دانشگاه صنعتی اصفهان

دانشکده مهندسی برق و کامپیوتر

تسریع یادگیری مشارکتی در سیستم‌های چند عاملی با بهره گیری از کوتاهترین مسیر تجربه شده

پایان‌نامه کارشناسی ارشد مهندسی کامپیوتر-هوش مصنوعی و رباتیک

محمدعلی میرزایی بادیزی

استاد راهنما:

دکتر مازیار پالهنگ

۱۳۹۵

دانشگاه صنعتی اصفهان

دانشکده مهندسی برق و کامپیوتر

پایان‌نامه کارشناسی ارشد مهندسی کامپیوتر-هوش مصنوعی محمدعلی میرزایی

تحت عنوان

تسریع یادگیری مشارکتی در سیستم‌های چند عاملی بهره گیری از کوتاهترین مسیر تجربه شده

در تاریخ ۱۳۹۴/۲/۱ توسط کمیته تخصصی زیر موردبررسی و تصویب نهایی قرار گرفت:

۱- استاد راهنمای پایان‌نامه — دکتر مازیار پالهنگ

۲- استاد داور (اختیاری) — دکتر مهران صفایانی

۳- سرپرست تحصیلات تکمیلی دانشکده — دکتر محمد رضا تابان

شکر شایان نثار **ایزد منان** که توفیق را رفیق راهم ساخت تا این پایان‌نامه را به پایان برسانم . از استاد فاضل و اندیشمند **جناب آقای دکتر مازیار پالهنگ** به‌عنوان استاد راهنما که همواره نگارنده را مورد لطف و محبت خود قرار داده‌اند و **خانواده** دلسوز و مهربانم که داشته‌هایم را مدیون آن‌هایم ،کمال تشکر را دارم .

کلیه حقوق مادی مترتب بر نتایج مطالعات، ابتکارات و نوآوری‌های ناشی از تحقیق موضوع این پایان‌نامه متعلق به دانشگاه صنعتی اصفهان است.

تقدیم به ساحت مقدس آخرین ذخیره الهی،

حضرت ولی عصرارواحنا فداه

فهرست مطالب

عنوان **صفحه**

فهرست تصاویر

فهرست جداول

چکیده:

می‌توان گفت بشر از ساخت اولین سیستم کامپیوتری، به دنبال سیستم‌های هوشمند بود. این رویا با فعالیت در شاخه هوش مصنوعی روزبه‌روز به واقعیت نزدیک‌تر شده است. اصلی‌ترین معیار هوشمندی قابلیت یادگیری است که بر همین اساس در هوش مصنوعی زیرشاخه یادگیری ماشین پدید آمد و روزبه‌روز بیشتر موردتوجه قرار گرفت. بعدها با ترکیب یادگیری ماشین با سیستم‌های توزیع‌شده، یادگیری در سیستم‌های چند عاملی باهدف افزایش سرعت و کیفیت مورد بررسی قرار گرفت. یادگیری در سیستم‌های چندعاملی می‌تواند به‌صورت رقابتی و یا مشارکتی باشد. در سیستم‌های چندعاملی مشارکتی عامل‌ها سعی دارند با همکاری پاداش گروهی خود را افزایش دهند. بر خلاف آن، در سیستم‌های رقابتی عامل‌های خودخواه در تلاش برای افزایش سود فردی خود بوده که این افزایش ممکن است به قیمت کاهش سود دیگران باشد.

ترکیب اطلاعات را می‌توان بزرگ‌ترین چالش در روش‌های یادگیری مشارکتی دانست که از یادگیری تقویتی استفاده می‌نمایند. در پژوهش پیش رو باهدف بهبود یادگیری مشارکتی در سیستم‌های چندعاملی روش‌های ارائه‌شده موردبررسی قرار گرفت که نتیجه این بررسی شناسایی سه نقطه بحرانی بود. اولین نقطه بحرانی که بررسی شد انتخاب عمل در یادگیری مستقل بود که با ارائهٔ معیاری به نام کوتاه‌ترین فاصلهٔ تجربه‌شده و بهره‌گیری از این معیار به‌عنوان یک مکاشفه در انتخاب عمل، منجر به بهبود یادگیری تقویتی شد. نقطه بحرانی دوم، بخش ترکیب داده‌های عامل‌ها است؛ در جهت بهبود این ترکیب داده‌ها ابتدا معیار جدیدی به نام شوک ارائه‌شده، سپس با ترکیب این معیار با معیار حداقل فاصله تجربه‌شده، ترکیب مؤثری ایجاد شده است. در آخر موضوعی که کمتر مورد بررسی قرارگرفته تقسیم‌کار بین عامل‌ها باهدف کاهش اعمال تکراری و افزایش سرعت است. انجام آزمایش‌ها نشان داد این سه عمل در کنار هم می‌تواند بهبود چشم‌گیری در یادگیری مشارکتی ایجاد نماید.

کلمات کلیدی:۱. سیستم‌های چندعاملی ۲. یادگیری مشارکتی ۳. یادگیری تقویتی

فصل ۱

مقدمه

یادگیری ماشین[1] یکی از شاخه‌های بسیار مهم و پرکاربرد در هوش مصنوعی است که سعی در ایجاد سیستم‌هایی با هوشمندی و قابلیت یادگیری دارد. معمولاً تحقیقاتی که امروزه در این زمینه صورت می‌گیرد با هدف بهبود کیفیت و افزایش سرعت یادگیری است. یکی از ایده‌های مطرح‌شده در این زمینه ترکیب یادگیری با سیستم‌های چندعاملی است. سیستم‌های چند عاملی[2] به‌تنهایی توانایی حل بسیاری از مسائل را دارند؛ ترکیب سیستم‌های چند عاملی با یادگیری ماشین می‌تواند به‌صورت مؤثری کارایی این سیستم‌ها را افزایش دهد [۱،۲]. با توجه به روابطی که می‌توان بین عامل‌ها در سیستم‌های چندعاملی تعریف کرد، دو مفهوم یادگیری رقابتی و یادگیری مشارکتی ایجاد می‌شود.

در یادگیری رقابتی عامل‌های خودخواه[3] سعی دارند تا از یادگیری برای افزایش کارایی خود بهره ببرند. گاهی این بالا بردن سود شخصی عامل‌ها به معنی کاهش سود دیگر عامل‌ها است؛ این بزرگ‌ترین دلیل رقابتی نامیدن این مبحث است. در این سیستم‌ها توجه زیادی به پویایی محیط و تعادل پایدار می‌شود. ساده‌ترین حالت از یادگیری رقابتی را می‌توان بازی‌های دونفره در نظر گرفت که عامل‌های آن قابلیت یادگیری داشته باشند.

عامل‌های یادگیری مشارکتی[4] برای رسیدن به کارایی بالا باهم مشارکت نموده و از تجربه‌ی یکدیگر برای رسیدن به‌سرعت و کیفیت بالا بهره می‌برند. در ادامه این فصل پس از تشریح دقیق مفاهیم یادگیری ماشین و سیستم‌های چندعاملی به تشریح دقیق‌تری از یادگیری مشارکتی در سیستم‌های چندعاملی خواهیم پرداخت.

---

[1] Machine learning
[2] Multi agent systems
[3] Seltish Agent
[4] Cooperative learning

۱-۱- یادگیری ماشین

تقریباً از ایجاد اولین رایانه‌ها انسان سعی داشته، بتواند با آموزش دادن، از رایانه بهره ببرد. از این رو به امید رسیدن به سیستم‌هایی که با آزمایش و کسب تجربه به هوشمندی می‌رسند تلاش‌های زیادی انجام داده است. میشل[1] در مقدمه کتاب خود [۳] تعریف ۱-۱ را برای یادگیری ماشین ارائه نموده است.

تعریف ۱-۱: زمانی گفته می‌شود که یک برنامه کامپیوتری از تجربه E[2] در مورد کار T[3] برحسب معیار کارایی P[4] یادگیری دارد که کارایی‌اش بعد از تجربه‌ی E برای کار T بهبود بیابد [۳].

امروزه می‌توان نمود این تفکر را به‌آسانی دید. میشل هم‌چنین به کارهایی که با موفقیت درزمینه یادگیری ماشین انجام‌شده اشاراتی داشته است. سیستم‌هایی مثل تشخیص صوت، تشخیص دست خط، سیستم‌هایی که می‌توانند مثل بازیکنان حرفه‌ای بازی کنند و سیستم‌های داده‌کاوی[5] که از طریق یادگیری کار می‌کنند نمونه‌هایی از کارهای موفق در زمینه یادگیری ماشین است که به‌خوبی نیز عمل می‌نمایند[۳].

اما هنوز نمی‌توان گفت که به هدف خود رسیده‌ایم زیرا سیستم‌های تهیه‌شده همه برای انجام کارهای خاص بوده و هنوز نتوانسته ایم به سیستمی برسیم که در تمام اهداف کارایی بالا داشته باشد. چالش‌های زیادی در این شاخه وجود دارد که محققان برای رفع آن‌ها تلاش می‌نمایند.

تقسیم‌بندی‌های فراوانی برای یادگیری ماشین وجود دارد که یکی از متداول‌ترین آن‌ها تقسیم‌بندی به سه دسته یادگیری با نظارت[6] و یادگیری بدون نظارت[7] و یادگیری نیمه نظارتی است. در یادگیری با نظارت یک مجموعه داده‌های برچسب دار وجود دارد و یادگیری بر اساس همین داده‌ها صورت می‌گیرد؛ اما در یادگیری بدون نظارت داده‌ها برچسب ندارند. در نتیجه خروجی این روش‌ها تا حدودی با هم متفاوت است. با توجه به یادگیری بانظارت و بدون نظارت می‌توان به ماهیت داده‌ای محیط‌های نیمه نظارت نیز پی برد. در این محیط‌ها بخشی از دادها دارای برچسب و بخش دیگر داده‌ها بدون برچسب هستند[۳]. با این حال رده‌ای از مسائل وجود دارند که خروجی مناسب که یک سیستم یادگیری تحت نظارت نیازمند آن است، برای آن‌ها موجود نیست. این نوع از مسائل چندان قابل جوابگویی با استفاده از یادگیری تحت نظارت نیستند. یادگیری تقویتی مدلی برای مسائلی از این قبیل فراهم می‌آورد. در یادگیری تقویتی، سیستم تلاش می‌کند تا تقابلات خود با یک محیط پویا را از طریق آزمون و خطا بهینه نماید. یادگیری تقویتی

---

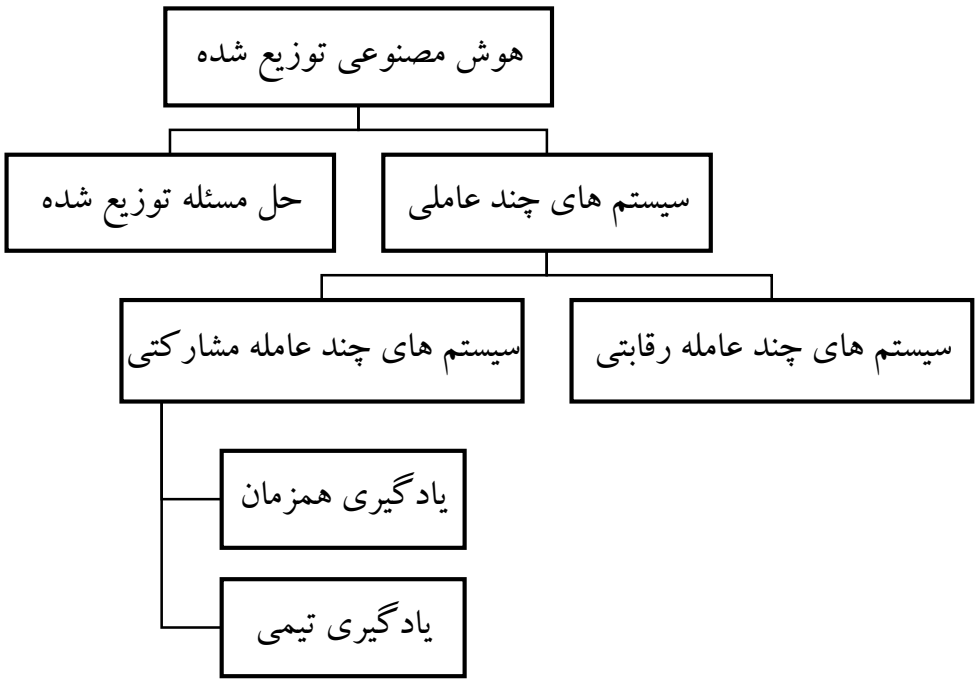
[1] Mitchell
[2] Experiece
[3] Task
[4] Performance
[5] Data mining
[6] Supervised learning
[7] Unsupervised learning

مسئله‌ای است که یک عامل که می‌بایست رفتار خود را از طریق تعاملات آزمون و خطا با یک محیط پویا فرا گیرد، با آن مواجه است. در یادگیری تقویتی هیچ نوع زوج ورودی- خروجی ارائه نمی‌شود. به جای آن، پس از اتخاذ یک عمل، حالت بعدی و پاداش بلافصل به عامل ارائه می‌شود. هدف اولیه برنامه‌ریزی عامل‌ها با استفاده از تنبیه و تشویق است بدون آنکه ذکری از چگونگی انجام وظیفه آن‌ها شود.

۱-۲- سیستم‌های چندعاملی[1]

در علوم مختلف معمولاً برای حل مشکلات از طبیعت الگوبرداری زیادی می‌شود که گاهی نیز بسیار موفقیت‌آمیز عمل می‌نمایند. از نمونه‌های این الگوبرداری می‌توان الگوریتم ژنتیک[2] و شبکه عصبی[3] را نام برد که به‌خوبی نیز برای حل مسائل عمل می‌نمایند.

یکی از دلایل پیشرفت در انسان و حیوانات زندگی گروهی آن‌ها است. می‌توان گفت سیستم‌های چندعاملی نیز الگوبرداری از همین زندگی گروهی انسان‌ها است. یکی از تقسیم‌بندی‌هایی که برای این سیستم‌ها ارائه‌شده، نشان می‌دهد که هوش مصنوعی توزیع‌شده از ترکیب سیستم‌های توزیع‌شده با هوش مصنوعی به وجود آمده‌اند و سیستم‌های چندعاملی زیرمجموعهٔ هوش مصنوعی توزیع‌شده است [۴]. در شکل ۱-۱ جایگاه سیستم های چند عاملی نسبت به هوش مصنوعی آمده است.



[1] Multi-agent Systems
[2] Genetic algorithm
[3] Neural Network

شکل ۱-۱: جایگاه یادگیری مشارکتی[۴]

سیستم‌های چندعاملی را می‌توان به دودسته سیستم‌های چندعاملی همگن و سیستم‌های چندعاملی ناهمگن تقسیم نمود. در سیستم‌های چندعاملی همگن عامل‌های یکسانی در حال فعالیت می‌باشند حال‌آنکه در سیستم‌های ناهمگن عامل‌ها از توانایی‌های متفاوتی برخوردار هستند. معمولاً به مجموعه عامل‌های سیستم‌های ناهمگن گروه نیز اطلاق می‌شود[۵].

در سیستم‌های همگن عامل‌ها می‌توانند باهدف یادگیری یک کار مستقل یا یک کار گروهی تلاش کنند. این تلاش از آنجایی ناشی می‌شود که هیچ‌یک از عامل‌ها به تنهایی قادر به انجام کار مورد نظر نیستند؛ و فقط با همکاری می‌توانند کار مورد نظر را انجام دهند. در این سیستم‌ها معمولاً سعی می‌شود از مرکزگرایی پرهیز شود چراکه ممکن است در زمان اجرای کار عامل‌هایی از دست بروند؛ اما در سیستم‌هایی که به دنبال یادگیری یک کار مستقل هستند عامل‌ها در محیط‌های مجزا گذاشته‌شده و یادگیری را انجام می‌دهند. این عامل‌ها در طول یادگیری باهم در ارتباط بوده و از اطلاعات یکدیگر برای سرعت بخشیدن به فرایند یادگیری بهره می‌برند. در پایان یادگیری باید تمام عامل‌ها به‌اندازه قابل قبولی توانایی داشته باشند. سیستم‌های چند عاملی ناهمگن در اکثر مواقع همانند دسته اول به دنبال یادگیری یک کار گروهی هستند و با همکاری سعی در رسیدن به هدف نهایی دارند.

سیستم‌های چندعاملی برای اولین بار در [۴][۶] با یادگیری ترکیب شدند که باهدف افزایش سرعت یادگیری تقویتی بودند. همان طور که گفته شد، یادگیری در سیستم‌های چندعاملی را می‌توان به‌صورت مشارکتی و رقابتی تقسیم‌بندی نمود. این تقسیم‌بندی بر اساس روابطی است که می‌توان بین آن‌ها پیدا نمود. یادگیری رقابتی به این صورت است که عامل سعی دارد با یادگیری از عامل‌های دیگر سود خود را افزایش دهد.

۱-۳- یادگیری مشارکتی در سیستم‌های چندعاملی

در علوم اجتماعی مشارکت را یک عمل جمعی برای رسیدن به سود متقابل تعریف می‌نمایند. سیستم‌های چندعاملی مشارکتی شرایطی شبیه زندگی انسان دارند. انسان‌ها پند می‌گیرد، مشورت می‌کنند، اعمال دیگران را می‌بینند و فرایند یادگیری خود را بهبود می‌بخشند. می‌توان گفت افراد مشارکت می‌کنند تا بیشتر و بهتر یادگیری نمایند.

یادگیری مشارکتی را می‌توان از جهات مختلف بررسی نمود. در یک دیدگاه کلی می‌توان یادگیری مشارکتی را به یادگیری هماهنگ و بهبود یادگیری تقسیم نمود. در یادگیری هماهنگ عامل‌ها در یک محیط قرار می‌گیرند و به دنبال روشی هستند تا به‌طور گروهی به هدف مشترکی برسند. باید بر مشترک بودن هدف در یادگیری هماهنگ تأکید نمایم[۷].

در بهبود یادگیری، عامل‌ها در محیط جداگانه قرارگرفته و یک کار یکسان را یاد می‌گیرند و با انتقال یادگیری‌های خود به یکدیگر به فرآیند یادگیری کل گروه سرعت می‌بخشند. در بهبود یادگیری ارسال اطلاعات می‌تواند به‌صورت مستقیم و یا غیرمستقیم انجام شود. اگر انتقال به‌خوبی انجام شود فرایند یادگیری بهبود خواهد یافت. عملکرد این سیستم‌ها وابسته به انتقال اطلاعات و نحوه استفاده از آن است. می‌توان چالش‌های این زمینه را به‌صورت سؤالاتی مطرح نمود:

- چه اطلاعاتی ارسال شود؟
- چه زمانی اطلاعات ارسال شود؟
- چگونه از اطلاعات ارسال‌شده استفاده نماییم؟
- اطلاعات دریافتی از کدام عامل اعتبار بیشتری دارند؟

استفاده از داده‌های تولید شده توسط چند عامل می‌تواند باعث فرار از بهینه محلی، تسریع در فرایند یادگیری و تنظیم پارامترهای یادگیری شود. در اکثر روش‌های یادگیری مشارکتی از یادگیری تقویتی[1] استفاده می‌شود. معمولاً روش‌های پیشنهادی برای یادگیری مشارکتی روش‌های ساده‌ای هستند که از علوم اجتماعی الهام گرفته‌شده‌اند روش‌هایی مثل مشورت کردن، پند دادن به عامل‌های کم‌تجربه، اندازه‌گیری خبرگی عامل‌ها و ... که با هدف بالا بردن سرعت و کیفیت یادگیری مطرح‌شده‌اند. در این روش‌ها تأکید بر نحوه ارسال داده‌ها و استفاده از آن‌هاست. روش‌هایی نیز هستند که بر اساس نظریه بازی استوار شده‌اند و از پیچیدگی بیشتری برخوردار هستند.

۱-٤- اهداف و نوآوری‌های پژوهش

هدف اصلی این پژوهش افزایش سرعت و دقت یادگیری مشارکتی تعریف شده است. در جهت میل به این هدف در این پژوهش سعی شد با بررسی دقیق روش‌های مشارکتی، نقاط بحرانی روش‌های ارائه شده برای یادگیری مشارکتی پیدا شده و تمرکز پژوهش در این نقاط قرار گیرد. بعد از بررسی‌های انجام‌شده سه نقطه اصلی شناسایی شد:

- انتخاب عمل در یادگیری مستقل
- ترکیب داده‌ها در مرحله همکاری
- تقسیم‌کار عامل‌ها در طول روال یادگیری

[1] Reinforcement learning

بهبود هر یک از این روش‌ها می‌تواند در بهبود یادگیری مشارکتی بسیار مؤثر باشد. قبل از بهبود این روش‌ها سعی شد معیارهایی برای نمایش برتری عامل‌ها نسبت به هم ارائه شود. ارائه چنین معیارهایی قبلاً هم صورت گرفته؛ روش‌هایی مثل تقلید، پند و خبرگی معیارهایی را مورد استفاده قرار داده بودند. در این پژوهش با ارائه معیار شوک و حداقل فاصله تجربه‌شده سعی شده تا سه نقطه بحرانی شناسایی‌شده، بهبود داده شود.

جهت بهبود نقطه بحرانی اول یک پارامتر اضافه تعریف شده تا بتواند نحوه تأثیرپذیری از معیار حداقل فاصله تجربه‌شده کنترل شود. این معیار بسیار وابسته به هدف یادگیری و جنس محیط یادگیری است. نقطه دوم با بهره‌گیری از هر دو معیار معرفی‌شده به شکلی بهبود داده شد که نقطه سوم را نیز پوشش خواهد داد. نهایتاً در پایان سعی شده با انجام آزمایش‌هایی عملکرد روش پیشنهادی و خصوصیات آن روشن شود. به طور واضح میتوان نوآوری های این پژوهش به صورت زیر تعریف نمود.

- ارائه معیار های جهت بررسی برتری عامل ها.
- ارائه یک تابع مکاشفه و بهره گیری از این تابع در راستای افزایش سرعت یادگیری.
- بهره گیری از معیارهای ارزیابی ارائه شده در راستای ترکیب داده ها و تقسیم کار بین عامل ها.

۱-۵- ساختار پایان‌نامه

در ادامه مروری بر روش‌های اراده شده در انتقال داده در یادگیری مشارکتی در فصل دوم صورت خواهد گرفت. در فصل سوم پیش‌نیازهای روش پیشنهادی بیان شده؛ فصل چهارم روش پیشنهادی بیان‌شده و جهت مشخص شدن مزایا و معایب روش، نتایج آزمایش‌های انجام شده آورده شده است. در فصل پنجم نیز یک جمع‌بندی و نتیجه‌گیری آورده شده و نهایتاً پیشنهادهایی برای کارهای آینده ارائه شده است.

فصل ۲

مروری بر کارهای گذشته

انسان‌ها به‌صورت منزوی زندگی نمی‌کنند چراکه اگر این‌گونه بود شاید یادگیری انسان هنوز در حد اولین انسان‌های روی زمین بود. گروهی زندگی کردن توانسته انسان را به چیزی که هست برساند. یادگیری در انسان بیشتر از آن‌که بر اساس مکاشفه باشد بر اساس انتقال اطلاعات است. یادگیری مشارکتی نیز برگرفته از همین واقعیت بوده است؛ معمولاً روش‌هایی که برای یادگیری مشارکتی ارائه می‌شود برگرفته از زندگی و روابط گروهی انسان‌ها است. کارهای انجام‌گرفته در یادگیری مشارکتی را به‌طورکلی می‌توان به دودسته تقسیم کرد. در دسته اول به روش‌های انتقال اطلاعات پرداخته‌شده و به دنبال روشی برای انتقال صحیح اطلاعات هستند. در این کارها محیط را به‌عنوان یک مدل مارکوف ساده می‌بینند. در دسته دوم نقش محیط را پررنگ‌تر می‌بینند و به‌صورت یک بازی تصادفی در نظر می‌گیرند که عامل‌ها به دنبال رسیدن به یک تعادل در بازی هستند. معمولاً این روش‌ها را بر مبنای نظریه بازی فعالیت می‌نمایند.

همان‌طور که در فصل قبل گفته شد اکثر روش‌های یادگیری مشارکتی از یادگیری تقویتی بهره می‌برند. اولین تلاش‌ها در یادگیری مشارکتی که در [۴]، [۸] آمده‌اند صرفاً برای کاهش زمان اجرای یادگیری Q با پیش قدر که یکی از روشهای یادگیری تقویتی است ارائه شدند. الگوریتم‌های یادگیری تقویتی نوعی جستجو را برای یافتن سیاست بهینه انجام می‌دهند. مکانیزم‌های مشارکتی می‌توانند به کاهش زمان جستجو در این الگوریتم‌ها کمک نمایند.

در [۸] دو مکانیزم برای یادگیری مشارکتی پیشنهادشده است. مکانیزم اول یادگیری باوجود یک نقاد خارجی است که بر اساس اعمال عامل به او پاداش یا جریمه‌ای اختصاص می‌دهد و عامل این پاداش را جهت یادگیری سیاست بهینه که بتواند مجموع پاداش‌های او را افزایش دهد، استفاده می‌نماید. در مکانیزم دوم که یادگیری بر اساس مشاهده است عامل بر اساس مشاهده عامل‌های دیگر به یادگیری می‌پردازد. در مکانیزم دوم نیازی به عامل دانشمند بیرونی نبوده و حتی در محیط‌هایی که تمام عامل‌ها خام هستند نیز می‌تواند به‌خوبی عمل نماید. پیچیدگی زمانی مکانیسم‌های پیشنهادی خطی است درصورتی که پیچیدگی یادگیری Q بدون پیش قدر نمایی به عمق تعداد حالت‌ها است. این خود نشان‌دهنده اهمیت استفاده از یادگیری مشارکتی دارد.

۲-۱- مشارکت به‌وسیله اشتراک‌گذاری

تن[1] در [۹] سه روش انتقال اطلاعات را معرفی می‌نماید. نویسنده به این موضوع می‌پردازد که آیا یک تعداد عامل در حالت یادگیری مشارکتی بهتر از یادگیری مستقل عمل می‌نمایند؟ نتیجه مهمی که در این آزمایش‌ها به‌دست‌آمده نشان می‌دهد که اگر مشارکت به‌خوبی پیاده‌سازی شود هر عامل می‌تواند از تجربیات عامل‌های دیگر استفاده بهینه نماید. در این مقاله یادگیری مشارکتی به سه رویکرد، اشتراک گذاری سیاست، اشتراک گذاری ادراک و اشتراک گذاری واقعیت‌ها تقسیم شده و عملکرد را نسبت به حالتی که عامل‌ها به‌طور مستقل کار می‌کنند می‌سنجد. تن اشتراک گذاری خود را که به‌روزرسانی جدول Q به‌وسیله میانگین‌گیری از جدول کل عامل‌ها است. معدل گیری ساده(SA[2])نامیده است.

نتایج نشان می‌دهد که اشتراک گذاری سیاست و واقعیت می‌تواند سرعت یادگیری را افزایش دهد و اشتراک گذاری ادراک نیز اگر به‌خوبی اجرا شود مفید است. همچنین نشان داده شده که برای اعمال مشترک یادگیری مشارکتی خیلی مفید است هرچند که سرعت یادگیری در شروع پائین است.

۲-۲- تقلید[3]

انسان‌ها از طریق تقلید کردن، یادگیری زیادی انجام می‌دهند و این واقعیت در کودکان به‌وضوح دیده می‌شود. کودکان با تقلید از اطرافیان، بسیاری از مسائل ابتدایی را یادگیری می‌نمایند اما این یادگیری یک‌طرفه است و تأثیری بر معلم ندارند.

با ایده گرفتن از همین موضوع می‌توان روشی برای یادگیری مشارکتی ارائه کرد. نکته مهمی که در این نوع یادگیری وجود دارد این است که عامل چه زمان و از چه کسی تقلید نماید تا یادگیری بهتر انجام گیرد. بر این اساس

---
[1] Tan
[2] Simple Averaging (SA)
[3] Imitation

یادگیری بر اساس تقلید به سه حالت تقلید ساده، تقلید شرطی و تقلید انطباقی تقسیم می‌شود. در تقلید ساده عامل‌ها از همسایگان خود یادگیری می‌نمایند. در این روش عامل‌های همسایه همیشه منتظر یکدیگر می‌مانند. همسایگی در این روش نزدیکی ازنظر فاصله محیطی است که با این توجیه که عامل‌های نزدیک در شرایط مشابهی قرار دارند و عملکرد یکسان کارایی را افزایش می‌دهد، ارائه می‌شود.

نوع دیگر تقلید، تقلید شرطی است. در تقلید شرطی مشکل انتظار به این صورت حل می‌شود که عامل‌های با کارایی پایین از عامل‌های با کارایی بالاتر تقلید می‌نمایند. محاسبه کارایی نیز بر اساس پاداش‌هایی است که عامل کسب نموده است.

تقلید انطباقی شبیه به تقلید شرطی است با این تفاوت که نرخ تقلید قابل تنظیم است. معمولاً نرخ تقلید بر اساس کارایی عامل‌های همسایه محاسبه می‌شود. با این روش وزن دهی می‌توان به یک حد بین یادگیری تقویتی و یادگیری مشارکتی رسید که می‌تواند بسیار مفید باشد؛ زیرا در زمان‌هایی از یادگیری اگر عامل به یادگیری مستقل بپردازد بعداً می‌تواند به یادگیری مشارکتی کمک شایانی نماید[۱۰].

۲-۳- حافظه جمعی[1]

در علوم اجتماعی شناخت توزیع‌شده را این‌گونه تعریف می‌نمایند که شناخت در اجتماع تنها در یک فرد صورت نمی‌گیرد بلکه بین افراد توزیع‌شده و هرکس شناخت خود را دارد. در [۱۱] با الهام از شناخت توزیع‌شده ایده حافظه جمعی مطرح می‌شود. در یک گروه از عامل‌ها شناخت عامل‌های پرتجربه می‌تواند عامل‌های خام و کم‌تجربه را برای فعالیت‌های مؤثرتر هدایت نماید. این کار برای عامل‌های پرتجربه نیز مفید خواهد بود زیرا ممکن است عامل کم‌تجربه راهی برای رد راهکار ارائه‌شده توسط عامل باتجربه پیدا نماید.

برای حل مشکلات یادگیری مشترک می‌توان از حافظه جمعی استفاده نمود. با این کار تعداد تلاش‌های عامل و تعداد انتقالات اطلاعات بین عامل‌ها کمتر می‌شود. حافظه جمعی را معمولاً با دو دیدگاه یادگیری رویه‌های مشترک و یادگیری قابلیت‌های عامل‌ها پیاده‌سازی می‌نمایند.

در یادگیری رویه‌های مشترک از حافظه جمعی صرفاً برای به یادآوردن الگوهای حل مسئله استفاده می‌شود و یادگیری قابلیت‌های عامل‌ها با تهیه یک ساختار درختی که نشان‌دهنده اعمال و احتمال موفقیت آن‌هاست به فرآیند یادگیری کمک می‌نماید. این احتمالات بر اساس تلاش عامل‌ها بروز رسانی می‌شوند.

[1] Procedural Knowledge

ایده‌ی حافظه جمعی را می‌توان به‌صورت حافظه متمرکز و یا حافظه توزیع‌شده در عامل‌ها پیاده‌سازی نمود. همچنین نکته دیگری که در مورد این ایده وجود دارد این است که حافظه جمعی را می‌توان هم‌زمان با روش‌های دیگر یادگیری مشارکتی مورداستفاده قرارداد؛ که نحوه ترکیب حافظه جمعی با روش‌های دیگر خود چالش‌هایی را ایجاد می‌نماید[۱۱].

۲-٤- پند

ایده‌ی دیگری که از علوم اجتماعی در یادگیری مشارکتی استفاده‌شده، ایده‌ی پند گرفتن است که اولین بار توسط نوییس [1] در [۱۲] آمد. معمولاً انسان در زندگی اجتماعی خود در برخورد با مشکلات از افرادی که قبلاً در شرایط مشابهی بوده‌اند پند می‌گیرد و یادگیری خود را از طریق پند گرفتن انجام می‌دهد. پیاده‌سازی این ایده در سیستم‌های چندعاملی به این صورت است که عامل‌های کم‌تجربه بعد از درک موقعیت محیط از عامل‌هایی که قبلاً در موقعیت مشابهی بوده‌اند درخواست پند می‌نماید و از پند عامل‌های باتجربه مانند برچسب‌های یادگیری با نظارت بهره می‌برد.

در پیاده‌سازی ایده پند، حتی می‌توان برای عامل‌ها از روش‌های یادگیری متفاوتی استفاده کرد. نوییس بعد از [۱۲] در [۱۳] به اصلاح ایده خود پرداخته است. بعد از آن در [۱۴] شرایطی در نظر گرفته‌شده که عامل‌ها در یک محیط یکسان تعامل دارند. نتایج حاصل نشان‌دهنده این بوده که روال پند دهی می‌تواند فرایند یادگیری را سرعت ببخشد. در توسعه پند دهی عامل‌های باتجربه، مفاهیمی چون اعتمادبه‌نفس تعریف‌شده است. در کار انجام‌شده در زمان مبادله پند یک جفت حالت و عمل به‌علاوه میانگین عملکرد عامل و بهترین عملکرد او ارسال می‌شود. عامل پند گیرنده با استفاده از این اطلاعات تصمیم می‌گیرد که آیا به پند گرفتن نیاز داشته یا خیر و اگر نیاز داشته پند ارسالی از جانب چه کسی برای او مفید است.

۲-۵- یادگیری مشارکتی مبتنی بر خبرگی

ایده بهره بردن از میزان خبرگی عامل‌ها که آن نیز برگرفته از جوامع انسانی است برای اولین بار در [۱۵] با عنوان مشارکت وزن دار سیاست (WSS [2]) با بررسی مشکلات SA مطرح شد. در SA از یک میانگین‌گیری در جداول عامل‌ها به یک جدول می‌رسیم اما بعد از گذشت مدتی از یادگیری معمولاً عامل‌ها ازنظر توانایی در یک سطح نیست و انتساب یک ضریب یکسان برای تمام عامل‌ها منطقی نیست. در WSS ابتدا معیارهایی برای اندازه‌گیری خبرگی عامل‌ها ارائه‌شده تا بتوان میانگین‌گیری را به‌صورت وزن‌دار و بر اساس خبرگی عامل‌ها از جدول Q انجام داد.

---

[1] Nunes
[2] Weighted Strategy Sharing (WSS)

فعالیت عامل‌ها در WSS به دو فاز تقسیم می‌شود در فاز اول عامل‌ها به یادگیری مستقل پرداخته و سعی در کسب تجربه دارند سپس در فاز دوم به یادگیری مشارکتی می‌پردازند و تجربیات خود را بر اساس میزان خبرگی خود به اشتراک می‌گذارند. در [۱۵] سه معیار برای اندازه‌گیری خبرگی معرفی می‌شود همچنین برای WSS پارامتری به نام بازه مشارکت تعریف می‌شود که هر چه بیشتر باشد تعداد تلاش‌ها در فاز یادگیری مستقل بیشتر خواهد بود که به این معناست که عامل تجربه بیشتری کسب خواهد کرد؛ اما تنظیم این پارامتر بسیار مهم است زیرا اگر خیلی زیاد شود یادگیری از مشارکتی به یادگیری تقویتی تبدیل می‌شود و اگر زمان خیلی کم باشد عامل فرصت کسب تجربه را پیدا نخواهد کرد.

اما WSS نیز دارای نقایصی است، ازجمله این‌که عامل‌ها علاوه بر میزان خبرگی، در محیطی که به خبرگی رسیده‌اند هم متفاوت هستند. در [۱۶] روشی برای اعمال این واقعیت ارائه می‌شود. بعد از آن نیز کارهای فراوانی در این زمینه صورت گرفته ازجمله ایده افزودن احتمال شرکت تجربیات عامل‌ها که توسط ریتپاوت [1] و همکاران در [۱۷] آمده است. این احتمال نسبت مستقیمی با تفاوت وزن عامل‌ها دارد و اگر میزان این تفاوت از حدی کمتر باشد دیگر فاز یادگیری مشارکتی انجام نمی‌شود. معیار دیگری که در [۱۷] آورده شده است پشیمانی نام دارد؛ که مقدار آن بر اساس تفاوت بین مقدار اولیه بهترین عمل و دومین بهترین عمل محاسبه می‌شود.

در [۱۵] علاوه بر میزان خبرگی ضریبی به نام ضریب تأثیرپذیری نیز در محاسبه اضافه شده است که یک عدد ثابت برای هر عامل است. بررسی‌های انجام‌شده در [۱۸] نشان می‌دهد که در صورت تنظیم صحیح، این پارامتر می‌توان تأثیر زیادی در بهتر نمودن مشارکت داشته باشد. شش معیار خبرگی که در [۱۵]آورده شده‌اند عبارت‌اند از:

- معیار خبرگی معمولی[2]: این معیار عاملی را برتر می‌شناسد که درگذشته پاداش پیروزی بیشتری داشته است.در رابطه ۲-۱ روش محاسبه این معیار آورده شده است.در این رابطه $r_i(t)$ پاداش دریافتی عامل در گام t است.

$$e_i^{Nrm} = \sum_{t=1}^{New} r_i(t) \qquad (۱-۲)$$

- معیار خبرگی مثبت[3]: این معیار بر اساس پیروزی‌ها قضاوت می‌کند و فقط پاداش‌های مثبت را شمارش می‌کند. در رابطه ۲-۲ r_i^+ پاداش‌های دریافتی مثبت عامل است.

[1] Ritthipravat
[2] Normal
[3] Positive

$$ei_i^{Pos} = \sum_{t=1}^{New} r_i^+(t)\,, \qquad r_i^+(t) = \begin{cases} 0 & if\ r_i() \le 0 \\ r_i(t) & otherwise \end{cases} \tag{۲-۲}$$

- معیار خبرگی قدر مطلق[1]: در معیار قدر مطلق به پاداش و جریمه بهاداده می‌شود. استدلالی که بر این روش وجود دارد بر اساس اثری است که جریمه بر یادگیری دارد. مسلماً همان‌طور که پاداش‌ها خانه‌هایی را که عامل باید در آن باشد را به عامل می‌آموزد جریمه هم‌خانه‌هایی را به عامل می‌آموزد که باید از آن پرهیز شود.در رابطه ۳-۲ $r_i(t)$ نشان از پاداش دریافتی عامل در گام t است.

$$e_i^{Abs} = \sum_{t=1}^{New} |r_i(t)| \tag{۳-۲}$$

- معیار خبرگی منفی[2]: این معیار برتری عامل‌ها را بر تجربیات ناموفق آن‌ها می‌داند. در رابطه ۲-۲ r_i^- پاداش‌های دریافتی منفی عامل است.

$$e_i^{Neg} = \sum_{t=1}^{New} r_i^-(t)\,, \qquad r_i^-(t) = \begin{cases} 0 & if\ r_i() > 0 \\ -r_i(t) & otherwise \end{cases} \tag{۴-۲}$$

- معیار خبرگی گرادیان: این معیار همانند معیار معمولی عمل می‌کند با این تفاوت که سنجش عامل‌ها را صرفاً بر اساس آخرین چرخه یادگیری آن‌ها محاسبه می‌نماید. C نشان‌دهنده زمان شروع آخرین چرخه یادگیری و $r_i(t)$ نشان از پاداش دریافتی عامل در گام t است.

$$e_i^{Nrm} = \sum_{t=c}^{New} r_i(t) \tag{۵-۲}$$

- معیار خبرگی متوسط تعداد قدم‌ها: این معیار نیز معکوس تعداد قدم‌های لازم برای رسیدن به هدف را نشان برتری عامل‌ها می‌داند. $trial$ نشان‌دهنده تعداد تلاش، n_{trial} تلاش فعلی و $m_i(tria)$ تعداد تلاش‌های لازم برای رسیدن هدف است.

$$e_i^{Am} = \left(\sum_{trial=1}^{n_{trial}} m_i(trial)/n_{trial} \right)^{-1} \tag{۶-۲}$$

[1] Absolute
[2] Negative

۲-٦- تخته‌سیاه

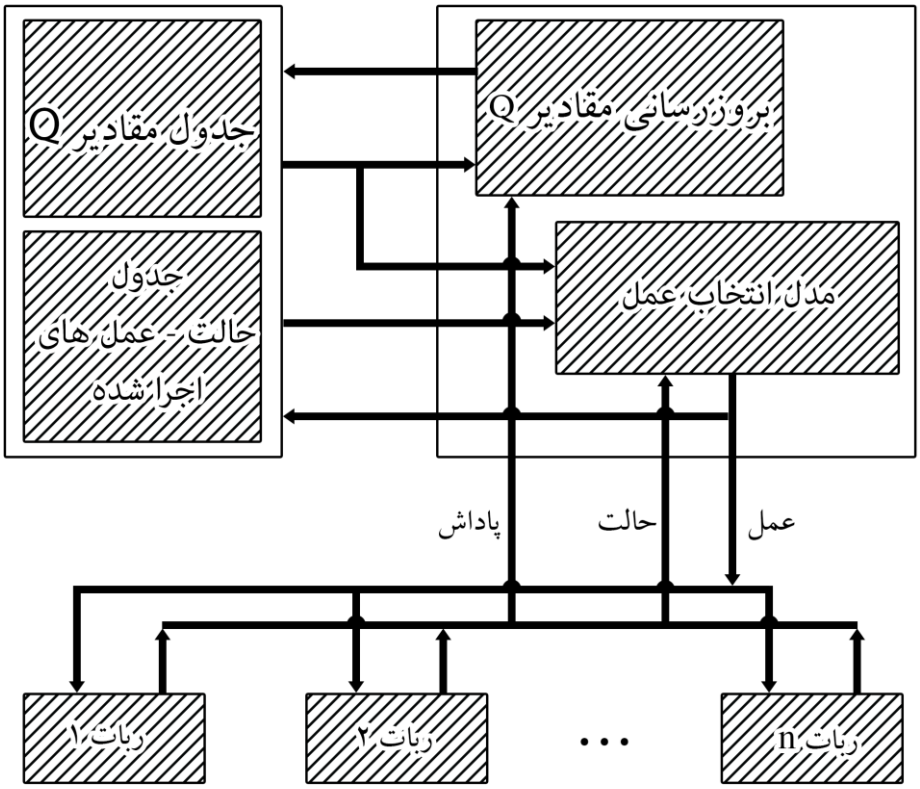
همان‌طور که تا به حال مطرح شد معمولاً یادگیری مشارکتی به دنبال روشی برای به اشتراک گذاشتن تجربه‌های عامل‌ها است. تحقیقات نشان داده که ارتباط درست می‌تواند کارایی سیستم‌های چندعاملی را افزایش دهد. ایده تخته‌سیاه که یک روش انتقال مؤثر در سیستم‌های توزیع‌شده است اولین بار در [۱۹] مطرح‌شده است.

تخته‌سیاه نوعی حافظه اشتراکی است که هر عامل می‌تواند به آن دسترسی داشته باشد. این دسترسی به‌صورت خواندن و نوشتن است. در این مکانیزم هیچ ارتباط مستقیمی بین عامل‌ها وجود ندارد و ارتباط تنها به‌صورت غیرمستقیم از طریق تخته‌سیاه صورت می‌گیرد.

یانگ [1] و همکاران در [۲۰] روش یادگیری Q بر مبنای این معماری پیاده‌سازی شده است و مکانیزم تخته‌سیاه مسئول بروز رسانی سیاست و عمل است. در این پیاده‌سازی هر عامل حالت جاری خود را به تخته‌سیاه ارسال نموده و پس از انجام عمل پیشنهادی از طرف تخته‌سیاه، پاداش دریافتی از محیط را به تخته‌سیاه اعلام می‌نماید. در این پیاده‌سازی عامل به‌عنوان منبع دانش است که مسئولیت اجرای اعمال و برگرداندن نتیجه را بر عهده دارد و تخته‌سیاه مسئول نگهداری فرستاده‌های عامل‌ها است. نتیجه حاصل از این آزمایش‌ها نشان می‌دهد که عملکرد روش تخته‌سیاه برای یادگیری وظیفه عدم برخورد به موانع از کیفیت بالایی برخوردار است.

در شکل ۲-۱ ساختار یادگیری مشارکتی بر اساس Q آورده شده است. ربات در این سیستم مسئولیت اجرای عملی که از طرف تخته‌سیاه به او داده شده و برگرداندن نتایج حاصل از اجرای عمل به تخته‌سیاه را بر عهده دارد. تخته‌سیاه نیز این اطلاعات را ذخیره کرده و بر اساس یک مدل در فاز انتخاب عمل برای عامل‌ها تصمیم‌گیری می‌کند. در این ساختار محلی برای نگهداری حالت و عمل‌ها وجود دارد و درصورتی‌که عامل‌ها حالت و عملی را تجربه نکرده باشند تخته‌سیاه فارغ از روش بولتزمن [2] عمل تجربه نشده را پیشنهاد می‌کند. در صورت تجربه شدن بر اساس تابع بولتزمن و جدول Q یادگیری انجام می‌شود.

[1] Yang
[2] Boltzmann



شکل ۲-۱: ساختار یادگیری مشارکتی بر مبنای تخته‌سیاه [۲۰]

۲-۷- یادگیری مشارکتی مبتنی بر پختگی سیاست

ایده دیگری که برای یادگیری مشارکتی ارائه شد یادگیری مشارکتی مبتنی بر پختگی سیاست است که به نحوی ترکیب شده یادگیری مشارکتی مبتنی بر خبرگی و تخته‌سیاه است. این روش که برای بهبود یادگیری در محیط‌های چند رباتی ارائه شد از روش تخته‌سیاه بهره برده و برای یادگیری نیز از روش یادگیری Q استفاده می‌نماید.

در واقعیت ممکن است یک عامل به دلیل تصادفی بودن سیستم یادگیری به سیاست بهتر برسد که بهتر است آن را به اشتراک بگذارد. برای این کار برخلاف روش ارائه‌شده توسط یانگ در [۲۰] در اینجا عامل‌ها علاوه بر ارتباط غیرمستقیمی که از طریق تخته‌سیاه دارند می‌توانند ارتباط مستقیمی با عامل‌های باسیاست بهتر داشته باشند و سیاست بهتر را از آن‌ها بیاموزند.

نکته مهم در اینجا، پیدا کردن معیاری برای نشان دادن عامل‌های دارای سیاست بهینه است که در [۲۱] به این منظور از معیارهای یادگیری مشارکتی مبتنی بر خبرگی بهره می‌برد. مجموع پاداش‌های منفی را به‌عنوان پختگی عامل در نظر گرفته و رباتی را که مقدار این پارامتر در آن کمتر باشد دارای پختگی بالاتر می‌داند. باید به این نکته توجه

داشت که این روش نسبت به روش‌هایی چون تخته‌سیاه این مزیت را دارد که عامل سیاست را نه به‌صورت کورکورانه بلکه از عامل‌های دیگر می‌آموزد.

۲-۸- یادگیری مشارکتی بر مبنای خبرگی چند معیاره

پاکیزه و همکاران در [۲۲] روشی تحت عنوان یادگیری مشارکتی بر مبنای خبرگی چند معیاره [1]ارائه کردند. در این کار سعی شده معیارهای خبرگی که در [۱۵] معرفی‌شده ترکیب شوند. ایده نویسنده برگرفته از زندگی انسانی است. او معتقد است همان‌طور که برای سنجش یک فرد در نظر گرفتن یک معیار باعث انتخاب درست نمی‌شود؛ مثلاً برای انتخاب دانش‌آموز نمونه باید پارامترهای مختلفی را سنجید؛ در محیط‌های مشارکتی نیز مشخص کردن خبرگی صرفاً بر اساس یک معیار اطلاعات زیادی در بر نخواهد داشت. ایشان در ترکیب این معیارها از روش مکاشفه که در بخش ۳-۵- خواهد آمد بهره برده‌اند.

در این روش نیز همانند یادگیری مشارکتی مبتنی بر خبرگی فرایند یادگیری را به دو بخش چرخه یادگیری مستقل و چرخه همکاری تقسیم می‌نمایند. در فاز چرخه یادگیری مستقل به‌صورت معمول به یادگیری می‌پردازند؛ اما در چرخه همکاری است که باید ترکیب داده‌ها انجام شود. پاکیزه برای ترکیب معیارهای مختلف یک فرایند سه مرحله را پیشنهاد داده است.

گام اول: مشخص کردن عامل با خبرگی کمتر و تولید جدول مشارکتی این عامل

دلیل ساخت جدول مشارکتی توسط ضعیف‌ترین عامل تنها بهره‌گیری از جدول تمام عامل‌هاست. همان‌طور که در بخش ۲-۵- گفته شد عامل‌ها برای تولید جدول مشارکتی خود تنها از جدول مشارکتی عامل‌های قوی‌تر بهره می‌برند. در این روش نیز برای اینکه جدول تولیدشده شامل بیشترین داده‌ها باشد از جدول عامل ضعیف‌تر بهره برده شده تا ترکیبی از جدول تمام عامل‌ها باشد.

گام دوم: تهیه جدول مشارکتی بر اساس هر یک از معیارهای خبرگی.

در این گام بر اساس هر معیار ضعیف‌ترین عامل مشخص‌شده و ساخت جدول مشارکتی مربوط به آن معیار خبرگی را به عهده می‌گیرد. در پایان این گام ۶ جدول مشارکتی که هر یک بر اساس یکی از معیارهای مطرح‌شده در [۲۳] است تولید وجود خواهد داشت.

[1] MSC

گام سوم: ساخت جدول مشارکتی عامل‌ها و استفاده از آن

در این روش جدول مشارکت نهایی، مجموع تمام جدول‌های تولیدشده در گام قبل در نظر گرفته شده است در رابطه ۲-۷ CoQ_i جدول‌های Q مشارکتی تولید شده به وسیله معیار i ام را نشان می‌دهد.

$$CoQ_{MSC} = \sum_{i=1}^{6} CoQ_i \qquad (۷-۲)$$

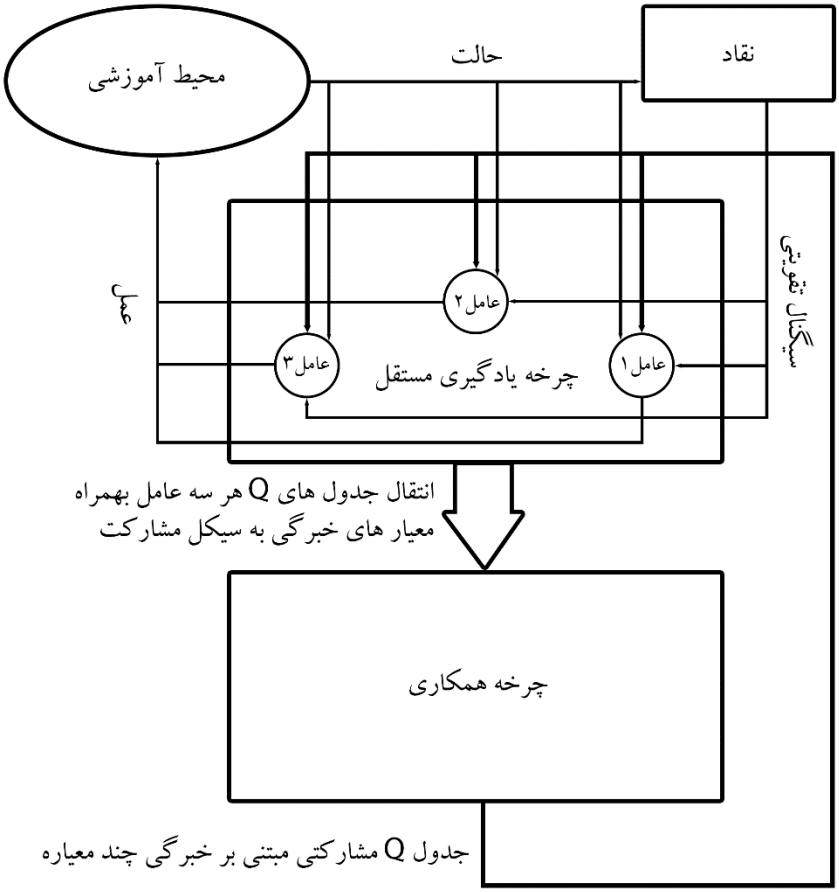
موضوع مهمی که وجود دارد نحوه بهره‌گیری از این جدول است؛ همان‌طور که از روش محاسبه این جدول نیز مشخص است نمی‌توان این جدول را با جدول Q عامل‌ها جایگزین کرد. چراکه وقتی در هر گام از مجموع تمام جدول‌ها جایگزین جدول Q شود یکی از مهم‌ترین شروط همگرایی یادگیری تقویتی که آستانه داشتن مقادیر هست را نقض نموده‌ایم؛ و باعث واگرایی در یادگیری می‌شویم.

راه‌حل ارائه‌شده توسط پاکیزه بهره‌گیری از مکاشفه در یادگیری است. از این جدول تنها برای انتخاب عمل بهره برده و جدول Q ثابت باقی می‌ماند. در رابطه ۲-۸ نحوه انتخاب عمل درروش MSC آورده شده است در این رابطه T پارامتری است جهت کنترل مقدار تصادفی بودن انتخاب عمل هرچه مقدار این پارامتر بیشتر باشد انتخاب تصادفی انتخاب شدن کاهش می‌یابد[۲۲].

$$\pi(s_t) = \arg\max_{a_t} \left(\frac{e^{\frac{CoQ_{MCE}(s_t,a_t)}{T}}}{\sum e^{\frac{CoQ_{MCE}(s_t,a_t)}{T}}} \right) \qquad (۸-۲)$$

پاکیزه و همکاران با ارائه این کار که ترکیب چهار معیار یادگیری بود در سال ۱۳۹۱ توانستند بهبود خوبی در روش‌های مبتنی بر یادگیری ایجاد کنند. در شکل ۲-۲ نمای کلی از روش خبرگی چند معیاره آورده شده است که به‌خوبی عملکرد این روش را نشان می‌دهد.پاکیزه در [۲۴] نیز همین روش را بر روی یادگیری سارسا[1] مورد آزمایش قرار داده‌اند.

[1] SARSA



شکل ۲-۲: نمای کلی از روش خبرگی چند معیاره [۷]

۲-۹- نتیجه‌گیری

در این فصل سعی شد با مروری بر روش‌های ارائه‌شده درزمینه یادگیری مشارکتی پیش‌زمینه‌ای جهت ارائه روش پیشنهادی فراهم شود. تمام‌کارهایی که در این فصل ارائه شد صرفاً روش‌هایی جهت ترکیب داده‌های عامل‌ها در سیستم‌های چند عامل مشارکتی مبتنی بر یادگیری تقویتی بود. معمولاً در تمام این روش‌ها یک فاز ترکیب در نظر گرفته‌شده بود؛ در کارهایی مثل خبرگی این فاز ترکیب جدای از فاز یادگیری مستقل و در کارهایی چون پند دهی وابسته به آن تعریف‌شده بود. موضوع دیگری که در اکثر روش‌ها مورد تأکید قرارگرفته معیاری‌هایی جهت نمایش برتری عامل‌ها نسبت به یکدیگر است.

فصل سوم:

پیش‌نیاز

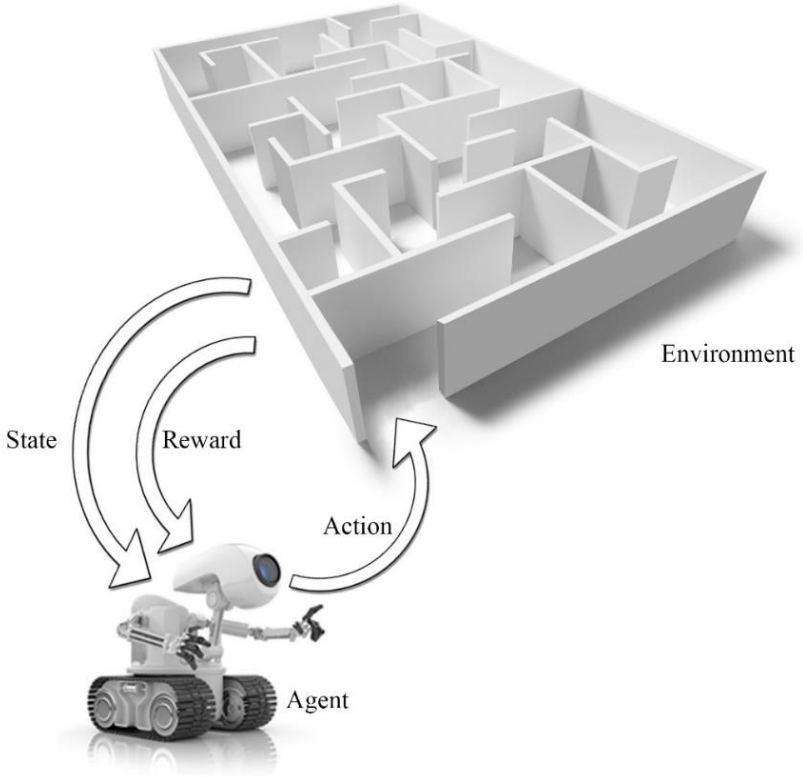
برای حل چالش‌های سیستم‌های چندعاملی و یادگیری مشارکتی ایده‌های فراوانی ارائه‌شده است که اکثر آن‌ها برگرفته از زندگی گروهی انسان‌ها بوده است. یادگیری انسان در اکثر موارد از طریق انتقال اطلاعات صورت می‌گیرند و به نسبت کمتری یادگیری از طریق اکتشاف است و این خود دلیلی بر یادگیری مشارکتی است.

معمولاً یادگیری مشارکتی در محیط‌هایی انجام می‌شود که داده‌ها بدون برچسب هستند و برای یادگیری نیاز به روش‌های یادگیری بدون نظارت است به همین دلیل یادگیری مشارکتی معمولاً به همراه یادگیری تقویتی که یک روش تقریباً بدون ناظر است مورداستفاده قرار می‌گیرند. در یادگیری تقویتی عامل باید بر اساس پاداش و جریمه‌ای که دریافت می‌نماید یادگیری کند. یادگیری تقویتی را می‌توان برای محیط‌های مختلف مورداستفاده قرارداد که برای آزمایش عامل‌ها، محیط‌های آزمایشی وجود دارد که می‌تواند نکات مثبت و منفی یادگیری را نشان دهد.

در بخش ۳-۱- به تشریح یادگیری تقویتی پرداخته و یادگیری Q را تشریح می‌شود بعدازآن در بخش ۳-۴- روش‌های انتخاب عمل در یادگیری تقویتی را تشریح خواهد شد. بخش ۳-۶- به معرفی محیط‌ها پرداخته و نهایتاً در بخش ۳-۷- چند محیط آزمایشی که برای سیستم‌های چندعاملی مورداستفاده قرار می‌گیرند ارائه خواهد شد.

۳-۱- یادگیری تقویتی

یادگیری تقویتی به مسائلی می‌پردازد که عامل حالت‌هایی را درک کرده و بر اساس درک خود عملی انجام دهد. بعد از انجام عمل از طرف محیط و یا یک معلم خارجی به عامل پاداش یا جریمه‌ای تعلق می‌گیرد. مثلاً در یک بازی ممکن است برای برد پاداش مثبت، برای باخت منفی و برای اعمال دیگر مقدار صفر باشد. عامل باید در این محیط از فرایند انجام عمل و دریافت پاداش یادگیری نماید تا بتواند مجموع پاداش‌های خود را در طول زمان افزایش دهد.



شکل ۳-۱ فرایند یادگیری تقویتی

یادگیری تقویتی حتی برای شرایطی که عامل هیچ اطلاعاتی درباره محیط ندارد هم می‌تواند مفید بوده و به‌خوبی عمل نماید. یادگیری تقویتی شباهت زیادی با الگوریتم‌های برنامه‌سازی پویا دارد. رباتی را در نظر بگیرید که در یک محیط قرار است یادگیری نماید. این ربات حسگرهایی برای درک و عملگرهایی برای انجام عمل در محیط دارد. هدف ربات یادگیری روش‌هایی است که او را به اهداف برسانند. اگر ربات تابعی را که حالت‌ها را به اعمال بهینه نگاشت کند یاد بگیرد می‌تواند به‌خوبی فرایند یادگیری را انجام دهد.

تساورا بازی TD_Gammon را معرفی نمود که با بهره‌گیری از یادگیری تقویتی و برای رسیدن به بازیکنان جهانی طراحی‌شده بود. او توانست بعد از یک و نیم میلیون بازی خودساخته به سطحی از یادگیری برسد که در مقابل بازیکنان جهانی بازی کند[۳]. معمولاً در یادگیری تقویتی بر اساس روش مخفی مارکف یادگیری انجام می‌شود.

۳-۲- فرآیند تصمیم‌گیری مارکف

در بخش ۳-۷- نشان خواهیم داد که محیط‌ها تا چه حد می‌توانند متفاوت باشند؛ اما نمی‌توان یادگیری تقویتی را وابسته به محیط کرد و به یک تعریف جامع برای آن نیازمند هستیم. میشل در [۳] با کمک فرایند تصادفی مارکف یادگیری تقویتی را به شکل ریاضی تعریف می‌نماید.

در فرایند تصمیم‌گیری مارکف[۱] (MDP) عامل مجموعه‌ای از حالت‌ها[۲] به نام S و مجموعه‌ای از اعمال[۳] به نام A را در اختیار دارد. در هرلحظه t، حسگرهای عامل حالت s_t را مشخص می‌کنند و عامل عمل a_t را انجام می‌دهد. محیط نیز مقدار پاداش[۴] $r_t = r(s_t,a_t)$ را به عامل می‌دهد و حالت $s_{t+1} = \delta(s_t,a_t)$ را ایجاد می‌نماید. در اینجا r و δ جزء محیط هستند و الزاماً برای عامل مشخص نیستند. در MDP توابع $\delta(s_t,a_t)$ و $r(s_t,a_t)$ فقط به حالت فعلی و عمل وابسته‌اند و به حالت‌ها و اعمال قبلی وابستگی ندارند [۳].

کار عامل یادگیری سیاستی[۵] مثل $\pi: S \rightarrow A$ است که بتواند حرکت a_t را بر اساس حالت فعلی به دست آورد. یکی از ساده‌ترین راه‌حل‌ها برای پیدا کردن خط‌مشی بهینه تعریف آن به صورتی است که تابع تجمّعی پاداش‌ها در طول زمان حداکثر شود. برای تعریف دقیق‌تر مقدار $V^{\pi}(s_t)$ را به فرم رابطه ۳-۱ تعریف می‌نماییم[۳].

$$V^{\pi}(st) \equiv r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots \equiv \sum_{i=0}^{\infty} \gamma^i r_{t+1} \qquad (۱-۳)$$

در اینجا r_{t+1} پاداش‌هایی هستند که از سیاست π و شروع از s_t به دست آمده‌اند. γ نیز که مقداری بین صفر و یک است، ثابتی برای پاداش تأخیری است. حال می‌توان مفهوم یادگیری عامل‌ها را دقیق مشخص نمود. باید سیاستی را پیدا نماییم که برای تمام حالت‌ها تابع $V^{\pi}(st)$ حداکثر شود. چنین سیاستی را سیاست بهینه می‌نامند و با π^* نشان می‌دهند[۳].

$$\pi^* \equiv \arg\max_{\pi} V^{\pi}(st), (\forall s) \qquad (۲-۳)$$

۳-۳- یادگیری Q

در بسیاری از الگوریتم‌های یادگیری مثال‌ها به‌صورت <s,a> داده می‌شوند اما همان‌طور که گفته شد در الگوریتم تقویتی برای محیط‌هایی است که در آن <s,a> مشخص نبوده و فقط $r(s_i,a_i)$ را داریم. الگوریتم یادگیری

---

[1] Markov decision processes
[2] State
[3] Action
[4] Reward
[5] Policy

Q که مهم‌ترین عضو خانواده یادگیری تقویتی است روشی است که در آن می‌توان از پاداش‌های دریافتی به یادگیری رسید [25]. یک انتخاب برای یادگیری V^{π^*} است و سیاست را می‌توان بر اساس رابطه ۳-۳ به دست آورد.

$$\pi^* \equiv \underset{\pi}{\operatorname{argmax}} \ [r(s,a) + \gamma V^{\pi^*}(\delta(s,a))] \qquad (۳-۳)$$

طبق معادله بالا با داشتن تابع δ و r می‌توان سیاست بهینه را پیدا نمود اما در بسیاری از محیط‌ها این توابع نامعلوم بوده و نمی‌توان یادگیری را بر این اساس انجام داد. در یادگیری Q برای رفع مشکل تابع تخمین Q(s,a) را تعریف می‌نماید که جمع مقدار پاداش لحظه‌ای و پاداش سیاست بهینه را نگه می‌دارد[۲۵].

$$Q(s,a) \equiv \ r(s,a) + \gamma V^{\pi^*}(\delta(s,a)) \qquad (۴-۳)$$

بنابراین می‌توان رابطه محاسبه سیاست بهینه را بازنویسی نمود [۱].

$$\pi^* \equiv \underset{a}{\operatorname{argmax}} \ Q(s,a) \qquad (۵-۳)$$

اهمیت این بازنویسی این است که دیگر بدون داشتن تابع δ و r می‌توان اعمال بهینه را پیدا کرد. کافی است با در نظر گرفتن تمام اعمال a در حالت s، عملی انتخاب شود که Q(s,a) را حداکثر نماید. حال که هدف یادگیری مقادیر Q از طریق پاداش‌های دریافتی است نگاهی دقیق‌تر به V^{π^*} می‌تواند کار را ساده‌تر سازد [۲۵].

$$V^{\pi^*} = \max_{a'} \mathrm{Q}(\mathrm{s}, \mathrm{a}') \qquad (۶-۳)$$

که با توجه به رابطه ۳-۷ خواهیم داشت.

$$Q(s,a) = r(s,a) + \gamma \max_{a'} Q(\delta(s,a), a') \qquad (۷-۳)$$

این رابطه بازگشتی کلید حل معمای یادگیری Q است. در الگوریتم یادگیری Q که یک الگوریتم برنامه‌نویسی پویا است از یک جدول بهره می‌برد که برای هر جفت حالت و عمل یک سلول دارد و در آغاز با مقادیر تصادفی پر می‌شود. در هرگام حالت s دیده‌شده و عمل a انجام می‌شود و بعد از دریافت پاداش r و رفتن به حالت 's جدول 'Q را به شکل زیر بروز می‌نماید.

$$Q'(s,a) = r + \gamma \max_{a'} Q(s', a') \qquad (۸-۳)$$

شبه کد الگوریتم Q را می‌توان در شکل ۳-۲ دید. باوجود این الگوریتم می‌توان تضمین کرد برای سیستم‌های تصمیم‌گیری مارکف مقدار 'Q به Q میل می‌کند با این فرض که تابع پاداش کران‌دار باشد و هر جفت حالت و عمل چند مرتبه دیده شود.

Q learning algorithm

1) For each s,a initialize the table entry $\hat{Q}(s, a)$ to zero.
2) Observe the current state s
3) Do forever:
    a. Select an action a and execute it
    b. Receive immediate reward r
    c. Observe the new state s'
    d. Update the table entry for $\hat{Q}(s, a)$ as follows:

$$\hat{Q}(s, a) \leftarrow r + \gamma \max_{á} \hat{Q}(ś, á)$$

    e. s←s

شکل ۳-۲: شبه کد یادگیری تقویتی[۳]

۳-٤- برقراری تعادل در اکتشاف و بهره‌برداری

برای رسیدن به یک یادگیری قابل‌قبول در روش‌های یادگیری تقویتی، باید به یک تعادل در اکتشاف و بهره‌برداری دست‌یافت [۱]. اکتشاف باعث شناخت تمام حالت‌ها و بهره‌بردار باعث تثبیت داده‌های فعلی می‌گردد. منظور از اکتشاف در یادگیری تقویتی عدم پیروی از داده‌های فعلی عامل و انتخاب تصادفی اعمال و منظور از بهره‌برداری تصمیم‌گیری بر اساس داده‌های فعلی عامل است. روش‌های متعددی برای ایجاد این تعادل ارائه‌شده که در ادامه دو روش از پرکاربردترین‌ها را که اکثر روش‌ها بر اساس آنها ایجاد شده‌اند خواهد آمد.

۳-٤-۱ ε - حریصانه

در روش ε - حریصانه انتخاب تصادفی عمل را برای یک عامل وابسته به یک پارامتر می‌نماییم. این پارامتر که عدد بین ۰ و ۱ است نشانگر دهنده احتمال انتخاب تصادفی عمل توسط عامل است [۲۳،۲۴].

$$\pi(s) = \begin{cases} random\ action\ from\ A(s) & if\ \xi < \varepsilon \\ argmax_{a \in A(s)} Q(s, a) & otherwise \end{cases} \quad (۹-۳)$$

در هر مرحله انتخاب عمل، $0 \leq \xi \leq 1$ به‌صورت یکنواخت تولید می‌شود و بر اساس رابطه ۳-۹ یک عمل انتخاب می‌شود. نکته مهمی که در این روش وجود دارد تنظیم ε است. اگر این مقدار زیاد در نظر گرفته شود عامل کمتر بهره‌برداری کرده و بهره‌برداری کمتر یعنی کم شدن سرعت یادگیری و اگر بسیار کوچک در نظر گرفته شود به معنی اکتشاف کم در یادگیری است که باعث پایین آمدن کیفیت یادگیری می‌شود. پس برقراری تعادل بین اکتشاف و بهره‌برداری به معنی رسیدن به کیفیت و سرعت یادگیری قابل‌قبول است که در این روش به ε سپرده‌شده است. اما این حساسیت در اینجا وجود دارد که آیا باید ε را در تمام مراحل یادگیری ثابت گذاشت؟ اگر پاسخ این سؤال مثبت است باید آن مقدار چه مقداری باشد و اگر منفی است ε باید چگونه تنظیم شود.

به‌طور ضمنی نیز می‌شود به این نکته رسید که در مراحل اول یادگیری اطلاعات زیادی برای بهره‌برداری وجود ندارد و در هرچه مراحل یادگیری بیشتری سپری می‌شود این اطلاعات جهت بهره‌برداری افزون می‌شود. پس با وابسته کردن ε به تعداد مراحل سپری‌شده، می‌توان به نتایج خوبی در یادگیری رسید به شکلی که هرچه مراحل یادگیری بیشتری سپری می‌شود باید ε را کاهش داد [۲۵]. اما اینکه دقیقاً با چه رابطه‌ای این مقدار باید کاهش پیدا کند مشخص نیست.

۳-٤-۲- بهره‌گیری از توزیع بولتزمن (Softmax)

اگرچه روش ε حریصانه روش مؤثری در برقرار کردن نقطه تعادلی در اکتشاف و بهره‌بردار است اما این روش خالی از اشکال هم نیست. یک اشکال مهم این روش این است که در آن غیر بهترین عملی که بر اساس داده‌ها استخراج می‌شود برای بقیه اعمال احتمال انتخاب یکسان است. برای رفع این مشکل بهترین روش این است که احتمال انتخاب اعمال بر اساس جدول Q تولید شود. این کار به‌سادگی بر اساس توزیع بولتزمن یا گیبز انجام می‌شود [۲۵].

$$\frac{e^{\frac{Q_t(a)}{\tau}}}{\sum_{b=1}^{n} e^{\frac{Q_t(b)}{\tau}}} \qquad (۱۰-۳)$$

τ پارامتر مثبتی است که برای مشخص کردن تفاوت احتمال انتخاب‌ها استفاده می‌شود. هر چه این پارامتر کوچک‌تر انتخاب شود تفاوت احتمال انتخاب‌ها بالاتر خواهد بود. اگر این مقدار ۰ در نظر گرفته شود روش انتخاب کاملاً حریصانه خواهد بود؛ و هرچه افزایش پیدا کند روش انتخاب به تصادفی میل می‌کند. در این روش هم همانند روش ε حریصانه یک پارامتر به پارامترهای سیستم اضافه شده است اما این روش وابستگی کمتری به تعداد مراحل سپری‌شده دارد و می‌توان با مشخص کردن یک τ ثابت به یک دقت خوب رسید [۲۵].

۳-۵- مکاشفه در یادگیری

بیاچی [1]در [۲۸] مفهوم مکاشفه را با هدف تسریع در سرعت یادگیری تقویتی پیشنهاد می‌دهد. در این روش خصوصیات مثبتی چون ضمانت همگرایی و انتخاب آزادانه اعمال یادگیری تقویتی حفظ شده است. به علاوه راهکاری برای حل مشکل سرعت پایین الگوریتم تقویتی ارائه نموده است.

در این روش که HAQL [2] نامیده شده یک سیاست برای عامل‌ها ایجاد می‌شود و در هر مرحله برای بهبود انتخاب اعمال مورداستفاده قرار می‌گیرد. درواقع این تابع مکاشفه بر روی انتخاب اعمال اثر می‌گذارد. این مکاشفه صرفاً برای انتخاب اعمال بوده و به عامل می‌گوید که کدام عمل را بدون در نظر گرفتن بقیه عمل‌ها در نظر بگیرد. قانون انتخاب عمل استفاده شده مطابق رابطه ۳-۱۱ و رابطه ۳-۱۲ تعریف می‌شود.

$$\pi(s_t) = \begin{cases} \arg\max_{a_t}[Q(s_t,a_t) + \varepsilon H_t(s_t,a_t)], & if\ q < p \\ a_{random} & otherwise \end{cases} \quad (۱۱-۳)$$

$$H(s_t,a_t) = \begin{cases} max_a\, Q(s_t,a) - Q(s_t,a_t) + 1, & if\ a_t \in \pi^H(s_t) \\ 0, & otherwise \end{cases} \quad (۱۲-۳)$$

در [۲۸] ضمن اثبات عملکرد روش مکاشفه شرایط تعریف تابع مکاشفه نیز اشاره شده است. در پژوهش پیش رو نیز از این روش بهره برده شده و برای انتخاب عمل از یک تابع مکاشفه بر اساس کوتاه‌ترین فاصله تجربه شده بهره خواهیم برد؛ که می‌توان درستی روش پیشنهادی را بر اساس روش HAQL اثبات کرد.

۳-۶- محیط‌های یادگیری

محیط‌ها در یادگیری می‌توانند به‌سادگی مثال دنیای شبکه که در [۳] توسط میشل ارائه‌شده‌اند و یا به پیچیدگی محیط رانندگی در بزرگ‌راه باشند. راسل در [۱] محیط‌ها را به پنج دیدگاه تقسیم‌بندی و بررسی نموده است. در ادامه به‌طور مختصر تقسیم‌بندی‌های محیط از دیدگاه راسل آورده شده است [۱].

- مشاهده پذیر و نیمه مشاهده پذیر: اگر عامل به کمک حسگر های خود توانایی تشخیص حالت محیط را داشته باشد محیط را مشاهده پذیر گویند. در غیر این صورت محیط نیمه مشاهده پذیر یا تا حدی قابل‌مشاهده نامیده می‌شود.

---

[1] Bianchi
[2] Heuristically Accelerated Q-Learning(HAQL)

- قطعی و غیرقطعی: اگر بتوان حالت بعدی محیط را بر اساس سابقه اعمال و حالت فعلی مشخص کرد محیط قطعی است و در غیر این صورت محیط غیرقطعی است.
- دوره‌ای یا غیر دوره‌ای : درصورتی‌که هر مرحله از مراحل دیگر مستقل باشد محیط را دوره‌ای می‌نامیم.
- ایستا و پویا: اگر محیط در مدت‌زمان بین درک و انتخاب عمل تغییر کند پویا و در غیر این صورت ایستا است.
- گسسته و پیوسته: اگر مشاهدات و اعمال به شکل جداگانه تعریف شوند محیط را گسسته گویند.

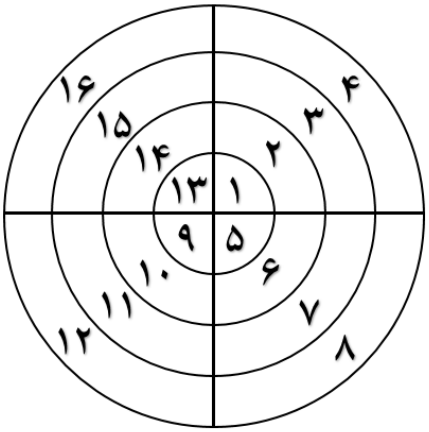
۳-۷- محیط‌های آزمایشی

محیط‌های آزمایشی زیادی برای آزمایش میزان کارایی روش‌های پیشنهادی در یادگیری مشارکتی پیشنهاد داده‌شده است. در اینجا دو محیط که بیشتر در پژوهش‌ها مورداستفاده قرارگرفته‌اند را شرح می‌دهیم. محیط صید و صیاد یک محیط پویا و پلکان مارپیچ محیطی ایستا است و می‌توان برای نمایش سازگاری روش پیشنهادی از هر دو بهره برد.

۳-۷-۱- صید و صیاد

محیط صید و صیاد یکی از مسائل کلاسیک در یادگیری است که جهت بررسی روش‌های یادگیری مناسب است. در این محیط دو نوع عامل وجود دارد. صیاد به دنبال پیدا کردن و شکار و صید که به دنبال فرار از دست صیاد است. صیاد با پیدا کردن، تعقیب کردن و پیش‌بینی حرکت صید است که او را شکار می‌نماید. مسائلی مانند نابودی اهداف متحرک در جنگ را می‌توان یک کاربرد از این سیستم دانست.

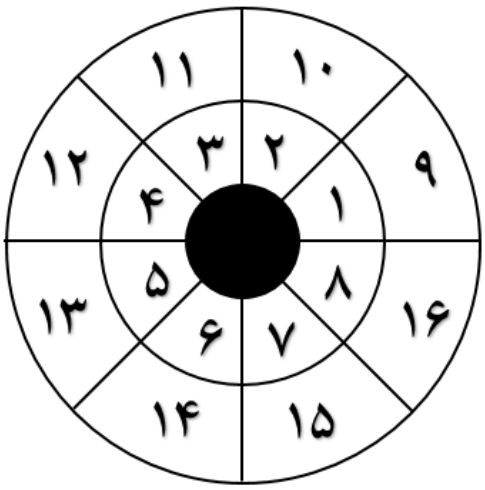
در این پژوهش محیط ۱۰×۱۰ در نظر گرفته‌شده که یک صید و یک صیاد در آن وجود دارد. در فاز یادگیری مستقل هر عامل در محیط های جداگانه به یادگیری می‌پردازند. سپس در فاز همکاری با انتقال اطلاعات به همکاری می‌پردازند. در این محیط صیاد با شکار صید پاداش ۱۰ را می‌گیرد در موارد دیگر برای هر حرکت ۰/۱ به عنوان جریمه دریافت می‌نماید. خصوصیات محیط و حرکت‌هایی که عامل‌ها می‌توانند داشته باشند در پیاده‌سازی بسیار مهم است. معمولاً صیاد یک محدوده دید دارد که عامل صید را در آن محدوده مشاهده می‌نماید. در این پایان‌نامه دامنه دید صیاد ۲ در نظر گرفته‌شده است. شکل ۳-۳ نمایی از محدوده دید عامل صیاد آمده است.



شکل ۳-۳: کانون دید صیاد

همان‌طور که مشخص است عامل صید نیز در این محیط باید از صیاد فرار کند؛ معمولاً حرکت‌های صید را به‌صورت تصادفی و صیاد با قابلیت یادگیری در نظر گرفته می‌شود. عامل‌ها در این محیط جهت حرکت از دو مؤلفه سرعت و زاویه حرکت استفاده می‌نمایند. هر صیاد می‌تواند با سرعتی بین ۰ و ۱ و هر صید با سرعت بین ۰ و ۰,۵ حرکت کند. سرعت صیاد بیشتر در نظر گرفته‌شده تا احتمال شکار افزایش پیدا کند. زاویه‌های حرکت عامل‌ها نیز بین ۰ تا ۳۶۰ درجه تعریف‌شده است. در شکل ۳-۴ به روشنی اعمال آورده شده است.

برای بهره‌گیری از روش‌های یادگیری تقویتی باید یک گسسته سازی در اعمال ایجاد می‌شد. بر همین اساس سرعت عامل صیاد را به دو حالت ۰,۵ و ۱ و زاویه انتخاب را به ۸ قسمت تقسیم نمودیم. به این شکل تعداد حرکت‌های قابل انجام برای عامل صیاد ۱۶ حرکت خواهد بود.



شکل ۳-۴: اعمال ممکن برای عامل صیاد

۳-۷-۲- پلکان مارپیچ

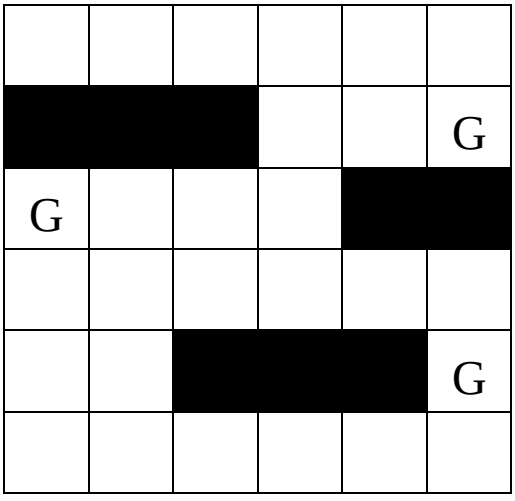
پلکان مارپیچ یک محیط ۶*۶ است که در بعضی از خانه‌های آن مانع قرار دارد. عامل باید خود را به اهداف برساند. هر تلاش عامل از قرار گرفتن تصادفی در خانه‌ای آغازشده و با رسیدن به هدف به اتمام می‌رسند. هر عامل می‌تواند در صورت نبود مانع به یکی از چهار جهت خود حرکت نماید. در زمان یادگیری در صورت برخورد به موانع میزان جریمه ۱ را دریافت می‌نماید و در صورت رسیدن به هدف میزان پاداش ۱۰ را می‌گیرد در موارد دیگر بر اساس فاصله‌ای که تا نزدیک‌ترین هدف دارد پاداش دریافت می‌نماید. در رابطه ۳-۱۳ می‌توان نحوه محاسبه پاداش را دید.

$$Reward = \frac{1}{distance\ between\ the\ agent\ and\ the\ goal} \qquad (۱۳-۳)$$

محیط پلکان مارپیچ یک محیط ایستا است. همان‌طور که گفته شد پاداش تمام خانه‌های در این محیط یکسان است. در محیط‌هایی که هیچ اولویتی بین اهداف نباشد معمولاً خانه‌های هدف را با پاداش یکسان در نظر می‌گیرند.

٦-۳-۲- پلکان مارپیچ تعمیم‌یافته

ازآنجایی‌که یکی از روش‌های مطرح‌شده در این پژوهش به طول محیط وابسته شده و مقدار پاداش‌ها را در نظر نمی‌گیرد لازم دانستیم که برای اثبات عملکرد روش پیشنهادی محیط جدیدی تعریف نماییم. این محیط کلیه شرایط پلکان مارپیچ به جر یکسان بودن پاداش خانه‌های جذب را دارد. پاداش خانه‌های هدف این محیط به ترتیب ۳۰،۱۵ و ۱۰ است.



شکل ۳-۵: نمایی از محیط پلکان مارپیچ

۳-۸- نتیجه‌گیری

می‌توان یادگیری Q را پایه‌ای ترین روش یادگیری تقویتی دانست. در این فصل سعی شد با تشریح یادگیری Q فهم روش‌های پیشنهادی تا حد امکان آسان‌تر شود. در تمام روش‌های بیان شده در فصل دوم از یادگیری Q استفاده شده است. در روش پیشنهادی نیز جهت یادگیری عامل‌ها از یادگیری Q استفاده شده است. معمولاً روش‌های یادگیری تقویتی مانند یادگیری Q نیاز به محیط‌هایی جهت ارزیابی نیز دارند که سعی شد با ارائه دو محیط استاندارد خواننده تا حد امکان با این محیط‌ها آشنا شود. در فصل آینده نیز جهت بررسی عملکرد روش پیشنهادی از همین محیط‌های استاندارد استفاده خواهد شد. نکته دیگری که باید یادآوری شود مکاشفه در یادگیری است که در روش‌های پیشنهادی نیز مورد استفاده قرار گرفته است.

فصل ٤

ارائه روش پیشنهادی

در راستای نیل به هدف این پژوهش که بهبود کارایی یادگیری مشارکتی تعریف‌شده بود نقاط بحرانی الگوریتم‌های خبرگی و خبرگی چند معیاره مورد تحلیل قرار گرفت. سه نقطه اصلی که اثرگذاری زیادی در کارایی الگوریتم دارند موردمطالعه قرار گرفتند تا با ارائه راه‌حل در این نقاط یک کارایی خوب در عملکرد الگوریتم مشارکتی حاصل شود.

نقطه بحرانی که ابتدا در نظر گرفته شد، یادگیری مستقل هر یک از عامل‌ها است. مسلماً درصورتی‌که یادگیری مستقل عامل‌ها بهبود یابد کارایی کل سیستم نیز افزایش می‌یابد. یکی از مهم‌ترین قسمت‌ها در یادگیری تقویتی انتخاب اعمال است که با ارائه روش‌هایی مانند روشSoftmax باعث تولید یک آستانه از اکتشاف و بهره‌برداری در این سیستم‌ها شده است. در این پژوهش نیز یک روش برای انتخاب اعمال ارائه شده است که در ابتدای این فصل به تشریح آن پرداخته خواهد شد.

دومین نقطه مورد بررسی نحوه ترکیب داده‌ها در یادگیری مشارکتی است که روش‌هایی چون تقلید و پند دهی و خبرگی برای این کار ارائه شده‌اند. در اکثر این روش‌ها معیاری برای نمایش برتری عامل‌ها نسبت به هم وجود دارد.

در پژوهش پیش رو نیز سعی شده معیاری برای ارزیابی عامل‌ها ارائه شود. سومین نقطه موردنظر که کمتر از دو مورد قبل مورد بررسی قرارگرفته موضوع تقسیم‌کار در الگوریتم‌های مشارکتی است؛ که سعی شده تا حدودی در این پژوهش در نظر گرفته شود.

در بخش اول این فصل معیارهای ارائه‌شده برای ارزیابی عامل‌ها تشریح و سپس پیشنهادهایی برای هر یک از نقاط بحرانی ارائه شده که مقدمات را برای ارائه یک پیشنهاد کلی مهیا می‌کند.

٤-١- معیارهای ارائه شده جهت ارزیابی عامل

همان‌طور که در فصل دوم بیان شد روش‌های متعددی برای یادگیری مشارکتی در سیستم‌های همگن ارائه‌شده است. در اکثر این روش‌ها نیاز به معیارهایی است تا بتوان برتری یک عامل را نسبت به عامل‌های دیگر نشان دهد. در تقلید شرطی، نیاز به معیارهایی است تا بتوان به کمک آن مشخص کرد که هر عامل از عامل‌هایی که دارای برتری نسبت به خود او هستند تقلید کند، درروش خبرگی نیاز به معیارهایی است تا بتوان اطلاعات عامل‌ها را بر اساس میزان خبرگی آن‌ها با هم ادغام کرد و یا درروش پند دهی نیاز بود عامل از کسانی پند بگیرد که دارای برتری نسبی از خود او باشند. در اینجا منظور از عاملی که دارای برتری بیشتر است مقدار نزدیکی اطلاعات عامل تا اطلاعات واقعی است و یا می‌توان گفت نسبت شناخت عامل‌ها از محیط است. در این پژوهش نیز سعی شده معیارهایی ارائه شود تا بتوان برای بالا بردن کارایی عامل‌ها مورداستفاده قرار بگیرند.

٤-١-١- شوک

ایده شوک بر اساس مهم‌ترین اصل یادگیری تقویتی استخراج‌شده است. مهم‌ترین موضوعی که در یادگیری تقویتی، عامل را به یادگیری می‌رساند انتقال پاداش حالت‌های جذب به حالت‌های دیگر است؛ که به‌طور ساده در هر چرخه یک مرحله حرکت می‌کند. بر این اساس می‌توان گفت زمان رسیدن اطلاعات به هر حالت رابطه مستقیمی با فاصله حالت تا نقطه جذب دارد. پس اگر قرار باشد معیاری برای سطح شناخت یک عامل از یک حالت و عمل داشته باشیم بدون شک وابستگی خاصی به زمان و تعداد دفعاتی دارد که اثر پاداش حالت هدفی در یک حالت دیده‌شده باشد. بر همین اساس معیار شوک معرفی‌شده که نمایش دهنده تعداد دفعاتی است که اثر پاداش حالت های جذب در یک حالت دیده شده است. این معیار در مقایسه با معیارهای دیگری که قبلاً ارائه‌شده دارای یک مزیت است و آن محلی بودن آن است. منظور از محلی بودن این است که این معیاری برای نمایش برتری عامل در هر حالت و عمل است درصورتی‌که معیارهای دیگر مقدار برتری عامل را در کل محیط نشان می‌دادند. برای محاسبه این معیار کافی

است یک جدول مشابه با جدول Q با مقدار اولیه ۰ ایجاد شده و پس از هر حرکت بر اساس رابطه ۴-۱ مقادیر این جدول بروزرسانی شوند.در این در این رابطه ś نمایش حالت جاری و s نمایش حالت قبل است.

$$\mathrm{shock}_t(\mathrm{s},\mathrm{a}) = \begin{cases} \mathrm{shock}_{t-1}(\mathrm{s},\mathrm{a}) + 1 & \mathrm{shock}\left(\acute{\mathrm{s}}, \underset{\acute{a}}{arg\max}\, Q(\acute{s},\acute{a})\right) > 0 \; or \; s \; is \; terminat \\ \mathrm{shock}_{t-1}(\mathrm{s},\mathrm{a}) & otherwise \end{cases} \quad (۱\text{-}۴)$$

٤-۱-۲- کوتاه‌ترین مسیر تجربه‌شده

معیار دیگری که در پژوهش فوق ارائه‌شده، کوتاه‌ترین مسیر تجربه شده است که می‌توان به صورت یک معیار برای سنجش برتری عامل‌ها نسبت به هم یا به‌عنوان اطلاعات مکاشفه‌ای در نظر گرفته شود. کوتاه‌ترین مسیر تجربه‌شده که جهت رعایت اختصار SEP[1] نامیده می‌شود؛ شامل یک جدول است که به تعداد حالت‌ها سطر و به تعداد اعمال عامل ستون دارد. در هر سلول این جدول طول کوتاه‌ترین مسیری قرار دارد که با اعمال آن عمل در آن حالت تجربه شده است. باید تأکید شود که در اینجا تفاوتی بین اهداف در نظر گرفته نشده و صرفاً کوتاه‌ترین فاصله تجربه‌شده ثبت می‌شود.

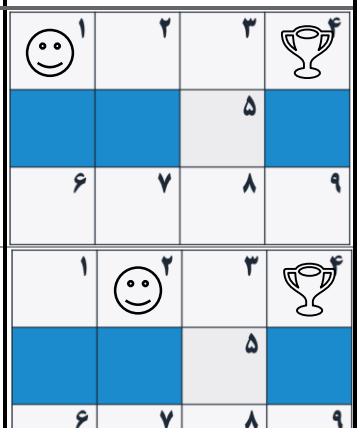
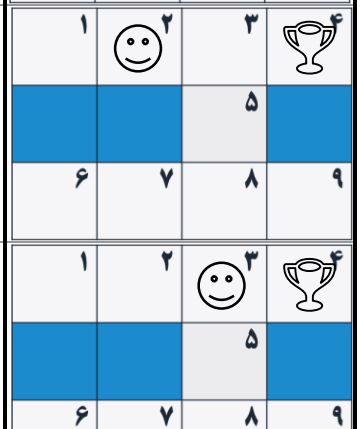
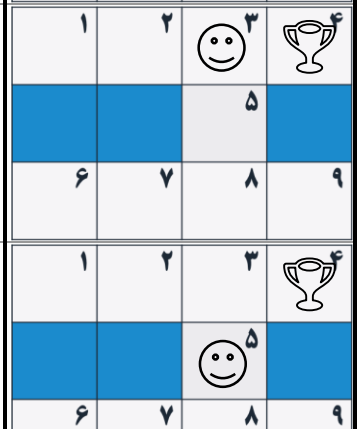
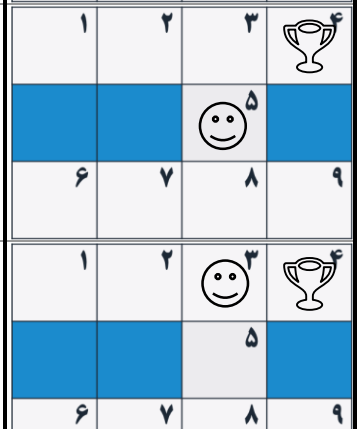
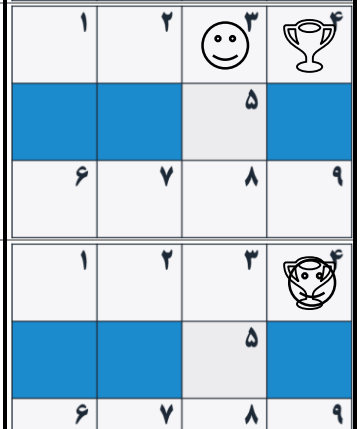
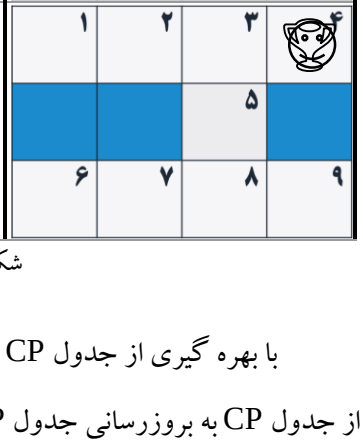
برای محاسبه SEP نیاز به یک جدول برای ثبت مسیر هر چرخه است؛ این جدول CP[2] نامیده می‌شود. در هر چرخه یادگیری یک مسیر طی می‌شود که از یک حالت تصادفی شروع و به یک حالت جذب ختم می‌شود.در یک چرخه ممکن است عامل چند مرتبه از یک حالت عبور کند. لازم است طول کوتاهترین میسر از آن حالت تا حالت جذب ثبت شود. به این سبب در جدول CP ، آخرین عملی که در آخرین حضور در خانه انجام شده،مقصد این عمل و شماره گام حرکت ثبت می‌شود. از روی این جدول (CP) می‌توان طول مسیر طی شده از این حالت تا حالت جذب را محاسبه کرد.

مقصد حالت nام	...	مقصد حالت دوم	مقصد حالت اول
آخرین عمل انجام شده در حالت nام	...	آخرین عمل انجام شده در حالت دوم	آخرین عمل انجام شده در حالت اول
آخرین گام ملاقات حالت nام	...	آخرین گام ملاقات حالت دوم	آخرین گام ملاقات حالت اول

شکل ۴-۱:نمایی از جدول *CP*

در شکل ۴-۲ سعی شده با ارائه مثالی ثبت مسیر با بهره گیری از جدول CP نمایش داده شده است. در مثال‌های ارائه شده در این بخش همانند پیاده سازی برای هر حرکت عددی در نظر گرفته شده است. در این جداول عدد ۱ نمایش حرکت به راست، عدد ۲ به معنی حرکت به پایین، عدد ۳ حرکت به چپ و ۴ حرکت به بالا است.

[1] Shortest Experienced Path
[2] Current Path

حالت بازی	حرکت	عمل	مقصد	جدول CP
	اول	راست	حالت شماره ۲	۱ ۲ ۳ ۴ ۵ ۶ ۷ ۸ ۹ ۰
	دوم	راست	حالت شماره ۳	۱ ۲ ۳ ۴ ۵ ۶ ۷ ۸ ۹ ۲ ۰ ۰ ۰ ۰ ۰ ۰ ۰ ۰ ۱ ۰ ۰ ۰ ۰ ۰ ۰ ۰ ۰ ۱ ۰ ۰ ۰ ۰ ۰ ۰ ۰ ۰
	سوم	پایین	حالت شماره ۵	۱ ۲ ۳ ۴ ۵ ۶ ۷ ۸ ۹ ۲ ۳ ۰ ۰ ۰ ۰ ۰ ۰ ۰ ۱ ۱ ۰ ۰ ۰ ۰ ۰ ۰ ۰ ۱ ۲ ۰ ۰ ۰ ۰ ۰ ۰ ۰
	چهارم	بالا	حالت شماره ۳	۱ ۲ ۳ ۴ ۵ ۶ ۷ ۸ ۹ ۲ ۳ ۵ ۰ ۰ ۰ ۰ ۰ ۰ ۱ ۱ ۲ ۰ ۰ ۰ ۰ ۰ ۰ ۱ ۲ ۳ ۰ ۰ ۰ ۰ ۰ ۰
	پنجم	راست	حالت شماره ۴	۱ ۲ ۳ ۴ ۵ ۶ ۷ ۸ ۹ ۲ ۳ ۵ ۰ ۳ ۰ ۰ ۰ ۰ ۱ ۱ ۲ ۰ ۴ ۰ ۰ ۰ ۰ ۱ ۲ ۳ ۰ ۴ ۰ ۰ ۰ ۰
	–	–	–	۱ ۲ ۳ ۴ ۵ ۶ ۷ ۸ ۹ ۲ ۳ ۴ ۰ ۳ ۰ ۰ ۰ ۰ ۱ ۱ ۱ ۰ ۴ ۰ ۰ ۰ ۰ ۱ ۲ ۵ ۰ ۴ ۰ ۰ ۰ ۰

شکل ۴-۲: مثال ثبت یک مسیر با بهره گیری از جدول CP

با بهره گیری از جدول CP در پایان هر چرخه یک مسیر ثبت شده وجود خواهد داشت که می‌توان با استفاده از جدول CP به بروزرسانی جدول SEP پرداخته شود. در شکل ۴-۳ الگوریتم بروزرسانی SEP آورده شد است.

UPDATE SEP MATRIX

دریافت جدول CurrentPath

۱) NZ = اندیس ستون بزرگترین عدد بزرگتر از صفر در سطر آخر جدول CurrentPath

۲) $SEP[NZ, CP(1, NZ)] = 1;$

۳) $CP(3, NZ) = -1;$

۴) NZ = اندیس ستون بزرگترین عدد بزرگتر از صفر در سطر آخر جدول CurrentPath

۵) تا زمانی که NZ تهی نشده ادامه بده

a. MD = کوچکترین مقدار در سطر CurrentPath(1,NZ) ام جدول SEP

b. $SEP[NZ, CP(2, NZ)] = MD + 1$

c. $CP(3, NZ) = -1;$

d. NZ = اندیس ستون بزرگترین عدد بزرگتر از صفر در سطر آخر جدول CurrentPath

شکل ۴-۳: بروزرسانی SEP

جهت فهم بهتر روش بروزرسانی مراحل بروزرسانی جدول SEP با توجه به CP ایجاد شده در شکل ۴-۲ در شکل ۴-۴ شکل ۴-۵ آمده است.

همان طور که از خروجی مثال قبل مشخص است با یک چرخه یادگیری ۶ خانه از جدول SEP بروزرسانی شده است؛ یعنی به تعداد حالت هایی که عامل در یک چرخه خواهد دید بروز رسانی در جدول SEP انجام می‌شود. که مشخصاً از مرتبه زمانی بیشتری بالاتری نسبت به جدول Q برخوردار نیست.

SEP				اندیس ستون بزرگ‌ترین گام در چرخه	CP								
∞	∞	∞	∞	۳	۱	۲	۳	۴	۵	۶	۷	۸	۹
∞	∞	∞	∞		۲	۳	۴	۰	۳	۰	۰	۰	۰
∞	∞	∞	∞		۱	۱	۱	۰	۴	۰	۰	۰	۰
∞	∞	∞	∞		۱	۲	۵	۰	۴	۰	۰	۰	۰
∞	∞	∞	∞										
∞	∞	∞	∞										
∞	∞	∞	∞										
∞	∞	∞	∞										
∞	∞	∞	∞										

شکل ۴-۴: محاسبه جدول کوتاه‌ترین مسیر تجربه شده بر اساس جدول مسیر شکل ۴-۲

SEP				اندیس ستون بزرگ‌ترین گام در چرخه	CP
∞	∞	∞	∞	۵	۱ ۲ ۳ ۴ ۵ ۶ ۷ ۸ ۹ ۲ ۳ ۴ ۰ ۳ ۰ ۰ ۰ ۰ ۱ ۱ ۱ ۰ ۴ ۰ ۰ ۰ ۰ ۱ ۲ ۰ ۰ ۴ ۰ ۰ ۰ ۰
∞	∞	∞	∞		
∞	∞	∞	۱		
∞	∞	∞	∞		
∞	∞	∞	∞		
∞	∞	∞	∞		
∞	∞	∞	∞		
∞	∞	∞	∞		
∞	∞	∞	∞	۲	۱ ۲ ۳ ۴ ۵ ۶ ۷ ۸ ۹ ۲ ۳ ۴ ۰ ۳ ۰ ۰ ۰ ۰ ۱ ۱ ۱ ۰ ۴ ۰ ۰ ۰ ۰ ۱ ۲ ۰ ۰ ۰ ۰ ۰ ۰ ۰
∞	∞	∞	∞		
∞	∞	∞	۱		
∞	∞	∞	∞		
۲	∞	∞	∞		
∞	∞	∞	∞		
∞	∞	∞	∞		
∞	∞	∞	∞		
∞	∞	∞	∞		
∞	∞	∞	∞	۱	۱ ۲ ۳ ۴ ۵ ۶ ۷ ۸ ۹ ۲ ۳ ۴ ۰ ۳ ۰ ۰ ۰ ۰ ۱ ۱ ۱ ۰ ۴ ۰ ۰ ۰ ۰ ۱ ۰ ۰ ۰ ۰ ۰ ۰ ۰ ۰
∞	∞	∞	۲		
∞	∞	∞	۱		
∞	∞	∞	∞		
۲	∞	∞	∞		
∞	∞	∞	∞		
∞	∞	∞	∞		
∞	∞	∞	∞		
∞	∞	∞	∞		
∞	∞	∞	۳	–	۱ ۲ ۳ ۴ ۵ ۶ ۷ ۸ ۹ ۲ ۳ ۴ ۰ ۳ ۰ ۰ ۰ ۰ ۱ ۱ ۱ ۰ ۴ ۰ ۰ ۰ ۰ ۰ ۰ ۰ ۰ ۰ ۰ ۰ ۰ ۰
∞	∞	∞	۲		
∞	∞	∞	۱		
∞	∞	∞	∞		
۲	∞	∞	∞		
∞	∞	∞	∞		
∞	∞	∞	∞		
∞	∞	∞	∞		
∞	∞	∞	∞		

شکل ۴-۵: ادامه محاسبه جدول کوتاه‌ترین مسیر تجربه شده بر اساس جدول مسیر شکل ۴-۲

٤-۲- افزایش کارایی در انتخاب عمل یادگیری تقویتی

همان‌طور که در بخش ۴-۱- گفته شد؛ روش Softmax یک روش مناسب برای ایجاد تعادل در اکتشاف و بهره‌برداری بوده که احتمال انتخاب هر عمل را بر اساس ارزش آن عمل در جدول Q مشخص می‌نماید. در روش بولتزمن احتمال انتخاب تصادفی اعمال با پارامتر τ که دما نامیده می‌شود کنترل می‌شود. در این پژوهش پیشنهاد می‌شود که از معیار کوتاه‌ترین فاصله تجربه‌شده در ترکیب با جدول Q استفاده شود. اگر انتخاب بر اساس جدول Q عامل‌ها محتمل‌ترین با ارزش‌ترین انتخاب دانسته شود، مطمئناً انتخاب بر اساس جدول SEP را می‌توان به‌عنوان محتمل‌ترین نزدیک‌ترین عمل تا هر هدفی دانست.روش پیشنهادی را SEPIL[1] مینامیم.

اما چرا باید از سیاست محتمل‌ترین کوتاه‌ترین مسیر تا هدف استفاده شود در صورتی که در یادگیری تقویتی به فاصله اهمیت زیادی داده نمیشود؟ مگر نه اینکه هدف یادگیری تقویتی رسیدن به با ارزشترین اهداف است پس بهره گیری از محتمل‌ترین با ارزش‌ترین برای یادگیر کفایت میکند. در تشریح دلیل بهره گیری از SEP باید در نظر داشت که این استفاده در جهت تسریع یادگیری است. معمولا در گام های اول یادگیری تقویتی حرکت های بی دلیل زیادی انجام می‌شود. در نتیجه استفاده از روشی که بتواند جلوی حرکت‌ها بی حاصل را بگیرد بسیار می‌تواند بر سرعت یادگیری تقویتی بیفزاید. در پژوهش پیش رو پیشنهاد می‌شود ترکیب جدول Q و جدولSEP براساس رابطه ۴-۲ انجام شود.

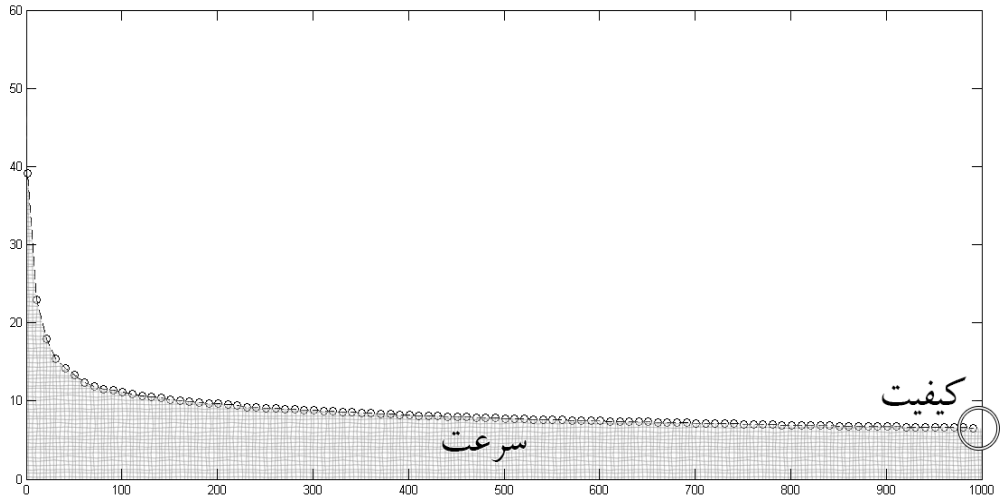
$$\pi(s)_t = (1-\mu)\ \text{Boltzmann}(Q_t, \tau_1) + \ \mu\ \text{Boltzmann}(\text{SEP}_t, \tau_2) \qquad (۲-۴)$$

برای ترکیب این دو معیار در این پژوهش پارامتر μ به سیستم اضافه می‌شود. μ که بین صفر و یک تنظیم می‌شود درصد بهره‌گیری از معیار SEP را مشخص می‌کند. ازآنجایی‌که رسیدن مقادیر جدول Q به مقدار واقعی وابسته به تکرار و رسیدن مقادیر جدول SEP نه وابسته به تکرار بلکه وابسته به مشاهده است این ترکیب می‌تواند اثربخشی بالایی در هدفمند کردن اکتشاف و بهره‌برداری داشته باشد. باید دقت داشت در رابطه ۴-۲ پارامتر دما برای جدول Q و جدول SEP متفاوت قرارگرفته است؛ این تفاوت بسیار مهم است چراکه مقادیر جدول SEP از نوع صحیح و مقادیر جدول Q اعشاری هستند. درنتیجه باید مقدار پارامتر دما برای جدول SEP بسیار کوچک‌تر تعریف شود.در این پژوهش پیشنهاد میشود مقدار پارامتر μ از معیار شوک و بر اساس رابطه ۴-۳ محاسبه شود.

$$\mu(s) = e^{-shock(s)} \qquad (۳-۴)$$

[1] shortest experimened path indevdual learning

در ادامه، این روش در آزمایش‌هایی مورد ارزیابی قرار خواهد رفت تا کارایی، نقاط قوت و ضعف آن مشخص شود؛ برای نشان دادن عملکرد روش پیشنهادی نیاز است که آزمایش‌هایی طراحی می‌شد. قبل از تعریف آزمایش‌ها باید معیارهای اصلی یادآوری شوند. در روش‌های یادگیری مشارکتی منظور از بهبود بالا بردن سرعت و دقت یادگیری است. برای سنجش این معیارها همانند پژوهش‌های پیشین در آزمایش‌ها از میانگین مجموع تعداد قدم‌های یادگیری استفاده‌شده که با این معیار استخراج کیفیت و سرعت یادگیری به‌سادگی قابل محاسبه است. سنجش سرعت یادگیری با محاسبه مساحت زیر نمودار و کیفیت یادگیری را نقطه نهایی نمودار به دست میاید. در شکل ۴-۶ جایگاه دقیق سرعت و دقت در نمودار به روشنی مشخص شده است که نمودار افقی نمایش چرخه های یادگیری و نمودار عمودی نمایش تعداد قدم های عامل در هر چرخه میباشد.



شکل ۴-۶: معیارهای ارزیابی

در آزمایش‌هایی که به‌منظور مقایسه روش های پیشنهادی با روش‌های دیگر انجام‌گرفته از دو محیط صید و صیاد و پلکان مارپیچ استفاده شد؛ در دیگر آزمایش‌ها ازآنجایی‌که صرفاً هدف بررسی اثر پارامترها بوده تنها در محیط پلکان مارپیچ آزمایش‌ها انجام شد. شرح کامل این محیط‌ها و نحوه پیاده‌سازی آن‌ها در بخش ۳-۷- آمده است.

پارامترهای الگوریتم یادگیری در حالت پیش‌فرض به صورتی که در جدول ۴-۱ آمده تنظیم‌شده است. علت تنظیم پارامترها با این مقادیر استفاده مکرر از این مقادیر روش های پیشین بوده است. در بعضی از آزمایش‌ها که هدف بررسی یک پارامتر خاص باشد مقدار پارامتر ذکر خواهد شد.

جدول ۴-۱: مقدار پیش‌فرض پارامترهای یادگیری

پارامتر	مقدار
μ	۰,۵
β	۰,۰۱
τ_1	۰,۴
gamma	۰,۹
τ_2	۰,۰۵

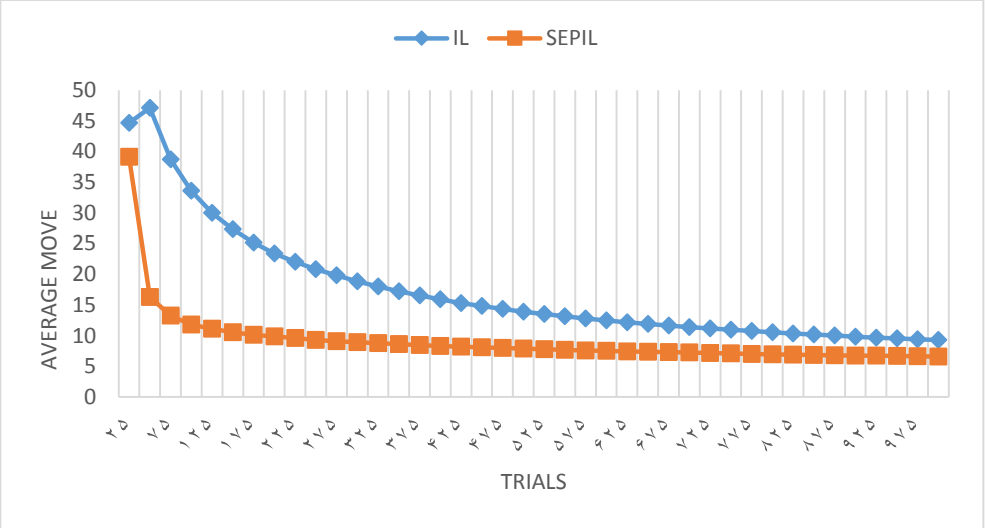
٤-۲-۱- آزمایش اول: بررسی و مقایسه روش پیشنهادی با روش یادگیری تقویتی

از ماهیت روش SEPIL پیداست عملکرد این روش در محیط‌های متفاوت متغیر خواهد بود. چیزی که عملکرد روش پیشنهادی را تحت تأثیر قرار می‌دهد مقدار پاداش حالت‌های هدف است. در محیط‌هایی که پاداش حالت‌های هدف یکسان باشند دو طرف رابطه ۴-۲ باهم تناقض ندارند اما در محیط‌هایی که پاداش حالت‌های هدف متفاوت باشند تناقضی ایجاد می‌شود. این تناقض از جایی ناشی می‌شود که یکی از حالت‌های هدف با سود کم در نزدیکی یک حالت باشد حال‌آنکه در فاصله دورتر هدفی با پاداش بالاتر وجود دارد. بر اساس همین موضوع آزمایش اول به بررسی روش SEPIL در دو محیط متفاوت پرداخته است. در محیط اول پاداش تمام حالت های هدف یکسان و در محیط دوم این پاداش‌ها متفاوت در نظر گرفته‌شده است.

در شکل ۴-۷ نمودار اجرای الگوریتم پیشنهادی در مقایسه با روش یادگیری تقویتی آورده شده است؛ در این نمودار محور افقی نمایش تلاش‌های یادگیر و محور عمودی نمایش میانگین تعداد حرکت‌های چرخه یادگیری است. لازم به ذکر است که کلیه نتایج ارائه شده بر اساس میانگین‌گیری از ۲۰ آزمایش است. همان‌طور که مشخص است روش پیشنهادی هم در سرعت و هم در کیفیت عملکرد بهتری داشته است. در شکل ۴-۸ و جدول ۴-۳ نیز حاصل اجرا در محیط با پاداش متفاوت آمده است. که در این محیط نیز عملکرد خوبی از روش پیشنهادی دیده می‌شود. نکته جالب‌توجه نزدیکی درصد بهبود در دو محیط است که دلیل این موضوع در آزمایش بعد تشریح خواهد شد.

جدول ۴-۲: حاصل اجرای روش پیشنهادی در محیطی که پاداش اهداف برابر در نظر گرفته‌شده

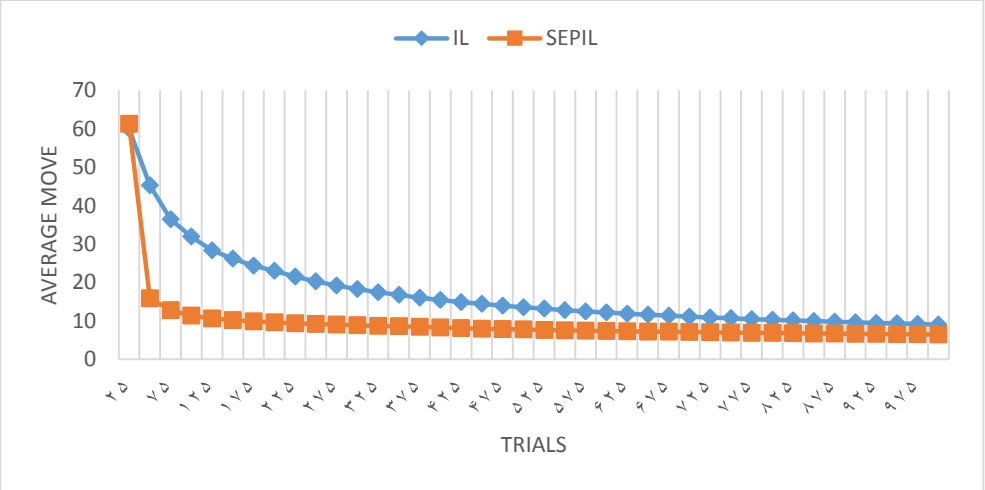
	IL	SEPIL	درصد بهبود
کیفیت	۹,۱۷	۶,۵۵	٪۳۹
سرعت	۱۷۲۰۸	۸۶۷۳	٪۹۸



شکل ۴-۷: حاصل اجرای روش پیشنهادی در محیطی که پاداش اهداف برابر در نظر گرفته شده

نتایج حاصل از آزمایش‌ها نشان از بهبود یادگیری با اعمال جدول SEP است. نکته جالب‌توجه این است که روش پیشنهادی حتی در محیطی که دارای پاداش‌های متفاوتی برای اهداف بوده هم بهبود چشم‌گیری ایجاد نموده است. یکی از اهداف بررسی این روش اثبات مؤثر بودن بهره‌گیری از روش‌های مکاشفه‌ای در یادگیری تقویتی است. حتی اگر هدف بهره‌گیری از این روش در محیط‌هایی با پاداش متفاوت حالت‌های هدف در نظر گرفته شود کافی است معیار مکاشفه بیشترین پاداش تجربه‌شده با SEP جانشین شود که مطمئناً با هدف نیز همخوانی بیشتری خواهد داشت.

نکته خاص دیگری که در روش SEP وجود دارد اثربخشی بالای این روش بر روی سرعت یادگیری است که تا حدود دو برابر این سرعت را افزایش داده است؛ که این بسیار مؤثر خواهد بود چراکه در بسیار از سیستم‌ها این افزایش سرعت حتی اگر به قیمت از دست رفتن کمی از کیفیت باشد بازهم ارزشمند است.



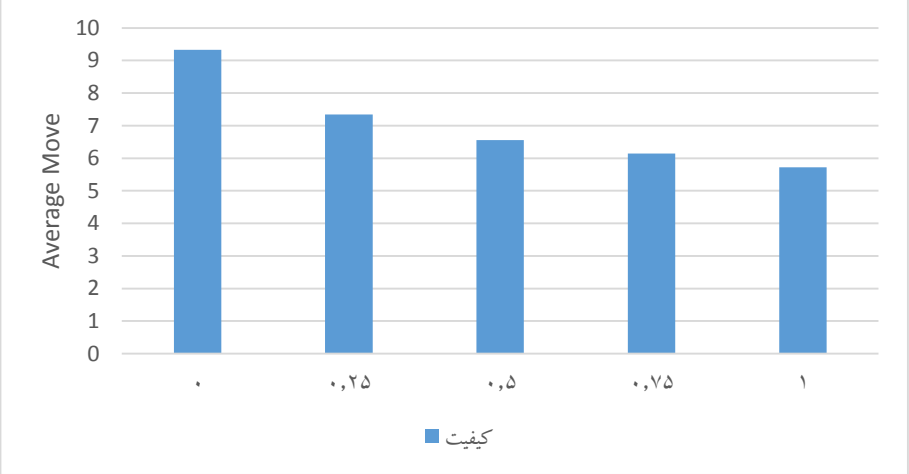
شکل ۴-۸: حاصل اجرای روش پیشنهادی در محیط با پاداش‌های متفاوت

جدول ۴-۳: حاصل اجرای روش پیشنهادی در محیط با پاداش‌های متفاوت

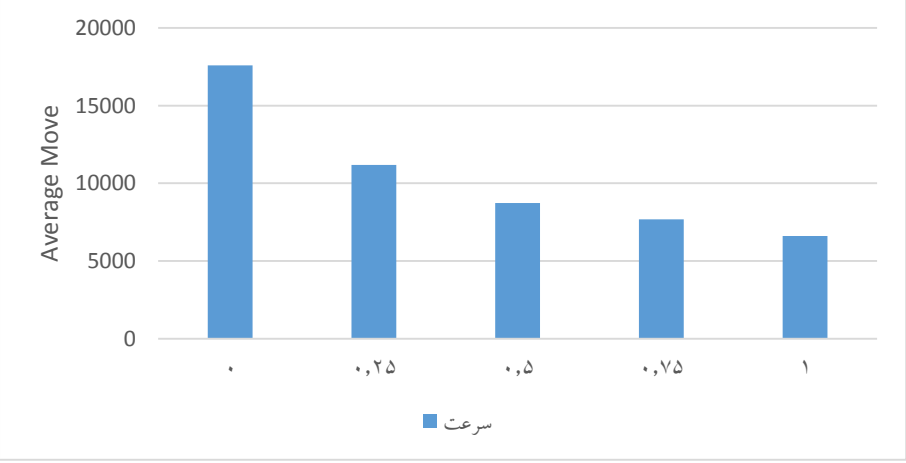
	IL	SEPIL	درصد بهبود
کیفیت	۸,۹۴	۶,۴۱	٪۳۹
سرعت	۱۶۴۹۹	۸۵۰۲	٪۹۴

٤-۲-۲- آزمایش دوم: بررسی حساسیت روش پیشنهادی در برابر پارامتر μ

در شکل ۴-۹ و شکل ۴-۱۰ عملکرد روش با مقادیر متفاوت برای μ آورده شده است. ازآنجایی که μ اصلی‌ترین پارامتر روش SEPIL است تحلیل حساسیت این پارامتر بسیار مهم است. همان‌طور که مشخص است مقدار این پارامتر تا هراندازه هم که افزایش داشته در معیارهای یادگیری اثر مؤثری داشته است. ازآنجایی‌که در تابع مکاشفه ارائه شده مبنا فاصله است و پاداش‌ها در نظر گرفته نمی‌شود آزمایش در محیط پلکان مارپیچ تکرار شد تا عملکرد روش به‌طور کامل مشخص شود.



شکل ۴-۹: بررسی حساسیت روش پیشنهادی در محیط پلکان مارپیچ بر کیفیت یادگیری

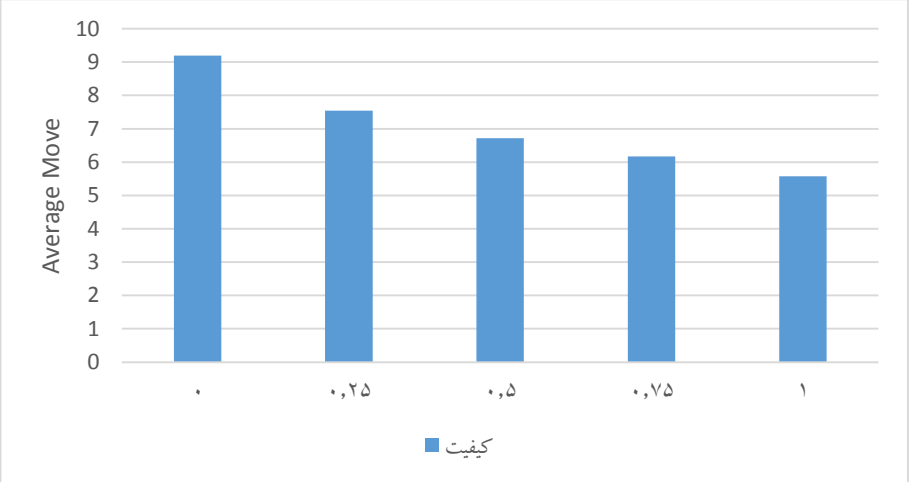


شکل ۴-۱۰: بررسی حساسیت روش پیشنهادی در محیط پلکان مارپیچ بر سرعت یادگیری

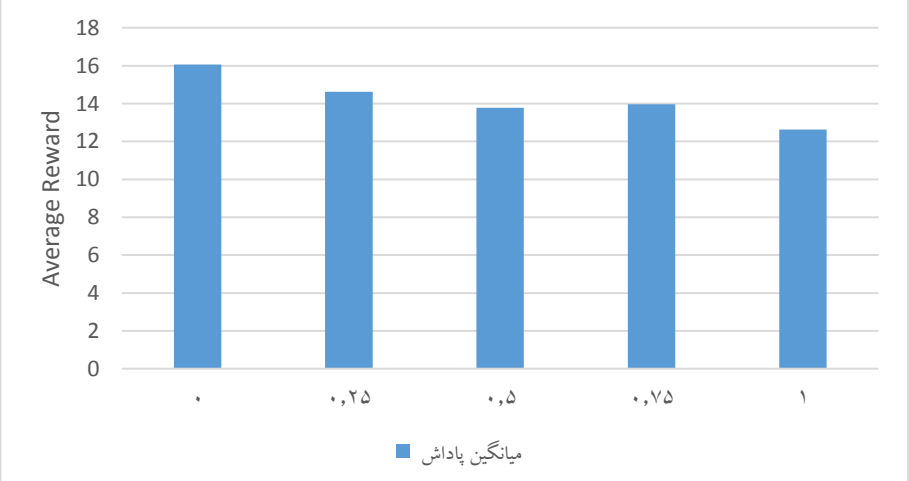
جدول ۴-۴: بررسی حساسیت روش پیشنهادی در برابر پارامتر μ

μ	۰,۰۰	۰,۲۵	۰,۵۰	۰,۷۵	۱,۰۰
کیفیت	۹,۳۳	۷,۳۴	۶,۵۶	۶,۱۴	۵,۷۲
سرعت	۱۷۶۰۱	۱۱۱۸۲	۸۷۲۴	۷۶۸۲	۶۶۰۹

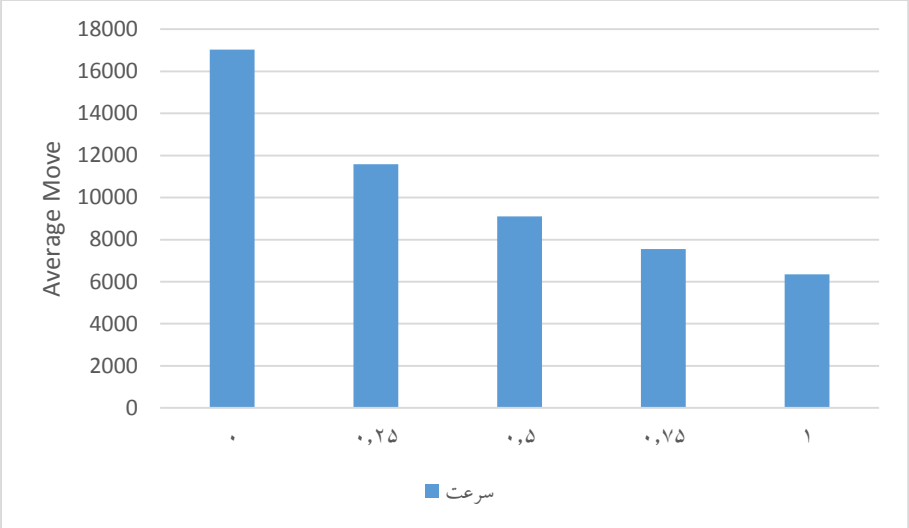
در راستای مشخص شدن دقیق‌تر روش پیشنهادی آزمایش تحلیل حساسیت پارامتر μ در محیط پلکان مارپیچ تعمیم‌یافته نیز تکرار شد. در این آزمایش به‌جز کیفیت و سرعت یادگیری، معیار میانگین پاداش‌های عامل در طول یادگیری نیز محاسبه شده است. این معیار که برعکس دو معیار دیگر هرچقدر بیشتر باشد نشان از عملکرد بهتر روش پیشنهادی است می‌تواند به‌خوبی جایگاه روش پیشنهادی را نمایش دهد. باید تأکید شود که دلیلی برای محاسبه این معیار در پلکان مارپیچ نیست؛ چراکه تمام اهداف در آن محیط پاداش یکسانی دارند.



شکل ۴-۱۱: بررسی حساسیت μ بر کیفیت یادگیری روش پیشنهادی در محیط پلکان مارپیچ تعمیم‌یافته



شکل ۴-۱۲: بررسی حساسیت μ بر میانگین پاداش یادگیری روش پیشنهادی در محیط پلکان مارپیچ تعمیم‌یافته



شکل ۴-۱۳: بررسی حساسیت μ بر میانگین سرعت روش پیشنهادی در محیط پلکان مارپیچ تعمیم‌یافته

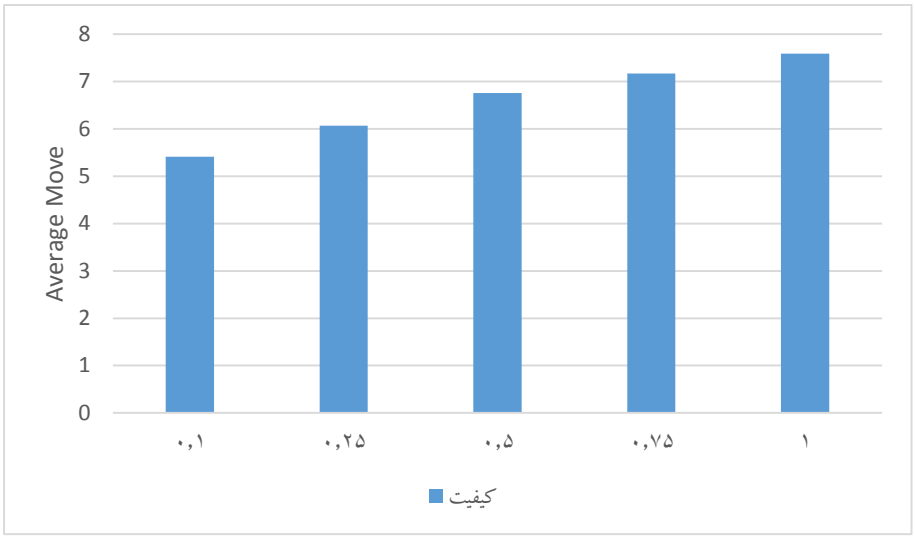
بررسی نتایج این آزمایش که تحلیل حساسیت پارامتر μ در محیط پلکان مارپیچ تعمیم‌یافته است در شکل ۴-۱۱، شکل ۴-۱۲ و شکل ۴-۱۳ آمده است.همانند آزمایش قبل با توجه به نتایج این آزمایش می‌توان به بهبود سرعت و کیفیت یادگیری پی برد؛ اما شکل ۴-۱۲ نشان دهنده تأثیر منفی بالا بردن μ در میانگین پاداش‌های عامل است. اثر منفی افزایش پارامتر μ بر میانگین پاداش‌ها در کنار اثر مثبت افزایش پارامتر μ بر سرعت و کیفیت یادگیری نشان‌دهنده این است که تنظیم پارامتر μ وابسته به هدف یادگیری است.

جدول ۴-۵ :بررسی حساسیت μ بر یادگیری روش پیشنهادی در محیط پلکان مارپیچ تعمیم‌یافته

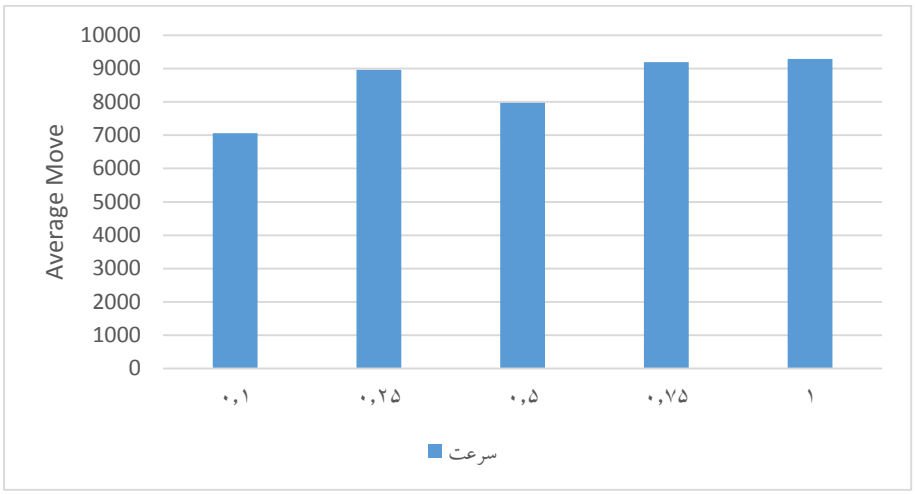
μ	۰,۰۰	۰,۲۵	۰,۵۰	۰,۷۵	۱,۰۰
کیفیت	۹,۱۹	۷,۵۴	۶,۷۲	۶,۱۷	۵,۵۷
سرعت	۱۷۰۳۴	۱۱۵۸۰	۹۱۰۲	۷۵۵۸	۶۳۴۳
میانگین پاداش	۱۶,۰۷	۱۴,۶۲	۱۳,۷۸	۱۳,۹۶	۱۲,۶۳

٤-۲-۳- آزمایش سوم: بررسی حساسیت روش پیشنهادی در برابر پارامتر τ_1

پارامتر دما در انتخاب عمل وظیفه مشخص کردن میزان تصادفی انتخاب شدن اعمال را بر عهده دارد [۲۹]. هر چه مقدار این پارامتر بیشتر باشد احتمال رفتار تصادفی از عامل بیشتر خواهد بود. درروش پیشنهادی ترکیبی از خروجی دو تابع بولتزمن جهت انتخاب عمل مورداستفاده قرارگرفته است. τ_1 جهت مشخص کردن میزان بهره برداری از اطلاعات جدول Q است. تحلیل عملکرد این روش نشان از حساسیت پایین آن نسبت به این پارامتر دارد.



شکل ۴-۱۴: بررسی حساسیت τ_1 بر کیفیت یادگیری روش پیشنهادی



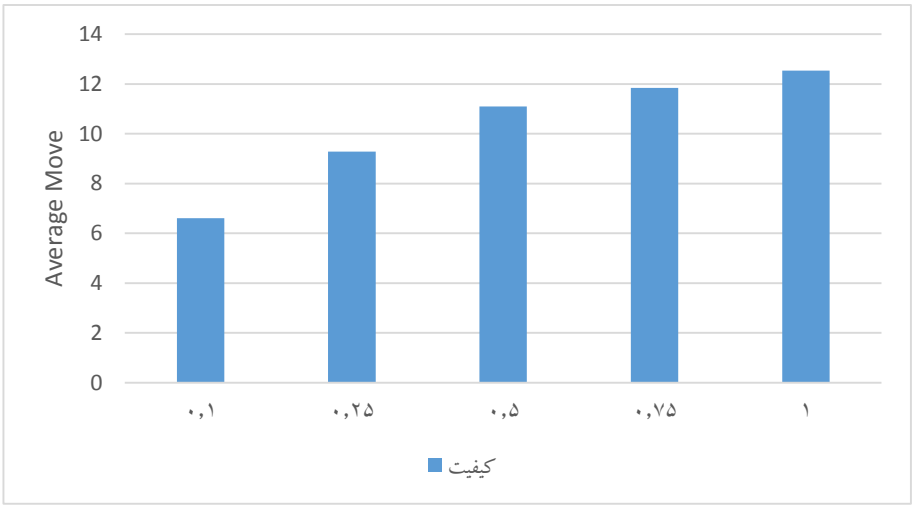
شکل ۴-۱۵: بررسی حساسیت τ_1 بر سرعت یادگیری روش پیشنهادی

جدول ۴-۶: بررسی حساسیت روش پیشنهادی در برابر پارامتر τ_1

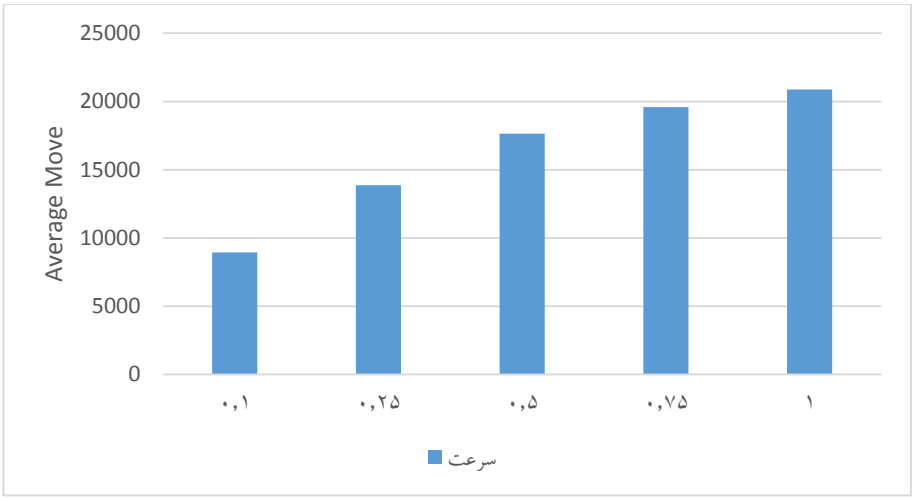
τ_1	۰,۰۱	۰,۲۵	۰,۵۰	۰,۷۵	۱,۰۰
کیفیت	۵,۴۱	۶,۰۷	۶,۷۶	۷,۱۷	۷,۵۹
سرعت	۷۰۵۹	۸۹۶۰	۷۹۷۳	۹۱۹۳	۹۲۹۱

٤-۲-٤- آزمایش چهارم: بررسی حساسیت روش پیشنهادی در برابر پارامتر τ_2

نتایج حاصل از آزمایش حاکی از آن است که حساسیت روش پیشنهادی در برابر τ_2 بسیار بالاتر از τ_1 است.تا جایی که افزایش این پارامتر کیفیت یادگیری را نصف و سرعت یادگیری را بیشتر از ۲ برابر کرده است. دلیل این موضوع را می‌توان در نوع داده‌های SEP جستجو کرد.



شکل ۴-۱۶: بررسی حساسیت $\tau 2$ بر کیفیت یادگیری روش پیشنهادی



شکل ۴-۱۷: بررسی حساسیت $\tau 2$ بر سرعت یادگیری روش پیشنهادی

از آنجایی مقادیر SEP همه مقادیر صحیح هستند نسبت اختلاف آن‌ها بسیار کم خواهد بود. درحالی‌که روش بولتزمن بیشتر بر اساس همین اختلاف عمل می‌کند. درنتیجه هراندازه که τ_2 بیشتر در نظر گرفته شود نمی‌تواند کم بودن اختلاف نسبت مقادیر SEP را جبران نماید؛ همین باعث می‌شود که روش پیشنهادی حرکت‌های تصادفی بیشتری انجام دهد.

جدول ۴-۷: بررسی حساسیت روش پیشنهادی در برابر پارامتر $\tau 2$

τ_2	۰,۰۱	۰,۲۵	۰,۵۰	۰,۷۵	۱,۰۰
کیفیت	۶,۶۱	۹,۲۸	۱۱,۱۰	۱۱,۸۴	۱۲,۵۴
سرعت	۸,۹۳۳	۱,۳۸۶۵	۱,۷۶۴۲	۱,۹۵۸۳	۲,۰۸۷۷

٤-۳- بررسی و ارائه راهکار در یادگیری مشارکتی

در این پژوهش سعی شده با ارائه راه‌حلی هر دو بخش تقسیم‌کار و ترکیب دادها بهبود داده شود. در راستای میل به این هدف از هر دو پارامتر شوک و SEP استفاده‌شده است. درروش پیشنهادی همانند روش مبتنی بر خبرگی فرایند یادگیری را به دو فاز یادگیری مستقل و فاز همکاری تقسیم‌شده است. در فاز یادگیری مستقل عامل‌ها بر اساس روش‌های یادگیری تقویتی به یادگیری می‌پردازند. سپس در فاز همکاری، داده‌های جمع‌آوری‌شده باهم ترکیب می‌شوند. درروش پیشنهادی ابتدا جداول SEP عامل‌ها با هم ترکیب می‌شوند. برای ترکیب این جداول نیازی به معیاری جهت مشخص کردن برتری عامل‌ها نسبت به هم وجود نداشته و می‌توان با یک کمینه ساده بهترین ترکیب ممکن را ایجاد کرد.

$$SEP\ (s,a) = \min_{agent} SEP\ _{agent}(s,a) \qquad (۴-۴)$$

پس از ترکیب SEP جهت تقسیم‌کار عامل‌ها در فاز همکار به دودسته تقسیم می‌شوند:

- عامل‌هایی که سیاست گرفته‌شده از SEP و جدول Q در آن‌ها هم‌خوانی دارد.
- عامل‌هایی که سیاست گرفته‌شده از SEP و جدول Q در آن‌ها تناقض دارد.

این تقسیم‌بندی برای هر حالت از محیط جداگانه صورت می‌گیرد؛ دلیل انجام این دسته بندی این است که برای هر گروه مأموریتی در نظر گرفته شود. گروه اول شناخت سیاست پیشنهادی توسط SEP را بر عهده خواهد گرفت و گروه دوم سعی در شناخت ناشناخته‌ها خواهد داشت؛ به‌عبارت‌دیگر می‌توان گفت که این تقسیم‌کار بین عامل‌ها یک تعادل بین اکتشاف و بهره‌برداری از SEP است؛ و این اکتشاف می‌تواند بسیار مفید باشد مخصوصاً برای محیط‌هایی که پاداش حالت‌های هدف متفاوت در نظر گرفته شده است.

در فاز ترکیب جداول هر گروه از یک میانگین‌گیری وزن‌دار بر اساس پارامتر شوک استفاده‌شده است تا هر عامل به‌اندازه اثری که از حالت‌های جذب داشته در تولید جدول مشارکتی اثر داشته باشد در رابطه۴-۵ و ۴-۶ می‌توان روش ترکیب داده ها آورده شده است.

$$Q_{co}(s,G) = \sum_{i \in G} w_i(s,G) Q_i\ (s) \qquad (۵-۴)$$

$$w_i(s,G) = \frac{shock_i(s)}{\sum_{j \in G} shock_j(s)} \qquad (۶-۴)$$

٤-٤- تشریح کامل روش پیشنهادی

در روش پیشنهادی در پژوهش سعی شده از راهکارهای ارائه‌شده در بخش‌های قبل استفاده شود؛ این کار بهبود مؤثری در یادگیری مشارکتی ایجاد می‌نماید. برای بهره بردن از راه‌کارهای ارائه‌شده در بخش‌های قبل در این پژوهش روال یادگیری به دو فاز یادگیری مستقل و فاز همکاری تقسیم‌شده است. در فاز یادگیری مستقل عامل بر اساس روش SEPIL به یادگیری می‌پردازد؛ و در فاز همکاری همانند روش پیشنهادی در بخش ۳-۴ عمل می‌نماید.

این ترکیب توانسته به میزان خوبی یادگیری مشارکتی را بهبود بخشد. در تحلیل این روش می‌توان دلیل مهم افزایش سرعت بین عامل‌ها را بهره‌گیری از SEP دانست که این اساس این روش را SEP می‌نامیم. در بخش انتخاب عمل معمولاً عامل معیار فاصله را در نظر نگرفته و صرفاً بر اساس داده‌ها، جدول Q انتخاب عمل انجام می‌دهد. این انتخاب عمل بر اساس جدول Q بهترین کاری است که می‌توان انجام داد. چراکه اطلاعات جدول Q شامل همه‌ی معیارهای هدف یادگیری هستند. دلیل بهره‌گیری اثرگذاری SEP را می‌توان در خام بودن داده‌های جدول Q در مراحل اول یادگیری دانست؛ در مراحل اول یادگیری عامل کاملاً به‌صورت تصادفی عمل می‌نماید و درصورتی‌که ممکن است قبلاً همان مسیر را بررسی کرده باشد؛ اما SEP خیلی زودتر از جدول Q اطلاعات می‌گیرد. به‌طور مثال در اولین چرخه یادگیری فقط یکی از خانه‌های جدول Q اطلاعات حالت‌های جذب را دریافت می‌کند درصورتی‌که SEP برای تمام حالت‌هایی که عامل سپری کرده اطلاعات جمع‌آوری می‌کند. درنتیجه پیشنهاد بهره‌گیری از SEP به معنی این نیست که انتخاب عمل به‌وسیله جدول Q درست نباشد بلکه درروش پیشنهادی از SEP به‌عنوان یک ابزار کمکی در انتخاب عمل استفاده می‌شود. موضوع دیگری که باید در نظر داشت ایجاد تعادل بین اکتشاف و بهره‌برداری است؛ در اینجا نیز این موضوع به‌اندازه زمانی که انتخاب عمل بر اساس جدول Q بود مهم است و بایستی با تنظیم درست دما این تعادل برقرار شود.

الگوریتم روش پیشنهادی (SEP)

۶) تولید جداول SEP,Q و CurrentPath

۷) تا پایان یادگیری انجام بده

a. در فاز یادگیری مستقل انجام بده

i. مشاهده حالت

ii. انتخاب عمل به روش SEP

iii. دریافت پاداش عمل انتخابی بر اساس حالت فعلی

iv. بروزرسانی جدول Q

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

v. بروزرسانی حالت فعلی

vi. بروزرسانی CurrentPath

vii. اگر در حالت نهایی قرار گرفتی

۱. بروزرسانی جدول SEP بر اساس تابع UpdateSEP

b. در فاز همکاری

i. ایجاد SEP مشارکتی با حداقل گیری از جداول SEP عامل ها

ii. برای هر حالت انجام بده

۱. پیدا کردن عامل های هم گروه با عامل بر اساس رابطه جدول Q با SEP

۲. ایجاد جدول Q مشارکتی بر اساس معیار شوک

$$w_i(s,G) = \frac{shock_i(s)}{\sum_{j\in G} shock_j(s)}$$

$$Q_{co}(s,G) = \sum_{i\in G} w_i(s,G) Q_i(s)$$

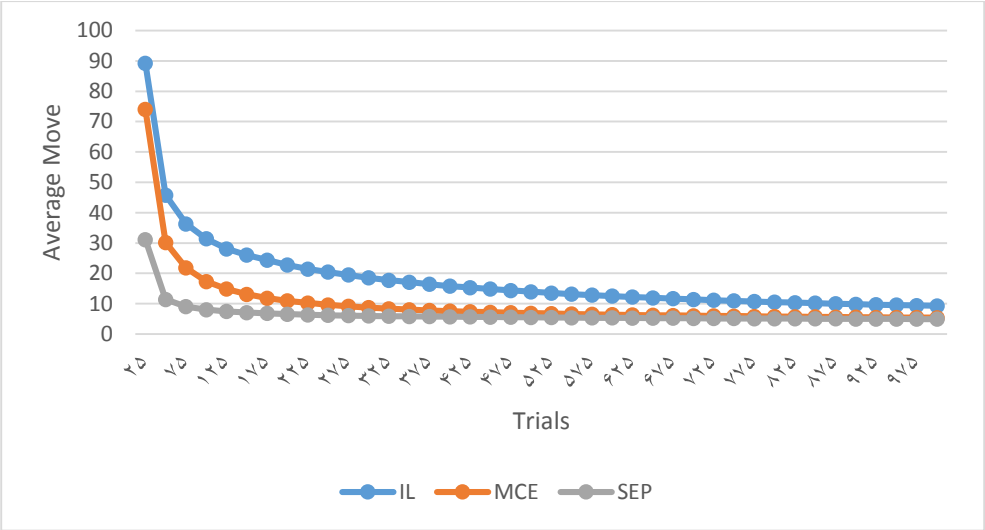
شکل ۴-۱۸: شبه کد الگوریتم یادگیری مشارکتی با بهره گیری از کوتاهترین فاصله تجربه شده

در شکل ۴-۱۸ شبه کد روش پیشنهادی آورده شده است در تشریح فاز همکاری از دو معیار شوک و SEP استفاده‌شده است. بهره‌گیری از معیار SEP باهدف تقسیم‌کار است. در هر حالت از محیط بسته به اینکه سیاست استخراج‌شده از جدول Q عامل‌ها باسیاست استخراج‌شده از SEP هماهنگ باشد یا نه عامل‌ها به دو گروه تقسیم می‌شوند. این تقسیم‌کار به این منظور است که عده‌ای به سیاست جدول Q بها دهند و عده دیگر بیشتر به سیاست استخراج‌شده از SEP؛ اما بعد از تقسیم عامل‌ها همان‌طور که در بخش ۴-۳- آمده است ترکیب داده‌ها بر اساس معیار شوک انجام می‌شود. این بهره‌گیری از معیار شوک به دلیل ماهیت این معیار است. ماهیت این مقدار ارزشمند بودن

مقادیر جدول Q را نشان می‌دهد. درنتیجه ترکیب بر اساس این معیار می‌تواند ترکیب مؤثری باشد. در ادامه با ارائه نتایج آزمایش‌ها اثر روش پیشنهادی را در عمل نیز بررسی می‌نماییم. در این آزمایش‌ها از محیط‌های تشریح شده در بخش ۳-۷- بهره برده شده است.

٤-٤-١- آزمایش اول: بررسی عملکرد روش پیشنهادی در مقایسه باکارهای گذشته

آزمایش اول باهدف نمایش عملکرد روش پیشنهادی در مقایسه به روش‌های دیگر انجام‌شده است. این آزمایش در دو محیط صید و صیاد و پلکان مارپیچ انجام‌شده است. آزمایش‌ها در محیط پلکان مارپیچ از ۲۰۰ چرخه یادگیری تشکیل‌شده‌اند؛ تعداد تلاش عامل‌ها در هر چرخه یادگیری مشترک ۵ در نظر رفته شده که مجموعاً عامل‌ها ۳۰۰۰ چرخه یادگیری انجام خواهند داد.



شکل ۴-۱۹: نمودار اجرا در محیط پلکان مارپیچ با تعداد تلاش برابر عامل‌ها

در شکل ۴-۱۹ می‌توان نمودار میانگین مجموع تعداد قدم‌های عامل‌ها آمده است همان‌طور که مشخص است روش پیشنهادی از سرعت و دقت بالاتری برخوردار است. و این اختلاف در سرعت و دقت را می‌توان در مراحل اول آموزش دید. ازآنجایی‌که روش‌های دیگر در فازهای اولیه یادگیری بدون هیچ اطلاعاتی حرکت می‌کنند زمان زیادی را از دست می‌دهند که این از دست رفتن زمان باعث شده سرعت و کیفیت یادگیری کمتری نسبت به روش SEP داشته باشند. جدول ۴-۸ نتایج دقیق این آزمایش به همراه درصد بهبود نسبت به روش تک عاملی آورده شده است.

جدول ۴-۸: اجرا در محیط پلکان مارپیچ با تعداد تلاش برابر عامل‌ها

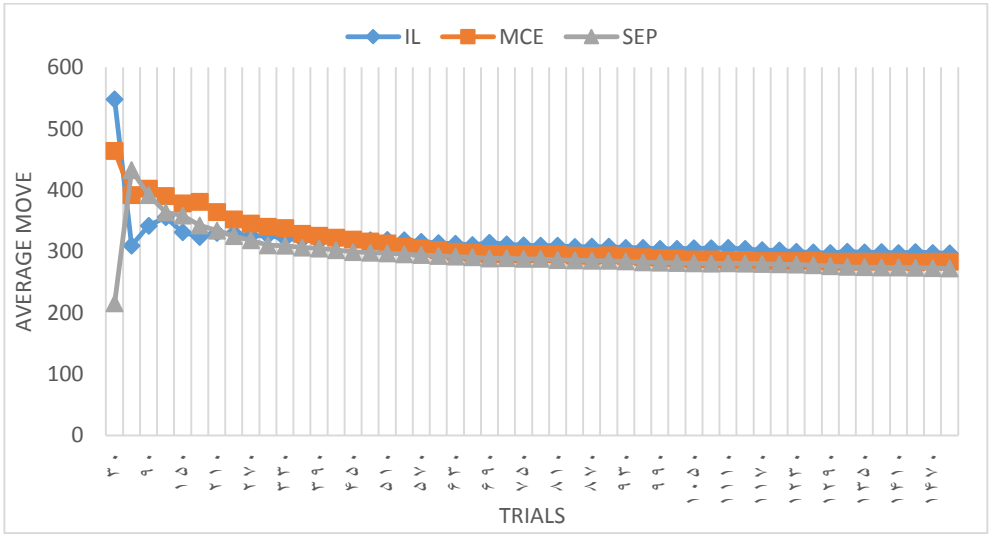
	IL	MCE	SEP
کیفیت	۹,۱۶	۵,۳۷	۴,۹۰
درصد بهبود	–	٪۶۰	٪۷۰
سرعت	۱۷۶۰۱	۱۱۱۸۲	۸۷۲۴
درصد بهبود	–	٪۸۳	٪۱۰۰

برای اطمینان از عملکرد روش پیشنهادی، در محیط صید و صیاد نیز موردبررسی قرار گرفت. تعداد چرخه‌های یادگیری در این روش ۱۵۰ در نظر گرفته‌شده است؛ تعداد چرخه‌های یادگیری مستقل ۱۰ در نظر گرفته‌شده است؛ یعنی مجموعاً سه عامل ۴۵۰۰ چرخه یادگیری خواهند داشت.

جدول ۴-۹: اجرای روش پیشنهادی در محیط صید و صیاد

	IL	MCE	SEP
کیفیت	۲۹۵,۶۴	۲۸۳,۳۳	۲۷۲,۱۲
درصد بهبود	–	٪۴	٪۸
سرعت	۴۶۵۳۲	۴۶۳۲۸	۴۴۵۲۸
درصد بهبود	–	٪۰,۴	٪۴

در جدول ۴-۹ نتیجه اجرا در محیط صید و صیاد آورده شده است. باید تأکید شود نمودار برحسب چرخه‌های یادگیری مشارکتی رسم شده است. همان‌طور که در شکل ۴-۲۰ مشخص است روش پیشنهادی نسبت به روش‌های دیگر دارای برتری است اما درصد بهبود روش نسبت به محیط پلکان مارپیچ به‌شدت پایین ترست. این پایین‌تر بودن ناشی از پیچیدگی محیط است. مشخصاً برای یادگیری دقیق این محیط به چرخه‌های بسیاری نیاز خواهد بود.



شکل ۴-۲۰: نمودار اجرای روش پیشنهادی در محیط صید و صیاد

٤-٤-٢- آزمایش دوم: بررسی عملکرد روش پیشنهادی با تعداد تلاش‌های متفاوت

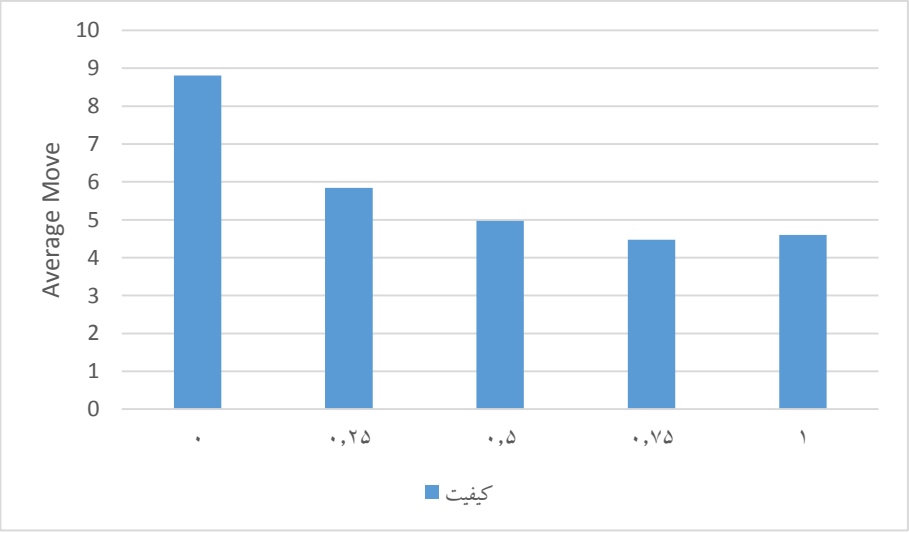
در جدول ۴-۱۰ می‌توان نتایج اجرای روش پیشنهادی را با تعداد تلاش‌ها متفاوت دید؛ همان‌طور که مشخص است روش پیشنهادی نسبت به تعداد تلاش‌ها مقاوم‌تر از روش‌های دیگر است. دلیل این موضوع را می‌توان در نحوه فعالیت روش پیشنهادی جستجو کرد. ترکیب SEP در فاز همکار تنها با یک میانگین‌گیری انجام می‌شود و این میانگین‌گیری و ترکیب اثر تعداد تلاش‌های متفاوت را رعایت می‌کند. به این صورت که عامل با تعداد تلاش بیشتر مقدارهای کوچک‌تری را شناسایی کرده پس اثر بیشتری در SEP نهایی خواهد داشت. پس در تولید جدول SEP تعداد تلاش‌ها اثری نخواهد داشت و این بی‌اثر بودن تعداد تلاش‌ها باعث عملکرد بهتر این روش بوده است.

جدول ۴-۱۰: اجرا در محیط پلکان مارپیچ با تعداد تلاش متفاوت عامل‌ها

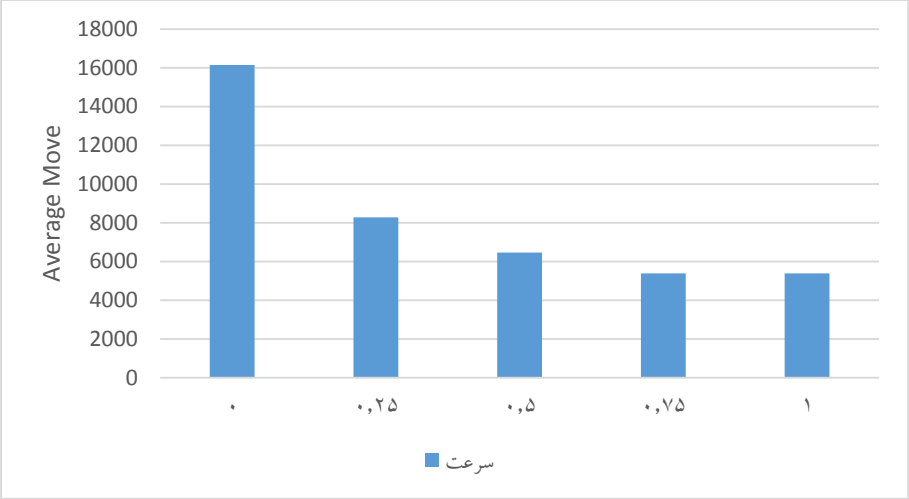
	IL	MCE	SEP
کیفیت	۱۲,۴۴	۱۰,۰۶	۴,۳۸
درصد بهبود	–	٪۲۳	٪۱۵۷
سرعت	۴۳۸۹	۲۴۷۰	۱۲۰۹
درصد بهبود	–	٪۷۷	٪۲۶۳

٤-٤-٣- آزمایش سوم: بررسی اثر افزایش پارامتر μ در عملکرد روش پیشنهادی

μ مهم‌ترین پارامتری است که در این روش به یادگیری مشارکتی اضافه‌شده است. تنظیم این معیار بسیار وابسته به محیط و هدف یادگیری است. در آزمایش‌های مربوط به یادگیری مستقل تحلیل این پارامتر انجام‌شده است. در اینجا باید تأکید شود که این معیار در یادگیری مشارکتی هم همانند یادگیری مستقل عمل می‌کند.



شکل ۴-۲۱: اثر افزایش پارامتر μ در کیفیت روش پیشنهادی



شکل ۴-۲۲: اثر افزایش پارامتر μ در شرعت روش پیشنهادی

جدول ۴-۱۱ :اثر افزایش پارامتر μ در عملکرد روش پیشنهادی

μ	۰,۰۰	۰,۲۵	۰,۵۰	۰,۷۵	۱,۰۰
کیفیت	۸,۸۱	۵,۸۴	۴,۹۷	۴,۴۷	۴,۶۰
سرعت	۱۶۱۴۶	۸۲۷۸	۶۴۵۵	۵۳۸۰	۵۳۸۲

٤-٤-٤- آزمایش چهارم: بررسی اثر افزایش دمای تابع بولتزمن در عملکرد روش پیشنهادی

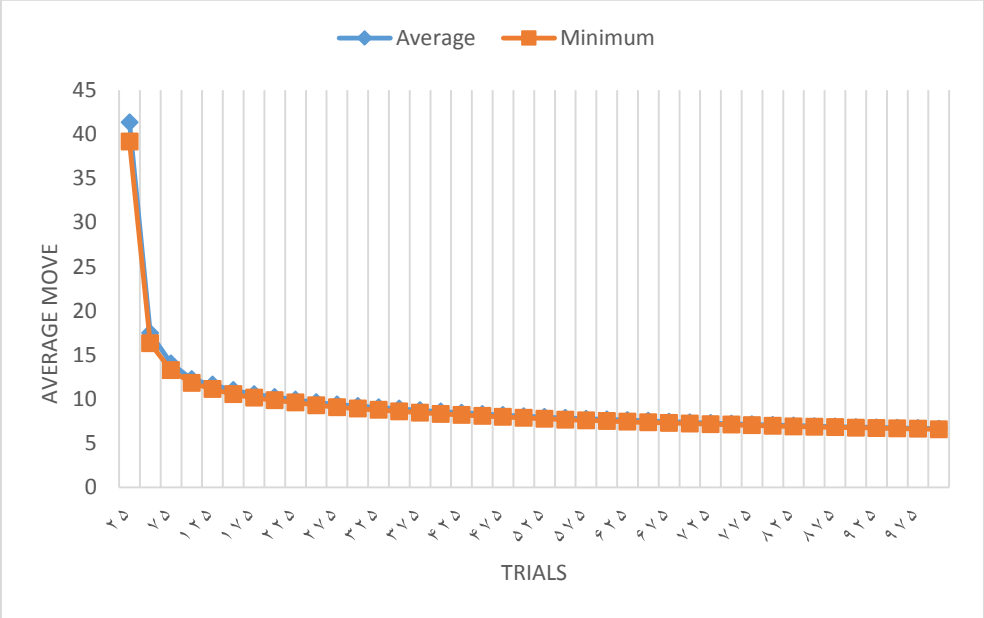
در جدول ۴-۱۲ نتایج اجرای روش پیشنهادی در محیط پلکان مارپیچ با مقادیر مختلف پارامتر دما آمده است. همان‌طور که مشخص است حساسیت روش پیشنهادی نسبت به τ_2 بشدت بیشتر از حساسیت روش نسبت به τ_1 است. این حساسیت ناشی از پایین بودن نسبت اختلاف بین مقادیر جدول SEP است. پیشنهاد می‌شود این مقدار معکوس تعداد حالت‌های محیط در نظر گرفته شود. به این معنی که هر چه تعداد حالت‌های محیط بیشتر باشند مقدار τ_2 کمتر در نظر گرفته شود. تحلیل کامل حساسیت پارامتر دما در بخش یادگیری مستقل آورده شده است.

جدول ۴-۱۲: اثر افزایش دمای تابع بولتزمن در عملکرد روش پیشنهادی

		۰,۱۰	۰,۲۵	۰,۵۰	۰,۷۵	۱,۰۰
τ_1	کیفیت	۴,۵۲	۴,۴۷	۵,۰۶	۵,۲۳	۵,۴۰
	سرعت	۵۸۱۵	۶۰۰۲	۶۶۵۳	۶۵۲۹	۶۷۳۳
τ_2	کیفیت	۶,۳۹	۸,۸۳	۱۰,۶۶	۱۱,۹۱	۱۲,۲۸
	سرعت	۸۴۷۵	۱۲۸۱۹	۱۷۳۹۰	۱۹۹۴۸	۲۰۶۲۸

٤-٤-٥- آزمایش پنجم: بررسی معیار میانگین فاصله تجربه شده

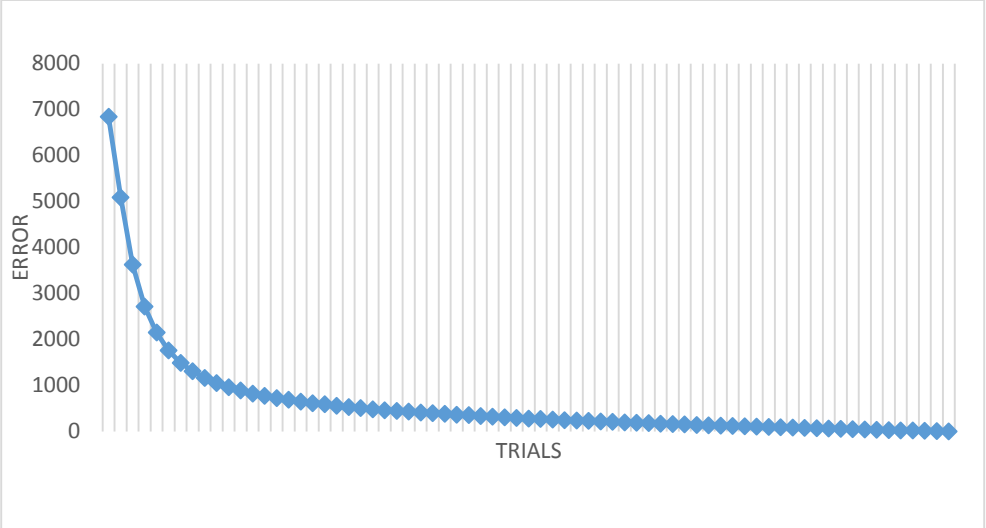
می‌توان معیار های مشابه با کوتاهترین فاصله تجربه شده نیز ارائه کرد. معیاری که در اینجا مورد بررسی قرار گرفت میانگین فاصله تجربه شده است. نتیجه آزمایش نشان می‌دهد که اثر این معیار نیز بسیار نزدیک به معیار کوتاهترین فاصله تجربه شده است. دلیل این موضوع در سرعت بروزرسانی دو روش است؛ سرعت بروزرسانی در این دو معیار برابر و سریع‌تر از بروزرسانی جدول Q است و این باعث جلوگیری از اعمال بی‌اثر در گام های اولیه یادگیری می‌شود.



شکل ۴-۲۳: بررسی معیار میانگین فاصله تجربه شده

٤-٤-٦- آزمایش ششم: همگرایی روش پیشنهادی

موضوع همگرایی روش پیشنهادی بر اساس تعریف یادگیری تقویتی اثبات شده است چرا که در روش پیشنهادی فقط بر روی روش اکتشاف و بهره برداری کار شده و نحوه بروزسانی جدول Q تغییری نداشته است با این وجود در یک آزمایش میزان اختلاف جدول Q عامل در طول یادگیری با جدول Q نهایی مورد بررسی قرار گرفت همانطور که در شکل ۴-۲۴ مشخص است این نمودار به خوبی همگرایی روش پیشنهادی به Q واقعی را نشان میدهد.



شکل ۴-۲۴:همگرایی روش پیشنهادی

٤-٥- نتیجه‌گیری

بیاچی در [۲۸] مفهوم مکاشفه در یادگیری را مطرح نمودند. ایشان اثبات درستی این موضوع را نیز ارائه نمودند. پاکیزه در سال ۱۳۹۱ در ادامه با بهره‌گیری از مکاشفه توانست روش کارایی در جهت بهبود یادگیری مشارکتی ارائه کنند. در این پژوهش نیز مهم‌ترین موضوعی که بهبود چشم‌گیری در روال کار ایجاد نموده همین مکاشفه در یادگیری است. تابع مکاشفه‌ای که درروش پیشنهادی استفاده شده بر اساس کوتاه‌ترین مسیر تجربه شده عمل می‌نماید. این تابع در ترکیب با جدول Q می‌تواند اثر بالایی در بهبود یادگیری تقویتی داشته باشد. نکته مهمی که باید در نظر گرفته شود اضافه شدن دو پارامتر جدید به یادگیری مشارکتی است. پارامتر اول میزان بهره‌گیری از تابع مکاشفه را کنترل می‌کند و با توجه به هدف یادگیری می‌تواند تنظیم شود. پارامتر دوم دمای تابع بولتزمن بر روی جدول مکاشفه را کنترل می‌کند این پارامتر از حساسیت بالایی برخوردار است و در صورت تنظیم نادرست این پارامتر عملکرد روش پیشنهادی به‌شدت کاهش پیدا می‌کند. بخش دیگری که درروش پیشنهادی سعی کرده تا حدی حساسیت‌های پارامترها را کنترل کند تقسیم‌کار بین عامل‌هاست. در این روش عامل‌هایی که اطلاعات جدول Q و جدول SEP آن‌ها یکسان هست در یک گروه و عامل‌های دیگر در گروه دیگری قرار می‌گیرند. به این صورت می‌توان گفت هر دسته از عامل‌ها برای بررسی یک دسته از اطلاعات فرستاده خواهند شد. آزمایش‌ها انجام شده نشان از عملکرد مؤثر روش پیشنهادی داشته است.

فصل ۵

نتیجه‌گیری

در این پژوهش روشی مؤثر جهت بهبود یادگیری مشارکتی در سیستم‌های چند عاملی ارائه شد. در این پژوهش سعی شد با شناخت نقاط بحرانی یادگیری مشارکتی و ارائه روش‌هایی هر یک از نقاط را تا حدودی بهبود بخشیده شود. نقطه اول که بهبود داده شد انتخاب اعمال در یادگیری مستقل بود؛ این کار با بهره‌گیری از مکاشفه در یادگیری انجام گرفت. بیانچی در [۲۸] مکاشفه در یادگیری را مطرح کرده و اثبات می‌کند که اگر مقادیر تابع مکاشفه در یک بازه محدود شده باشند یادگیری مشارکتی به‌خوبی عمل می‌کند و شروط همگرایی یادگیری تقویتی حفظ می‌شود. جهت اثبات روش پیشنهادی می‌توان از کار بیانچی وام گرفت.

تابع مکاشفه استفاده شده در این پژوهش تابعی است که با معیار کوتاه‌ترین مسیر تجربه شده عمل می‌کند. این معیار در بازه ۰ تا تعداد حالت‌ها تعریف‌شده است؛ و این محدود شدن در جهت رعایت شرط بهره‌گیری از مکاشفه در یادگیری توسط بیانچی استفاده‌شده است. نقطه دوم که در این کار مورد بررسی و بهبود قرار گرفت ترکیب داده‌ها است. در این فاز نیز تنها یک میانگین‌گیری هدفمند بین جدول‌های Q عامل‌ها انجام‌شده که تناقضی با شرایط الگوریتم‌های مشارکتی ندارد. نقطه سوم مورد بررسی تقسیم‌کار بین عامل‌هاست که در این پژوهش تقسیم‌کار بر اساس عامل‌های موافق با SEP و عامل‌های مخالف SEP انجام‌شده است. جمع‌بندی تمام این روش‌ها توانسته اثر مؤثری در سرعت و دقت یادگیری مشارکتی داشته باشد.

در ادامه تحلیل روش پیشنهادی ارائه می‌گردد، سپس پس از آن کارهای ناموفق انجام‌شده تشریح می‌شود تا توسط افراد دیگر تکرار نشود؛ نهایتاً در بخش نهایی پیشنهادهایی جهت کارهای آتی ارائه خواهد شد.

۵-۲- نوآوری‌های تحقیق

در این پژوهش با ارائه راه‌حل‌هایی سعی شد تا حد امکان کارایی روش‌های یادگیری مشارکتی افزایش داده شود. هدف اصلی این پژوهش بالا بردن سرعت و کیفیت یادگیری تعریف شد بود. در جهت میل به این مهم دو معیار جدید معرفی شد. معیار اول که شوک نامیده شد یک معیار محلی است که تعداد دفعاتی که در بروز رسانی یک خانه جدول Q از اثر حالت‌های هدف تأثیر گرفته را نشان می‌دهد. ازآنجایی‌که زمان رسیدن اثر پاداش حالت‌های جذب رابطه مستقیم با فاصله هر حالت از حالت جذب دارد، این معیار می‌تواند به‌خوبی نشان‌دهنده میزان ارزشمندی دادهای داخل هر حالت جدول Q باشد.

معیار دیگری که ارائه شد حداقل فاصله تجربه‌شده بود که در روش پیشنهادی دومرتبه مورداستفاده قرار گرفت. از این معیار می‌توان جهت اندازه‌گیری میزان خبرگی عامل‌ها و ورودی تابع مکاشفه بهره برد. در این پژوهش در قسمت انتخاب عمل به عنوان ورودی تابع مکاشفه مورداستفاده قرار گرفت؛ طبق آزمایش‌هایی که در محیط تک عامل صورت گرفت بسیار مؤثر بود. این مکاشفه با یک پارامتر کنترل می‌شود؛ می‌توان با تنظیم میزان بهره‌برداری از تابع مکاشفه الگوریتم یادگیری را با محیط تطبیق داد. استفاده دیگری که از این معیار شد در بخش ترکیب داده‌ها بود که باعث ایجاد یک تقسیم‌کار مؤثر بین عامل‌ها شد.

۵-۳- نتایج نهایی

همان‌طور که گفته شد یادگیری مشارکتی توسط بیانچی جهت بالا بردن سرعت یادگیری که بزرگ‌ترین عیب یادگیری تقویتی است ارائه شد. بعد از آن کارهای دیگری چون تقلید، پنددهی و خبرگی ارائه شد که هر یک در عین مؤثر بودن در بالا بردن سرعت و کیفیت یادگیری خود پیش‌زمینه‌ای برای کارهای بعدی شدند. آخرین کار در این زمینه در سال ۱۳۹۱ توسط پاکیزه و همکاران صورت گرفت که بسیار مؤثر بود. ایشان توانست با ترکیب معیارهای خبرگی یک راهکار مؤثر در ترکیب اطلاعات ارائه کند.

شروع کار این پژوهش نیز باهدف رسیدن به روش جدیدی در ترکیب داده‌ها بود که بعداً با شناسایی سه‌نقطه بحرانی در یادگیری مشارکتی به‌جز ترکیب داده‌ها، تقسیم‌کار بین عامل‌ها و یادگیری مستقل نیز بررسی و بهبود داده شد که مجموع این سه بهبود منجر به روش SEP شد.

در این روش دو پارامتر به یادگیری مشارکتی افزوده شد. پارامتر اول که میزان بهره‌برداری از تابع مکاشفه را کنترل می‌کند از حساسیت پایینی برخوردار است و با توجه به جنس محیط تنظیم می شود؛ پارامتر دوم که پارامتر دمای

تابع مکاشفه است از حساسیت بالایی برخوردار است. این حساسیت بالا به دلیل نوع دادهای جدول SEP است. ازآنجایی‌که این داده‌ها از نوع صحیح هستند نسبت اختلاف آن‌ها کم بوده و این باعث نزدیکی احتمال بهترین حرکت و حرکت‌های دیگر می‌شود. در آزمایش‌ها نشان داده شد که این پارامتر را باید بر اساس تعداد حالت‌های محیط تنظیم نمود. در این پژوهش آزمایش‌هایی نیز باهدف نمایش خصوصیات روش صورت گرفت که نشان از تأثیر مثبت این روش بر یادگیری مشارکتی دارد.

۵-٤- تجربه‌های ناموفق

معمولاً در هر پژوهش ایده‌هایی در نظر گرفته می‌شود که قبل از پیاده‌سازی بسیار کارا به نظر می‌رسد اما بعد از اجرا نشان‌دهنده تأثیر قابل‌قبول در جهت هدف پروژه نیستند. این عدم موفقیت‌ها با وجود اینکه فرایند پژوهش را به تأخیر می‌اندازد باعث شناخت هر چه بیشتر مسئله توسط پژوهشگر است. در این پژوهش نیز ایده‌هایی در نظر گرفته‌شده که به نتیجه‌ی قابل‌قبول منجر نشده‌اند. در ادامه سعی شده با ارائه این ایده‌ها از تکرار آن‌ها جلوگیری شود.

۵-٤-۱- استفاده از معیار شوک در یادگیری مشارکتی مبتنی بر خبرگی

روش WSS که توسط نیلی احمدآبادی و همکاران در [۱۵] ارائه شد. در این روش ترکیب اطلاعات عامل‌ها بر اساس یک معیار خبرگی سنجیده می‌شد. ایشان شش معیار جهت سنجش خبرگی عامل ارائه نموده اند. در فاز همکاری WSS عامل‌ها جدول Q تمام عامل‌های خبره‌تر از خود را دریافت نموده و یک میانگین‌گیری وزن‌دار بین جداول انجام داده و جدول تولیدشده را جایگزین جدول خود می‌نمایند.

در این پژوهش نیز سعی شد از پارامتر شوک به‌عنوان جایگزین معیارهای خبرگی استفاده شود؛ که نتایج آزمایش‌ها نشان از عدم تأثیر این پارامترها بود. نتایج نشان داده که استفاده از این معیار حداکثر می‌تواند کیفیت یادگیری در حد معیارهای خبرگی دیگر داشته باشد.

۵-٤-۲- استفاده از معیار شوک جهت میانگین‌گیری محلی

درروش WSS برای ترکیب کل اطلاعات عامل‌ها از یک ضریب خبرگی که بر اساس مقدار خبرگی عامل در کل محیط یادگیری ایجاد می‌شد استفاده شده بود. یکی از اقدامات ناموفقی که در این پژوهش انجام شد بهره‌گیری محلی از معیار شوک است؛ به این صورت که در بخش ترکیب داده‌ها اطلاعات هر حالت به‌صورت جداگانه

میانگین‌گیری می‌شد. ضرایب این میانگین‌گیری نیز بر اساس پارامتر شوک تهیه شد؛ اما متأسفانه این روش نیز برای بالا بردن کارایی یادگیری مشارکتی مؤثر نبود.

۵-٤-۳- استفاده از معیار کوتاه‌ترین فاصله تجربه‌شده در WSS

تجربه دیگری که در جهت بهبود یادگیری مشارکتی در این پژوهش انجام شد بهره‌گیری از SEP در WSS است. ازآنجایی‌که مقادیر SEP را نمی‌توان به‌صورت پارامتری برای عامل در نظر گرفت معیاری تحت عنوان میزان هماهنگی در نظر گرفته شد؛ که برای هر عامل میزان هماهنگی جدول Q و جدول SEP محاسبه می‌شد. سپس از این معیار به‌عنوان معیار خبرگی عامل در WSS بهره برده شده که متأسفانه این راهکار نیز در جهت بهبود یادگیری مشارکتی مؤثر نبود.

۵-۵- پیشنهادهایی جهت کارهای آتی

همانند هر طرح پیشنهادی پژوهش پیش رو نیز چالش‌های و ایده جدیدی ایجاد کرده که هر یک می‌توانند هدف مناسبی برای پژوهش‌های آتی باشند. در ادامه سعی شده است با ارائه پیشنهادهایی پژوهش‌های این زمینه را به سمت رشد هدایت کند.

۵-۵-۱- پیشنهاد اول: تعادل در بهره‌گیری از حداقل فاصله تجربه‌شده

بزرگ‌ترین چالش در یادگیری تقویتی را می‌توان رسیدن به یک تعادل در اکتشاف و بهره‌برداری در نظر گرفت. در این پژوهش نیز سعی شده با ارائه معیاری به برآورده شدن این هدف کمک شود؛ اما این پیشنهاد چالش جدیدی در یادگیری مشارکتی ایجاد کرده که می‌تواند هدف مناسبی برای پژوهش‌های آتی باشد. پیشنهاد می‌شود هدف کارهای آینده را مشخص کردن حدی برای بهره‌گیری از معیار حداقل فاصله تجربه‌شده و اطلاعات عامل تعریف نمود.

این کار با مشخص کردن مقدار پارامتر μ انجام خواهد شد. می‌توان این پارامتر را متغیر در نظر گرفت. پیشنهاد می‌شود که از معیار شوک در جهت مقداردهی به این متغیر استفاده شود. چراکه با مقدار شوک می‌توان به مقدار مفید بودن اطلاعات عامل پی برد؛ و احتمالاً برقراری یک رابطه معکوس بین شوک و μ می‌تواند در بالا بردن کارایی روش‌ها مؤثر باشد.

۵-۵-۲- پیشنهاد دوم: تقسیم‌کار مناسب

موضوع مهمی که در یادگیری مشارکتی کمتر در نظر گرفته‌شده است موضوع تقسیم‌کار بین عامل‌هاست که می‌تواند به مقدار زیادی از انجام کارهای تکراری جلوگیری نماید در این پژوهش سعی شد تا حدودی این موضوع در نظر گرفته شود. مستورعشق نیز در [۳۰] سعی کرده با تقسیم‌بندی فیزیکی محیط بین عامل‌ها تقسیم‌کار را انجام دهد که نتایج خوبی هم داشته است؛ اما در تقسیم‌بندی فیزیکی نیاز به شناخت محیط است که در محیط‌هایی که از یادگیری مشارکتی استفاده می‌شود معمولاً شناختی از محیط وجود ندارد.

پیشنهاد می‌شود جهت تقسیم‌کار بین عامل‌ها از یک تقسیم‌بندی پویا استفاده شود. این تقسیم‌بندی پویا می‌تواند بر اساس حالت‌های محیط و یا اعمال عامل باشد. در تقسیم‌بندی محیط کاری که می‌توان کرد پویا کردن احتمال شروع عامل‌ها از هر حالت است یعنی در شروع هر چرخه احتمال قرار گیری عامل در هر حالت برابر نباشد. در حالت عادی در هر چرخه یادگیری عامل با احتمال مساوی می‌تواند از هریک از حالت‌ها چرخه یادگیری را شروع کند. پیشنهاد می‌شود که احتمال شروع به‌صورت پویا تنظیم شود؛ اما این تنظیم احتمال می‌تواند بر اساس مسیرهای استخراج‌شده از حداقل فاصله تجربه‌شده باشد.

روش دیگر دسته‌بندی که می‌توان به‌عنوان یک پژوهش در نظر گرفته شود تقسیم‌کار بر اساس اعمال است که باعث می‌شود هر عامل نسبت به عواقب هر عمل آشنایی پیدا کند. جهت انجام این کار پیشنهاد می‌شود که با استفاده از یک تابع مکاشفه احتمال انتخاب اعمال توسط هر عامل کنترل شود.

۵-۵-۳- پیشنهاد سوم: تولید معیاری جهت سنجش میزان شک و یقین در عامل

شک را می‌توان یکی از مهم‌ترین دلایلی دانست که انسان را به پژوهش وامی‌دارد و یقین را نیز می‌توان دلیلی برای اتمام کار هر پژوهشی در نظر گرفت. جهت کارهای آتی پیشنهاد می‌شود که با وام گرفتن از شک و یقین در انسان در جهت تولید دو معیار محلی جهت سنجش معیار شک و یقین عامل‌ها کار شود. با داشتن چنین معیارهایی می‌توان به‌خوبی ترکیب داده‌ها و تقسیم‌کار را در عامل‌ها هدفمند کرد.

۵-۵-٤- پیشنهاد چهار: تهیه معیارهایی مشابه معیار کوتاهترین فاصله تجربه شده

یک معیار که به دفعات در روش پیشنهادی مورداستفاده قرارگرفته؛ اطلاعات مکاشفه‌ای را تولید می‌نمود که بسیار هم در روش پیشنهادی مورد استفاده قرار گرفت. می‌توان معیارهای مشابه را نیز در یادگیری مورداستفاده قرارداد. معیارهایی مثل بیشترین پاداش تجربه‌شده ممکن است اثرات قابل توجهی داشته باشد. این معیار در محیط می‌تواند همانند معیار کوتاه‌ترین فاصله تجربه‌شده بسیار مفید باشد.

مراجع:

[1] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed., vol. 25. Citeseer, 2010.

[2] M. Wooldridge, *An introduction to multiagent systems*. John Wiley & Sons, 2009.

[3] T. Mitchell, *Machine learning*. McGraw-Hill, Inc., New York, NY, 1997.

[4] S. D. Whitehead and D. H. Ballard, *A study of cooperative mechanisms for faster reinforcement learning*. University of Rochester, Department of Computer Science Rochester, NY, 1991.

[5] L. Nunes and E. Oliveira, "Advice-exchange amongst heterogeneous learning agents: Experiments in the pursuit domain," *poster Abstr. Auton. Agents Multiagent Syst.*, 2003.

[6] H. R. Berenji and D. Vengerov, "Cooperation and coordination between fuzzy reinforcement learning agents in continuous state partially observable Markov decision processes," in *FUZZ-IEEE'99. 1999 IEEE International Fuzzy Systems. Conference Proceedings (Cat. No.99CH36315)*, 1999, vol. 2, pp. 621–627 vol.2.

[7] ع. پاکیزه حاجی یار, "یادگیری مشارکتی بر مبنای خبرگی چند معیاره در سیستم های چند عامله," دانشگاه صنعتی اصفهان, ۱۳۹۰.

[8] S. D. Whitehead, "A Complexity Analysis of Cooperative Mechanisms in Reinforcement Learning.," in *AAAI*, 1991, pp. 607–613.

[9] M. Tan, "Multi-Agent Reinforcement Learning: Independent vs. Cooperative Agents," *Proc. Tenth Int. Conf. Mach. Learn.*, pp. 330–337, 1993.

[10] T. Yamaguchi, M. Miura, and M. Yachida, "Multi-agent reinforcement learning with adaptive mimetism," in *Proceedings 1996 IEEE Conference on Emerging Technologies and Factory Automation. ETFA '96*, 1996, vol. 1, pp. 288–294.

[11] A. Garland and R. Alterman, "Multiagent learning through collective memory," in *Adaptation, Coevolution and Learning in Multiagent Systems: Papers from the 1996 AAAI Spring Symposium*, 1996, pp. 33–38.

[12] L. Nunes and E. Oliveira, "On Learning by Exchanging Advice," *Proc. Artif. Intell. Simul. Behav. Conv. Symp. Adapt. agents multi-agent Syst. (AISB/AAMAS-II), Imp. Coll. london*, vol. cs.LG/0203, pp. 583–599.

[13] L. Nunes and E. Oliveira, "Cooperative Learning Using Advice Exchange," E. Alonso, D. Kudenko, and D. Kazakov, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 33–48.

[14] L. Nunes and E. Oliveira, "Advice-Exchange Between Evolutionary Algorithms and Reinforcement Learning Agents: Experiments in the Pursuit Domain," D. Kudenko, D. Kazakov, and E. Alonso, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 185–204.

[15] M. N. Ahmadabadi, M. Asadpur, S. H. Khodanbakhsh, and E. Nakano, "Expertness measuring in cooperative learning," in *Proceedings. 2000 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2000) (Cat. No.00CH37113)*, 2000, vol. 3, pp. 2261–2267.

[16] T. Yamaguchi, Y. Tanaka, and M. Yachida, "Speed up reinforcement learning between two agents with adaptive mimetism," in *Proceedings of the 1997 IEEE/RSJ International*

Conference on Intelligent Robot and Systems. Innovative Robotics for Real-World Applications. IROS '97, 1997, vol. 2, pp. 594–600.

[17] P. Ritthipravat, T. Maneewarn, J. Wyatt, and D. Laowattana, "Comparison and Analysis of Expertness Measure in Knowledge Sharing Among Robots," M. Ali and R. Dapoigny, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 60–69.

[18] M.-R. Akbarzadeh-T, H. Rezaei-S, and M. B. Naghibi-S, "A fuzzy adaptive algorithm for expertness based cooperative learning, application to herding problem," in *22nd International Conference of the North American Fuzzy Information Processing Society, NAFIPS 2003*, 2003, pp. 317–322.

[19] N. Carver and V. Lesser, "Evolution of blackboard control architectures," *Expert Syst. Appl.*, vol. 7, no. 1, pp. 1–30, Jan. 1994.

[20] Y. Yang, Y. Tian, and H. Mei, "Cooperative Q Learning Based on Blackboard Architecture," in *2007 International Conference on Computational Intelligence and Security Workshops (CISW 2007)*, 2007, pp. 224–227.

[21] J. W. McManus and W. L. Bynum, "Design and analysis techniques for concurrent blackboard systems," *IEEE Trans. Syst. Man, Cybern. - Part A Syst. Humans*, vol. 26, no. 6, pp. 669–680, 1996.

[22] E. Pakizeh, M. Palhang, and M. M. Pedram, "Multi-criteria expertness based cooperative Q-learning," *Appl. Intell.*, vol. 39, no. 1, pp. 28–40, Jul. 2013.

[23] M. N. Ahmadabadi and M. Asadpour, "Expertness based cooperative Q-learning," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 32, no. 1. pp. 66–76, 2002.

[24] E. Pakizeh, M. M. Pedram, and M. Palhang, "Multi-criteria expertness based cooperative method for SARSA and eligibility trace algorithms," *Appl. Intell.*, vol. 43, no. 3, pp. 487–498, 2015.

[25] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 1998.

[26] R. Patrascu and D. Stacey, "Adaptive exploration in reinforcement learning," in *Neural Networks, 1999. IJCNN'99. International Joint Conference on*, 1999, vol. 4, pp. 2276–2281.

[27] M. Tokic, "Adaptive ε-Greedy Exploration in Reinforcement Learning Based on Value Differences," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 6359 LNAI, 2010, pp. 203–210.

[28] R. A. C. Bianchi and A. H. R. Costa, "The use of heuristics to speedup reinforcement learning," *Bol. Interno, No. BT/PCS*, vol. 409, pp. 125–144, 2004.

[29] C. J. C. H. Watkins and P. Dayan, "Q-learning," *Mach. Learn.*, vol. 8, no. 3–4, pp. 279–292, 1992.

[30] S. M. Eshgh and M. N. Ahmadabadi, "An extension of weighted strategy sharing in cooperative q-learning for specialized agents," in *Neural Information Processing, 2002. ICONIP'02. Proceedings of the 9th International Conference on*, 2002, vol. 1, pp. 106–110.

speed-up cooperative learning in multi-agent systems using shortest experimented path

Mohammad ali mirzaei badizi

ma.mirzaei@ec.iut.ac.ir

Department of Electrical and Computer Engineering

Isfahan University of Technology, Isfahan 84156-83111, Iran

Degree: M.Sc. Language: Farsi

Supervisor: Prof. Maziyar palhang (*palhang@cc.iut.ac.ir***)**

Abstract

Intelligent world and systems were a dream in the past,but by growing artificial intelligence field it in becoming a reality. the main factor in a intelligent system is learning and artificial intelligence makes it possible. mutiagent system learning becomes more accurate and faster by combining machine learning methods with them. multi agent learning include Cooperative and Competitive methods. In Competitive learning agents try to increase their utility however other's utility may be decreased. In cooperative learning agents try to increase utility of all agents simultaneously.

In recent years many works have been performed in cooperative learning, Most of these methods used reinforcement learning for learning. While these methods have a main challenge in how to combine the knowledge of agents.

In this thesis we have addressed some of those challenges to improve Cooperative learning methods. to achieve this objective , some main points of cooperative learning have been detected, the first point is action selection in reinforcement learning which we used a new heuristic function to select actions, the second and third points are combining knowledge and task division by two criteria "shortest experienced path" and "Shock". By using these two criteria combine the knowledge has been improved. Overall experiments showed improve in quality and learning speed.

Key Words: Cooperative learning, Multi-agent system,Reinforcement learning

Isfahan University of Technology

Department of Electrical and Computer Engineering

speed-up cooperative learning in multi-agent systems using shortest experimented path

By

Mohammad ali mirzaei badizi

Evaluated and Approved by the Thesis Committee, on March 21, 2015

1. Maziar Palhang, Associate Prof. (Supervisor)

2. , Prof. (Examiner)

3. . Prof (Examiner)

Mohammad Ali Khosravifard, Department Graduate Coordinator