از طریق چهار مکانیسم کلیدی اتفاق می افتد[۴۰].

۱. مشاهده: افراد با مشاهده رفتار دیگران و پیامدهای آن رفتار می آموزند. این می تواند مشاهده مستقیم یا از طریق رسانه هایی مانند فیلم ها یا رسانه های اجتماعی باشد.
۲. الگوسازی: مردم اغلب از رفتار دیگرانی که آنها را تحسین می کنند یا به آنها احترام می گذارند تقلید می کنند، به ویژه اگر آن رفتار را موفق یا مؤثر بدانند.
۳. بازخورد: افراد از بازخورد دریافتی از دیگران، چه مثبت یا منفی، یاد می گیرند. این بازخورد می تواند از طرف همسالان، مربیان یا منتورها باشد.
۴. همکاری: افراد از طریق همکاری با دیگران برای حل مشکلات یا دستیابی به اهداف یاد می گیرند. یادگیری مشارکتی می تواند اشکال مختلفی داشته باشد، از جمله پروژه های گروهی، فعالیت های تیمی و یادگیری همتا به همتا.

یادگیری اجتماعی در فرآیند تصمیم گیری

برخی از تصمیمات در سطح خرد اغلب شامل دو یا چند طرف است. مثلا خریدار- فروشنده، کارمند- کارفرما و .... بخشی از هزینه دیگری است که عمدتاً در تضاد منافع است و به نفع طرف مقابل است. اطلاعات نقش مهمی در تصمیم گیری ها و ارتباطات آنها دارد. در بسیاری از موارد، اطلاعات مورد نیاز ناقص یا محدود است. حال در این مقاله به بررسی نقش اساسی دانش و اطلاعات در تصمیم گیری اجتماعی می پردازیم[۴۰]. وقتی مردم می خواهند کاری انجام دهند به امید دریافت پاداش (نتیجه) تصمیم می گیرند. اما هر تصمیمی لزوماً منجر به دریافت آن پاداش نمی شود، بلکه تصمیمی وجود دارد که افراد را در دستیابی به پاداش مورد نظر پیروز می کند. برای یافتن این تصمیم، جمع آوری اطلاعات ضروری است. بنابراین نتیجه (پاداش) تصمیمات نامعلوم است، شاید جمع آوری اطلاعات پراکنده در بین افراد در میان جمعیت منجر به پاداش شود. در این مورد، تصمیم هر فرد به سیگنال های شخصی دریافت شده و اطلاعات ارسال شده از دیگران بستگی دارد. می تواند اشکال مختلفی برای انتقال اطلاعات از دیگران وجود داشته باشد[۴۴]:

۱- افراد ممکن است تمام اطلاعات دیگران را بدانند.
۲-علائم شخصی دریافتی (اطلاعات شخصی) افراد دیگر را مشاهده کنند.
۳- تصمیمات مبتنی بر اطلاعات شخصی افراد را مشاهده کنند و بر اساس آنها تصمیم گیری نمایند.

در این میان مطمئن ترین و معتبرترین راه مشاهده افراد و استخراج اطلاعات بر اساس آن است.

لایه های شناختی به سطوح مختلف تفکری که افراد برای پردازش اطلاعات و تصمیم گیری استفاده می کنند، اشاره دارد. این لایه ها شامل ادراک، توجه، حافظه، زبان و استدلال است[۲۵]. هر لایه در نحوه درک و تفسیر اطلاعات، قضاوت و تصمیم گیری افراد نقش دارد. عواملی که می توانند بر تصمیم گیری تأثیر بگذارند عبارتند از ارزش ها و باورهای شخصی، احساسات، هنجارها و انتظارات اجتماعی، سوگیری های شناختی و عوامل موقعیتی. این عوامل می توانند بر نحوه پردازش اطلاعات، ارزیابی گزینه ها و انتخاب افراد تأثیر بگذارند. درک این عوامل می تواند به افراد کمک کند تا تصمیمات آگاهانه و مؤثرتری بگیرند[۵]. مدل‌سازی لایه‌های شناختی در تصمیم‌گیری اجتماعی شامل درک چگونگی پردازش و تفسیر اطلاعات در موقعیت‌های اجتماعی و تأثیر این امر بر تصمیم‌گیری افراد است. این را می توان از طریق روش های مختلفی مانند بررسی، آزمایش یا شبیه سازی انجام داد. یک رویکرد استفاده از مدل های شناختی است که نحوه پردازش اطلاعات و تصمیم گیری افراد بر اساس عوامل مختلف مانند هنجارهای اجتماعی، احساسات و سوگیری های شناختی را شبیه سازی می کند. این مدل‌ها را می‌توان برای پیش‌بینی چگونگی رفتار افراد در موقعیت‌های مختلف اجتماعی و شناسایی عواملی که ممکن است بر تصمیم‌گیری آن‌ها تأثیر بگذارد استفاده کرد[۳۲]. رویکرد دیگر، مطالعه چگونگی تصمیم گیری افراد در موقعیت های اجتماعی، از طریق نظرسنجی یا آزمایش است. این می تواند بینشی در مورد فرآیندهای شناختی درگیر در تصمیم گیری، و همچنین عواملی که بر این فرآیندها تأثیر می گذارد، ارائه دهد. به طور کلی، مدل‌سازی لایه‌های شناختی در تصمیم‌گیری اجتماعی برای درک چگونگی تصمیم‌گیری افراد در زمینه‌های اجتماعی و برای توسعه استراتژی‌هایی برای ارتقای تصمیم‌گیری آگاهانه‌تر و مؤثرتر مهم است[۲۴].

یادگیری تقویتی و روش های موجود برای تصمیم گیری

در یادگیری تقویتی، عامل با محیط در یک سری مراحل زمانی به صورت پنهانی تعامل دارد. در زمان t ، عامل حالت سیستم را که با s_t نشان داده می شود دریافت نموده و کنش a_t را از بین کنش های مجاز که با $A(s_t)$ نشان داده می شود انتخاب و به محیط اعمال می کند. عامل پاداش فوری a_{t+1} را دریافت می کند و به حالت S_t+1 می رود. یک خط مشی بین هر حالت و احتمال انتخاب هر اقدام ممکن، یک نگاشت، یک نقشه برداری ایجاد می کند. در یادگیری تقویتی، هدف به حداکثر رساندن تابع پاداش در آینده است که معمولا آن را تابع ارزش می نامیم. تابع ارزش ممکن است با متوسط پاداش، پاداش تخفیف یافته و ... مشخص شود. مقدار تابع ارزش با پاداش تخفیف‌یافته در حالت s هنگامی که سیاست π دنبال می‌شود و نرخ تخفیف γ است با رابطه زیر بدست می آید[۲].

$$V^{\pi}(s) = E(\mathrm{r}_{t+1} + \gamma \mathrm{r}_{t+2} + \gamma^2 \mathrm{r}_{t+3} + \ldots | \mathrm{s}_t) = s, \pi$$

تابع مقدار بهینه تابع مقدار خط مشی است که در هر مورد بیشترین مقدار را دارد و از رابطه زیر که به رابطه بلمن معروف است به دست می آید.

$$V^*(s) = \max_{\alpha} [R(s,a) + \gamma \sum_{s'} p(s'/,a) V^*(S^*)]$$

γ< <0 1 نرخ تخفیف نامیده می شود و $V^*(s)$ مجموع پاداش تخفیف یافته حالت S با اجرای سیاست بهینه می باشد. اگر در حالت S باشیم و کنش a را انتخاب کنیم و سپس سیاست بهینه را در پیش بگیریم، می‌توانیم رابطه بلمن را برای یک تابع ارزش-کنش تعریف کنیم که در یادگیری تقویتی با Q^* نشان داده می‌شود.

$$\mathrm{Q}_{k+1}(s,a) = (1-\alpha_k)\mathrm{Q}_k(s,a) + \alpha_k [r + \gamma \max_{\alpha'} \mathrm{Q}_k(s',a')]$$

اخیراً از روش های یادگیری تقویتی در تلاش برای حل مسائل با فضای حالت بسیار بزرگ در حوزه تصمیم گیری استفاده شده است. چارچوب های مختلفی برای یادگیری تقویتی سلسله مراتبی ارائه شده است[۱]. فریمورک های معروف در این زمینه عبارتند از: گزینه [۸] و MAXQ [۱۰]. برای استخراج ساختار سلسله مراتبی و تصمیم گیری ها، از روش اختیار بیش از سایر روش ها استفاده می شود و به همین دلیل در این تحقیق از روش اختیاری نیز استفاده شده است.

هر گزینه از سه جز تشکیل شده است. یک سیاست به صورت $\pi: S \times A \to [0,1]$، یک دامنه بصورت $I \subseteq S$ و یک شرط خاتمه به صورت $\beta: S^+ \to [0,1]$. I مجموعه حالت هایی است که گزینه می تواند از آن ها آغاز شود و ادامه پیدا کند و S مجموعه حالت های ممکن در محیط می باشد. گزینه $< I, \pi, \beta >$ در حالت s قابل دسترسی است اگر و فقط اگر s I∈ . اگر گزینه اجرا شود کنش ها بر اساس π انتخاب می شوند تا زمانی که گزینه به صورت تصادفی بر اساس β خاتمه یابد[۱۵].

روش های مختلفی برای ایجاد شرایط تصمیم گیری در مطالعات قبلی استفاده شده است. در ادامه این بخش به معرفی برخی از این روش ها می پردازیم. به طور کلی روش های تصمیم گیری در یادگیری تقویتی را می توان به دو گروه اصلی تقسیم کرد. در گروه اول، عامل اهداف فرعی را تعیین می کند تا مشکل را به یک سری مسائل فرعی تقسیم کند و یک ساختار فرآیند سلسله مراتبی ایجاد کند. در گروه دوم، عامل مستقیماً یک ساختار سلسله مراتبی بدون شناسایی اهداف فرعی ایجاد می کند [۱۰].

روش پیشنهادی

در این مقاله یک روش یادگیری تقویتی جدید مبتنی بر یادگیری اجتماعی ارائه شده است که به تدریج از عوامل مختلف یادگیری اجتماعی برای هدایت مکانیسم اکتشاف و کشف اهداف فرعی استفاده می شود. ما از حقایق زیر در مورد فرآیند یادگیری در انسان استفاده می کنیم تا چارچوب جدیدی را برای فرآیند یادگیری ارائه دهیم:

۱- دانش به صورت تدریجی در طول عمر از تولد تا مرگ آموخته می شود. آن ها از اطلاعات ساده تر برای یادگیری و تصمیم گیری امور پیچیده تر استفاده می‌کنند[۳۹].

۲- دانش و اطلاعات می‌توانند در حین انجام فعالیت های گروهی مانند بازی یا در حال انجام یک وظیفه خاص یاد گرفته شوند. بنابراین دانش و اخذ نظرات دیگران برای اتخاذ تصمیم می‌توانند وابسته به وظیفه یا مستقل از آن باشند.

۳- برای انجام یک وظیفه خاص، انسان ها قادر به انتخاب میان دانش و اطلاعات از پیش یادگرفته شده می‌باشند که برای اتخاذ تصمیم مورد نیاز است [۷].

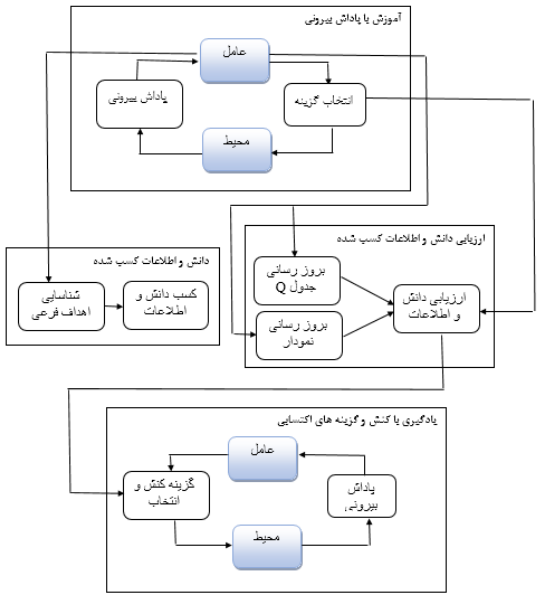
۴- افکار و نظرات درونی شده با انجام فعالیت های جدید و ورود دانش جدید بهبود می یابند.

۵- نظرات دیگران را می توان براساس انگیزه های متفاوتی مانند کنجکاوی، تازگی، خلاقیت، سبب بودن و ... یاد گرفت و در تصمیم های خود از آن استفاده کرد [۱۴].

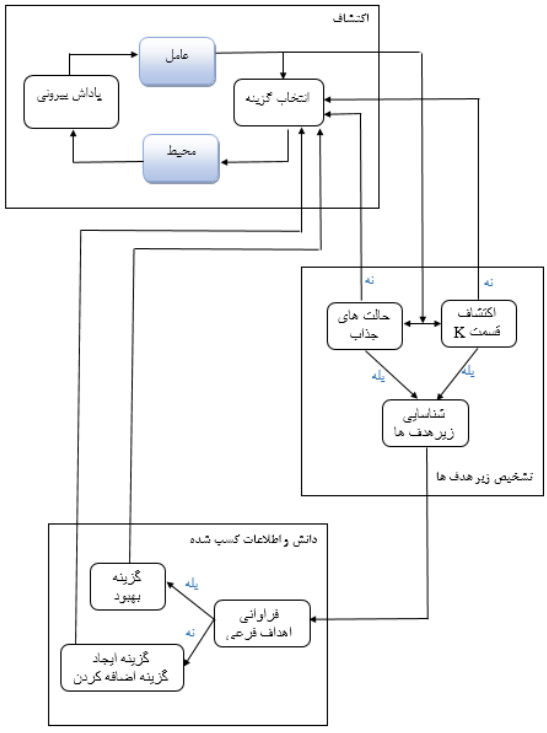
شکل ۲، مدل ارائه شده در این بررسی، فرآیند یادگیری را در یک عامل مصنوعی که برخی از توانایی های یادگیری را در انسان شبیه سازی می کند، نشان می دهد. در این مدل فرآیند یادگیری به دو مرحله تقسیم می شود: مرحله تکامل و مرحله حل وظیفه

ورودی. در مرحله اول، فرض می شود که قبل از مواجهه عامل با یک وظیفه ورودی، یک مرحله تکامل وجود دارد. در این مرحله، از یادگیری اجتماعی برای انتخاب کنش ها استفاده می شود و عامل دانش و اطلاعاتی را می آموزد که می تواند در تصمیم گیری در آینده کمک کند. در مرحله دوم، عامل بر اساس وظیفه محول شده، اطلاعات و دیدگاه های مغرضانه را ارزیابی می کند و مواردی را که برای حل مسئله و تصمیم گیری مفید است انتخاب می کند. بنابراین مرحله اول مستقل از تکلیف است و مرحله بعدی به آن بستگی دارد. دانش و بینش به‌دست‌آمده در مرحله اول را می‌توان در میان سایر وظایف با ماهیت مشابه (اما با عملکرد پاداش خارجی متفاوت) به اشتراک گذاشت.

شکل ۱ در قسمت اول، معماری مرحله افزایش را در سه بخش نشان می‌دهد: ۱) اکتشاف محیط، ۲) تشخیص زیرهدف ها و ۳) اتخاذ تصمیم ها. با کمک یک معیار یادگیری تقویت شده، عامل محیط را کاوش می کند و به دنبال موقعیت های جالب می گردد. سپس اهداف فرعی را بر اساس جذابیت آنها مشخص می کند و اطلاعات لازم برای رسیدن به اهداف فرعی را به دست می آورد. در این مقاله، رابطه بین عامل و محیط در طول مرحله تکامل با یک مدل فرآیند تصمیم گیری مارکوف [۴]مدل شد. عامل برای انتخاب بین اقدامات و تعیین اهداف فرعی از انگیزه های مختلفی استفاده می کند. بر اساس فاکتورهای یادگیری اجتماعی که برای تشخیص هدف فرعی استفاده می‌شود، کنترل موقعیت‌ها برای یافتن موقعیت‌های جذاب می‌تواند پس از تعامل هر عامل با گروه انجام شود، یا می‌توان آن را تا انجام چندین دور کاوش به تأخیر انداخت. هنگامی که یک هدف فرعی شناسایی شد، اطلاعات مربوطه باید به دست آید تا عامل را در دستیابی به آن هدف راهنمایی کند. محدوده و خط مشی گزینه ها باید در این مرحله مشخص شود، زیرا دانش و اطلاعات به دست آمده از گروه با کادر گزینه نشان داده می شود. اگر هدف خارجی بر عامل تحمیل شود، در مرحله افزایش باقی می ماند، محیط را کاوش می کند و دانش و بینش جدیدی از اعضای گروه به دست می آورد. پس از تحمیل یک وظیفه به عامل ، عامل وارد مرحله حل وظیفه ورودی[۵] می‌شود. در این مرحله عامل باید سیاست بهینه برای حل مساله مطرح شده را یاد بگیرد. در این مرحله، عامل می تواند از دانش و اطلاعاتی که در طول مرحله توسعه کسب کرده است، استفاده کند.



۱) مرحله افزایش



۲) مرحله حل وظیفه ورودی

شکل ۲) مدل تحقیق

مرحله افزایش:

در طول مرحله افزایش، عامل می تواند بدون نیاز به حل تکلیف ورودی، کاملاً بر روی تصمیم گیری تمرکز کند و مجموعه ای از قابلیت ها را به دست آورد که می تواند برای حل وظایف مختلف مورد استفاده قرار گیرد. مرحله افزایش نشان داده شده در شکل به سه قسمت اکتشاف محیط زیست، استخراج اهداف فرعی و

[۴] MDP

[۵] Solving External Task

کسب دانش و اطلاعات تقسیم می شود که در این قسمت به آنها خواهیم پرداخت.

اکتشاف محیط:

بدلیل عدم وجود پاداش خارجی برای راهنمایی عامل در مرحله افزایش، لازم است از یک مکانیسم داخلی برای هدایت رفتار آن استفاده شود. در این مقاله، مکانیزم یادگیری اجتماعی مبتنی بر کنجکاوی و تمایل به کسب اطلاعات برای تصمیم گیری ارائه شده است. در این ساز و کار، عامل یک جریمه بعد از انجام هر عمل دریافت می‌کند و به این صورت تشویق می شود به مکان هایی برود که تا کنون نرفته است. ، برای راهنمایی عامل در مرحله اکتشاف، یک MDP[۶] و تابع پاداش مبتنی بر انگیزه کسب اطلاعات در راستای تصمیم گیری تعریف می شود و یک تابع ارزش-کنش Q^I بعد از انجام هر عمل با رابطه زیر به روز می‌شود:

$$Q^I_{k+1}(s, a) = (1 \in As" - \alpha_I)Q^I_k(s, a) + \alpha_I[r_i + \gamma_I \max Q^I_k(s', a')]_a \quad (۱-۳)$$

که α_I نرخ یادگیری، $0 < \gamma_I < 1$ فاکتور تخفیف، Q^I_k تابع ارزش-کنش در گام k و r_i جریمه است. عامل، اعمال بعدی را براساس سیاست اپسیلون- حریصانه انتخاب می‌کند و به این صورت زیرهدف ها را بعد از هر تراکنش یا بعد از چند مرحله تراکنش با محیط پیدا می کند[۳۱].

استخراج زیرهدف ها

در بسیاری از مطالعات قبلی در مورد اهداف فرعی، پس از اینکه عوامل محیط را برای چند مرحله پردازش کرده و اطلاعات مربوط به محیط را جمع آوری کردند، از یک الگوریتم هدف گذاری فرعی استفاده می شود و اطلاعات و دانش لازم برای دستیابی به این اهداف فرعی به دست می آید. در این مقاله، اهداف فرعی با یک روش واحد و تنها در زمان معین شناسایی می‌شوند، اما همان طور که قبلاً گفته شد، افراد می‌توانند همزمان اطلاعات و دانش مورد نیاز برای تصمیم‌گیری‌های مختلف را بیاموزند. این رفتار را می توان در دستور کار مصنوعی تقلید کرد. هر روش تشخیص هدف فرعی می تواند اهداف فرعی با ویژگی های خاص را شناسایی کند و ممکن است همه اهداف فرعی مفید را شناسایی نکند. استفاده از دو یا چند روش تشخیص هدف فرعی به عامل کمک می کند تا اهداف فرعی با ویژگی های مختلف را شناسایی کند و فرآیند را سرعت می بخشد. دانش و اطلاعات به دست آمده می تواند به تصمیم گیری های پیچیده تر کمک کند و از مشاهدات جدید می توان برای بهبود تصمیم گیری استفاده نمود. به این صورت عامل قادر می باشد اطلاعات را به سرعت وارد فرآیند تصمیم گیری کند و آن را با اطلاعات جدید توسعه دهد. در اینجا، از سه مفهوم برای تشخیص زیرهدف ها در یادگیری اجتماعی استفاده می شود:۱) جدید بودن ۲) انگیزه دادن ۳) تقلید[۱۲]، که در ادامه این بخش به شرح آنها می‌پردازیم.

۱) جدید بودن

پیش از این جدید بودن با معیارهای مختلفی در یادگیری اجتماعی نشان داده شده است. مشابه مفهوم تازگی و تازگی نسبی (BN) که توسط شیمشک ارائه شده است، جدید بودن را با تعداد دفعاتی که یک موقعیت دیده می شود، و تازگی نسبی را با تازگی موقعیت های قبل و بعد از آن نشان می دهیم. همچنین، تعریف شیمشک از تازگی نسبی را با در نظر گرفتن فاصله هر حالت قبلی یا بعدی تا حالت مورد نظر اصلاح کردیم و آن را با معادله زیر تعریف کردیم[۳۷]:

$$BNN_{s_t} = \frac{(N_{s_t} + d\,N_{s_{t+1}} + d^2 N_{s_{t+2}} + \cdots + d^{l-1}N_{s_{t+l-1}})^{-2}}{(N_{s_{t-1}} + d\,N_{s_{t-2}} + d^2 N_{s_{t-3}} + \cdots + d^{l-1}N_{s_{t-l}})^{-2}}$$

۲) انگیزه دادن

یکی دیگر از فاکتورهای یادگیری اجتماعی که برای تعیین اهداف فرعی استفاده می کنیم، مفهوم لذت است که به عنوان دلیل کشف و کسب دانش انسان در نظریه رشد کودک پیاژه ارائه شده است. مفهوم مرکزیت بینابینی که توسط newsman معرفی شده است می تواند نشان دهنده مفهوم یادگیری اجتماعی باشد. مرکزیت بینابینی گره v با رابطه زیر نشان داده می‌شود[۴۱]:

$$BetweennessCentrality(v) = \sum_{s \neq t \neq v} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

۳) تقلید

مردم عاشق تقلید هستند و این یکی از دلایل تداوم هنجارهای اجتماعی است. انجام کارها به طور همزمان با افراد دیگر، مانند تماشای یک رویداد ورزشی یا آواز خواندن، حس مثبت و خوبی به افراد می دهد. اگر چند دقیقه با کسی صحبت کنید، سرعت صحبت کردن شما به تدریج با او یکی می شود و بدن شما حالت بدن طرف مقابل را می گیرد. انطباق را می توان برای جلوگیری از رد، توجه یا تایید انجام داد.

[۶] Markov Decision Process

کسب دانش و اطلاعات (دریافت تجربه از افراد گروه)

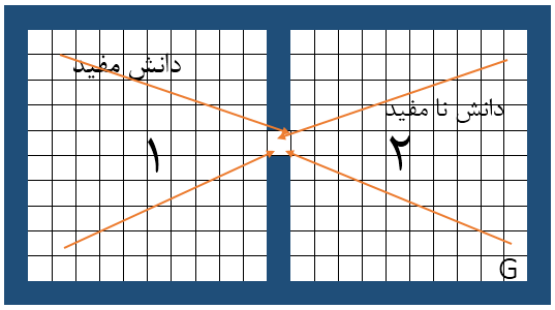
پس از تعریف هر هدف فرعی باید اطلاعات لازم برای انتقال عامل به آن هدف فرعی ایجاد شود. در این تحقیق دانش و اطلاعات با یک چارچوب انتخاب نمایش داده می شود. بنابراین باید محدوده و خط مشی هر گزینه مشخص شود. دامنه هر گزینه برابر با تعداد وضعیت هایی که قبل از زیرهدف مشاهده شده اند و کمترین فاصله را تا زیرهدف در گراف انتقال دارند[۲۸]. در محیط‌های پیوسته که از الگوریتم تشخیص ارتباط مرکز لبه برای یافتن اهداف فرعی استفاده می‌کنیم، عامل گزینه‌ای برای پیمایش از هر ارتباط به هر یک از انجمن‌های همسایه ایجاد می‌کند. خط مشی هر گزینه بر اساس نحوه نمایش و توسعه تجربه در معاملات بعدی با گروه تعیین می شود[۳۴].

مرحله حل وظیفه ورودی

در طول مرحله حل وظیفه ورودی، از اطلاعات به دست آمده در طول مرحله رشد می توان برای حل تکلیف خارجی استفاده کرد. علاوه بر این، عامل می تواند اطلاعات و دانش جدیدی در مورد وظیفه خارجی در طول مرحله حل وظیفه ورودی بیاموزد. یادگیری و کسب دانش نباید در هیچ مرحله ای از زندگی متوقف شود. بنابراین، در این مرحله عامل می تواند در یک فرآیند تصمیم گیری مارکوف جدید با انگیزه پاداش های خارجی و اهداف فرعی شرکت کند و دانش و بینش جدید استخراج شود. در این مرحله تابع ارزش-عمل در Q یاد گرفته می شود و برای ارزیابی اطلاعات به دست آمده از گروه استفاده می شود. در این بخش، چهار روش پیشنهادی برای ارزیابی دانش و اطلاعات در این بررسی ارائه شده است.

ارزیابی اطلاعات و دانش کسب شده از افراد گروه

در حالی که دانش و تجربه اعضای گروه گاهی سرعت تصمیم گیری را تکامل می دهد، گاهی نتیجه عکس دارد. به عنوان مثال، دو نوع اطلاعات در محیط دو اتاقه نشان داده شده در شکل زیر تعریف شده است. اطلاعات ۱ عامل را از اتاق سمت چپ به راهروی وسط دو اتاق هدایت می کند و اطلاعات ۲ عامل را از اتاق سمت راست به همان مکان هدایت می کند. اگر عامل از اتاق چپ شروع کند و هدف در اتاق سمت راست باشد، اطلاعات ۱ مفید و اطلاعات ۲ بی فایده و مضر است. اطلاعات ۲ عامل را از هدف منحرف می کند و سرعت تصمیم گیری را کاهش می دهد. شکل ۳-۵ تعداد تعاملات عامل با محیط برای دستیابی به هدف اصلی را در دو مورد نشان می دهد. در حالت اول، هر دو اطلاعات به مجموعه اولیه اقدامات اضافه می شوند و در حالت دوم، فقط اطلاعات خوب (مفید) اضافه می شوند. در هر دو مورد، اطلاعات پس از ۱۰ مرحله از برخورد عامل با محیط اضافه شد[۱۸].



شکل ۳) دانش مفید و دانش نامفید

این نتایج نشان داد که اخذ اطلاعات و دانش و افزودن آنها به وظیفه اصلی در تصمیم گیری می تواند پیچیدگی تصمیم گیری را تکامل دهد. بنابراین، هرگونه اطلاعات و تأثیر آن بر اثربخشی تصمیم گیری نیاز به ارزیابی دارد. انجام این کار یعنی اندازه گیری اثربخشی عامل تصمیم گیری که دانش و اطلاعات به دست آمده از اعضای گروه است ساده نیست[۱۹].

تشخیص دانش و اطلاعات نامفید دریافتی از یک گروه اجتماعی

در این قسمت روشی برای تشخیص اطلاعات زائد ارائه می دهیم که در روش اول تابع مقدار یک گزینه با تابع ارزش بهترین اقدامات مقایسه می شود. در روش دوم، تأثیر یک اختیار بر توزیع ارزش بررسی می شود.

ارزش دانش و اطلاعات دریافتی از گروه

با استفاده از سیاست فرآیند های تصمیم گیری مشابه مارکوف[۷]، در هر وضعیت عامل می‌تواند بر اساس سیاست تعیین شده از میان دانش و اطلاعات دریافتی و اعمال پایه موجود در آن وضعیت انتخاب انجام دهد. این روند در شکل ۳-۶ نمایش داده شده است. تابع ارزش-کنش بهینه برای هر عمل پایه یا دانش با استفاده از معادله بلمن برای Q^*، از رابطه زیر محاسبه می‌گردد.

$$Q^*(s,o) = R(s,o) + \sum_{s',k} P(s',k|s,o)\gamma^k max_{a'} Q^*(s',a')$$

[۷] SMDP

زمان انتظار با متغیر تصادفی k نمایش داده می شود. با شروع عمل o از وضعیت s اگر k برابر یک باشد، o یک عمل پایه ای است و در غیر اینصورت o یک دانش است.

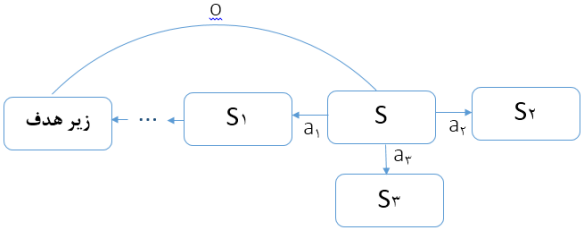
$P(s', k|s, o)$ احتمال انتقال از وضعیت s به وضعیت s' پس از k گام پس از اجرای عمل o است. $\gamma \in [0,1]$ نرخ تخفیف است. $Q^*(s, o)$ مقدار تابع ارزش عمل بهینه برای وضعیت s و عمل o می‌باشد.

$R(s, o)$ پاداش تخفیف یافته مورد انتظار عامل در زمان اجرای عمل o در وضعیت s است و از رابطه زیر بدست می آید:

رابطه(۸-۳)

$$R(s,o) = E[r_1 + \gamma r_2 + \cdots + \gamma^{k-1} r_k]$$

r_1 پاداش بدست آمده در گام اول اجرا،r_2 پاداش بدست آمده در گام ۲ و r_k پاداش بدست آمده در گام k است.



شکل ۴) اطلاعات و کنش های پایه ی قابل اجرا در یک حالت

در هر مورد از دامنه گزینه، اگر ارزش گزینه از همه اقدامات اساسی بیشتر باشد، آن گزینه با ارزش تلقی می شود. در این حالت، معیار مورد نظر برای سودمندی گزینه o در دامنه گزینه های DO معادل است با:

$$\forall s \in D_0, a \in A_s \; Q^*(s,o) \geq Q^*(s,a)$$

اگر رابطه بالا برای یک گزینه درست نباشد، این گزینه مورد علاقه یک سیاست حریصانه نخواهد بود، به این معنی که بی فایده خواهد بود. از آنجایی که سعی بر ارزیابی گزینه ها در حین یادگیری وظیفه فعلی است، تابع مقدار عمل بهینه تخمین زده می شود و ممکن است مقدار واقعی آن را هنوز یاد نگیرد؛ بنابراین،نمی توان انتظار داشت این معادله برای همه موارد حوزه صادق باشد. پس از افزودن اطلاعات و دانش به انتخاب های عامل، به خصوص در عملیات یادگیری اولیه نمی توان چنین انتظاری داشت[۲۶].

ارائه نتایج و پیشنهادات جهت ارزیابی روش های پیشنهادی

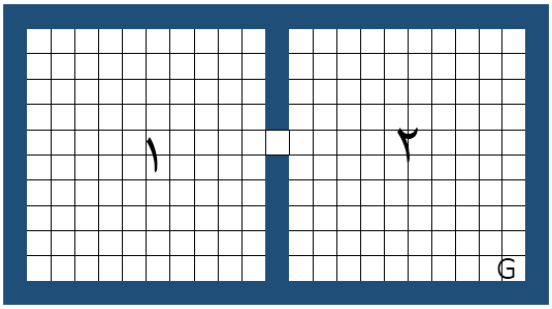
این بخش به بررسی نتایج آزمون های انجام شده برای ارزیابی روش های پیشنهادی و تأثیر آنها بر تصمیم گیری می پردازد. در بخش اول، محیط های آزمایشگاهی که در آن آزمایش ها انجام می شود، معرفی می شود. محیط های آزمایشگاهی منتخب در معروف ترین آثار این حوزه مورد استفاده قرار گرفت. در قسمت‌های بعدی این بخش، روش‌های کسب دانش و اطلاعات و اثربخشی روش‌های ارزیابی این اطلاعات، تأثیر استفاده از آن ها در تصمیم‌گیری و نتایج آزمون‌های انجام‌شده به منظور مقایسه روش پیشنهادی با مطالعات قبلی ارزیابی خواهد شد.

۴-۱ محیط‌های آزمایشگاهی

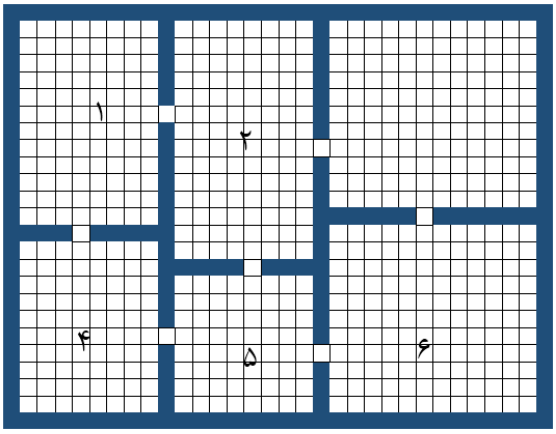
به منظور ارزیابی الگوریتم های تصمیم گیری اجتماعی در یادگیری تقویتی، مجموعه ای از تنظیمات تجربی در تحقیقات قبلی معرفی شده است. محیط هایی که نتایج تجربی در این مقاله در آنها نشان داده خواهد شد به شرح زیر است: محیط اتاق های مشبک، محیط اتاق بازی و محیط **MDP** های تصادفی ایجاد شده. محیط های نامبرده قابل مشاهده، غیرقطعی و ایستا هستند[۸]. عدم قطعیت به این معنی است که انجام یک عمل مشابه در یک موقعیت در دو زمان مختلف می تواند به موقعیت های مختلف و پاداش های متفاوت منجر شود. مشاهده پذیری به این معنی است که عامل می تواند موقعیتی را که در آن قرار دارد مشاهده کند و به آن محیط ایستا گفته می شود. در محیط ایستا مدل رابطه بین عامل و محیط را می توان با فرآیند مارکوف شبیه سازی کرد. این محیط ها در زیر توضیح داده شده است.

۱) محیط مشبک

محیط مشبک دو اتاقه و شش اتاقه که در شکل ۳ و ۴ نشان داده شده است، توسط **saten** برای ارزیابی الگوریتم های یادگیری تقویتی ارائه شد. محیط های مشبک از تعدادی اتاق تشکیل شده است که هر اتاق دارای یک کف از کاشی است[۴۲]. بین برخی اتاق ها دربی وجود دارد. هر یک از کاشی های این اتاق ها یک حالت محسوب می شود. عامل در یکی از حالت های انتخاب شده تصادفی از این محیط ها قرار می گیرد و سپس از وی خواسته می شود به خانه مورد نظر که یکی دیگر از حالت های این محیط ها است برسد. عامل می تواند چهار کنش پایه‌ای "حرکت به شمال"، "حرکت به جنوب"، " حرکت به شرق" و "حرکت به غرب" را انجام دهد. در نمودار مربوط به این محیط، هر خانه یا حالت نشان دهنده گره ای است که توسط یک یال به خانه های همسایه متصل است[۳۰].



شکل ۳) محیط مشبک دو اتاقه



شکل ۴) محیط مشبک شش اتاقه

۲) محیط اتاق بازی

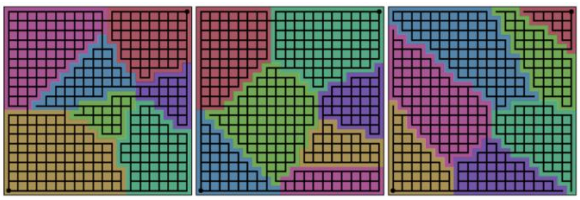
در محیط اتاق بازی که در شکل۵ نشان داده شده است‌یک عامل با اشیاء اطراف خود در تعامل است. اشیاء این محیط عبارتند از: کلید برق ، توپ ، زنگ ، دو کلید برای روشن و خاموش کردن موسیقی، یک میمون. عامل نیز دارای چشم، دست و ماژیک است. اگر هر دو چشم به جسمی نگاه می کنند و دست روی آن است، می توان عملیات مخصوص آن شی را انجام داد. مثلاً اگر چشم و دست روی کلید پاور متمرکز باشد، می توان عمل کلید پاور را انجام داد و اگر چشم و دست روی توپ متمرکز باشد، عمل ضربه زدن به توپ را می توان انجام داد. قبل از روشن کردن موسیقی باید چراغ روشن شود. اگر توپ به زنگ برخورد کند، زنگ شروع به زدن می کند.در صورت نواخته شدن موزیک، فشار زنگ در به طور همزمان و تاریکی اتاق میمون گریه می کند و اگر موسیقی خاموش باشد میمون ساکت خواهد بود. وظیفه عامل این است که میمون را به گریه بیاندازد و سپس میمون را ساکت کند[۳۳].

مجموعه ای از متغیرهای حالت برای نمایش این محیط استفاده می شود که می توان در منابع به آنها دسترسی پیدا کرد.

شکل ۵) محیط اتاق بازی

محیط فرآیندهای تصمیم گیری مارکوف تصادفی

این محیط آزمایشی توسط متزن به کار گرفته شده است. فرآیند های تصمیم گیری مارکوف به صورت تصادفی ایجاد می‌شوند[۴۱]. فضای حالت همه فرآیندها زیرمجموعه‌ای از محیط مشبک دو بعدی است که شامل ۴۰۰ حالت $(S=\{۰,۱,\ldots,۱۹\}^۲)$ و فضای کنش‌ها شامل چهار کنش گسسته $(A=\{(-۰,۱),(۰,۱),(۰,-۱),(۱,۰)\})$ است.نحوه ایجاد هر فرآیند به صورت زیر است: ۷ حالت به صورت تصادفی از مجموعه **S** انتخاب می‌گردند. هر وضعیت دیگر متعلق به **S** به گروه نزدیک ترین مرکز آن اختصاص داده می شود. یک یال گراف بین هر جفت حالت همسایه اضافه می شود (همسایگی-۴). یک یال اضافی دیگر به طور تصادفی بین دو گره همسایه در یک رابطه همسایه اضافه می شود. در شکل ۴-۴ سه نمونه از این نمودارها نشان داده شده است. پریودها از حالت اولیه $s = (۰,۰)$ شروع شده و در $s_t = (۱۹,۱۹)$ پایان می‌یابند. یک پاداش ۱- به هر وضعیت منتسب می شود. برای هر زوج وضعیت/عمل **s** و **a**، وضعیت بعدی به صورت قطعی برابر با $d(s,a) = s+a$ می‌باشد اگر $(s, s+a)$ یک یال باشد. در غیر این صورت این مقدار برابر با $d(s,a)=s$ است. مشابه پژوهش **Metzen**، $P_a^{ss'}=1-(8/9)X$ برای $s'=d(s,a)$ و $P_a^{ss'}=(1/9)X$ برای هر حالت دیگر s' در بین ۹ همسایه خود قرار دارد.



شکل ۶) سه مثال از ۵۰ گراف متصل ام دی پی های تصادفی ایجاد شده

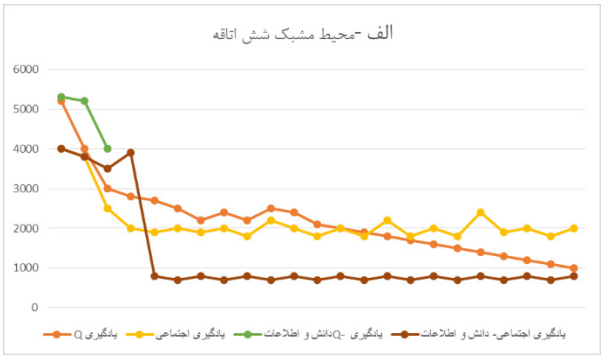
نتایج آزمایشات انجام شده و محاسبات انجام شده برای نشان دادن اثر روش های ارائه شده بر فرآیند تصمیم گیری

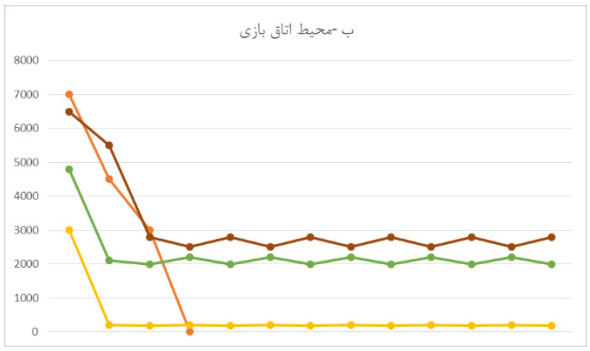
در این بخش، برخی از نتایج تجربی به دست آمده برای نشان دادن مزایای روش های ارائه شده در این مقاله ارائه شده است. در بیشتر مطالعات مرتبط، نرخ همگرایی به خط مشی بهینه یا تقریباً بهینه اغلب برای اندازه گیری اثربخشی یک الگوریتم یادگیری تقویتی استفاده می شود[۳۵]. برای این منظور تعداد اقدامات انجام شده توسط عامل برای رسیدن به هدف مشخص می شود. از منظر دیگر برای اندازه گیری میزان همگرایی، پاداش دریافت شده توسط عامل پس از چندین مرحله یادگیری محاسبه می شود. بنابراین، ما از این معیارها برای مقایسه نرخ یادگیری الگوریتم های مختلف استفاده می کنیم.

در شکل۷ تعداد گام های لازم برای ۴ عامل به منظور رسیدن به هدف در محیط های شش اتاقه و اتاق بازی نشان داده شده است. عامل اول از از هیچ مکانیزم یادگیری اجتماعی استفاده نمی‌کندو از یادگیری-Q استفاده می‌کند. عامل دوم از هیچ دانش و اطلاعاتی استفاده نمی کند بلکه از عناصر برای یادگیری اجتماعی بهره مند می شود که در آن اعمال براساس سیاست - εحریصانه تابع ارزشی انتخاب می‌گردند. عامل های سوم و چهارم از دانش و اطلاعات دریافتی در طی ۶ پریود اکتشاف محیطی در طول مرحله تکامل استفاده می کنند. در مراحل های اکتشاف، تفاوت عملکرد عامل سوم و چهارم در انتخاب تابع ارزش آنهاست. به این ترتیب که، عامل سوم اعمال را براساس سیاست - εحریصانه تابع ارزشی به روز رسانی شده با پاداش خارجی انتخاب نموده و عامل چهارم اعمال را بر اساس سیاست - εحریصانه تابع ارزشی که با جریمه به روز رسانی شده انتخاب می‌کند. مقادیر اولیه Q صفر، نرخ یادگیری (α)، فاکتور تخفیف (γ) و مقدار ε برابر با ۰/۱، ۰/۹ و ۰/۰۵ قرار داده شده است. هر منحنی یادگیری از میانگین ۵۰ اجرای مستقل به دست آمده است.

نتایج بررسی ها نشان می‌دهد که در مرحله های اولیه یادگیری، استفاده ازروش یادگیری پیشنهادی برای انتخاب اعمال نسبت به یادگیری-Q سبب می شود عامل زودتر به هدف دست یابد. همچنین، استفاده از دانش و اطلاعات دریافتی از محیط در مرحله تکامل، بازدهی تصمیم گیری را بیشتر می کند. همچنین نتایج استفاده از دانش و اطلاعات به دست آمده پس از کاوش در محیط را با جریمه و دانش و اطلاعات به دست آمده پس از کاوش در محیط را با پاداش خارجی مقایسه کردیم. بنابراین بهتر است در مراحل اولیه شناخت محیط از مکانیسم یادگیری اجتماعی استفاده کرد و بعداً به دانش و اطلاعات رسید که نتایج در محیط های شش اتاق و بازی نشان می دهد. شکل ۸ زمان اعمال عوامل آزمون در هر مرحله را نشان می دهد.

شکل ۹ تعداد گام های انجام شده برای رسیدن به هدف در مرحله حل وظیفه ورودی[۸] توسط سه عامل در محیط‌های شش اتاقه و بازی را نشان می‌دهد. همه عوامل در مرحله افزایش ۶ مرحله را طی می کنند و در این مرحله ها دانش و تجربه کسب می کنند. عامل اول از تمام دانش و اطلاعات کسب شده در طول مرحله تکامل استفاده می کند؛ عامل دوم از یک روش برای از بین بردن اطلاعات و دانش نادرست[۹] و عامل سوم از روش دیگری برای حذف اطلاعات اضافی[۱۰] استفاده می کند. همان طور که در این شکل مشاهده می شود، پایش دانش و اطلاعات و انتخاب موارد مناسب برای حل مسئله و تصمیم گیری می تواند کارایی تصمیم گیری را تکامل دهد.



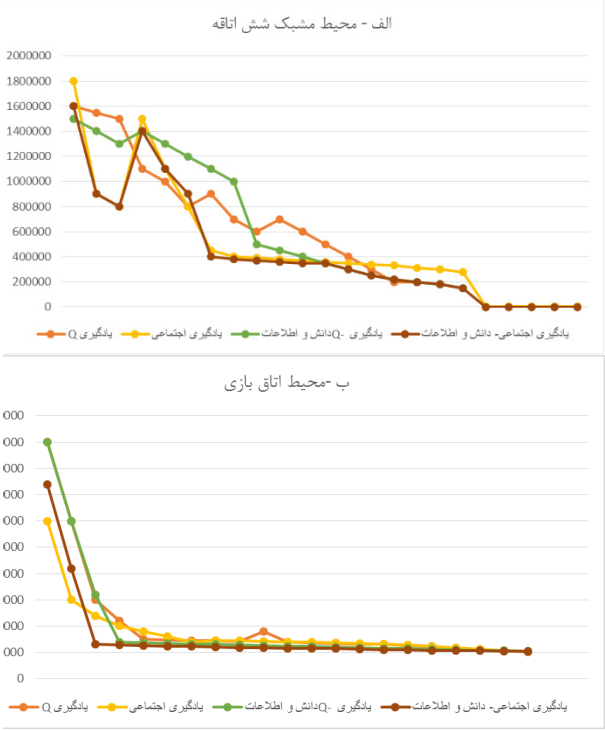


شکل ۷) مقایسه چهار عامل که اولی از یادگیری-Q، دومی از پارامترهای یادگیری اجتماعی و کنجکاوی با کمک جریمه، سومی از دانش و اطلاعات کسب شده در زمان اکتشاف با یادگیری-Q و چهارمی از یادگیری اجتماعی و دانش و اطلاعات کسب شده در زمان اکتشاف با کنجکاوی استفاده می‌کنند. آزمایش ها در محیط شبکه شش اتاقه و محیط اتاق بازی انجام شد و میانگین و انحراف معیار از میانگین در ۵۰ تکرار نشان داده شد.

[۸] SET

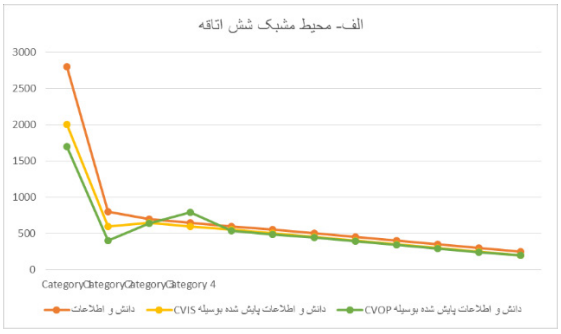
[۹] CVIS

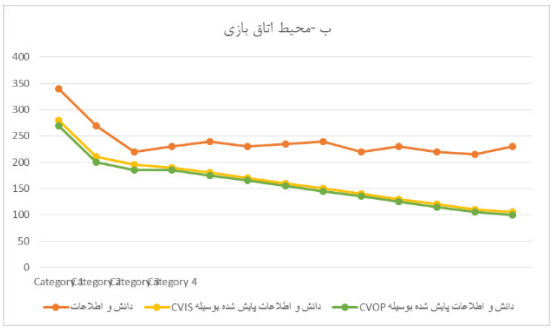
[۱۰] CVOP



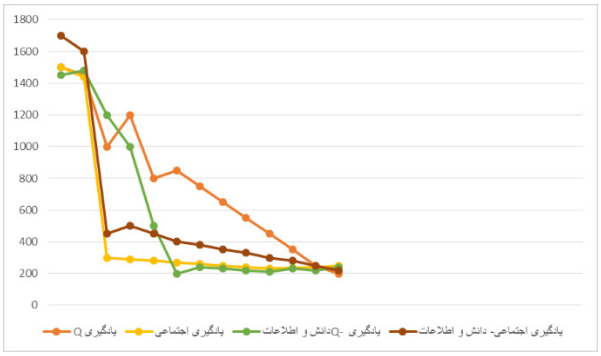
شکل۸) مقایسه زمان اجرای چهار عامل با روشهای یادگیری نامبرده

در آزمون بعدی دانش و قابلیت های اطلاعات دریافتی به مقایسه روش های ارائه شده برای ارزیابی اطلاعات و دانش به دست آمده پرداخته شد. شکل های ۱۰ نمایانگر نتایج در محیط مشبک دو اتاقه و شش اتاقه می باشند. عامل اول از یادگیری-Q ، عامل دوم از نمایش تجربه پس از ۴ مرحله در یک محیط دو اتاقه و ۶ مرحله در یک محیط ۶ اتاقه (بدون ایجاد گزینه)، عامل سوم از همه گزینه های تولید شده و عامل چهارم از گزینه های پایش شده با یکی از روشهای پایش دانش و اطلاعات استفاده می کند.





شکل ۹ مقایسه سرعت یادگیری سه عامل در مرحله حل وظیفه ورودی[۱۱]: اولی از تمام دانش و اطلاعات استفاده می کند و بقیه از یکی از روش های ارزیابی اطلاعات و دانش در تنظیمات اتاق شش و اتاق بازی استفاده می کنند. همه عوامل ۶ مرحله را در مرحله رشد سپری می کنند، بنابراین نمودارها از ۷ مرحله شروع می شوند.



شکل ۱۰) مقایسه میان یادگیری های نامبرده در ۱۰۰ تکرار

مقایسه روش ارائه شده برای تصمیم گیری اجتماعی با روش های دیگر

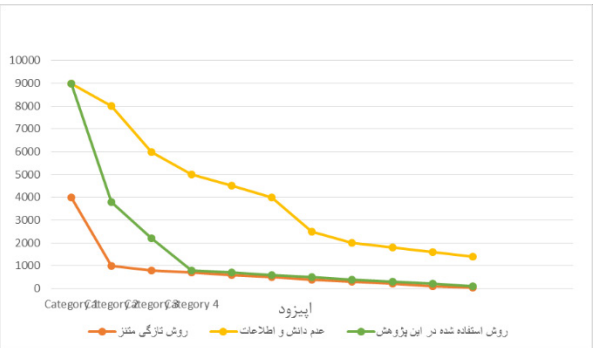
مشابه روش Barto در این رساله، از روش های یادگیری اجتماعی برای ایجاد مجموعه ای از دانش و اطلاعات قابل استفاده مجدد استفاده شده است[۶]. همچنین، از پارامترهای یادگیری اجتماعی برای کشف اهداف فرعی استفاده شده و فرض بر این است که اهداف فرعی از پیش تعیین شده اند. سایر روش های مرتبط با مطالعه عبارتند از: Bonarini و Metzen. شباهت مطالعه اول به روش ارائه شده در این بررسی این است که هر دو یک چارچوب یادگیری آبشاری ارائه می‌کنند. عامل، موقعیت های جذاب در محیط را تشخیص می دهد و دانش و اطلاعات لازم برای رسیدن به آن ها را می آموزد. تفاوت این روش با روش ارایه شده این است که الگوریتم پیشنهادی همه اطلاعات به دست آمده را ارزیابی می کند، در حالی که در این روش، اطلاعات و دانش ارزیابی نمی شوند. این دو روش در عوامل یادگیری اجتماعی مورد استفاده نیز متفاوت هستند.

روش Metzen از منظر تقسیم فرآیند یادگیری به دو مرحله مشابه روش پیشنهادی در این پایان‌نامه است: مرحله رشد که در

[۱۱] SET

آن دانش و اطلاعات از طریق مکانیسم‌های یادگیری اجتماعی کسب می‌شود و مرحله حل وظیفه ورودی. عامل به عنوان یک واسطه، دانش و اطلاعات آموخته شده را به کار می گیرد. علاوه بر این شباهت، این دو روش از بسیاری جهات با هم تفاوت دارند[۳۳]. تفاوت مهم این دو روش در این است که در پژوهش Metzen و Kirchner مکانیزم های یادگیری اجتماعی برای تعیین اینکه در یک زمان خاص بر روی کدام دانش و اطلاعات یادگیری می‌بایست تمرکز کند استفاده می‌شود؛ اما در روش ارائه شده در این پایان نامه از مکانیسم های یادگیری اجتماعی برای انتخاب کنش ها و تعیین اهداف فرعی استفاده می شود. علاوه بر این، عوامل یادگیری اجتماعی مورد استفاده در این دو مطالعه متفاوت است. همچنین، در این رساله چهار مکانیزم ارزیابی دانش و اطلاعات ارائه شده ولی در پژوهش متزن و کیچر به مسئله ارزیابی دانش و اطلاعات اخذ شده از گروه پرداخته نشده است و همه دانش و اطلاعات یادگرفته شده بدون هرگونه ارزیابی به انتخاب های عامل اضافه می‌شوند[۴۱].

شکل ۱۱ نشان دهنده هزینه ۱۲ وظیفه در محیط دوبعدی چند دره ای است که توسط روش پیشنهادی در مرحله حل وظیفه ورودی بازگردانده شده است. در این آزمایش، دانش و اطلاعات در ۵۰۰۰۰ گام مرحله تکامل بدست آمده اند. در روش پیشنهادی، دانش واطلاعات کسب شده توسط عامل به ازای هر وظیفه در طول مرحله توسعه ارزیابی شده و تنها دانش و اطلاعات مفید انتخاب می شود، اما در پژوهش متزن از تمام اطلاعات کسب شده در پژوهش Metzen و Kirchner استفاده می شود. Metzen و همکاران دو مکانیزم یادگیری اجتماعی متفاوت را استفاده کردند: روش جدید و روش خطای پیش بینی. به این ترتیب، نتایج روش جدید نشان می دهد که خروجی بهتر از روش خطای پیش‌بینی است. علاوه بر این، اثربخشی عاملی که از دانش و تجربه استفاده نمی کند و مرحله تکاملی ندارد، به این شکل قابل مشاهده است. نتایج به طور متوسط با پنجره ای به طول ۵۰ و در ۱۰ مطالعه مستقل محاسبه گردیده است.



شکل ۱۱) مقایسه میانگین هزینه وظیفه‌ها در محیط دوبعدی چند دره ای

نتیجه‌گیری

در بسیاری از تصمیماتی که توسط افراد جامعه اتخاذ می گردد گروه یا جمعیتی در یک موقعیت تصمیم قرار دارند. اغلب اطلاعات لازم برای تصمیم صحیح را هیچ کدام از افراد جمع به طور کامل در اختیار ندارند. شاید بتوان با جمع کردن اطلاعات تک تک افراد، یک مجموعه اطلاعات کامل را برای اتخاذ یک تصمیم صحیح فراهم آورد[۳۳]. لذا در این مقاله بیان می گردد که افراد نه تنها اطلاعات شخصی خود که ممکن است اصلاً صحیح هم نباشد، بلکه اطلاعاتی که از تصمیم دیگران استنتاج می کنند نیز برای تصمیم گیری مد نظر قرار می دهند. یکی از نکات نشان داده شده در این مقاله رخداد تقلید اطلاعاتی است که در آن افراد اطلاعات شخصی خود را نادیده گرفته و تحت تاثیر کامل تصمیم دیگران قرار می گیرند.

همچنین، در این مقاله به ارائه مدلی جهت یادگیری اجتماعی جهت اخذ تصمیم های بهینه متاثر از دانش، اطلاعات و تجربه اعضای گروه پرداختیم. در محیط های چندعاملی، اغلب با فضای حالت بزرگی مواجه هستیم و از طرفی دیگر اگر محیط پویا و غیرقطعی نیز باشد، یادگیری عامل ها یک مسئله چالش برانگیز خواهد شد. استفاده از راهکار انتقال و تشریک دانش، همانطور که نشان داده شده است، می تواند در راستای افزایش کارایی و تصمیم گیری موثر واقع شود برای نتیجه‌گیری، زمانی که مدل‌های یادگیری تقویتی به درستی مورد استفاده قرار گیرند، می‌توانند فرآیندهای تصمیم گیری اجتماعی را کشف کنند. هنگام به کارگیری مدل های یادگیری تقویتی در تصمیم گیری اجتماعی، به حداقل رساندن تصور نادرست و سوء تعبیر مهم است[۱۳]. این پیشنهادات به منظور کمک به مطالعات آینده در تفسیر و بازگشایی مکانیسم‌های رفتارها و تصمیم های اجتماعی است. در یادگیری تقویتی، ایجاد تعادل بین اکتشاف و بهره برداری یک چالش بزرگ است و در نتیجه برای تصمیم سازی افراد باید بر روی عامل اکتشاف در یادگیری تقویتی تمرکز نماییم و از گزینه های پاداش جهت تقویت اکتشاف استفاده نماییم. یک عامل یادگیری تقویتی باید اقداماتی را از تجربه خود انجام دهد و آن اقداماتی است که از تجربه گذشته خود گرفته و برای کسب پاداش بهینه شده است[۲۰]. در این مقاله دریافتیم که ممکن است در برخی شرایط ما دچار دام های محاسباتی در اتخاذ تصمیم های بهینه درمواجه با شرایط نامطمئن شویم. همچنین بررسی کردیم که چگونه با استفاده از یادگیری تقویتی می توانیم برای تصمیم‌گیری آگاهانه‌تر اقدام کرد. رویکردها را می توان برای درک بهتر و بر اساس کاربرد آنها طبقه بندی کرد و در نهایت، به مسائل و چالش‌های باز در مدل‌های یادگیری تقویتی فعلی برای تصمیم‌گیری در محیط های اجتماعی پرداخت.

در همین راستا در این پژوهش تعدادی محیط آزمایشگاهی معرفی گردید و پس از آن نتایج اعمال روش ارایه شده در این فضاها و تاثیر استفاده در بهبود فرآیند یادگیری و تصمیم گیری نشان داده شد. نتایج بدست آمده نشان می‌دهد که برای تکامل سرعت یادگیری مکانیزم های یادگیری اجتماعی باید در مرحله های ابتدایی شناخت محیط استفاده شود و پس از آن عامل دانش و اطلاعات را از محیط اخذ کند. بررسی تاثیر روش های پایش دانش و اطلاعات کسب شده از اعضای گروه در سرعت یادگیری، نشان می دهد که پایش دانش و اطلاعات و انتخاب دانش و اطلاعات مناسب برای تصمیم گیری می‌تواند باعث بهبود فرآیند تصمیم گیری شود.

همچنین یکی از تفاوت های این پژوهش با تحقیقات پیشین، ارزیابی و پایش دانش و اطلاعات کسب شده از افراد گروه قبل از استفاده آن ها در فرآیند تصمیم گیری بود که در هیچ کدام از کارهای پیشین اشاره ای به ان نشده بود.

سپس، مدل یادگیری اجتماعی پیشنهاد شده با تعدادی از روش های قبلی مقایسه شد و نتایج اعمال روش پیشنهادی با تحقیقات گذشته مقایسه شد. ویژگی‌های این مدل یادگیری مانند امکان استفاده از چند پارامتر یادگیری اجتماعی برای اخذ دانش و اطلاعات کسب شده در مرحله تکامل و پایش دانش و اطلاعات اخذ شده در مرحله حل وظیفه ورودی، نقطه قوت روش ارایه شده در این مقاله ایجاد می‌کند. بررسی های انجام شده نشان داد اقدامات انجام شده در این پژوهش تایید کننده برتری روش های ارائه شده نسبت به پژوهش های پیشین می باشد.

همچنین در این پژوهش مدل جدیدی برای کسب دانش و اطلاعات (تصمیم گیری اجتماعی) و استفاده از آن در یادگیری تقویتی بر اساس یادگیری اجتماعی ارائه شد که در آن فرآیند یادگیری به دو مرحله تکامل و حل وظیفه ورودی تقسیم می شود.

در طول مرحله ورودی، از عوامل یادگیری اجتماعی برای انتخاب اقدامات و کسب دانش و اطلاعات از اعضای گروه استفاده می شود. در مرحله بعدی، بر اساس وظیفه محول شده، عامل، دانش و اطلاعات قبلی را ارزیابی و پایش می کند و دانش و اطلاعات مفید را برای اخذ تصمیم انتخاب می کند. استفاده از چنین مدل یادگیری به عامل کمک می کند تا محیط را کشف کند و دانش و اطلاعاتی را کسب کند که می تواند فرد را در تصمیم گیری صحیح بهینه راهنمایی کند.

منابع

[1] Asada, M., Noda, S., Tawaratsumida, S., and Hosoda, K. (1996). “Purposive Behavior Acquisition for a Real Robot by Vision-Based Reinforcement Learning.” In: Machine Learning 23.2-3, pp. 279–303.

[2] A. Bonarini, A. Lazaric, M. Restelli, and P. Vitali, “Self-development framework for reinforcement learning agents,” in *Proceedings of the Fifth International Conference on Development and Learning, Bloomington, IN, USA*, 2006.

[3] Asadi, M. and Huber, M. (2005). “Accelerating Action Dependent Hierarchical Reinforcement Learning Through Autonomous Subgoal Discovery.” In: In Proceedings of the ICML 2005 Workshop on Rich Representations for Reinforcement Learning.

[4] Bacon, P.-L. and Precup, D. (2013). “Using label propagation for learning temporally abstract actions in reinforcement learning.” In: Proceedings of the Workshop on Multiagent Interaction Networks (MAIN 2013), held in conjunction with AAMAS 2013. Saint Paul, Minnesota, USA.

[5] Baird, L. (1995). “Residual Algorithms: Reinforcement Learning with Function Approximation.” In: In Proceedings of the 12th International Conference on Machine Learning. Morgan Kaufmann, pp. 30–37.

[6] Barto, A. G. and Mahadevan, S. (2003). “Recent Advances in Hierarchical Reinforcement Learning.” In: Discrete Event Dynamic Systems 13.4, pp. 341–379.

[7] Bieberstein, J. (2006). “Evaluierung der Dekomposition von Reinforcement Learning Problemen mittels der Subgoal-Utility-Erweiterung von Q-Learning.” de. Diploma Thesis. University of Bremen.

[8] Botvinick, M. M., Niv, Y., and Barto, A. G. (2009). “Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective.” In: Cognition 113.3, pp. 262–280.

[9] Bradtke, S. J. and Duff, M. O. (1994). “Reinforcement Learning Methods for Continuous- Time Markov Decision Problems.” In: Advances in Neural Information Processing Systems. MIT Press, pp. 393–400.

[10] Dayan, P. and Hinton, G. E. (1993). “Feudal Reinforcement Learning.” In: Advances in Neural Information Processing Systems 5, pp. 271–278.

[11] Botvinick, M. M., Niv, Y., and Barto, A. G. (2009). “Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective.” In: Cognition 113.3, pp. 262–280.

[12] Dietterich, T. G. (2000a). “An Overview of MAXQ Hierarchical Reinforcement Learning.” In: Abstraction, Reformulation, and Approximation. Lecture Notes in Computer Science 1864. Springer Berlin Heidelberg, pp. 26–44.

[13] Dietterich, T. G. (2000c). “State Abstraction in MAXQ Hierarchical Reinforcement Learning.” In:

Advances in Neural Information Processing Systems 12. MIT Press, pp. 994–1000.

[14]G. Baldassarre and M. Mirolli, *Intrinsically Motivated Learning in Natural and Artificial Systems*. Springer Berlin Heidelberg, 2013.

[15] Ghafoorian, M., Taghizadeh, N., and Beigy, H. (2013). "Automatic Abstraction in Reinforcement Learning Using Ant System Algorithm." In: 2013 AAAI Spring Symposium Series.

[16] Gullapalli, V. and Barto, A. (1992). "Shaping as a method for accelerating reinforcement learning." In: Proceedings of the 1992 IEEE International Symposium on Intelligent Control, pp. 554–559.

[17] Hengst, B. (2002). "Discovering Hierarchy in Reinforcement Learning with HEXQ." In: Proceedings of the 19th International Conference on Machine Learning. Morgan Kaufmann, pp. 243–250.

[18] I. X. Y. Leung, P. Hui, P. Lio, and J. Crowcroft, "Towards real-time community detection in large networks," *Physical Review E*, vol. 79, no. 6, 2009, p. 066107.

[19] J. H. Metzen and F. Kirchner, "Incremental learning of skill collections based on intrinsic motivation," *Frontiers in Neurorobotics*, vol. 7, no. July, 2013, pp. 1–12.

[20] J. Murata, "Controlled Use of Subgoals in Reinforcement Learning," in *Robotics, Automation and Control, Book*, October 2008, pp. 167–182.

[21] J. H. Metzen, "Learning the Structure of Continuous Markov Decision Processes," PhD thesis, Universität Bremen, 2014.

[22] Jong, N. K., Hester, T., and Stone, P. (2008). "The utility of temporal abstraction in reinforcement learning." In: Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems, pp. 299–306.

[23] Kaelbling, L. P., Littman, M. L., and Moore, A. W. (1996). "Reinforcement Learning: A Survey." In: Journal of Artificial Intelligence Research 4, pp. 237–285.

[24] Kazemitabar, S. J. and Beigy, H. (2009). "Using Strongly Connected Components as a Basis for Autonomous Skill Acquisition in Reinforcement Learning." In: Proceedings of the 6th International Symposium on Advances in Neural Networks. ISNN '09. Berlin, Heidelberg: Springer-Verlag, pp. 794–803.

[25] Kirchner, F. (1995). "Automatic Decomposition of Reinforcement Learning Tasks." In: Proceedings of the AAAI 95 Fall Symposium Series on Active Learning. Cambridge, MA, USA: AAAI Press, pp. 56–59.

[26] Kirchner, F. and Richter, C. (2000). "Q-Surfing: Exploring a World Model by Significance Values in Reinforcement Learning Tasks." In: Proceedings of the European Conference on Artificial Intelligence, pp. 311–315.

[27] Konidaris, G. and Barto, A. (2007). "Building portable options: Skill transfer in reinforcement learning." In: Proceedings of the 20th International Joint Conference on Artificial Intelligence, pp. 895–900.

[28] Konidaris, G. and Barto, A. G. (2009). "Skill Discovery in Continuous Reinforcement
Learning Domains using Skill Chaining." In: Advances in Neural Information
Processing Systems. Vol. 22, pp. 1015–1023.

[29] Konidaris, G., Kuindersma, S., Barto, A., and Grupen, R. (2010). "Constructing Skill
Trees for Reinforcement Learning Agents from Demonstration Trajectories." In:
Advances in Neural Information Processing Systems (NIPS). Vol. 23, pp. 1162–
1170.

[30] Lewis, F. L. and Vrabie, D. (2009). "Reinforcement learning and adaptive dynamic
programming for feedback control." In: IEEE Circuits and Systems Magazine 9.3,
pp. 32–50.

[31] Lopes, M., Lang, T., Toussaint, M., and Oudeyer, P.-Y. (2012). "Exploration in Modelbased
Reinforcement Learning by Empirically Estimating Learning Progress." In:
Advances in Neural Information Processing Systems (NIPS), pp. 206–214.

[32] Mannor, S., Menache, I., Hoze, A., and Klein, U. (2004). "Dynamic abstraction in
reinforcement learning via clustering." In: Proceedings of the 21st International
Conference on Machine Learning. ACM, pp. 560–567.
[40] McGovern, A. and Barto, A. G. (2001). "Automatic Discovery of Subgoals in Reinforcement
Learning using Diverse Density." In: In Proceedings of the 18th International
Conference on Machine Learning, pp. 361–368.

[33] Metzen, J. H. (2013). "Learning Graph-based Representations for Continuous Reinforcement
Learning Domains." In: Proceedings of the European Conference
on Machine Learning and Principles and Practice of Knowledge Discovery in
Databases (ECML PKDD). Ed. by H. Blockeel, K. Kersting, S. Nijssen, and F.
Zelezny. Springer Berlin Heidelberg, pp. 81–96.

[34] M.-J. Lee, S. Choi, and C.-W. Chung, "Efficient algorithms for updating betweenness centrality in fully

dynamic graphs," *Information Sciences*, vol. 326, 2016, pp. 278–296.

[35] M. E. J. Newman, "Fast algorithm for detecting community structure in networks," *Physical Review E*, vol. 69, no. 6, 2004, p. 066133.

[36] O. Simsek and A. G. Barto, "Using relative novelty to identify useful temporal abstractions in reinforcement learning," in *machine learning international workshop then conference*, 2004, vol. 21, p. 751.

[37] O. Simşek, "Behavioral building blocks for autonomous agents: description, identification, and learning," PhD Thesis, University of Massachusetts Amherst, 2008.

[38] Parr, R. and Russell, S. (1997). "Reinforcement Learning with Hierarchies of Machines."
[48] In: Advances in Neural Information Processing Systems 10. MIT Press,
pp. 1043–1049.

[39] P. Jensen, M. Morini, M. Karsai, and T. Venturini, "Detecting global bridges in networks," Journal of Complex, vol. 4.3, 2015, pp. 319–329.

[40] S. Singh, A. G. Barto, and N. Chentanez, "Intrinsically motivated reinforcement learning," *Advances in neural information processing systems*, 2005, pp. 1281–1288.

[41] R. S. Sutton, D. Precup, and S. Singh, "Intra-option learning about temporally abstract actions," in *Proceedings of the Fifteenth International Conference on Machine Learning*, 1998, pp. 556–564.

[42] R. S. Sutton and a. G. Barto, "Reinforcement Learning: An Introduction," *IEEE Transactions on Neural Networks*, vol. 9, no. 5, Sep. 1998, pp. 1054–1054.

[43] S. Singh, A. G. Barto, and N. Chentanez, "Intrinsically motivated reinforcement learning," *Advances in neural information processing systems*, 2005, pp. 1281–1288.

[44] U. N. Raghavan, R. Albert, and S. Kumara, "Near linear time algorithm to detect community structures in large-scale networks," *Physical Review E*, vol. 76, no. 3, 2007, p. 036106.

[45] مرادی، استخراج خودکار مهارت در یادگیری تقویتی با استفاده از عاملهای خودمختار، پایان نامه دکتری، دانشکده علوم ریاضی و کامپیوتر، دانشگاه صنعتی امیرکبیر،1389

Modeling social decision making using reinforcement learning

Abstract

In recent years, there have been a significant increase in the use of reinforcement learning (RL) models in cognitive science. However, the increased use of relatively complex computational approaches has led to potential misconceptions and misinterpretations. Here, we present a comprehensive framework for investigating social decision-making using a reinforcement learning approach.
We also monitor the knowledge and information received from the group and provide practical suggestions. A review of the studies conducted on these topics shows that decision-making is a time-consuming process and a person cannot make more than one decision at a time. Emotions are an essential component in regulating interactions between environmental conditions and the human decision-making process. Emotional systems provide valuable implicit and explicit knowledge for quick and rational decision making. Finally, decisions are cognitive in nature, and therefore the findings of cognitive science can help strengthen decision-making theories for a more complete understanding of people's choice process and provide more complete and realistic views of decision-making. In this research, the functional role of reinforcement learning in social decision-making will be examined first, and then the proposed model will be presented. . Our goal in this research is to provide simple, scalable explanations and practical instructions and extract sub-goals and obtain collective information and knowledge to help social decision-making based on the reinforcement learning approach. The results of this research show that the use of several parameters of social learning to obtain knowledge and information creates significant advantages for the proposed method in this article.

Keywords:
Social decision making, computational modeling, reinforcement learning.