دانشگاه صنعتی اصفهان

دانشکده مهندسی برق و کامپیوتر

یادگیری مشارکتی بر مبنای خبرگی چندمعیاره در سیستم‌های چندعامله

پایان‌نامه کارشناسی ارشد مهندسی کامپیوتر- هوش مصنوعی و رباتیک

عصمت پاکیزه حاجی‌یار

استاد راهنما

دکتر مازیار پالهنگ

۱۳۸۹

بسم الله الرحمن الرحیم

هست کلید درِ گنج حکیم

دانشگاه صنعتی اصفهان

دانشکده مهندسی برق و کامپیوتر

یادگیری مشارکتی بر مبنای خبرگی چندمعیاره در سیستم‌های چندعامله

پایان‌نامه کارشناسی ارشد مهندسی کامپیوتر- هوش مصنوعی و رباتیک

عصمت پاکیزه حاجی‌یار

استاد راهنما

دکتر مازیار پالهنگ

۱۳۸۹

دانشگاه صنعتی اصفهان

دانشکده برق و کامپیوتر

پایان‌نامه‌ی کارشناسی ارشد رشته‌ی مهندسی کامپیوتر-هوش مصنوعی و رباتیک خانم عصمت پاکیزه حاجی‌یار تحت عنوان

یادگیری مشارکتی بر مبنای خبرگی چندمعیاره در سیستم‌های چندعامله

در تاریخ ۸۹/۱۲/۲۴ توسط کمیته‌ی تخصصی زیر مورد بررسی و تصویب نهایی قرار گرفت.

۱- استاد راهنمای پایان‌نامه	دکتر مازیار پالهنگ
۲- استاد مشاور پایان‌نامه	دکتر میر محسن پدرام
۳- استاد داور	دکتر عبدالرضا میرزایی
۴- استاد داور	دکتر محمدرضا احمدزاده
سرپرست تحصیلات تکمیلی دانشکده	دکتـــــــــــــــــــــــــــــــر

سپاس

تقدیم به:

فهرست مطالب

فهرست اشکال

عنوان **صفحه**

فهرست جداول

عنوان **صفحه**

چکیده

مشارکت کلید اصلی رسیدن به موفقیت در سیستم‌های طبیعی و مصنوعی به شمار می‌رود و از این رو مشارکت در سیستم‌های چندعامله به منظور رسیدن به راه‌حل‌های بهتر ضروری به نظر می‌رسد. اکثر تحقیقات در حوزه یادگیری ماشین نیز بر دو پایه اصلی بهبود کیفیت و کاهش زمان یادگیری متمرکز هستند. انتظار می‌رود که یادگیری مشارکتی چندعامله در مقایسه با یادگیری مستقل عامل‌ها، به دلیل دارا بودن دانش و منابع اطلاعاتی بیشتر به نتایج بهتری از نظر کیفی و سرعت یادگیری دست یابد.

استفاده از استراتژی‌های مشارکت بهتر منجر به افزایش سرعت و کیفیت یادگیری می‌شود. امروزه بیشتر تحقیقات در حوزه یادگیری مشارکتی چندعامله از یادگیری تقویتی به عنوان روش یادگیری پایه خود استفاده می‌کنند. یادگیری تقویتی به دلیل ساختار یادگیری فاقد نظارت و قابلیت یادگیری پیوسته‌اش حتی در محیط‌های پویا، یکی از معتبرترین تکنیک‌های یادگیری ماشین به شمار می‌رود. استفاده از این نوع یادگیری در سیستم‌های چندعامله مشارکتی به هر عامل مستقل این اجازه را می‌دهد که علاوه بر این که از تجربیات خود می‌آموزد، از سایر عامل‌های حاضر در سیستم نیز بیاموزد و بدین ترتیب سرعت یادگیری افزایش یابد.

انسان در طول دوره زندگی تجربیات مختلفی را در بازه‌های زمانی متفاوتی از زندگی‌اش می‌آموزد. گاهی تجربیات فرد به طور کامل موفقیت آمیز هستند و گاهی شکستی کامل محسوب می‌شوند. شخصیت یک فرد بر اساس در نظر گرفتن همه تجربیاتش در کنار هم شکل می‌گیرد. در واقع تصمیم‌های فرد راجع به آینده‌اش بر اساس تجربیات مختلفی که در طول زمان و در جنبه‌های مختلف زندگی بدست آورده است، شکل می‌گیرد. این واقعیت قابل تعمیم به دنیای عامل‌های مصنوعی نیز می‌باشد و ایده اصلی این پایان‌نامه نیز بر این اساس شکل گرفته است. در این پایان‌نامه روش یادگیری مشارکتی جدیدی مبتنی بر خبرگی چندمعیاره معرفی می‌شود که به منظور مشارکت بهتر از همه معیارهای خبرگی تعریف شده برای عامل‌های مصنوعی استفاده می‌کند. برای ارزیابی روش پیشنهادی از دو محیط آموزشی معتبر پلکان مارپیچ و صید و صیاد استفاده شده است. نتایج آزمایش‌ها پتانسیل بالای روش پیشنهادی در تولید یادگیری مشارکتی بهتر را تایید می‌کنند.

کلمات کلیدی: ۱- سیستم‌های چندعامله ۲-یادگیری مشارکتی ۳-خبرگی چندمعیاره ۴-یادگیری تقویتی ۵-انتقال دانش

فصل اول
مقدمه

۱-۱ تعریف مساله و اهمیت آن

سیستم‌های چندعامله یکی از مفاهیم نوپا و توانمند در حوزه هوش مصنوعی به شمار می‌رود که توانایی حل دسته مختلفی از مسائل را دارد. افزودن قابلیت یادگیری به این سیستم‌ها به طرز محسوسی عملکرد سیستم را بهبود می‌بخشد[۲-۱]. یادگیری در این سیستم‌ها، با آن‌چه که به طور معمول در یادگیری ماشین وجود دارد، دارای تفاوت-هایی است و از این رو تحقیقات در این زمینه مورد توجه بسیاری از محققان رشته یادگیری ماشین قرار دارد. یادگیری در این سیستم‌ها به دو دسته کلی یادگیری مشارکتی و یادگیری رقابتی تقسیم می‌شود[۳].

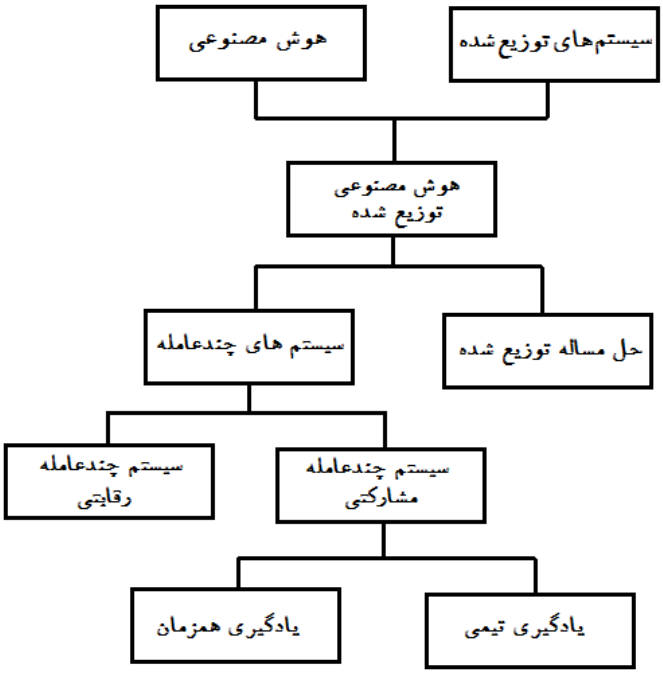
دسته‌ی اول در محیط‌های مشارکتی تعریف می‌شود. در این محیط‌ها، عامل‌ها برای رسیدن به اهدافشان با یکدیگر همکاری می‌کنند و سعی بر این است که در حین این همکاری عملکرد عامل‌ها بهبود یابد. از این گونه یادگیری به یادگیری مشارکتی[1] تعبیر می‌شود. به عبارت دیگر یادگیری مشارکتی بین عامل‌ها گونه‌ای از یادگیری است که به‌واسطه آن عامل‌ها قادر خواهند بود تا در کنارهم و به منظور رسیدن به اهدافی مشترک، با استفاده از تجربیات یکدیگر عملکرد خود را بهبود بخشند.

---

[1] Cooperative Learning

دسته‌ی دوم در محیط‌های رقابتی تعریف می‌شود که در آن هر عاملِ خودخواه[1] تلاش می‌کند تا با یادگیری رفتار سایر عامل‌ها مقدار سودمندی[2] خود را بیشینه کند. این گونه یادگیری، یادگیری رقابتی نامیده می‌شود. در محیط‌های رقابتی به پویایی محیط و این نکته که آیا عامل‌ها به یک تعادل پایدار می‌رسند، توجه خاصی می‌شود. در ساده‌ترین حالت یادگیری رقابتی به صورت بازی‌های تکراری با عامل‌های یادگیرنده[3] است.

تا کنون تقسیم‌بندی‌های مختلفی که نشان‌دهنده مبدا پیدایش مفهوم سیستم‌های چندعامله و هم‌چنین جایگاه یادگیری مشارکتی در آن هستند، ارائه شده است. یکی از معتبرترین آن‌ها، تقسیم‌بندی‌ای است که در [۴] معرفی شده و در شکل۱-۱به صورت ساده نشان داده شده است.



شکل ۱-۱- جایگاه یادگیری مشارکتی در سیستم‌های چندعامله

از دیدگاه‌های مختلفی می‌توان یادگیری مشارکتی در یک سیستم چندعامله را بررسی کرد. انتخاب دیدگاه به طرز تفکر طراح و نحوه بهره برداری از یادگیری در سیستم بستگی دارد. دو دیدگاه کلی موجود در حوزه یادگیری مشارکتی در سیستم‌های چندعامله به صورت زیر است[۴]:

- مشارکت به منظور یادگیری هماهنگی و همکاری برای رسیدن به یک هدف مشترک، در این حالت همه عامل‌ها در یک محیط قرار دارند و مشارکت در یادگیری به منظور آموختن شیوه‌های هماهنگی عامل‌ها برای رسیدن به یک هدف مشترک تعریف شده است.

---

[1] Selfish Agent
[2] Utility
[3] Repeated Games with Learning Agents

- مشارکت به منظور بهبود یادگیری هر یک از عامل‌ها در آموختن یک کار یکسان، در این حالت هر یک از عامل‌ها در محیطی جداگانه و به طور مستقل فرآیند یادگیری یک کار یکسان را انجام می‌دهند و با استفاده از انتقال اطلاعات بین یکدیگر یادگیری اعضای گروه بهبود داده می‌شود.

در دسته دوم عامل‌ها بوسیله ارتباطات مستقیم و یا غیر مستقیم، اطلاعات حاصل از فرآیند یادگیری خود را با دیگران به اشتراک می‌گذارند. در صورتی که انتقال اطلاعات و به کارگیری آن‌ها در طول یادگیری به خوبی تعریف شده باشد، مشارکت قادر است که یادگیری اعضای گروه را بهبود بخشد. در این دسته از روش‌ها عملکرد روش‌های یادگیری مشارکتی به اطلاعاتی که بین عامل‌ها مبادله می‌شود و روشی که بر مبنای آن اطلاعات مبادله شده در طول یادگیری مورد استفاده قرار می‌گیرند، بستگی دارد. انتظار می‌رود در صورتی که اطلاعات مبادله شده کیفیت بهتر و محتوای غنی‌تری داشته باشند، مشارکت در یادگیری در یک گروه از عامل‌های یادگیرنده به نحو چشمگیری، یادگیری مستقل[1] عامل‌ها را بهبود بخشد. به عبارت دیگر در یادگیری مشارکتی بر مبنای انتقال اطلاعات[2] تلاش بر این است که به دو سوال پاسخ مناسبی داده شود: "چه اطلاعاتی باید بین عامل‌ها مبادله شود و عامل‌ها چگونه باید از اطلاعات مبادله شده استفاده کنند؟" دیدگاه‌ها و پاسخ‌های متفاوت محققان به این دو سوال، مبدا پیدایش روش‌های مختلف یادگیری مشارکتی بر مبنای انتقال اطلاعات شده است.

۱-۲ یادگیری مشارکتی در سیستم‌های چندعامله

در جوامع انسانی هیچ فردی همه چیز را از ابتدا و به تنهایی، به صورت درست یاد نگرفته است. در حقیقت انسان‌ها پندپذیر هستند، با یکدیگر مشورت می‌کنند، اطلاعات پردازش نشده دریافت می‌کنند، دیگران را در عمل نگاه می‌کنند و بدین ترتیب با مشاهده فعالیت‌ها و تجربیات سایرین، فرآیند یادگیری خود را به انجام می‌رسانند. در حقیقت می‌توان گفت که افراد مشارکت می‌کنند تا بیشتر و بهتر یاد بگیرند. در علوم اجتماعی مشارکت به عنوان یک عمل جمعی به منظور سود متقابل تعریف شده است[۵].

تعریف ۱-۴- یادگیری یعنی این‌که یک عامل رفتارش را بر اساس تجربه‌های گذشته‌اش، عوض می‌کند[۶].

تعریف ۱-۵- سیستمی که در آن چندین عامل برای رسیدن به یک هدف و یا انجام یک وظیفه‌ی مشترک به هم همکاری می‌کنند، سیستم چندعامله مشارکتی[3] نامیده می‌شود[۴].

---

1 Individual Learning
2 Knowledge Transfer
3 Cooperative Multi-Agent System

تعریف ۱-۶- بهره‌گیری از دانش سایر عامل‌های موجود در یک سیستم چندعامله به‌واسطه مبادله صریح یا ضمنی قوانین یادگرفته شده و اطلاعات جمع‌آوری شده به منظور بهبود عملکرد عامل‌ها، یادگیری مشارکتی نامیده می‌شود[۷].

در سیستم‌های چندعامله مشارکتی شرایطی نظیر جوامع انسانی برقرار است، چندین عامل برای رسیدن به یک هدف و یا انجام یک وظیفه‌ی مشترک با هم همکاری می‌کنند. به عبارت بهتر می‌توان گفت زمانی که چندین فرد یا سازمان مختلف با اهداف و اطلاعات مختلف وجود داشته باشند، برای مدیریت تعاملات بین آن‌ها به یک سیستم چندعامله[1] نیازمندیم. هم‌چنین همیشه در دنیای واقعی با سیستم‌های چندعامله‌های همگن مواجه نخواهیم بود و معمولاسیستمی با حضور تعدادی عامل ناهمگن که هر یک دارای توانایی‌های متفاوتی هستند، در اختیار داریم.

یادگیری جزو توانایی‌های ضروری برای عامل‌های هوشمند به شمار می‌رود و کیفیت عملکرد سیستم چندعامله را به نحو چشمگیری افزایش می‌دهد. افزایش سرعت یادگیری و بهبود کیفیت دانش یاد گرفته شده، جزو اهداف اساسی رشته یادگیری ماشین به شمار می‌روند. ایده‌ی اصلی یادگیری مشارکتی نیز تلاشی در جهت رسیدن به اهداف یادگیری ماشین در حوزه سیستم چندعامله است. بهره‌گیری از دانش سایر عامل‌های موجود به‌واسطه مبادله صریح یا ضمنی قوانین یادگرفته شده، اطلاعات جمع‌آوری شده و ...، به بهبود عملکرد عامل‌ها در سیستم چندعامله منجر می‌شود. گاهی انجام کار در یک محیط و به صورت همزمان روی می‌دهد. یعنی همه عامل‌ها همزمان در یک محیط قرار دارند و هدف آن‌ها انجام یک کار مشترک[2] و رسیدن به یک هدف به صورت تیمی است. در چنین شرایطی از یادگیری مشارکتی به منظور بهبود نحوه همکاری بین عامل‌ها و ایجاد هماهنگی بین آن‌ها برای رسیدن به هدف مشترک استفاده می‌شود. گاهی نیز عامل‌ها صرفا به منظور انتقال دانش و استفاده از تجربیات یکدیگر به منظور بهبود عملکردشان در انجام یک کار یکسان با یکدیگر مشارکت می‌کنند. در چنین حالتی عامل‌ها هر کدام در یک محیط جداگانه و مستقل مشغول به یادگیری کاری یکسان هستند. مشارکت در یادگیری[3] نیز یکی از مفاهیمی است که در سیستم چندعامله بیان می‌شود. اگر عامل‌ها قادر باشند از آزمایش‌های خودشان و از دانش و خبرگی سایر عامل‌ها استفاده کنند، آن‌گاه گفته می‌شود که در یادگیری، مشارکت صورت گرفته است. به‌کارگیری روشی مناسب برای مشارکت، می‌تواند کارآیی یادگیری را تا حد زیادی افزایش دهد.

---

[1] Multi-Agent System
[2] Joint Task
[3] Cooperation in Learning

۱-۳ چالش‌های موجود در بررسی یادگیری مشارکتی

مهمترین سوالی که در یک سیستم چندعامله مشارکتی با آن مواجهیم این است که "چگونه یک تیم از عامل‌های یادگیرنده در حل یک مساله می‌توانند برای بهبود عملکردشان با یکدیگر همکاری کنند؟"[۷]

برای پاسخ به این سوال که چگونه عامل‌ها در یادگیری به مشارکت می‌رسند، به دو نکته زیر می‌بایست توجه کرد:

- ✓ اطلاعاتی که بین عامل‌ها مبادله می‌شوند.
- ✓ روشی که بر مبنای آن اطلاعات مبادله شده در طول یادگیری مورد استفاده قرار می‌گیرند.

اطلاعات بدست آمده از یادگیرنده‌های دیگر، می‌تواند به چندین روش زیرموجب بهبود عملکرد شود:

- تسریع فرآیند یادگیری
- قادر ساختن عامل به فرار از بهینه‌های محلی برای معیاری که قصد بهبود آن را داریم.
- کمک به عامل برای یافتن پارامترهای یادگیری مناسب

یکی دیگر از نکاتی که در یادگیری مشارکتی مطرح است، چگونگی ساختار تیم است. در اکثر منابع واژه‌ی تیم برای گروهی از عامل‌های یادگیرنده ناهمگن بکار برده شده است. یکی از مهمترین دلایل این انتخاب این است که چنین گروهی می‌تواند از عامل‌هایی تشکیل شده باشد که تکنیک‌های اکتشافی[1]متفاوتی دارند و بنابراین قادرند به راه‌حل‌های متفاوتی برسند و یا این‌که حداقل در بهینه‌های محلی یکسانی به دام نمی‌افتند.

دیدگاه‌های بسیار متفاوتی برای حل مساله یادگیری مشارکتی وجود دارند. پاسخ به سوالاتی نظیر آن‌چه در ادامه می‌آید، کمک زیادی به روشن شدن کیفیت دیدگاه انتخابی دارد[۷].

۱) چه اطلاعاتی می‌بایست بین عامل‌ها مبادله شود؟ پارامترهای یادگیری، حالت مساله، پارامترهای کیفی بدست آمده توسط دیگر عامل‌ها.

۲) چه زمانی اطلاعات می‌بایست مبادله شوند؟ با توجه به درخواست عاملی که به این اطلاعات نیاز دارد، زمانی که یک عامل اطلاعات مناسبی را به‌دست آورده است، دنباله‌ای از واقعه‌های[2] خیلی خوب یا خیلی بد، می‌توان گام‌های زمانی مشخصی نیز تعریف کرد که در آن‌ها عامل‌ها به مشارکت روی بیاورند.

۳) چه میزانی از اطلاعات می‌بایست مبادله شود؟ میزان اطلاعات با توجه به توازن بین هزینه ارتباطات و بهبود عملکرد یادگیری، انتخاب می‌شود. نیاز عامل‌ها به اطلاعات نیز در تعیین میزان آن، نقش دارد.

---

[1] Exploration
[2] Episode

۴) در چه منبعی باید به جست‌وجوی اطلاعات پرداخت؟ نظارت انسانی[1]، عامل‌های خبره و یا عامل-هایی که در یادگیری راه‌حل برای مسائل مشابهی استفاده شده‌اند.

۵) عامل چگونه می‌بایست انتخاب کند که در حالت جاری کدام یک از اطلاعات موجود اهمیت بیشتری دارد؟ با تخمین میزان اعتماد به اطلاعاتی که تولید/فرستاده/دریافت شده است و یا با تخمین میزان ارتباط هر بخش از اطلاعات به هر راه حل خاص.

۶) عامل چگونه باید از اطلاعات بدست آمده استفاده کند؟ تعویض پارامترهایی که فرآیند یادگیری را تنظیم می‌کنند.

مشورت[2]، رای‌گیری، رقابت، راهنمایی، تقلید و مبادله اطلاعات جزء روش‌هایی هستند که فرآیند انجام الگوریتم‌های مختلف یادگیری را تسهیل و سرعت‌دهی می‌کنند. اکثر این روش‌ها با یادگیری تقویتی ترکیب شده‌اند و بخش عمده‌ی استفاده از این الگوریتم یادگیری، به‌خاطر ذات انعطاف‌پذیر، عمومی و تقریبا فاقد نظارت[3] آن است. یادگیری تقویتی موجب می‌شود تا عامل یادگیرنده با حفظ تعادل بین اکتشاف و بهره‌برداری[4] دانشی که از منابع مختلف و روش‌های متفاوت به‌دست می‌آورد، بازخوردش از محیط را بیشینه کند. از این رو در سال‌های اخیر در حوزه‌ی عمومی سیستم‌های چندعامله، یادگیری تقویتی مورد توجه قرار گرفته است. به نظر می‌رسد زمانی که یادگیری تقویتی در همکاری بین عامل‌ها استفاده می‌شود، نتایج قابل قبولی در پی خواهد داشت.

۱-۴ اهداف و نوآوری‌های پایان‌نامه

مهمترین هدف این پایان‌نامه ارائه روشی موثر برای مشارکت در یادگیری بین عامل‌های حاضر در یک سیستم چندعامله و بهبود کیفیت به همراه کاهش زمان یادگیری است. رسیدن به این هدف به اطلاعاتی که بین عامل-ها مبادله می‌شود و روشی که بر مبنای آن اطلاعات مبادله شده در طول یادگیری مورد استفاده قرار می‌گیرند، بستگی دارد. در این پایان‌نامه مفهوم جدیدی به نام خبرگی چندمعیاره معرفی شده است که قادر است به خوبی اطلاعات همه‌جانبه‌ای را در مورد عامل‌ها و آن چه که تا کنون یاد گرفته‌اند، ارائه دهد. در روش پیشنهادی جدول مشارکتی که بر اساس خبرگی چندمعیاره ساخته شده است، به عنوان اطلاعات بین عامل‌ها مبادله می‌شود. آزمایش‌های انجام شده به خوبی بهبود یادگیری به دلیل کیفیت بالای اطلاعات مبادله شده را نشان می‌دهند. هم‌چنین در این پایان‌نامه روش

---

[1] Human Supervision
[2] Consultation
[3] Semi-Unsupervised
[4] Exploitation

جدیدی برای استفاده از اطلاعات مبادله شده در طول یادگیری ارائه شده که در آن جدول مشارکتی به عنوان نوعی راهنما در انتخاب عمل به کار گرفته می‌شود. در حالی‌که در سایر روش‌های موجود در یادگیری مشارکتی، بهبود در یادگیری با تغییر محتوای یادگرفته شده‌ی عامل‌ها توسط اطلاعات مبادله شده بدست می‌آید، در روش پیشنهادی محتوای یادگیری جمعی عامل‌ها به عنوان راهنمایی برای انتخاب بهتر اعمال عامل‌ها استفاده می‌شود و بهبود در یادگیری بدون تغییر محتوای یادگرفته شده و فقط با راهنمایی عامل‌ها در انتخاب عمل بر اساس تجربیات جمعیشان انجام می‌پذیرد. برای ارزیابی عملکرد روش پیشنهادی هفت نوع آزمایش متفاوت طراحی شده است و اهمیت همه-جانبه بودن تجربیات مبادله شده بین عامل‌ها در بهبود کیفیت و زمان یادگیری، به خوبی در نتایج حاصل از انجام آزمایش‌ها مشهود است.

۱-۵ ساختار پایان‌نامه

در این بخش ساختار پایان نامه معرفی می‌شود. در فصل ۲، روش‌های ارائه شده در مسئله‌ی یادگیری مشارکتی چندعامله که در آن‌ها مبنای مشارکت بر اساس انتقال اطلاعات است، مورد بررسی قرار می‌گیرند. در فصل ۳ مفاهیم علمی مورد نیاز برای درک روش پیشنهادی به طور دقیق توضیح داده خواهد شد. روش پیشنهادی و جزییات مربوط به آن در فصل ۴ بیان می‌شود. در فصل ۵ آزمایش‌هایی برای بررسی عملکرد روش پیشنهادی و مقایسه‌ی آن با سایر روش‌ها انجام می‌شود. در فصل ۶، به بیان خلاصه‌ای از کارهای انجام شده در این پایان نامه و نتایج حاصل از آن پرداخته می‌شود و در انتها نیز پیشنهاداتی برای کارهای آینده بیان خواهد شد.

فصل دوم
مروری بر کارهای گذشته

۲-۱ مقدمه

در حالی‌که یادگیری مشارکتی رشته نسبتا جدیدی محسوب می‌شود ولی حجم مقالات و کارهای انجام شده در آن زیاد است. اکثر روش‌های پیشنهاد داده شده مبتنی بر ایده‌هایی ساده، برگرفته از زندگی جمعی حیوانات و انسان‌ها هستند. جامعه‌ای از عامل‌های انسانی هوشمند را در نظر بگیرید: عامل‌های هوشمند موجود در جامعه، در انزوا قرار ندارند، بلکه در یک جامعه خیراندیش[1] از این عامل‌ها به منظور ساختار دهی و هدایت یادگیری استفاده می‌شود. انسان‌ها به‌وسیله مشاهده دیگران، سخن گفتن و با دریافت انتقادها و تشویق‌ها، فرآیند یادگیری خود را انجام می‌دهند. می‌توان گفت در اکثر موارد یادگیری بیشتر از این‌که انجام اکتشاف باشد، یک انتقال است. به طور مشابه، نمی‌توان انتظار داشت که ربات‌های هوشمند وظایف پیچیده دنیای واقعی را در تنهایی و انزوا و با فرآیند آزمون-خطا بیاموزند. برای یادگیری بهتر، ربات‌ها می‌بایست در یک محیط مشارکتی قرار بگیرند و می‌بایست الگوریتم‌هایی به منظور تسهیل انتقال دانش بین آن‌ها توسعه داده شود. از این رو مطالعه روش‌های یادگیری مشارکتی در یک محیط چندعامله، دارای اهمیت زیادی است.

---

[1] Benevolent

یکی از ساده‌ترین راه‌های افزایش سرعت یادگیری در یک سیستم چندعامله، یکسان سازی رفتار همه عامل‌های موجود در سیستم است، ولی در واقعیت این روش همیشه جواب خوبی نمی‌دهد. اکثر مسائل دنیای واقعی قابل تقسیم به اجزای موازی نیستند، از این رو یکی از مهمترین مسائل در یادگیری مشارکتی، نحوه تقسیم و به اشتراک گذاری دانش یادگیری شده توسط تک تک عامل‌ها بین اعضای تیم است.

کارهای انجام گرفته را می‌توان بر حسب چگونگی در نظر گرفتن محیط در دو دسته‌بندی کلی مطالعه کرد. اولین دسته کارهایی است که منظور از مشارکت در آن‌ها صرفا چگونگی انتقال دانش و روش‌های متفاوت آن است. در این کارها محیط در قالب یک مدل مارکوف ساده بررسی می‌شود. در دسته دوم محیط نقش مهم‌تری را بر عهده بردارد. در این کارها محیط غالبا به صورت یک بازی تصادفی[1] در نظر گرفته می‌شود که در آن عامل‌ها نحوه رسیدن به تعادل در بازی را می‌آموزند. در واقع روش‌های دسته دوم بر مبنای نظریه بازی‌ها استوار شده‌اند. لازم به ذکر است که اکثر روش‌های موجود در حوزه یادگیری مشارکتی، یادگیری Q را به عنوان روش یادگیری پایه خود استفاده می‌کنند و عمده تحقیقات در این رشته با عنوان یادگیری Q مشارکتی چندعامله صورت می‌پذیرد. روش‌های مختلف یادگیری مشارکتی بر مبنای انتقال اطلاعات بر اساس دیدگاه‌ها و پاسخ‌های متفاوت محققان به دو سوال: "چه اطلاعاتی باید بین عامل‌ها مبادله شود و عامل‌ها چگونه باید از اطلاعات مبادله شده استفاده کنند؟" بنیاد نهاده شده‌اند. در ادامه این فصل روش‌های مختلف مطرح شده در حوزه یادگیری مشارکتی مبتنی بر انتقال اطلاعات طبق ترتیب زمان پیشنهاد شدن آن‌ها بررسی می‌شوند.

۲-۲ معرفی مکانیزم‌های مشارکتی

یکی از اولین تلاش‌ها برای انجام یادگیری مشارکتی مطالعه‌ای بود که در [۹-۸] انجام شد. در این تحقیقات، برای اولین بار به استفاده از مکانیزم‌های مشارکتی در یادگیری اشاره شده و این مکانیزم‌ها به عنوان روشی برای کاهش پیچیدگی زمانی یادگیرنده Q فاقد پیش‌قدر[2] معرفی شده‌اند. یادگیری Q فاقد پیش‌قدر به یادگیری Qای که در آن جدول Q فاقد هرگونه مقداردهی اولیه باشد، اطلاق می‌شود. پیدایش مکانیزم‌های مشارکتی با الهام از ماهیتِ محیط‌های اجتماعی مشارکتی صورت پذیرفته است. در یک محیط اجتماعی مشارکتی، عامل‌های هوشمند به ساختاردهی و هدایت یادگیری کمک شایانی می‌کنند. در چنین محیط‌هایی ماهیت یادگیری به همان اندازه که به اکتشافِ عامل یادگیرنده با استفاده از آزمون-خطا مرتبط است، با انتقال اطلاعات نیز در ارتباط است.

---

[1] Stochastic Game
[2] Bias

الگوریتم‌های یادگیری تقویتی نوعی جستجوی افزایشی را برای یافتن سیاست تصمیم بهینه انجام می‌دهند. پیچیدگی زمانی این جستجو به اندازه و ساختار فضای حالت به علاوه دانش زمینه‌ای[1] که در مقادیر اولیه پارامترهای یادگیرنده‌ها وجود دارد، بستگی دارد. وقتی هیچ دانش زمینه‌ای وجود نداشته باشد، جستجو بدون پیش‌قدر و مفرط[2] خواهد بود. مکانیزم‌های مشارکت به کاهش جستجو با فراهم آوردن بازخوردهای مفیدتر و منابع تجربی اضافی کمک می‌کنند.

در [۸] دو نوع مکانیزم مشارکتی پیشنهاد شده است: یادگیری با حضور یک نقاد خارجی[3] و یادگیری با مشاهده[4]. ایده‌ی یادگیری با حضور یک نقاد خارجی مبتنی بر حضور یک مربی است که یادگیرنده را می‌بیند و متناظر با اعمالش، پاداش‌هایی را به وی تخصیص می‌دهد. از این پاداش جهت بهبود استراتژی کنترل یادگیرنده استفاده می‌شود. در یادگیری با مشاهده، عامل با مشاهده رفتار سایر عامل‌ها نسبت به خودش تجربه بدست می‌آورد. در مکانیزم اول الگوریتم به تعامل با عامل‌های دانشمند نیازمند است. در این مکانیزم منظور از عامل دانشمند، نقادی است که خارج از سیستم قرار دارد و اعمال عامل‌های داخل آن را ارزیابی می‌کند[۹]. در صورتی که نقاد دارای دانش خوبی نباشد (به اصطلاح خام و بی‌تجربه باشد)، یادگیرنده بازخوردهای ضعیفی از نقاد دریافت خواهد کرد و در نتیجه سیاست‌های خوبی را فرا نمی‌گیرد. در مکانیزم اول نقاد یک عامل خارج از محیط در نظر گرفته شده است.

در مکانیزم دوم حتی زمانی که دو طرف یک تعامل خام و بی‌تجربه هستند نیز مفید واقع می‌شود. در این مکانیزم یادگیرنده تجربیات اضافی‌اش را با مشاهده رفتار سایرین بدست می‌آورد. سایرین می‌توانند هم عامل‌هایی باشند که رفتار مناسبی دارند و هم عامل‌هایی که خام و بی‌تجربه هستند. در هر صورت یادگیری با مشاهده از رفتارهای همه عامل‌های موجود در سیستم استفاده می‌کند. پیچیدگی زمانی مکانیزم‌های مشارکتی پیشنهادی در مقایسه با یادگیری Q فاقد پیش‌قدر جالب توجه است. زمان جست‌وجو برای یادگیری Q فاقد پیش‌قدر زمانی نمایی به عمق فضای حالت است، در حالی‌که دو مکانیزم ارائه شده زمانی خطی متناسب با اندازه فضای حالت دارند.

مکانیزم یادگیری با حضور نقاد خارجی در بدترین حالت در نهایت پیچیدگی زمانی خطی متناسب با اندازه فضای حالت دارد. لازم به ذکر است که در این مکانیزم به دلیل این که ارزیابی نقاد پس از انجام کار دریافت می‌گردد و در زمان دریافت نظر نقاد، یادگیرنده ممکن است خود را در حالت دیگری ببیند، ممکن است مشکل‌هایی بوجود بیاید. در واقع عامل پس از انجام عمل، نظر نقاد را نسبت به عمل دریافت می‌کند در حالی که از حالتی که این بازخورد را به دلیل حضور در آن دریافت کرده است، گذر کرده است . در چنین موقعیتی بازخورد دریافتی از نقاد

---

[1] Prior Knowledge
[2] Excessive
[3] Learning with an External Critic (LEC)
[4] Learning By Watch (LBW)

نقشی فوری در بهبود انتخاب عمل نخواهد داشت و طبعا چنین بازخوردی اثرگذاری کمتری در رفتار عامل خواهد داشت. در چنین مواقعی بازگشت به حالتی که عامل قبلا در آن قرار داشته است نیز اثر منفی در زمان جستجو خواهد داشت. در چنین حالت‌هایی سیستم به عامل اجازه می‌دهد در صورتی که پس از طی مدت زمان خاصی موفق به انجام کار نشد، دوباره شروع کند یا این که از یک مدل معکوس برای بازگشت به حالت اولیه استفاده کند.

با توجه به موارد مطرح شده می‌توان گفت که استفاده از روش یادگیری با حضور نقاد خارجی زمانی موثرتر خواهد بود که سطح تصمیم نرم[1]باشد و یا سیاست را بتوان با استفاده از تعمیم توابع تقریب‌زن بدست آورد. نرم بودن سطح تصمیم بدین معنی است که در حالت‌های همسایه عمل بهینه یکسان باشد. بدین ترتیب در صورت نرم بودن سطح تصمیم، جابجایی عامل پس از انجام عمل به حالت همسایه و سپس دریافت بازخورد مشکل چندانی ایجاد نخواهد کرد زیرا حالت‌های همسایه در نهایت عمل‌های بهینه یکسانی دارند. نتایج آزمایش‌های انجام گرفته در [۹]، پایایی دو مکانیزم فوق نسبت به بازخورد حاوی اغتشاش[2] را بررسی و تایید می‌کنند.

۲-۳ مشارکت به‌وسیله به اشتراک گذاری ادراک – واقعه - سیاست

در [۱۰] سه روش انتقال اطلاعات مختلف برای مشارکت در حوزه یادگیری مشارکتی مبتنی بر انتقال دانش، مطرح شده است. نویسنده در مقاله خود به بررسی این موضوع می‌پردازد که آیا با استفاده از تعداد یکسان عامل‌های یادگیرنده تقویتی، عملکرد عامل‌های مشارکتی بهتر از عامل‌های یادگیرنده مستقل از هم هست یا خیر؟ هزینه مشارکت بین عامل‌ها چیست؟ مهمترین نتیجه حاصل از این پژوهش این است که اگر مشارکت به صورت مناسبی صورت بپذیرد، هر عامل می‌تواند از دانش آموخته شده توسط سایر عامل‌ها استفاده کافی ببرد. در این تحقیق عامل-های مشارکتی در سه حالت مختلف (به اشتراک‌گذاری ادراک‌ها[3](حالت‌های مشاهده شده) ، به اشتراک‌گذاری واقعه‌ها(سه‌تایی‌های حالت–عمل–کیفیت) و به اشتراک‌گذاری سیاست‌های آموخته شده (پارامترهای راه‌حل داخلی)) بررسی شده‌اند و عملکرد آن‌ها نسبت به حالتی که عامل‌ها فاقد مشارکت و مستقل از هم هستند، سنجیده شده و نتایج جالبی بدست آمده است. به اشتراک‌گذاری در این روش که SA[4] نامیده شده است، به صورت بهبود جدولQ یک عامل با استفاده از میانگین جداول Q سایر عامل‌ها صورت گرفته است. طبق آزمایش‌های انجام گرفته، ادراک-های اضافی از سایر عامل‌ها در صورتی که بتوان به طور موثر از آن‌ها استفاده کرد، مفید خواهند بود. به اشتراک-

[1] Smooth
[2] Noise
[3] Sharing Sensation
[4] Simple Averaging

گذاری سیاست‌ها یا واقعه‌ها بین عامل‌ها، سرعت یادگیری را افزایش می‌دهد و البته هزینه ارتباطات را هم در بردارد. برای انجام اعمال مشترک، استفاده از امکان مشارکت بین عامل‌ها به صورت قابل توجهی عملکرد را نسبت به حالتی که عامل‌ها مستقل از هم هستند، افزایش می‌دهد، البته در این حالت ممکن است سرعت یادگیری در آغاز کار پایین باشد.

۲-۴ یادگیری مشترک

[۱۱] نیز جزو اولین مراجعی به شمار می‌رود که به بررسی مزایای مشارکت بین عامل‌های یادگیرنده تقویتی در یک از تیم از عامل‌ها پرداخته است. در واقع چارچوب ارائه شده برای اثبات مزیت مشارکت بر حالت استقلال عامل‌ها از[۸] قوی‌تر است و نتایج حاصل قابل تعمیم به حالت‌های بیشتری هستند. در [۱۲] مفهومی با عنوان یادگیری مشترک[1] معرفی شده است. یادگیری مشترک به فرآیند یادگیری عامل‌هایی اطلاق می‌شود که از یک سیاست مشترک استفاده می‌کنند و به صورت مشترک آن را به‌روز رسانی می‌کنند. نویسندگان این مقاله مطالعه‌ای روی یک نوع فازی از یادگیری تقویتی چندعامله انجام داده‌اند. در این مطالعه عامل‌ها در طول یادگیری با به روز رسانی یک جدول مشترک با یکدیگر مشارکت می‌کنند. نتایج این مطالعه به صورت نظری و آزمایشی این نکته را تایید کردند که در دسته گسترده‌ای از مسائل، یک گروه از یادگیرنده‌های مشارکتی نتایج بهتری را در مقایسه با یک گروه از عامل‌های مستقل به دست می‌آورند.

۲-۵ تقلید

یکی از روش‌های یادگیری در انسان‌ها، به خصوص در مراحل رشد و در دوره کودکی تقلید از رفتار بزرگسالان است[۱۳]. ایده تقلید[2] نیز یکی از روش‌های یادگیری مشارکتی در سیستم‌های کامپیوتری محسوب می‌شود. در این روش، یادگیرنده‌ها اعمال یک معلم را می‌بینند، آن‌ها را یاد می‌گیرند و در شرایط مشابه آن‌ها را تکرار می‌کنند. این روش بر روی عملکرد معلم تاثیری ندارد و روش به راهنمایی وابسته نیست. مثلا در [۱۴] یک ربات حرکت ساده یک انسان را دریافت می‌کند و سعی می‌کند تا با جداسازی بخش‌های با معنی حرکت آن را بیاموزد و در محیط‌های مختلف آن را تکرار کند. در [۱۵] یک سیستم تقلید رباتیک را برای یادگیری ربات‌های هل دهنده

[1] Joint Learning
[2] Imitation

توپ[1]توسعه داده شده است. در این سیستم، عامل‌ها ابتدا به صورت فردی مرحله یادگیری را انجام می‌دهند و سپس با استفاده از روش‌های تقلید ساده[2]، تقلید شرطی[3] و تقلید انطباقی[4] از یکدیگر تقلید می‌کنند. یکی از نکات مهم در تقلید این است که یک عامل چه زمانی و از چه کسی تقلید کند؟ پاسخ به این سوال مبدا تفاوت سه روش تقلید ذکر شده است. در تقلید ساده عامل‌ها همیشه از عامل‌های همسایه‌شان تقلید می‌کنند و این تقلید با استفاده از تقویت‌های محیطی صورت می‌گیرد. در این روش دو عامل همسایه همیشه منتظر یکدیگر هستند. لازم به ذکر است که منظور از همسایگی بین عامل‌ها نزدیکی آن‌ها از نظر مکانی است. عامل‌هایی که در مکان‌هایی نزدیک به هم قرار دارند، در حال یادگیری رفتارهای مشابهی هستند و لذا تقلید در چنین شرایطی موثر خواهد بود.

در تقلید شرطی، مساله انتظار تقریبا حل شده است. در این روش دو عاملی که عملکرد پایین‌تری از سایرین داشته‌اند، از سایرین تقلید می‌کنند. سنجش عملکرد توسط مجموعه پاداش‌ها و تنبیه‌هایی که عامل تا کنون دریافت کرده است، صورت می‌پذیرد. تقلید انطباقی نیز شبیه تقلید شرطی است ولی با این تفاوت که نرخ تقلید قابل تنظیم کردن است. در این روش نرخ تقلید با توجه به اختلاف عملکرد دو عامل همسایه تنظیم می‌شود. رفتار تقلید تمایل به میرایی و پایداری رفتار دارد در حالی که رفتار مبتنی بر یادگیری تقویتی تمایل به یافتن بهترین راه‌حل دارد. در تقلید انطباقی با وزن‌دار کردن تقلید به عامل این اجازه داده می‌شود که بر حسب شرایط روش یادگیری خود را به تقلید یا به یادگیری تقویتی تغییر دهد[۱۶].

۲-۶ حافظه جمعی

یکی دیگر از ایده‌های قابل قبول برای یادگیری مشارکتی در سیستم‌های چندعامله، ایده‌ی استفاده از حافظه جمعی[5] است که در [۱۷] مطرح شد. این روش با الهام از ایده شناخت توزیع‌شده[6] در علوم اجتماعی شکل گرفته است[۱۸]. شناخت توزیع شده بر این نکته مهم تکیه دارد که در یک اجتماع شناخت تنها در یک فرد صورت نمی‌گیرد بلکه در افراد اجتماع توزیع می‌شود و هر کدام از افراد دارای شناختی خاص خود هستند. در یک گروه مشارکتی، زمانی که یک عامل تازه‌وارد با عامل‌های پرتجربه‌تر مشارکت می‌کند، درخواست‌ها و پاسخ‌های عامل-های پرتجربه‌تر، تازه‌وارد را به سمت الگوهای فعالیت موثرتر هدایت می‌کند. این مشارکت به نوعی برای عامل‌های

---

1 Ball-Pusher Robots
2 Simple Mimetism
3 Conditional Mimetism
4 Adaptive Mimetism
5 Collective Memory
6 Distributed Cognition

پرتجربه‌تر نیز سودمند است، ممکن است عامل تازه‌وارد در بی‌خبری خود گاهی ابتکاراتی برای رد کردن راه‌حل قدیمی ارائه دهد که موثرتر از راه‌حل قدیمی باشد. حافظه دسته‌جمعی به وسعت دانش رویه‌ای[1] که اجتماع به‌واسطه تجربیات حاصل از تعامل اعضای آن با یکدیگر و تعامل‌شان با دنیا بدست آورده، اطلاق می‌شود.

حافظه دسته‌جمعی مکانیزمی است که در آن اجتماعی از عامل‌ها، تجربه‌هایشان را در یک منبع مشترک قرار داده و سپس از آن به منظور بهبود حل مسائل مشارکتی در تعامل با یکدیگر بهره می‌برند. با استفاده از حافظه جمعی، تعداد تلاش عامل‌ها و هم‌چنین نیاز به برقراری ارتباط بین عامل‌ها به منظور حل مسائل مشارکتی کاهش می‌یابد. دانشی که در این روش به عنوان حافظه جمعی معرفی می‌شود قادر است تا بسیاری از مشکلات چندعامله را پاسخ-گویی کند، ولی تا کنون تمرکز بر روی استفاده از این دانش در دو دیدگاه زیر مورد بررسی قرار گرفته است:

- ✓ یادگیری رویه‌های مشارکتی: با استفاده از حافظه جمعی عامل‌ها الگوهای موفق حل یک مساله مشارکتی را که در آن حضور داشته است را به یاد می‌آورد و می‌تواند از آن‌ها به عنوان پایه‌ای برای انجام تعامل‌های آینده‌اش استفاده کند.
- ✓ یادگیری قابلیت‌های عامل‌ها: از یک ساختار درختی برای نگهداری تخمین احتمال موفقیت عملگرهای اجرایی عامل استفاده می‌شود. با استفاده از این ساختار می‌توان فهمید که هر عامل در کدام‌یک از اعمالش موفق‌تر عمل می‌کند. از این نکته می‌توان برای بهبود طراحی سیستم و کاهش ارتباطات استفاده کرد.

حاظه جمعی را می‌توان به‌وسیله یک حافظه متمرکز یا حافظه‌های توزیع‌شده در تک تک عامل‌ها و یا در یک نوع حافظه ترکیبی نظیر حافظه سازمانی، پیاده‌سازی نمود. در [۱۹] نحوه به روز رسانی حافظه جمعی به صورت کامل توضیح داده شده است.

۲-۷ پند

پندپذیری[2] و مبادله پند[3] بین عامل‌ها نیز جزو ایده‌هایی است که از علوم اجتماعی به دنیای چندعامله وارد شده است و اولین بار در [۲۰] مطرح شده است. پیاده‌سازی ایده مبادله پند بین عامل‌ها، یعنی قادر ساختن یک عامل به درخواست بازخورد اضافی از سایر عامل‌هایی که در حال حل مسائل مشابهی هستند. تکنیک مبادله پند از نوعی یادگیری نظارت‌شده[4] بهره می‌برد که سیگنال تقویت در آن لزوما از محیط به عامل انتقال نمی‌یابد، بلکه این

---

[1] Procedural Knowledge
[2] Advice-Taking
[3] Advice-Exchange
[4] Supervised Learning

سیگنال می‌تواند بر مینای پندی باشد که عامل از سایر عامل‌هایی که عملکرد بهتری داشته‌اند، دریافت کرده است. از مکانیزم مبادله پند می‌توان برای ارتقا عملکرد یک گروه از عامل‌های یادگیرنده به نحو شایانی بهره برد. عامل‌های یادگیرنده با مسائل مشابهی مواجهند و در محیطی قرار گرفته‌اند که فقط سیگنال تقویت در دسترس است. هر کدام از عامل‌های یادگیرنده می‌توانند از روش‌های یادگیری متفاوتی استفاده کند.

ایده‌پردازان مبادله پند در [۲۱] ایرادهای وارد به ایده اولیه خود را تحلیل و بررسی کرده‌اند. در مقاله‌های بحث شده، تمرکز بر روی عامل های یادگیرنده‌ای قرار دارد که بر روی مسائل مشابه ولی جداگانه‌ای کار می‌کنند. در [۲۲] شرایطی بررسی می‌شود که عامل‌ها در یک محیط یکسان با هم تعامل دارند. در چنین محیطی، مکانیزم مبادله پند به صورت زیر انجام می‌شود: حالت جاری محیط که توسط پندپذیرنده[1] دیده شده است، به عاملی که عملکرد بهتری در مسائل مشابه دارد، ارائه می‌شود و سپس از عمل پیشنهاد داده شده توسط پنددهنده به عنوان پاسخ نوعی یادگیری نظارت شده استفاده می‌شود. این مکانیزم عامل را قادر می‌سازد که هم از تقویت محیط و هم از همکارانی که بیش از سایرین موفق هستند و به عنوان معلم/نقاد عمل می‌کنند، یاد بگیرد. توسعه ایده مبادله پند نیازمند طرح مفاهیم جدیدی نظیر اعتماد به‌نفس، اعتماد و ارجحیت پنددهنده است که در [۲۳] معرفی شده‌اند.

در [۷] مطالعه‌ای بر روی مساله تعویض اطلاعات واقعه‌ای[2] بنا به درخواست عاملی که به اطلاعات نیازمند است، انجام شده است. اطلاعات مبادله شده شامل جفتهای حالت/عمل، میانگین عملکرد جاری عامل و بهترین عملکرد آن است. اطلاعات در طول فرآیند یادگیری و در زمانی که یک عامل به این درک برسد که عملکرد جاری‌اش در مقایسه با سایرین پایین است، انجام می‌شود. در هر عامل مقدار اطلاعاتی که قرار است مبادله شود با مقداردهی پارامتری به نام اعتماد به نفس کنترل می‌شود. در واقع از پارامتر اعتماد به نفس به این منظور استفاده می‌شود که عامل بفهمد آیا برای هر حالت جدید نیازمند گرفتن یک پند هست یا خیر. هم‌چنین انتخاب مکانی که اطلاعات باید جمع‌آوری شود بر اساس اطلاع از نتایج کار سایر عامل‌ها است. در درالگوریتم ۲-۱ شبه کد روند کلی الگوریتم‌های مبتنی بر مبادله پند نشان داده شده است.

[1] Advisee
[2] Episodic

While not train finished
  Broadcast:
  cq_i: relative current quality
  bq_i: relative best quality, for $i \in$ Agents
  While not epoch finished
    1. Get state s for evaluation.
    2. If *best quality not good enough* or
       current quality not good enough or
       uncertain/confused concerning state s
       2.1 Select the best advisor (k).
       2.2 Request advice to agent k for state s
       2.3 Agent k: process request of agent i
           producing advised action (a)
       2.4 Process advised action (a)
    3. Evaluate state s and produce response (r)
    4. Receive reward for action taken
  End epoch loop
  Update cq_i, bq_i, $trust_{ij}$ and sc_i (*self-confidence*).
End train loop

شکل ۲-۱- شبه کد روند کلی الگوریتم‌های مبتنی بر مبادله پند[۲۲]

۲-۸ یادگیری مشارکتی بر مبنای خبرگی

یکی دیگر از ایده‌هایی که در یادگیری مشارکتی مطرح شده، ایده استفاده از میزان خبرگی عامل‌ها در یادگیری است. این ایده نیز برگرفته از خصوصیات جامعه انسانی است و اولین بار در [۲۴] تحت عنوان الگوریتمی به نام WSS[1] که مشکلات الگوریتم SA پیشنهادی [۱۰] را حل می‌نماید، مطرح شد. در روش SA، برای ایجاد یادگیری مشارکتی یک عامل جدول Q خود را با میانگینی از جداول Q سایر عامل‌ها بهبود می‌بخشد. در روش جدید به این نکته دقت شده است که عامل‌ها پس از مرحله یادگیری از نظر توانایی‌ها یکسان نیستند و برخی از عامل-ها خبره‌تر از سایرین هستند. در روش WSS، ابتدا چند معیار برای محاسبه خبرگی پیشنهاد داده شده است و سپس هر عامل با وزنی متناظر با خبرگی سایر عامل‌ها از جداول Q آن‌ها برای بهبود جدول خود استفاده می‌کند.

در WSS عامل‌ها دو فاز یادگیری مستقل و یادگیری مشارکتی دارند. در واقع الگوریتم به تعدادی گام مشارکت تقسیم شده است. در طول هر گام ابتدا هر یک از عامل‌ها به صورت جداگانه فاز یادگیری مستقل خود را طی می‌کند و سپس در پایان هر گام، عامل‌های موجود در سیستم وارد فاز یادگیری مشارکتی می‌شوند و جدول‌های Q خود را با دیگران به اشتراک می‌گذارند. اشتراک و نحوه وزن‌دهی جدول‌ها بر اساس مقادیر معیارهای خبرگی که در طول فاز یادگیری مستقل به روز رسانی شده‌اند، صورت می‌پذیرد. در [۲۴] سه روش وزن‌دهی متفاوت مطرح شده

[1] Weighted Strategy Sharing

است که هر یک مزایا و معایب خود را دارد و در واقع نوع خاصی از یادگیری مشارکتی بین عامل‌ها را سبب می‌شود. به دلیل این که روش پیشنهادی در این پایان‌نامه توسعه‌ای بر روش WSS محسوب می‌شود، در فصل بعد این روش به صورت کامل و دقیق بررسی خواهد شد وفعلا از ذکر جزییات چشم‌پوشی می‌شود.

در WSS تعداد تلاش‌های یادگیری که در طول هر گام مشارکت انجام می‌پذیرد به عنوان پارامتر بازه مشارکت تعریف شده است. هر چقدر بازه مشارکت بزگتر باشد، عامل‌ها تعداد تلاش‌های یادگیری مستقل بیشتری را به انجام می‌رسانند و تجربه بیشتری بدست می‌آورند. در صورتی که بازه مشارکت خیلی کوچک باشد، عامل‌ها فرصت کافی برای بدست آوردن تجربه را در اختیار ندارند و لذا مشارکت تاثیر کمتری بر بهبود یادگیری خواهد داشت. تعیین مناسب مقدار این پارامتر نقش موثری در بهبود کیفیت یادگیری خواهد داشت. در [۲۵] مطالعه‌ای بر روی مقداردهی موثر پارامترهای روش WSS انجام شده است.

مهمترین نکته‌ای که در WSS، در نظر گرفته نشده این است که عامل‌ها علاوه بر این‌که از نظر سطح خبرگی با هم متفاوت هستند، از نظر نوع و دامنه ناحیه‌ای که در آن خبره شده‌اند، نیز با هم متفاوتند. در [۲۶] روشی پیشنهاد شد که در یادگیری مشارکتی علاوه بر سطح خبرگی آن‌ها، مهارت‌های متفاوت‌شان نیز برای بهبود جدول Q را در نظر می‌گیرد. در روش پیشنهادی ناحیه خبرگی هر عامل با استفاده از انتقال حالات عامل‌ها و تاریخچه یادگیری آن‌ها محاسبه می‌شود. یکی از محدودیت‌های روش مذکور این است که محاسبه خبرگی سطح حالت عامل نیاز به ثبت سیگنال‌های تقویت دریافتی برای هر حالت توسط هر عامل است و این نیاز در حالت عملی چندان قابل تامین نخواهد بود. علاوه بر آن با استفاده از این روش عامل‌ها محدود به مشارکت فقط با عامل‌هایی خواهند شد که تاریخچه تقویت‌های دریافتی خود را ذخیره کرده‌اند. برای حل این مشکل در [۲۸-۲۷] روشی برای شناسایی ناحیه خبرگی عامل‌ها پیشنهاد داده شده است.در این روش به‌جای تکیه بر مشاهده رفتار سایر عامل‌ها یا تاریخچه دقیق یادگیری خود عامل، از تابعی که بر روی مقادیر Q تعریف شده، استفاده می‌شود. البته در صورت در دسترس بودن تاریخچه یادگیری، عامل‌ها از آن نیز می‌توانند استفاده کنند.

در [۲۹] روشی به نام اشتراک‌وزن‌دار استراتژی وفقی[1]معرفی شده است که سعی در برطرف کردن ایرادهای موجود در روش WSS دارد. در روش پیشنهاد شده عامل‌ها پس از تعیین وزن و اعلام آمادگی برای به اشتراک گذاری دانششان با در نظر گرفتن احتمالی اطلاعاتشان را به اشتراک می‌گذارند.

احتمال اشتراک اطلاعات با نسبت تفاوت بین وزن‌های عامل‌ها نسبت مستقیم دارد. هر چقدر وزن تخصیص یافته به هر یک از دو عاملی که قصد اشتراک اطلاعات را دارند بیشتر باشد، احتمال اشتراک اطلاعات بین آن‌ها

[1] Adaptive WSS

افزایش می‌یابد. اگر تفاوت بین وزن‌ها از یک آستانه تعریف شد کمتر باشد، دو عامل اطلاعاتی را با یکدیگر به اشتراک نخواهند گذاشت. هم‌چنین در [۲۹] معیار خبرگی جدیدی به نام پشیمانی[1] ارائه شده است که مقدار آن با توجه به تفاوت بین مقادیر Q اولین بهترین عمل و دومین بهترین عمل محاسبه می‌شود. از مزیت‌های این معیار عدم وابستگی آن به پارامتر بازه مشارکت است. عملکرد معیار پیشنهادی با دو معیار از معیارهای تعریف شده در [۲۴] مورد مقایسه قرار گرفته است.

۲-۹ یادگیری مشارکتی بر مبنای معماری تخته سیاه[2]

در سیستم‌هایی که دارای چند ربات هستند، وجود ارتباطات برای به اشتراک گذاری تجربیات، پارامترها و سیاست‌های کنترل ضروری است. تحقیقات نشان می‌دهد که ارتباطات صحیح قادر است عملکرد سیستم‌های چند رباته را به نحو چشمگیری افزایش دهد. معماری تخته سیاه یک مکانیزم ارتباطی موثر در هوش مصنوعی توزیع شده است که اولین بار در [۳۱-۳۰] مطرح شده است. تخته سیاه نوعی ناحیه ذخیره‌سازی اشتراکی است. هر ربات می‌تواند به تخته سیاه دسترسی داشته باشد و همین‌طور قادر است اطلاعات دلخواه خود را بر روی تخته سیاه بنویسد. در این معماری هیچ نوع ارتباطات مستقیمی بین عامل‌ها وجود ندارد. انتقال اطلاعات بین ربات‌ها به صورت غیرمستقیم و از طریق تخته سیاه صورت می‌پذیرد.

در [۳۲] روش یادگیری مشارکتی Q بر مبنای این معماری پیشنهاد شده است که در آن معماری تخته سیاه نحوه به روز رسانی مقادیر Q، تغییر کنترل سیاست و عمل هر کدام از عامل‌ها را مدیریت می‌کند. روال کار در الگوریتم پیشنهادی بدین صورت است که هر عامل حالت فعلی‌اش را به تخته سیاه می‌فرستد و پس از انجام عملی که توسط تخته‌سیاه پیشنهاد داده شده است، تقویت دریافتی از محیط را به تخته‌سیاه می‌فرستد. در سیستم تخته‌سیاه دو جدول وجود دارد: یکی جدول Q ای است که مقادیر Q سیستم یادگیرنده فعلی را ذخیره می‌کند(جدول ۲-۱) و دیگری جدولی که جفت حالت-عمل‌های اجرا شده برای همه ربات‌ها را در خود نگهداری می‌کند(جدول ۲-۲).

[1] Regret
[2] Blackboard Architecture

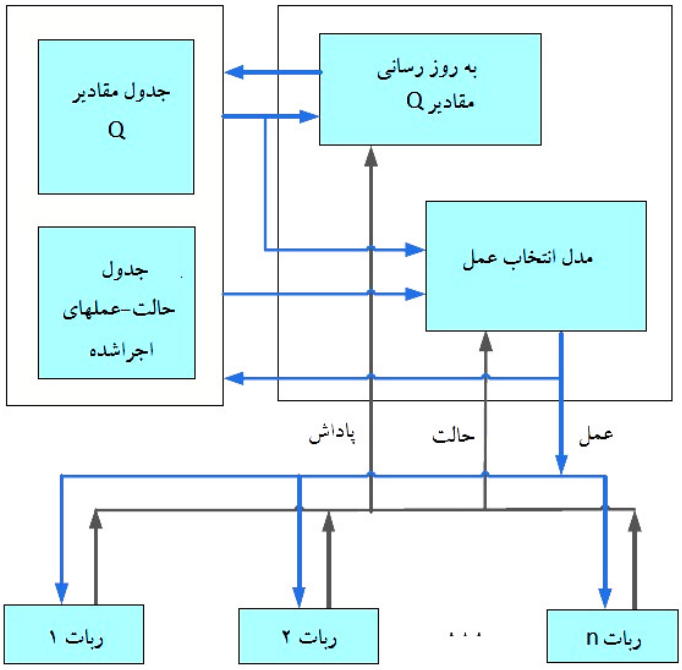
State \ Action	a_1	a_2	...	a_m
s_1	$Q(s_1,a_1)$	$Q(s_1,a_2)$	...	$Q(s_1,a_m)$
s_2	$Q(s_2,a_2)$	$Q(s_2,a_2)$	...	$Q(s_2,a_m)$
⋮	⋮	⋮	⋮	⋮
s_n	$Q(s_n,a_1)$	$Q(s_n,a_2)$	...	$Q(s_n,a_m)$

جدول ۲-۱- جدول Q سیستم یادگیرنده فعلی[۳۲]

Robot ID	*State*	*Action*	*Q(s, a)*
1	s_{1i}	a_{1j}	$Q(s_{1i}, a_{1j})$
2	s_{2i}	a_{2j}	$Q(s_{2i}, a_{2j})$
⋮	⋮	⋮	⋮
n	s_{ni}	a_{nj}	$Q(s_{ni}, a_{nj})$

جدول ۲-۲- جدول مقادیر حالت- عمل‌های اجرا شده[۳۲]

ساختار سیستم یادگیری مشارکتی Q بر مبنای تخته‌سیاه در شکل ۲-۳ نشان داده شده است. در این ساختار ربات نوعی منبع دانش به حساب می‌آید و مسئولیت اجرای عملی که توسط تخته‌سیاه به او نسبت داده شده است و سپس باز فرستادن پاداش دریافتی از محیط به تخته‌سیاه پس از انتقال حالت در نتیجه اجرای عمل را بر عهده دارد. تخته‌سیاه حالت‌ها و دانش فرستاده شده توسط همه ربات‌ها را ذخیره می‌کند. مکانیزم کنترل شامل ماژول به‌روزرسانی مقادیر Q و ماژول انتخاب عمل است که وظیفه به‌روزرسانی مقادیر Q و انتخاب عملی که هر ربات باید انجام دهد را بر عهده دارد.



شکل ۲-۲- ساختار سیستم یادگیری مشارکتی Q بر مبنای معماری تخته‌سیاه[۳۲]

لازم به ذکر است که در روش پیشنهادی از جدول حالت–عمل‌های اجرا شده برای کنترل نرخ اکتشاف محیط استفاده می‌شود. در هر انتخاب عمل این جدول بررسی می‌شود و در صورتی که جفت حالت–عملی وجود داشته باشد که تا کنون توسط ربات‌ها انجام نشده باشد، عمل مربوطه به ربات پیشنهاد داده می‌شود و در غیر این صورت عمل پیشنهادی توسط تابع انتخاب عمل بولتزمن و بر مبنای مقادیر Q انتخاب شده و به عامل پیشنهاد داده می‌شود. نتایج آزمایش‌های انجام شده در [۳۲]، کیفیت خوب عملکرد روش پیشنهادی در یادگیری وظیفه جلوگیری از برخورد با موانع در یک محیط را نشان می‌دهند.

۲-۱۰ یادگیری مشارکتی بر مبنای پختگی سیاست[1]

روش پیشنهاد داده شده در [۳۳] ترکیبی از روش‌های یادگیری مشارکتی بر مبنای تخته سیاه و یادگیری مشارکتی بر مبنای خبرگی است. این روش به منظور بهبود یادگیری در محیط‌هایی که دارای چند ربات هستند، طراحی شده است. در یادگیری مشارکتی بر مبنای پختگی سیاست همانند روش قبلی از معماری تخته‌سیاه استفاده شده است. روش پیشنهادی بر اساس یادگیری کلی مقادیر Q ربات‌های دیگر استوار است. در واقع به دلیل تفاوت حالت و موقعیت ربات‌ها و ذات تصادفی موجود در سیستم یادگیری، ممکن است برای بعضی حالت‌ها سیاست بهتری نسبت به بقیه بدست آید و سایر ربات‌ها بتوانند سیاست بهتری برای حالت‌های دیگر بدست بیاورند. لذا ربات‌ها می‌توانند از طریق ارتباطات از یکدیگر بیاموزند.

تفاوت اصلی روش پیشنهادی با روش موجود در [۳۲] در این است که در روش قبلی هر کدام از ربات‌ها بدون توجه به سایرین صرفا با تخته‌سیاه در ارتباط است و مقادیر Q خود را به صورت ارتباط غیر مستقیم با دیگران از طریق تخته‌سیاه به روز رسانی می‌کند. در روش پیشنهادی [۳۳]، ربات علاوه بر ارتباط با تخته‌سیاه و تاثیر غیر مستقیم از دیگران، از سایر ربات‌هایی که سیاست‌های بهتری دارند، نیز به طور مستقیم می‌آموزد. مفهوم پختگی سیاست نیز با استفاده از معیارهای خبرگی تعریف شده در [۲۴]، برای ارزیابی سیاست ربات‌ها مورد استفاده قرار گرفته است. مجموع پاداش‌های منفی که ربات در طول یادگیری خود دریافت کرده است به عنوان مفهوم پختگی تعریف شده است و هر چقدر مقدار مجموع کمتر باشد، پختگی ربات بیشتر خواهد بود. رابطه به روزرسانی مقادیر Q هر ربات با توجه به جدول Q سایر عامل‌ها نیز بر اساس روابط موجود در [۲۴] که در فصل بعد به طور مفصل توضیح داده شده است، انجام می‌شود.

یادگیری مشارکتی بر مبنای پختگی سیاست این مزیت را دارد، که ربات سیاست بهینه را فقط از ربات‌هایی که میزان پختگی بیشتری نسبت به او دارند می‌آموزد و این امر موجب می‌شود که ربات سیاست بهتر را به صورت

[1] Maturity of the Policy

کورکورانه نیاموزد. همان‌طور که گفته شد روابط به روز رسانی در روش پیشنهادی [۳۳] مطابق با روش تخصیص وزن به صورت یادگیری از افراد خبره در [۲۴] است. در [۲۴] هر عامل ضریبی به نام ضریب تاثیر پذیری از دیگران را نیز در تخصیص وزن دخیل می‌کند که این ضریب به صورت یک عدد ثابت در نظر گرفته شده است. آزمایش‌های انجام گرفته در [۲۵]، نشان می‌دهند که تعیین درست مقدار این ضریب نقش موثری در بهبود مشارکت دارد.

نوآوری [۳۳] ارائه روشی برای مقداردهی موثرتر به این ضریب است. مقداردهی ثابت به ضریب تاثیرپذیری از دیگران از نظر منطقی کار درستی نیست. در آغاز یادگیری، سیاست ربات‌ها پخته نیست و اگر در چنین حالتی ربات‌ها به دانش سایر ربات‌ها اعتماد کنند، ممکن است سیستم به سوی واگرایی و کاهش سرعت یادگیری پیش برود. در روش پیشنهادی، ضریب تاثیر پذیری به صورت متغیر تعریف شده است. مقدار اولیه ضریب برابر صفر است و با افزایش مراحل آموزشی مقدار ضریب نیز افزایش می‌یابد. لازم به ذکر است که یکی از تفاوت‌های روش پیشنهادی با کاری که در [۲۴] انجام شده است، ارتباط دائمی عامل‌ها با تخته‌سیاه است و لذا در این روش مفهوم بازه مشارکت مورد استفاده قرار نمی‌گیرد. تعریف بازه مشارکت به عامل‌ها این اجازه را می‌دهد که در طول بازه تجربه بدست بیاورند و سپس وارد چرخه مشارکت شوند و از این رو مشکل مطرح شده در زمینه مقدار ضریب تاثیر پذیری کمرنگ‌تر از روش ارائه شده در [۳۳] است.

۲-۱۱ نتیجه گیری

در این فصل روش‌های یادگیری مشارکتی مبتنی بر انتقال اطلاعات از آغاز تا کنون مورد بررسی قرار گرفتند. اکثر روش‌های موجود در این حوزه بر مبنای آن چه که در جوامع انسانی وجود دارد، شکل گرفته‌اند. بیشتر روش‌های موجود سعی در بهبود مشارکت بر اساس مبادله اطلاعات موثر بین عامل‌ها دارند. نکته‌ای که در بیشتر روش‌ها دیده می‌شود استفاده از اطلاعات مبادله شده برای تغییر بهینه مقادیر Q است. در فصل بعد مفاهیم علمی مورد نیاز برای درک بهتر روش پیشنهاد شده در این پایان‌نامه مورد بحث و بررسی قرار خواهند گرفت.

فصل سوم
مفاهیم علمی مورد نیاز در روش پیشنهادی

۳-۱ مقدمه

در این فصل مقدمات علمی مورد نیاز روش پیشنهادی معرفی خواهد شد. در روش پیشنهادی ارائه شده در این پایان‌نامه از یادگیری Q که در دسته روش‌های یادگیری تقویتی قرار می‌گیرد، به عنوان روش یادگیری برای عامل‌های حاضر در سیستم استفاده شده، از این رو بخش آغازین این فصل به معرفی یادگیری تقویتی و روند کلی الگوریتم یادگیری Q خواهد پرداخت. در فصل قبل معرفی مختصری برای روش اشتراک وزن‌دار استراتژی که نوعی یادگیری مشارکتی بر مبنای خبرگی است، ارائه شد. به دلیل استفاده از این الگوریتم در روش پیشنهادی، بخش دوم به معرفی دقیق و همراه با جزییات روش اشتراک وزن‌دار استراتژی اختصاص یافته است. یکی دیگر از مفاهیمی که در روش پیشنهادی مورد استفاده قرار گرفته است، الگوریتم HAQL است که نوعی تسریع برای یادگیری Q با استفاده از مکاشفه محسوب می‌شود. در بخش پایانی این فصل مروری کلی بر خصوصیات این الگوریتم صورت خواهد گرفت.

۳-۲ یادگیری تقویتی

یادگیری تقویتی[1] یک نام عمومی برای خانواده‌ای از تکنیک‌هاست که در آن یک عامل تلاش می‌کند تا یک وظیفه را به‌وسیله تعامل مستقیم با محیط یاد بگیرد. ریشه‌های این روش در مطالعه رفتار حیوانات تحت تاثیر محرک خارجی است[۳۴]. یادگیری تقویتی در [۳۴] به عنوان "یادگیری می‌خواهد چه بکند؟ چگونه موقعیت‌ها را به اعمال نگاشت می‌دهد و این‌که یادگیری برای بیشینه کردن سیگنال پاداش عددی چه می‌خواهد بکند؟" مورد بحث قرار گرفته است. ظهور یادگیری تقویتی به عنوان یکی از الگوهای یادگیری ماشین، به روزهای آغازین سایبرنتیک و کار در آمار، روانشناسی، علوم عصبی و علوم کامپیوتر باز می‌گردد. یادگیری تقویتی به این پرسش که چگونه یک عامل خودمختار[2]- که توانایی ادراک و عمل کردن دارد- در یک محیط می‌تواند برای انتخاب اعمال بهینه برای رسیدن به اهدافش آموزش ببیند، پاسخ می‌دهد و برای محیط‌هایی مناسب است که یک عامل می‌بایست با آزمون و خطا در محیط آن را درک کند و عامل هیچ دانش زمینه‌ای در مورد محیط ندارد.

در یادگیری تقویتی، نظیر اکثر فرم‌های یادگیری ماشین، به یک یادگیرنده گفته نمی‌شود که چه اعمالی با چه پاداشی ممکن است انجام شود، بلکه یادگیرنده می‌بایست کشف کند که کدام اعمال بیشترین پاداش را در پی خواهند داشت. در بیشتر حالت‌ها ممکن است عامل پاداش‌های فوری نداشته باشند، بلکه بواسطه موقعیت‌های بعدی که در آن قرار می‌گیرد، پاداش‌دهی شود. "جست‌وجوی آزمون-خطا" و"پاداش تاخیری[3]" دو مشخصه‌ی مهم یادگیری تقویتی هستند. برای حل مساله یادگیری تقویتی دو استراتژی مهم وجود دارد:

استراتژی اول: جستجوی فضای رفتارها به منظور یافتن رفتاری که بهتر از سایرین در محیط عمل کند.

استراتژی دوم: استفاده از متدلوژی‌های برنامه‌نویسی پویا و آماری برای تخمین ارزش اعمال انجام شده در فضای حالت مساله

در یک مدل یادگیری تقویتی استاندارد، یک عامل به‌وسیله ادراک و عمل با محیط‌اش تعامل می‌کند. در هر قدم، عامل حالت جاری S محیط را به عنوان ورودی دریافت می‌کند و عمل a را برای تولید خروجی انتخاب می‌کند. انجام عمل الزامی برای تغییر حالت محیط نیست و محیط می‌تواند حالت خود را حفظ کند. نهایتا عامل، ارزش عمل خود را به عنوان تقویت[4]دریافت می‌کند. تقویت می‌تواند مثبت باشد(پاداش[5]) یا منفی (تنبیه[6]). سیستم کنترل

---

[1] Reinforcement Learning (RL)
[2] Autonomous
[3] Delayed Reward
[4] reinforcement
[5] reward
[6] punishment

اعمال عامل می‌بایست اعمالی را انتخاب کند که مجموع سیگنال‌های تقویتی را افزایش دهد. یک عامل می‌تواند انجام این فرآیند را در حین تعامل آزمون-خطای سیستماتیک با محیط یاد بگیرد.

۳-۳ فرآیند تصمیم‌سازی مارکوف و الگوریتم یادگیری Q[1]

تصمیم‌سازی دنباله‌ای یک عامل در یک دنیای تصادفی قابل مشاهده با مدل انتقال مارکوف[2]، یک فرآیند تصمیم‌سازی مارکوف[3] نامیده می‌شود[۳۴].

فرآیند تصمیم‌سازی مارکوف معین به صورت چهارتایی $<S, A, \delta, R>$ تعریف می‌شود که در آن:

S : مجموعه حالات محیط،

A: مجموعه اعمالی که عامل می‌تواند انجام دهد،

$\delta : S \times A \rightarrow S$: تابع انتقال حالت،

$R: S \times A \rightarrow \mathcal{R}$: تابع پاداش

در فرآیند تصمیم‌سازی مارکوف معین، عامل می‌تواند مجموعه حالات متمایز محیط را درک کند و هم‌چنین می‌تواند از مجموعه اعمال موجود در A عملی را انتخاب و آن را در محیط اجرا کند. محیط با دادن پاداش حاصل از اجرای عمل و انتقال به حالت بعدی (طبق تابع انتقال) به عامل پاسخ می‌دهد.

می‌توان مساله یادگیری تقویتی را با الگوریتم یادگیری Q که شبه کد آن در شکل ۳-۳ نشان داده شده است، حل کرد. الگوریتم Q نوعی یادگیری تقویتی فاقد مدل به شمار می‌رود که اولین بار در [۳۵] مطرح شده است.

---

[1] Q-learning algorithm
[2] Markovian Transition Model
[3] Markov Decision Process

1. $x \leftarrow$ the current state
2. Select an action a that is usually consistent with $\pi(x)$, but occasionally an alternate. For example, one might choose a according to the Boltzmann distribution: $p(a|x) = \frac{e^{Q(x,a)/T}}{\sum_{b \in A} e^{Q(x,b)/T}}$ where T is a temperature parameter that adjusts the degree of randomness.
3. Execute action a, and let y be the next state and r the reward received.
4. Update the action-value function:

$$Q(x,a) \leftarrow (1-\alpha)Q(x,a) + \alpha[r + \gamma U(y)],$$

where $U(y) = \max_{b \in A}[Q(y,b)]$.

5. Go to 1.

شکل ۳-۳- شبه کد الگوریتم یادگیری Q [۳۵]

در الگوریتم یادگیری Q، α نرخ یادگیری است که مقداری بین ۰ و ۱ دارد. T پارامتر دما است که میزان تصادفی بودن در انتخاب اعمال و در واقع نرخ اکتشاف را کنترل می‌کند. نرخ یادگیری بیان می‌کند که چقدر به پاداش‌های جدید در مقابل مقادیر قدیمی که یاد گرفته شده است، می‌بایست توجه کرد. یکی از مهمترین مزایای الگوریتم یادگیری Q اثبات همگرایی آن است. طبق قضیه‌ای که در [۳۶] اثبات شده است اگر نرخ اکتشاف و نرخ یادگیری به قدر کافی و به طور آرام کاهش یابند، این تضمین وجود دارد که یادگیری Q به یک سیاست بهینه همگرا شود. خصوصیت همگرا شدن باعث می‌شود که الگوریتم یادگیری Q به راحتی در بسیاری از محیط‌ها قابل استفاده باشد. در اکثر مراجع شبه کد ارائه شده در شکل ۳-۳ با عنوان یادگیری Q تک قدم[1] شناخته می‌شود که ساده‌ترین نسخه یادگیری Q به شمار می‌رود[۳۴].

۳-۴ روش اشتراک وزن‌دار استراتژی(WSS)

روش پیشنهادی توسعه‌ای بر روش اشتراک وزن‌دار استراتژی محسوب می‌شود. در فصل قبل این روش به صورت مختصر توضیح داده شد، ولی به دلیل اهمیت این روش و ماهیت آن در درک بهتر روش پیشنهادی، در این بخش ابتدا مفهوم خبرگی و لزوم استفاده از آن معرفی خواهد شد و پس از آن الگوریتم اشتراک وزن‌دار استراتژی به طور دقیق مطرح خواهد شد.

[1] One-Step Q-Learning

۳-۴-۱ مفهوم خبرگی و لزوم استفاده از آن

تا کنون کارهای زیادی در زمینه یادگیری مشارکتی در سیستم‌های چندعامله صورت گرفته است. یک دسته مهم از این کارها بر مبنای اندازه‌گیری میزان خبرگی عامل‌ها و نواحی خبرگی آن‌هاست که در[۲۸-۲۴] به طور دقیق بررسی شده است. یادگیری مشارکتی در سیستم‌های چندعامله در مقایسه با یادگیری فردی[1]، به خاطر دارا بودن منابع اطلاعاتی و دانش بیشتر یادگیری سریع‌تر و موثرتری را داراست.

در سیستم‌های یادگیری سطح پایین، یادگیری مشارکتی به خوبی وجود دارد. به عنوان مثال یک سیستم مبتنی بر اجتماع مورچه‌ها را در نظر بگیرید. اجتماع مورچه‌ای یک نوع سیستم چندعامله ساده به شمار می‌رود[۳۷]. عامل‌ها (مورچه‌ها) از محیط برای تبادل اطلاعات و دانش به منظور بدست آوردن یک راه حل سراسری استفاده می‌کنند. در چنین سیستمی، عامل‌ها با استفاده از تغییر یا ترکیب راه‌حل‌های فردی به یک راه حل سراسری دست پیدا می‌کنند. در کلونی مورچه‌ها از مفهوم فرومون[2] برای برقراری ارتباط بین عامل‌ها و تنظیم نحوه مشارکت آن‌ها استفاده شده است. حال اگر عامل‌ها دارای توانایی‌های متفاوتی باشند، سیستم به خوبی کار نخواهد کرد و حتی در مواردی هم منجر به شکست خواهد شد چرا که تفاوت موجود در توانایی عامل‌ها، در فرومون در نظر گرفته نشده است.

وقتی که عاملی با یک مساله جدید که تا کنون آن را ندیده است مواجه می‌شود، ابتدا مدت زمانی طول می‌کشد تا عامل متوجه شود که به تنهایی قادر به حل مساله نیست. پس از این تشخیص، عامل می‌بایست تصمیم بگیرد که چه کسی می‌تواند به وی کمک کند؟ حل این قضیه بدین جهت پیچیده است که عامل دقیقا نمی‌داند علت اصلی بروز مساله چیست، پس چگونه می‌خواهد راجع به علتی که نمی‌داند چیست، سوال بپرسد. پس از فهمیدن علت، عامل می‌بایست به دنبال عاملی بگردد که در زمینه مذکور خبره باشد. در چنین مواردی صرفا شناسایی افراد خبره برای حل مساله یک شخص یا جواب سوال کافی نیست. یک شخص می‌تواند به یک یا بیش از یک نفر برای دریافت کمک مراجعه کند. همین نکته منجر به مرحله انتخاب خبره می‌شود[۳۸]. می‌توان با استخراج نواحی خبرگی سایر عامل‌ها، در مرحله انتخاب خبره، عملکرد موثرتری داشت.

تعریف ۳-۱- فردی که دارای دانش، مهارت و توانایی باشد، خبره نامیده می‌شود. یک فرد خبره ممکن است سطوح خبرگی متفاوت و همین طور زمینه‌های خبرگی متفاوتی داشته باشد[۳۸].

خبرگی می‌تواند موضوعی و یا رویه‌ای[3] (وابسته به طرز عمل و رویه فرد) باشد. هم‌چنین می‌توان خبرگی را بر حسب تنظیمات اجتماعی و سازمانی، مقداردهی و ترتیب‌دهی کرد. در دیدگاهی عملی‌تر، عامل‌ها باید در یک

---

[1] Individual Learning
[2] Pheromone
[3] Procedural

محیط که شامل چندین هدف و یا وظیفه است، بررسی شوند. مثلا فرض کنید که عامل‌ها می‌توانند در دامنه‌های مختلفی و با سطوح مختلفی از خبرگی، خبره و کارآمد شوند. با توجه به خصوصیات محیط و شرایط آزمایش‌هایی که عامل در آن قرار دارد، یک عامل ممکن است در برخی نواحی نادیده گرفته شود و به نتایج جالب توجهی دست پیدا نکند و در برخی نواحی نیز بسیار خبره شود و راه‌های رسیدن به هدف را به خوبی بیاموزد. تفاوت در نواحی خبرگی[1] می‌تواند به عنوان یک فاکتور مثبت تلقی شود، مثلا می‌توان به هر کدام از اعضای تیم یادگیری یک زیر وظیفه خاص را محول نمود و در این صورت دانش آموخته شده توسط همه عامل‌ها مورد استفاده قرار بگیرد. نکته جالب توجه این است که حتی در سیستم‌های چندعامله همگن نیز تفاوت در نواحی خبرگی وجود دارد. علت این امر را می‌توان در نوع توزیع سیستم و یا چند هدفه بودن آن جست. همگی این شرایط منجر به این حقیقت می‌شود که یک عامل مفاهیم و مهارت‌های مختلفی را یاد بگیرد. در چنین سیستمی مهمترین سوالی که با آن مواجه هستیم به صورت زیر است:

"کدام بخش از دانش عامل می‌تواند به بهبود عملکرد کلی کمک کند؟ "

کاملا واضح و روشن است که همه‌ی دانش یک عامل قابل اعتماد نیست. در اکثر موارد بخش کوچکی از دانش یک عامل ارزشمند و قابل اعتماد است. در یک سیستم چندعامله، هر عامل می‌تواند در یک بخش خاص از دامنه‌ی دانش کلی مفید واقع شود. چنین عاملی یک عامل خاص‌شده[2] نامیده می‌شود[۲۷]. در تعاملات انسانی نیز زمانی که قصد داریم مطلبی را فرا بگیریم به سراغ کسی می‌رویم که از سایرین در زمینه مذکور خبره‌تر باشد. در واقع ما به دنبال یک فرد متخصص خواهیم گشت که بتواند سوالات ما را به خوبی پاسخ دهد.

استخراج نواحی خبرگی در عامل‌ها و دانش چکیده و ترکیب چنین دانشی، دو موضوع تحقیقاتی مهم در هوش مصنوعی نمادین[3] به شمار می‌روند. الگوریتم یادگیری مشارکتی بر مبنای خبرگی، در سه مرحله طراحی شده است: الف) انجام یادگیری فردی برای همه عامل‌های حاضر در سیستم، ب)اندازه‌گیری سطح خبرگی پ)ترکیب دانش بدست‌آمده از مرحله (ب) و انجام یادگیری مشارکتی.

در ادامه، معیارهایی برای اندازه‌گیری سطح خبرگی معرفی و بررسی خواهند شد. لازم به ذکر است که به منظور توسعه‌ی روش WSS، ایده استخراج ناحیه خبرگی در [۲۸-۲۷] مطرح شده است ولی به دلیل این که روش پیشنهادی صرفا از معیارهای خبرگی و روش WSS استاندارد استفاده می‌کند، در این پایان‌نامه از بررسی دقیق مباحث مربوط به ناحیه خبرگی و استخراج آن خودداری می‌شود.

---

[1] Area Of Expertise
[2] Specialized Agent
[3] Symbolic Artificial Intelligence

۳-۴-۲ معیارهای اندازه‌گیری خبرگی

پیش‌نیاز منطقی انجام یک یادگیری مشارکتی صحیح، ارزیابی دانش عامل‌ها به منظور ارزیابی سطح دانش و دامنه خبرگی آن‌ها است. در این بخش انواع معیارهایی که تا کنون برای اندازه‌گیری خبرگی پیشنهاد داده شده‌اند، معرفی می‌شوند. تا کنون معیارهای خبرگی صرفا برای عامل‌های یادگیرنده‌ی Q پیشنهاد داده شده‌اند ولی به نظر می‌رسد در کارهای آتی بتوان آن‌ها را به سایر الگوریتم‌های یادگیری نیز تعمیم داد.

همان‌طور که گفته شد، یک عامل می‌تواند سطوح خبرگی مختلفی داشته باشد و می‌توان در سطوح مختلفی، خبرگی عامل را ارزیابی کرد. اندازه‌گیری خبرگی در سطح عامل (مثلا جدول Q برای یک عامل یادگیرنده‌یQ)، سطح خبرگی کلی یک عامل و توانایی وی در خوب رفتار کردن را نشان می‌دهد. معیار خبرگی می‌تواند در سطح حالت نیز تعریف شود، در این صورت معیار خبرگی نشان‌دهنده این نکته خواهد بود که توانایی عامل در یافتن عمل بهینه در یک حالت خاص چقدر است. لازم به ذکر است که معیارهای پیش رو تنها سطح خبرگی را بیان می‌کنند و با تشخیص دامنه نواحی خبرگی یک عامل سر و کار ندارند.

معیارهای قضاوت در مورد خبرگی عامل‌ها به دو دسته کلی تقسیم می‌شوند[۲۴]:

دسته اول- معیارهای خبرگی مبتنی بر تاریخچه یادگیری: محاسبه معیارهای این دسته به اطلاعاتی فراتر از جدولQ احتیاج دارد. این اطلاعات می‌بایست در حین فاز یادگیری جمع‌آوری شوند. معیارهای خبرگی عادی، مطلق، مثبت و منفی به این دسته تعلق دارند.

۱. **خبرگی عادی**[1]- این معیار اعتبار را به کسانی می‌دهد که در گذشته موفقیت‌های بیشتر و خطاهای کمتر ی داشته‌اند.

$$e_i^{Nrm} = \sum_{t=1}^{now} r_i(t) \qquad ۱-۳$$

۲. **خبرگی مطلق**[2]- این معیار هم به تشویق و هم به تنبیه توجه می‌کند. در این معیار شکست و موفقیت با سیگنال‌های تنبیه و پاداش وزن‌دار شده‌اند و از این رو هر دو برای عامل ارزش‌مند هستند.

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

۳. **خبرگی مثبت**[1] – این معیار فقط تجربیاتی را مد نظر قرار می‌دهد که پاداش در پی داشته باشند.

[1] Normal
[2] Absolute

$$e_i^P = \sum_{t=1}^{now} r_i^+(t), r_i^+(t) = \begin{cases} 0 & if\ r_i(t) \le 0 \\ r_i(t) & otherwise \end{cases} \qquad ۳-۳$$

۴. **خبرگی منفی**[2]– این معیار فقط کوشش‌های ناموفق را در نظر می‌گیرد و به هر عاملی که بیشتر شکست خورده باشد، امتیاز بیشتری نسبت می‌دهد. در این معیار از این فلسفه بهره گرفته شده است که عاملی که بیشتر اشتباه می‌کند، راه‌های نرسیدن به پاداش را به خوبی یادگرفته است.

$$e_i^N = \sum_{t=1}^{now} r_i^-(t), r_i^-(t) = \begin{cases} 0 & if\ r_i(t) > 0 \\ -r_i(t) & otherwise \end{cases} \qquad ۴-۳$$

۵. **خبرگی مبتنی بر گرادیان** – این معیار تغییرات سیگنال از آخرین بازه مشارکت تا کنون را نشان می‌دهد که در آن c نشان‌دهنده زمان شروع آخرین چرخه یادگیری مستقل است.

$$e_i^{Gr} = \sum_{t=c}^{now} r_i(t) \qquad ۵-۳$$

۶. **خبرگی متوسط تعداد قدم‌ها**– این معیار معکوس تعداد قدم‌های لازم برای رسیدن به هدف را نشان می‌دهد که در آن trial، تعداد تلاش، n_{trial}، تلاش فعلی و $m_i(trial)$ تعداد قدم‌هایی که عامل برای رسیدن به هدف نیاز دارد.

$$e_i^{Am} = \left(\sum_{trial=1}^{n_{trial}} m_i(trial)/n_{trial}\right)^{-1} \qquad ۶-۳$$

دسته دوم – معیارهای خبرگی مبتنی بر جدول Q: خبرگی فقط بر اساس جدول Q محاسبه می‌شود و به هیچ اطلاعات اضافی دیگری نیاز نیست. در این حالت نیازی به دانش قبلی[3] وجود ندارد. قطعیت[4] و آنتروپی جزو این دسته هستند.

۱. **قطعیت**– احتمال انتخاب اعمال با بالاترین مقادیر ممکن برای Q. در حقیقت معیار اطمینان مشخص می‌کند که عامل تا چه حد به مقدار مورد انتظار عمل انتخابی‌اش

---

[1] Positive
[2] Negative
[3] Prior Knowledge
[4] Certainty

نسبت به سایر اعمال موجود مطمئن است. اگر استراتژی انتخاب اعمال از نوع بولتزمن باشد، اطمینان به صورت زیر محاسبه می‌شود که در آن T پارامتر دما است.

$$e^{\mathrm{Cer}}(\mathrm{x}) = \frac{\max_{a_k \epsilon \text{ actions}} \exp\left(\frac{Q(x,a_k)}{T}\right)}{\sum_{a_k \epsilon \text{ actions}} \exp\left(\frac{Q(x,a_k)}{T}\right)} \qquad ۷-۳$$

۲. **آنتروپی**– این معیار به درجه نسبی تصادفی بودن مربوط می‌شود. هر چه مقدار آنتروپی بالاتر باشد، اختلاف بین احتمالات برای انتخاب اعمال مختلف کمتر است. زمانی‌که احتمالات همه اعمال برای انتخاب شدن یکسان باشد، بیشینه آنتروپی بدست می‌آید. کمینه آنتروپی (آنتروپی=۰) هم زمانی رخ می‌دهد که یکی از اعمال به صورت واضح انتخابش بر سایرین ارجحیت داشته باشد.

$$e^{\mathrm{Ent}}(\mathrm{x}) = -\sum_{a \epsilon \text{actions}} \Pr(a|x)\, \mathrm{Ln}(\Pr(a|x)) \qquad ۸-۳$$

معیار اطمینان فقط به بهترین عمل توجه می‌کند در صورتی که آنتروپی به همه‌ی اعمال توجه می‌کند. اگر بخواهیم این معیار را به شرایطی که بیش از یک حالت مدنظر است گسترش بدهیم، میانگین آنتروپی روی حالت‌ها محاسبه می‌شود. هر چند جمع‌آوری معیارهای دسته نخست زمان‌بر است، اما این معیارها کیفیت رفتار عامل را به نحو دقیق‌تری بیان می‌کنند. در واقع هر کدام از معیارهای دسته اول نشان‌دهنده یکی از خصوصیات رفتاری عامل هستند. خصوصیاتی نظیر این که عامل در طول یادگیری تا چه اندازه موفق بوده یا این که متوسط تعداد مراحلی که برای رسیدن به پیروزی طی کرده، چقدر بوده است. در این پایان‌نامه برای معرفی مفهوم خبرگی چندمعیاره از معیارهای دسته نخست استفاده شده است.

۳-۴-۳ الگوریتم اشتراک وزن‌دار استراتژی

در الگوریتم وزن‌دار استراتژی (WSS) فرض شده است که n عامل همگن یادگیرنده Q که بر اساس الگوریتم یادگیری Q تک قدم در محیط‌های جداگانه مشغول یادگیری هستند. عامل‌ها در دو فاز یادگیری خود را انجام می‌دهند: فاز یادگیری مستقل و فاز یادگیری مشارکتی. شکل ۳-۴ روند کلی الگوریتم WSS را نشان می‌دهد.

در ابتدا همه عامل‌ها در فاز یادگیری مستقل قرار می‌گیرند. عامل iام تعداد t_i تلاش یادگیری انجام می‌دهد. هر تلاش یادگیری با قرار گرفتن عامل در یک حالت تصادفی شروع می‌شود و زمانی که عامل به هدف می‌رسد، پایان می‌پذیرد. بعد از طی شدن تعداد مشخصی تلاش یادگیری، همه عامل‌ها به فاز یادگیری مستقل منتقل می‌شوند.

در فاز یادگیری مشارکتی، هر عامل یادگیرنده به جدول Q سایر عامل‌ها با توجه به مقدار خبرگی‌شان وزن تخصیص می‌دهد. در واقع هر عامل میانگین وزن‌دار جدول‌های Qسایرین را محاسبه می‌کند و جدول حاصل از رابطه ۳-۸ را به عنوان جدول Q جدید خود در فاز یادگیری مستقل بعد استفاده می‌کند.

Algorithm 1-Weighted Strategy Sharing Algorithm for each agent A_i

(1) *Initialize*
(2) while not *EndOfLearning* do
(3) begin
(4)If *InIndividualLearningMode* then
(5)begin *{ Individual Learning}*
(6) x_i :=*FindCurrentState()*
(7) a_i := *SelectAction()*
(8) *DoAction*(a_i)
(9) r_i := *GetReward()*
(10)y_i := *GoToNextState()*
(11) $V(y_i) := Max_{b \in actions} Q(y_i, b)$
(12) $Q_i^{new}(x_i, a_i) := (1-\beta_i)Q_i^{old}(x_i, a_i) + \beta_i\ (r_i + \gamma_i\ V(y_i))$
(13)e_i := *UpdateExpertness*(r_i)
(14)end
(15)else *{Cooperative Learning}*
(16)begin
(17) for j:= *1* to *n* do
(18)e_j := *GetExpertness*(A_j)
(19) $Q_i^{new} := 0$
(20) for j:= *1* to *n* do
(21) begin
(22) W_{ij} := *ComputeWeights*($i,j,e_1 \ldots e_n$)
(23) Q_j^{old} := *GetQ*(A_j)
(24) $Q_i^{new} := Q_i^{new} + W_{ij} * Q_j^{old}$
(25) end
(26) end
(27) end

شکل ۳-۴- شبه‌کد الگوریتم اشتراک وزن‌دار استراتژی[۳۹]

$$Q_i^{new} \leftarrow \sum_{j=1}^{n}(W_{ij} * Q_j^{old}) \qquad ۹-۳$$

در [۴۰] چند مکانیزم مختلف برای تخصیص وزن پیشنهاد داده شده است که استفاده از هر کدام منجر به نوع متفاوتی از مشارکت می‌شود. مکانیزم اول بر اساس ایده یادگیری از همه شکل گرفته است. در این روش فرض بر این است که همه عامل‌ها دانشی ارزشمند برای یاد دادن به دیگران دارند، لذا هر عامل در ساختن جدول جدید خود از همه عامل‌های دیگر استفاده می‌کند. رابطه ۳-۱۰ مکانیزم وزن‌دهی در روش یادگیری از همه را نشان می‌دهد. در این روش تاثیر یک عامل بر همه عامل‌های دیگر یکسان است و همه جدول‌های Q پس از مشارکت مقدار یکسانی خواهند داشت.

$$W_{i,j} = \frac{e_j}{\sum_{k=1}^{n} e_k} \qquad ۱۰-۳$$

مکانیزم دوم بر اساس یادگیری از همه با وزن‌های مثبت شکل گرفته است.

اگر $e_{\min} = \min\{e_k | k = 1, \dots, n\}$ و $c > 0$ یک عدد ثابت باشد آن‌گاه $e_j - e_{min} + c > 0$ برقرار است. با توجه به این موارد رابطه ۳-۱۱ به عنوان روش تخصیص وزن یادگیری از همه با وزن‌های مثبت معرفی شده است که در آن وزن عاملی که خبرگی‌اش از سایرین کمتر است با رابطه ۳-۱۲ نشان داده شده است.

$$W_{i,j} = \frac{e_j - e_{min} + c}{\sum_{k=1}^{n} e_j - e_{min} + c} > 0 \qquad ۱۱-۳$$

$$W_{i,\min} = \frac{c}{\sum_{k=1}^{n} e_j - e_{min} + c} > 0 \qquad ۱۲-۳$$

در مکانیزم دوم اگر $c \to \infty$، آن‌گاه $W_{i,j} = 1/n$ و روش WSS به روش معدل‌گیری ساده همگرا می‌شود.

مکانیزم سوم بر اساس ایده یادگیری از افراد خبره شکل گرفته است. استفاده از این مکانیزم منجر به کاهش تعداد ارتباطات مورد نیاز برای مبادله‌ی جدول‌های Q بین عامل‌ها می‌شود زیرا در این حالت عامل فقط از عاملی که خبرگی بیشتری دارد، خواهد آموخت. یادگیرنده i بر اساس میزان تفاوت خبرگی‌اش با عامل jام با استفاده از رابطه ۳-۱۳ به جدول Q عامل jوزن تخصیص می‌دهد.

$$W_{i,j} = \begin{cases} 1 - \alpha_i & if\ i = j \\ \alpha_i \frac{e_j - e_i}{\sum_{k=1}^{n} (e_k - e_i)} & if\ e_j > e_i \\ 0, & otherwise \end{cases} \qquad ۱۳-۳$$

در رابطه ۳-۱۳ پارامتر α_i، ضریب تاثیرپذیری از دیگران نامیده می‌شود و مشخص می‌کند که یک عامل تا چه اندازه به دانش دیگران اعتماد می‌کند. مقدار این پارامتر بین صفر و یک تعریف می‌شود. اگر میزان خبرگی سایر عامل‌ها از عامل iام کمتر باشد، وزنی که به دانش آن‌ها اختصاص می‌یابد برابر صفر است.

در [۴۰] و [۲۴] آزمایش‌های مختلفی بر اساس هر سه مکانیزم وزن‌دهی انجام شده است و در اکثر موارد مکانیزم سوم عملکرد بهتری نسبت به سایر مکانیزم‌ها داشته است. از این رو در این پایان‌نامه منظور از آن چه با عنوان روش WSS مورد بررسی قرار می‌گیرد، الگوریتم اشتراک وزن‌دار استراتژی مبتنی بر استفاده از مکانیزم سوم – یادگیری از افراد خبره است.

۳-۵ الگوریتم HAQL: تسریع یادگیری Q با استفاده از مکاشفه

در [۴۱] برای اولین بار مفهوم استفاده از مکاشفه برای تسریع یادگیری تقویتی ارائه شده است. هدف از ارائه این دسته از روش‌ها حفظ خصوصیات مثبت یادگیری تقویتی نظیر تضمین همگرایی، انتخاب آزادانه اعمال و یادگیری فاقد نظارت است در حالی که سعی در برطرف کردن مهمترین عیب آن یعنی زمان طولانی لازم برای یادگیری دارد. در این مقاله دسته جدیدی از الگوریتم‌ها به نام الگوریتم‌های یادگیری Qمکاشفه‌ای تسریع‌یافته[1]معرفی شده‌اند که در آن‌ها عامل در مراحل اولیه یادگیری با استفاده از محیط یک سیاست مکاشفه می‌سازد و در مراحل بعد از سیاست مکاشفه ساخته شده برای بهبود انتخاب اعمال استفاده می‌کند. در واقع تابع مکاشفه H بر روی چگونگی انتخاب اعمال اثرگذار است. در [۴۱-۴۲] روش‌هایی برای استخراج خودکار تابع مکاشفه از دامنه مساله مورد مطالعه (دانش زمینه[2]) یا در طول یادگیری ارائه شده است.

در [۴۲] الگوریتم یادگیری Q مکاشفه‌ای تسریع‌یافته به نام HAQL مطرح شده است که در آن پس از ساخت مکاشفه در مراحل اولیه یادگیری، مطابق روابط ۳-۱۴ و ۳-۱۵ از مکاشفه در بهبود انتخاب عمل استفاده می‌شود. الگوریتم HAQL می‌تواند به عنوان راهی برای حل مساله یادگیری تقویتی در نظر گرفته شود که در آن با استفاده از یک تابع مکاشفه $H: S \times A \rightarrow R$ بر نحوه انتخاب عمل در طول فرآیند یادگیری اعمال اثر می‌شود. $H_t(s_t, a_t)$ مکاشفه‌ای را تعریف می‌کند که بر طبق آن اهمیت انجام عمل a_t زمانی که عامل در حالت s_t است، مشخص می‌شود. تابع مکاشفه اکیدا همراه با سیاست تعریف می‌شود، در واقع هر مکاشفه تعیین می‌کند که کدام عمل بدون در نظر گرفتن سایر اعمال می‌بایست انجام شود. به عبارت دیگر می‌توان گفت که تابع مکاشفه نوعی سیاست مکاشفه‌ای تولید می‌کند. لازم به ذکر است که تابع مکاشفه صرفا در قانون انتخاب عمل استفاده می‌شود و تعیین کننده این امر است که وقتی عامل در حالت s_t است چه عملی باید انجام شود. قانون انتخاب عملی که در HAQL استفاده شده است، انتخاب عمل به شیوه ε- حریصانه[3] است که مطابق رابطه ۳-۱۴ تعریف می‌شود.

$$\pi(s_t) = \begin{cases} \arg max_{a_t}[\hat{Q}(s_t, a_t) + \varepsilon H_t(s_t, a_t)], & if\ q < p \\ a_{random}, & otherwise \end{cases} \qquad ۱۴\text{-}۳$$

$$H(s_t, a_t) = \begin{cases} max_a \hat{Q}(s_t, a) - \hat{Q}(s_t, a_t) + 1, & if\ a_t \epsilon\ \pi^H(s_t) \\ 0, & otherwise \end{cases} \qquad ۱۵\text{-}۳$$

در روابط ۳-۱۴ و ۳-۱۵، ε متغیری با مقداری حقیقی است که برای وزن دادن به تاثیر مکاشفه در انتخاب عمل استفاده می‌شود، q مقداری تصادفی با احتمال یکنواخت از بازه [۰و۱] و p $(0 \leq p \leq 1)$ پارامتری است که نرخ

---

[1] Heuristically Accelerated Q-Learning (HAQL)
[2] Prior Knowledge
[3] ε-Greedy

اکتشاف/بهره‌برداری[1] را تعیین می‌کند، هر چه مقدار p بزرگتر باشد، احتمال انتخاب عمل تصادفی a_{random} کمتر خواهد بود. a_{random} نیز عملی تصادفی است که از بین اعمال ممکن در حالت s_t انتخاب می‌شود. شبه کد الگوریتم $HAQL$ به صورت کامل در شکل ۳-۵ نشان داده شده است.

Initialize $Q(s, a)$.
Repeat:
Visit the s state.
Select an action a using the action choice rule (equation 13-3).
Receive the reinforcement $r(s, a)$ and observe the next state s'.
Update the values of $H_t(s, a)$.
Update the values of $Q(s, a)$ according to:
$Q(s, a) \leftarrow Q(s, a) + \alpha[r(s, a) + \gamma \max_{a'} Q(s', a') - Q(s, a)]$.
Update the $s \leftarrow s'$ state.
Until some stop criteria is reached.

where: $s = s_t$, $s' = s_{t+1}$, $a = a_t$ e $a' = a_{t+1}$.

شکل ۳-۵- شبه کد الگوریتم HAQL [۴۲]

آن چه در مورد الگوریتم HAQL قابل ذکر است قضیه مطرح شده در [۴۱] است که ضمن اثبات صحت الگوریتم به نوعی شرایط تعریف مکاشفه مورد استفاده در ۳-۱۴ را نیز معرفی می‌کند. صورت قضیه به صورت زیر است:

قضیه – یک عامل یادگیرنده HAQL که در یک محیط مارکوف معین دارای مجموعه محدودی از حالت‌ها و اعمال، پاداش‌های محدود شده $(\exists c \in R; (\forall s, a), |R(s, a)| < c)$ و نرخ تخفیف $0 \leq \gamma < 1$ قرار دارد، زمانی که تابع مکاشفه مورد استفاده کران‌دار باشد، $\forall (s_t, a_t) h_{min} \leq H(s_t, a_t) \leq h_{max}$، مقادیر جدول Q چنین عاملی به Q^* همگرا خواهد شد و استفاده از مکاشفه موجب تولید مقادیر نامتناهی در جدول Q نخواهد شد.

قضیه ذکر شده همراه با لم‌ها و قضایای جانبی مورد استفاده در [۴۱] به طور کامل اثبات شده است. یکی از مهمترین نتایج قضیه مطرح شده که در فصل‌های آینده مورد استفاده قرار خواهد گرفت، ضابطه تعریف شده برای تابع مکاشفه است. بر طبق قضیه اگر تابع مکاشفه مورد استفاده کران‌دار باشد، استفاده از آن در قانون انتخاب عمل

[1] Exploration/Exploitation

موجب واگرایی در مقادیر Q نخواهد شد. یکی دیگر از ویژگی‌های ذکر شده برای HAQL در [۴۱]، قابلیت تعمیم اثبات آن به حالتی است که عامل به جای انتخاب عمل به شیوه ε- حریصانه از شیوه‌های دیگر انتخاب عمل تعریف شده در یادگیری Q نظیر تابع بولتزمن استفاده می‌کند.

۳-۶ نتیجه گیری

در این فصل مفاهیم علمی مورد نیاز در روش پیشنهادی ارائه شده در فصل ۴ مطرح شد. در روش پیشنهادی از الگوریتم یادگیری Q به عنوان روش یادگیری پایه استفاده می‌شود، لذا در بخش اول این فصل الگوریتم یادگیری Q معرفی شد. به دلیل این که روش پیشنهادی نوعی توسعه برای روش اشتراک وزن‌دار استراتژی به شمار می‌رود، جزییات روش اشتراک وزن‌دار استراتژی در بخش دوم ارائه شد. در بخش سوم نیز دسته‌ای از الگوریتم‌هایی که با استفاده از مکاشفه به تسریع یادگیری Q پرداخته‌اند، توضیح داده شده است. در فصل ۵ با استفاده از مقدمات بیان شده در این فصل، روش یادگیری مشارکتی Q مبتنی بر خبرگی چندمعیاره معرفی خواهد شد.

فصل چهارم
یادگیری مشارکتی بر مبنای خبرگی چندمعیاره

۴-۱ مقدمه

در این فصل نوآوری‌های ارایه شده در این پایان نامه مورد بحث و بررسی قرار خواهند گرفت. مهمترین ایده‌ای که در این پایان نامه مطرح شده است، معرفی مفهوم خبرگی چندمعیاره است. استفاده از این مفهوم جدید، توانسته است یادگیری مشارکتی چندعامله را به اندازه قابل توجهی بهبود بخشد. در این فصل ابتدا مفهوم خبرگی چندمعیاره و دلایل استفاده و همین‌طور ریشه‌های مرتبط با آن در علوم روانشناسی مورد بحث قرار خواهد گرفت و سپس یادگیری مشارکتی چندعامله مبتنی بر خبرگی چندمعیاره معرفی خواهد شد. در انتهای فصل نیز دلایل درستی استفاده از این روش تحت عنوان یک قضیه اثبات می‌شود.

۴-۲ خبرگی چندمعیاره و لزوم بررسی آن

اگر هر فرد در دنیای انسان‌ها را به مثابه‌ی یک عامل در دنیای مصنوعی در نظر گرفته شود، نگاشت‌های جالبی بین این دو مجموعه برقرار خواهد شد. هر فرد دارای یک شخصیت چند بعدی است و عوامل مختلفی بر

کیفیت یادگیری‌اش تاثیرگذارند. مثلا هنگام انتخاب یک دانش‌آموز نمونه، پارامترهای مختلفی نقش دارند و هر گونه نگاه تک‌بعدی به مساله انتخاب دانش‌آموز نمونه، از کیفیت انتخاب خواهد کاست. به عنوان مثال دانش‌آموزی که صرفا نمرات بسیار بالایی دارد و یا فقط عنوانهای ورزشی زیادی کسب کرده است، انتخاب جامعی برای معرفی به عنوان دانش‌آموز نمونه محسوب نمی‌شود. یک دانش‌آموز نمونه باید قادر باشد همزمان جمعی از معیارهای تعریف شده را به طورکارا برآورده سازد. علاوه بر شخصیت چندبعدی، هر فرد در طول زمان نیز تجارب متفاوتی را بدست می‌آورد. گاهی اوقات و در بازه‌ای از زندگی، تجربه فرد کاملا موفقیت‌آمیزاست و گاهی در بازه‌ای دیگر فرد شکستی نابهنگام را تجربه می‌کند. در واقع شخصیت هر فرد علاوه بر ویژگی‌های چند بعدی ذاتی‌اش در طول زمان نیز دستخوش تغییرات و تجارب متفاوتی خواهد شد و شخصیت هر فرد نیز بر مبنای تجربیاتی که کسب کرده است – فارغ از خوب و بد بودن – شکل می‌گیرد. این واقعیت قابل تعمیم به دنیای عامل‌ها است و ایده اصلی تحقیق حاضر نیز بر مبنای این واقعیت شکل گرفته است.

دانش عامل‌های یادگیرنده Q را می‌توان با معیارهای مختلفی ارزیابی کرد. یکی از مفاهیم معرفی شده در این زمینه، معیارهای خبرگی معرفی شده در [۲۴] هستند. در فصل قبل معیارهای خبرگی و دلایل معرفی این مفهوم به طور مفصل بررسی شد. در واقع خبرگی عامل تعیین می‌کند که سیاست جاری عامل تا چه اندازه درست بوده است. همان‌طور که در فصل‌های قبلی مطرح شد، در تحقیقاتی که تا کنون انجام گرفته است، از تعریف معیارهای خبرگی به منظور بهبود نحوه مشارکت و انتقال دانش بین عامل‌های همکار استفاده شده است. با نگاهی دقیق‌تر به معیارهای معرفی شده در [۲۴]، به نظر می‌رسد که هر یک از معیارها بیان‌کننده یکی از جنبه‌های رفتاری عامل هستند و ودر واقع در نظر گرفتن معیارهای مختلف در کنار یکدیگر ابعاد شخصیت رفتاری عامل یادگیرنده را بهتر نشان می‌دهد.

در روان‌شناسی نوین مفهوم جدیدی به نام خود پنداری[1] مطرح شده است[۴۴-۴۳] که بر اساس آن رفتار آینده یک فرد قابل توجیه است. خود پنداری یک مفهوم چندبعدی است که بر اساس ادارک فرد از جنبه‌های مختلف شخصیتی‌اش شکل گرفته است و تحقیقات روانشناسی نشان می‌دهد که اکثر تصمیمات فرد بر اساس دیدگاهی که نسبت به خودش دارد و در واقع بر اساس خود پنداری‌اش، گرفته می‌شود. در واقع در مفهوم خود پنداری نیز نوعی نگرش جامع به همه تجربیاتی که فرد از ابتدا و در جنبه‌های مختلف زندگی کسب کرده است، دیده می‌شود. از این رو می‌توان ایده خبرگی چندمعیاره در دنیای عامل‌ها را معادل مفهوم خودپنداری در دنیای انسانی در نظر گرفت.

[1] Self Concept

۴-۳ یادگیری مشارکتی Q بر مبنای خبرگی چند معیاره

همانطور که در بسیاری از منابع ذکر شده است[۴۶-۴۵و۲] ، شاخه سیستم‌های چندعامله ارتباط تنگاتنگ و تعامل دوطرفه‌ای با رشته روانشناسی برقرار کرده است و در طول سالیان اخیر بسیاری از ایده‌های جدید و تاثیرگذار در این شاخه از علوم روانشناسی نشات گرفته‌اند. در مقابل سیستم‌های چندعامله نیز کیفیت مطالعه بر روی مباحث روانشناسی را بهبود بخشیده‌اند[۴۷]. توجه به این حقیقت که بسیاری از واقعیت‌های دنیای انسانی قابل تعمیم به دنیای عامل‌ها است، نقش مهمی در چگونگی ایده‌یابی برای گسترش روش‌های موجود دارد.

هدف از انجام این تحقیق، یافتن روشی است که یادگیری مشارکتی بر مبنای خبرگی را بهبود ببخشد. در اکثر تحقیقات انجام گرفته در زمینه یادگیری مشارکتی بر مبنای خبرگی، روشهای ارائه شده بر مبنای محاسبه یک معیار خبرگی و سپس تشکیل جدولهای مشارکتی Q بر اساس معیار محاسبه شده، استوار هستند. به عبارت ساده‌تر، مشارکت بین عامل‌ها بر اساس سطح خبرگی آن‌ها انجام می‌شود و یکی از مباحث مهم در این روش، چگونگی انتخاب فرد خبره است.

به‌کارگیری ایده‌ی خبرگی چندمعیاره در یادگیری مشارکتی چندعامله بر اساس مشارکت بین انسان‌ها با در نظر گرفتن تجربیات متفاوت اعضای گروه شکل گرفته است. آن چه در مشارکت مهم است به اشتراک گذاشتن دانش عامل‌ها به نحو موثر بین یکدیگر است. تجربیات انسانی به خوبی نشان می‌دهند که در یک مشارکت موثر حتی توجه به تجربیات عضوی از گروه که شکست سختی را تجربه کرده نیز حائز اهمیت است. در این پایان‌نامه، از بررسی و تحلیل همزمان معیارهای خبرگی موجود به عنوان مکاشفه‌ای برای بهبود انتخاب عمل در یادگیری تقویتی استفاده شده است. روش پیشنهادی[1] MCE نام گرفته که در آن به جای محاسبه جدول مشارکتی Q بر اساس یک معیار خبرگی، به ازای همه معیارهای خبرگی تعریف شده در [۲۴]، جدول‌های Q مشارکتی محاسبه می‌شوند و سپس جدول‌های حاصل با یکدیگر ترکیب شده و جدول مشارکتی بر مبنای خبرگی چندمعیاره را تشکیل می‌دهند. از این جدول به عنوان ورودی تابع انتخاب عمل در یادگیری استفاده می‌شود. شبه کد الگوریتم پیشنهادی در شکل ۴-۱ نشان داده شده است. در زیر بخش بعدی الگوریتم پیشنهادی با جزییات دقیقتری توضیح داده خواهد شد.

[1] Multi-Criteria Expertness based Cooperative Q-learning

(1) Initialize Q
(2) $CoQ_{MCE} = Q$; { CoQ_{MCE} is used instead of Q in action selection}
(3) While not End Of Learning do
(4) Begin
  a. if In Individual Learning Mode then
  b. begin {Individual Learning}
    i. visit the state s
    ii. Select an action a based on Boltzmann function using CoQ_{MCE}.
    iii. $\pi(s_t) = \arg max_{a_t}(Boltzmann(CoQ_{MCE}(s_t, a_t))$
    iv. Receive the reinforcement $r(s,a)$ and notice the next state s'
    v. Update the values of $Q(s,a)$ according to the rule of updating $Q(s,a) \leftarrow Q(s,a) + \alpha[r(s,a) + \gamma\, max_{a'} Q(s',a') - Q(s,a)]$
    vi. Update the $s \leftarrow s'$ state.
    vii. Update Expertness Values of all agents
  c. End
  d. Else {Cooperative Learning}
  e. Begin
  f. For i=1:6
    i. Detect the agent which has the minimum value in i^{th} expertness measure
    ii. Calculate Cooperative Q-table (CoQ_i) based on less expert agent's view using WSS method.
  g. end
  h. $CoQ_{MCE} = \sum_{i=1}^{6} CoQ_i$;
  i. End
(5) End

شکل ۴-۱- شبه کد الگوریتم یادگیری مشارکتی Q بر مبنای خبرگی چندمعیاره

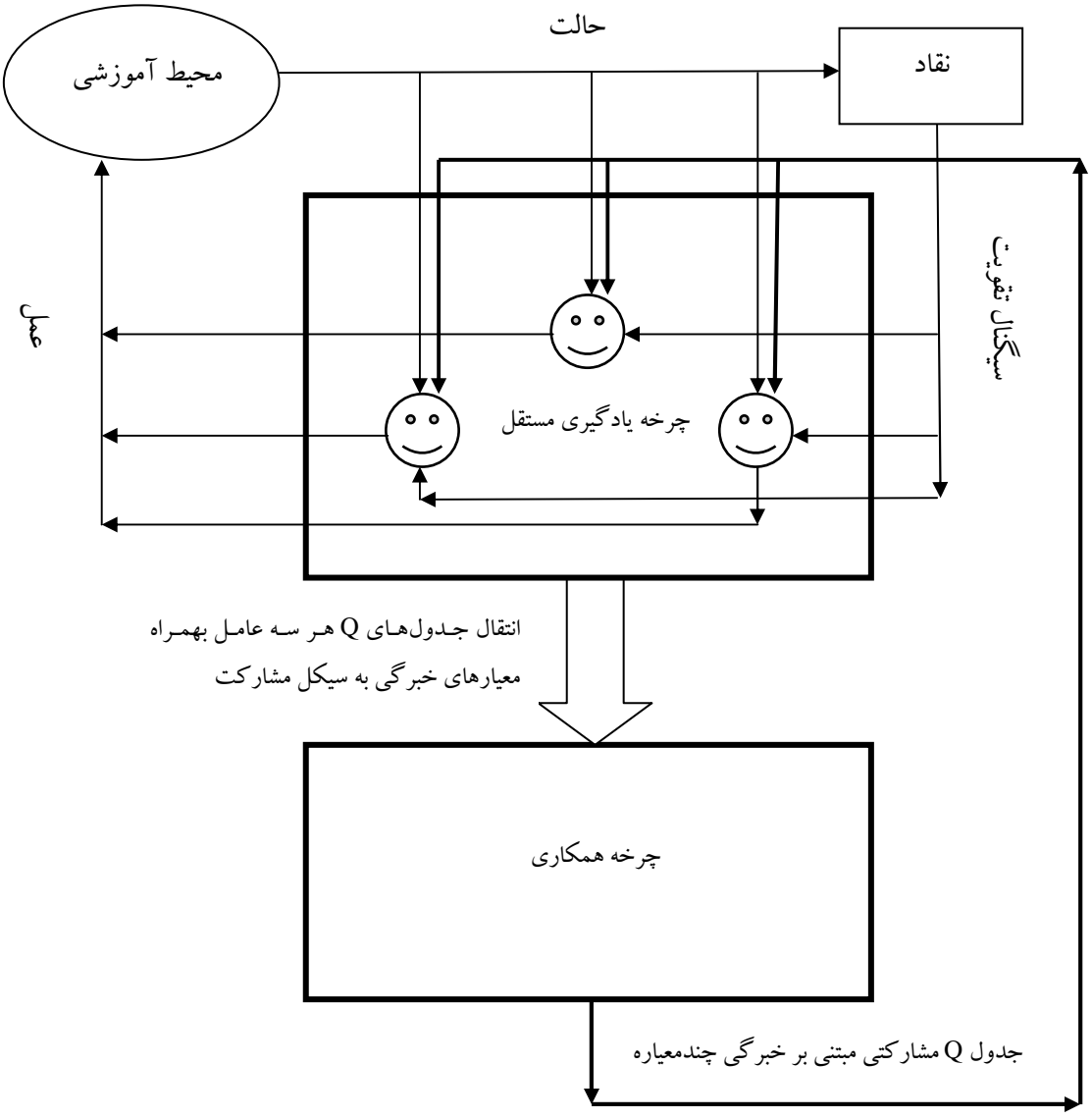
٤-۳-۱ جزییات الگوریتم پیشنهادی

الگوریتم MCE نوعی توسعه برای روش WSS که در [۲۴] پیشنهاد داده شده است، می‌باشد. نمای ساده-تری از الگوریتم در شکل ۴-۲ نشان داده شده است. این روش نیز همانند روش WSS دو چرخه کلی دارد: چرخه یادگیری مستقل(IL) و چرخه یادگیری مشارکتی(CL).

به طور خلاصه در این روش، N_s گام مشارکت تعریف شده است که هر گام از دو چرخه یادگیری مستقل و چرخه همکاری تشکیل شده است. در اولین گام مشارکت، جدول‌های Q تمامی عامل‌ها با صفر مقداردهی می‌شود. هر گام مشارکت با چرخه یادگیری مستقل شروع می‌شود که در آن عامل‌ها هر کدام به تنهایی و در محیط‌های جداگانه، انجام کاری یکسان را می‌آموزند. تعداد تلاش‌های عامل‌ها می‌تواند یکسان یا متفاوت باشد.در این تحقیق آزمایش‌ها در هر دو حالت تعداد تلاش یکسان و متفاوت انجام می‌شوند. در یک گام مشارکت، پس از اتمام چرخه یادگیری مستقل تمامی جدول‌های Q تولیدی حاصل از یادگیری مستقل عامل‌ها، به چرخه همکاری فرستاده می-

شوند. در چرخه همکاری، با استفاده از جدول‌های Q عامل‌ها، جدول مشارکتی مبتنی بر روش MCE ساخته می‌شود و در پایان گام به عنوان جدول Q مورد استفاده در انتخاب عمل در گام بعد به همه عامل‌های حاضر در سیستم نسبت داده می‌شود.

لازم به ذکر است که جدول Q حقیقی هر کدام از عامل‌ها بدون تغییر باقی می‌ماند و صرفا استفاده از جدول مشارکتی جدید موجب بهبود رفتار عامل‌ها در انتخاب عمل می‌شود و در واقع استفاده از این جدول به معنی اضافه شدن یک مکاشفه جدید به سیستم چندعامله است که موجب تسریع روند یادگیری مشارکتی خواهد شد. بدین ترتیب در پایان هر گام مشارکت، همه عامل‌های حاضر در سیستم دارای جدول‌های Q یکسانی در انتخاب عمل خواهندبود. هر کدام از چرخه‌های مذکور دارای زیر مراحلی هستند که در ادامه به طور دقیق‌تری توضیح داده خواهند شد.



شکل ۴-۲- نمایی کلی از روش پیشنهادی

۴-۳-۲ چرخه یادگیری مستقل

در این چرخه هر کدام از عامل‌ها با استفاده از الگوریتم یادگیری Q که در فصل پیش شرح داده شد، به صورت فردی و در محیط‌های جداگانه انجام یک کار یکسان را یاد می‌گیرند. لازم به ذکر است که در طول فاز یادگیری مستقل، مقادیر همه معیارهای خبرگی معرفی شده در فصل قبل نیز در سیستم نگهداری می‌شود. از این مقادیر در فاز یادگیری مشارکتی به منظور ساخت جدولهای مشارکتی مرتبط با هر کدام از معیارها و ساخت جدول Qمشارکتی نهایی استفاده می‌شود.

٤-۳-۳ چرخه همکاری

چرخه همکاری دارای ۳ مرحله اصلی است که در ادامه این بخش دقیق‌تر توضیح داده خواهند شد. البته لازم به ذکر است که چگونگی اجرای این سه مرحله در الگوریتم ۴-۱ مشخص شده است. ابتدا مراحل اول و دوم به ازای تمامی معیارهای خبرگی موجود انجام می‌شوند و سپس مرحله سوم انجام می‌شود.

- گام اول: انتخاب عاملی که سطح خبرگی‌اش از سایرین کمتر است.

اول از همه می‌بایست بین همه عامل‌ها، عاملی که از بقیه میزان خبرگی کمتری دارد، شناسایی شود. سوالی که مطرح می‌شود این است که شناسایی این عامل چه اهمیتی دارد؟ همانطور که در بخش‌های قبلی و در مورد روش WSS توضیح داده شد، در این روش هر یک از عامل‌ها به ساخت یک جدول مشارکتی از دیدگاه خود می‌پردازد که در آن از جدول Q سایر عامل‌هایی که از او خبره‌ترند، استفاده می‌کند. سه عامل را در نظر بگیرید. در فاز یادگیری مستقل عامل اول ۴ تلاش یادگیری و عامل‌های دوم و سوم هر کدام به ترتیب ۲ و ۱ تلاش یادگیری انجام می‌دهند. فرض کنید که هر چقدر عامل تعداد تلاش‌های یادگیری بیشتری را انجام دهد، خبره‌تر محسوب خواهد شد. عاملی که از سایرین خبره‌تر است، در ساخت جدول مشارکتی خود از جدول‌های سایرین هیچ استفاده‌ای نمی‌کند در حالی‌که عاملی که سطح خبرگی‌اش از سایرین کمتر است، در ساخت جدول مشارکتی‌اش از همه عامل‌های دیگر استفاده مطلوب را می‌برد. روابط ۴-۱، ۴-۲ و ۴-۳ جدول‌های مشارکتی حاصل از دیدگاه هر عامل را نشان می‌دهند.

$$CoQ_1 = w_{11}Q_1 \qquad ۴-۱$$

$$CoQ_2 = w_{21}Q_1 + w_{22}Q_2 \qquad ۴-۲$$

$$CoQ_3 = w_{31}Q_1 + w_{32}Q_2 + w_{33}Q_3 \qquad ۴-۳$$

به طور کلی می‌توان نتیجه گرفت که در پایان چرخه همکاری، عاملی که دارای سطح خبرگی کمتری است، جدول مشارکتی مفیدتری را ساخته است چرا که در ساخت جدول خود از تمامی افراد خبره بهره برده است. در مقابل جدول مشارکتی عاملی که از همه خبره‌تر است هیچ تغییری نمی‌کند و در واقع عامل با سایرین مشارکتی نداشته است. هر چند جدول Q عامل خبره‌تر ارزشمند است، ولی ترکیب آن با تجربیات سایرین – هر چند تجربیات سایرین موفقیت‌های چندانی در بر نداشته باشد – به بهبود کیفیت آن کمک شایانی خواهد کرد.

- گام دوم: ساخت جدول‌های مشارکتی متناظر با هر کدام از معیارهای خبرگی

در گام دوم، به ازای هر یک از معیارهای خبرگی با در نظر گرفتن جدول Q عاملی که از سایرین سطح خبرگی کمتری دارد، با استفاده از روش WSS برای معیار خبرگی مذکور، جدول Q مشارکتی محاسبه می‌شود. همان‌طور که قبلا ذکر شد در طول چرخه یادگیری مستقل، مقادیر همه معیارهای خبرگی در طول یادگیری مستقل محاسبه و در سیستم نگهداری می‌شوند و در گام دوم با استفاده از مقدار خبرگی و جدول Q عاملی که از بقیه خبرگی کمتری دارد، جدول مشارکتی بر مبنای روش WSS برای هر یک از معیارهای خبرگی محاسبه می‌شود(CoQ_i). همان‌طور که گفته شد بهتر است که جدول مشارکتی از دیدگاه عاملی که خبرگی‌اش از سایرین کمتر است، ساخته شود. لذا پیشنهاد می‌شود مانند شبه کد مطرح شده در شکل ۴-۱، به ازای هر یک از معیارهای خبرگی، عاملی که خبرگی کمتری دارد تعیین شود و سپس جدول مشارکتی مربوط به آن معیار خبرگی بر اساس دیدگاه همان عامل ساخته شود.

- گام سوم: ساخت جدول مشارکتی بر مبنای خبرگی چندمعیاره و استفاده از آن

هر کدام از جدول‌های مشارکتی بدست آمده در گام قبل بخشی از دانش عامل‌ها را نشان می‌دهند. مشارکت موثر رابطه مستقیمی با استفاده از قسمت‌های مختلف دانش عامل‌های حاضر در سیستم دارد. در گام سوم، با استفاده از مجموع جدول‌های مشارکتی مجزا، جدول مشارکتی چندمعیاره ساخته می‌شود(رابطه ۴-۴). انتظار می‌رود که جدول مشارکتی چندمعیاره تجربیات مختلف هر سه عامل را به خوبی در برداشته باشد. به عبارت دیگر جدول مشارکتی چندمعیاره نشان‌دهنده دانش جمعی عامل‌ها پس از گذران چرخه یادگیری مستقل است.

$$CoQ_{MCE} = \sum_{i=1}^{6} CoQ_i \qquad ۴-۴$$

مساله مهمی که وجود دارد چگونگی استفاده از دانش جمعی بدست آمده است. همان‌طور که در فصل پیش در مورد روش اشتراک وزن‌دار استراتژی توضیح داده شد در این روش پس از ترکیب جدول سه عامل، جدول

مشارکتی حاصل از ترکیب آن‌ها به عنوان جدول Q مورد استفاده در چرخه یادگیری مستقل بعدی استفاده می‌شود. به دلیل تفاوت‌هایی که در نحوه ساخت این دو جدول مشارکتی وجود دارد در روش پیشنهادی نمی‌توان مانند WSS جدول مشارکتی نهایی را جایگزین جدول‌های Q عامل‌ها کرد. در روش WSS، جدول مشارکتی بر اساس رابطه ۴-۵ ساخته می‌شود که در آن همه وزن‌ها مقادیری بین صفر و یک دارند و مجموع وزن‌ها نیز برابر یک است. لذا اگر فرض شود که در پایان یادگیری جدول Q همه عامل‌ها در نهایت به بهینه‌ترین مقدار خود یعنی Q^* همگرا خواهند شد، در نتیجه رابطه ۴-۶ نیز برقرار خواهد بود و جدول مشارکتی حاصل از اعمال روش WSS نیز در نهایت به Q^* همگرا خواهد شد.

$$CoQ_i = \sum_{j=1}^{n}(W_{ij} * Q_j) \qquad ۴-۵$$

$$CoQ_i = Q^* \sum_{j=1}^{n}(W_{ij}) = Q^* \qquad ۴-۶$$

مقادیر Q در طول یادگیری در بازه‌ای محدود قرار گرفته‌اند و در همه مسائل یادگیری با توجه به ماهیت محیط، مقادیر بیشینه و کمینه‌ای برای خانه‌های جدول Q وجود دارد. مقدار بیشینه و کمینه Q در یک مساله یادگیری که در آن تقویت‌های دریافتی از محیط بین دو مقدار $r_{\max}$ و $r_{\min}$ محدود شده‌اند و نرخ تخفیف نیز در رابطه ۴-۷ صدق می‌کند، توسط روابط ۴-۸ و ۴-۹ قابل محاسبه هستند. لازم به ذکر است که این دو رابطه تحت عنوان دو لم در [۴۱] به اثبات رسیده‌اند. مقادیر موجود در جدول Q^* نیز در محدوده بین مقدار بیشینه و کمینه محاسبه شده توسط روابط ۴-۸ و ۴-۹ قرار می‌گیرند.

$$0 \leq \gamma < 1 \qquad ۴-۷$$

$$\max Q(s_t, a_t) = \frac{r_{max}}{1-\gamma} \qquad ۴-۸$$

$$\min Q(s_t, a_t) = \frac{r_{min}}{1-\gamma} \qquad ۴-۹$$

عبور از مقدار بیشینه منجر به واگرا شدن یادگیری و گذر از مقدار کمینه منجر به تاخیر در همگرایی خواهد شد. به دلیل این که جدول‌های مشارکتی حاصل از اعمال روش WSS همواره در محدوده کمینه و بیشینه تعریف شده می‌گیرند، خللی در روند همگرایی در یادگیری بوجود نخواهد آمد.

واگرا شدن یادگیری با عبور مقادیر Q از مقدار بیشینه مجاز تعریف شده در رابطه ۴-۸ را می‌توان بر اساس ماهیت انتخاب اعمال توسط تابع انتخاب عمل بولتزمن نیز نشان داد. تابع انتخاب عمل بولتزمن تابعی نمایی است و در صورتی که مقادیر Q بیش از اندازه بزرگ شوند، این تابع با توجه به ذات نمایی‌اش قادر به نشان دادن اختلاف بین

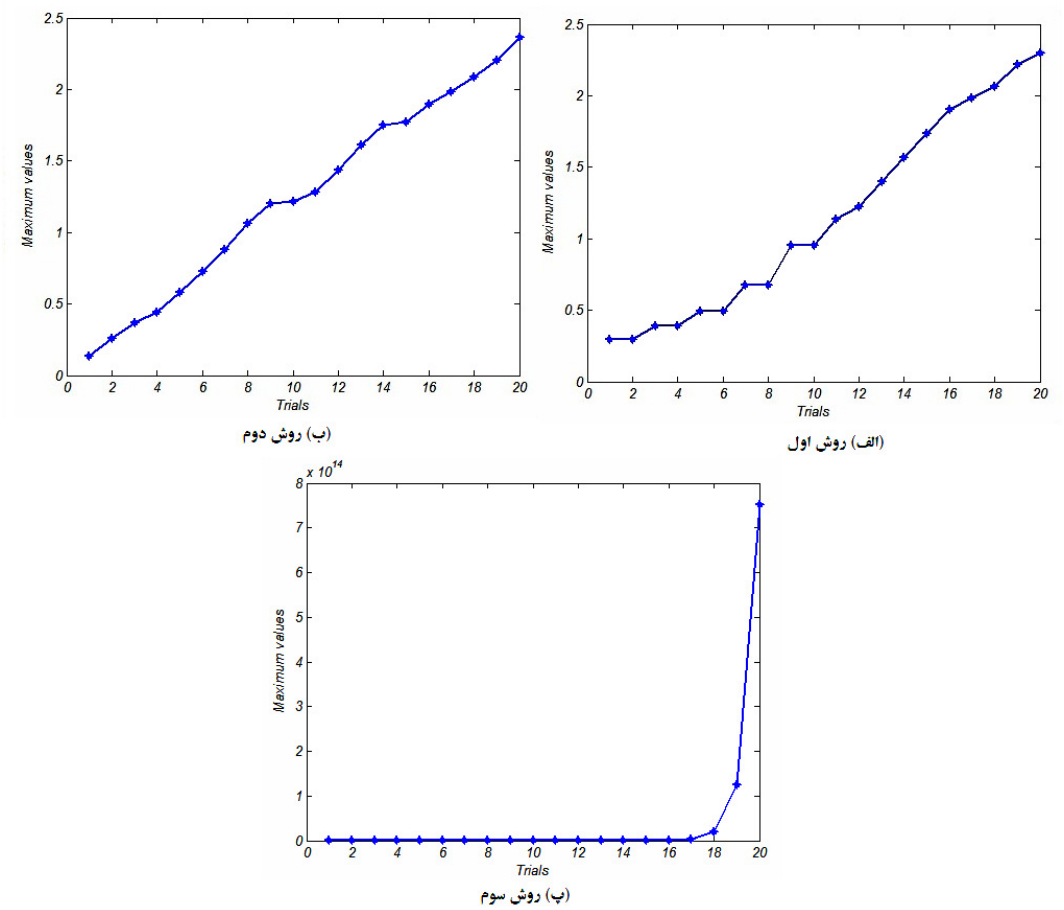
اهمیت اعمال نیست و در واقع با بزرگ شدن بیش از اندازه مقادیر Q، تابع انتخاب عمل بولتزمن توانایی خود در انتخاب صحیح عمل را به تدریج از دست می‌دهد. به عبارت دیگر رشد بیش از حد مقادیر Q منجر به ناتوانی عامل در کنترل رفتارش خواهد شد.

در ساخت جدول مشارکتی مبتنی بر خبرگی چندمعیاره طبق رابطه ۴-۱۰ جدول نهایی از محدوده‌های مجاز تعریف شده عبور می‌کند و لذا استفاده از این جدول به صورت جایگزینی آن با جدول Q مورد استفاده عامل‌ها منجر به واگرایی روند یادگیری خواهد شد. البته در گام‌های ابتدایی مشارکت، به دلیل این‌که مقادیر Qهنوز به اندازه کافی رشد نکرده‌اند، استفاده از جدول مشارکتی و جایگزینی آن صرفا به صورت تسریع روند همگرایی خود را نشان می-دهد، اما پس از گذشت چند گام مشارکت، مقدار بیشینه جدول مشارکتی از مقدار بیشینه مجاز تعریف شده برای مقادیر Q گذر خواهد کرد و یادگیری واگرا خواهد شد.

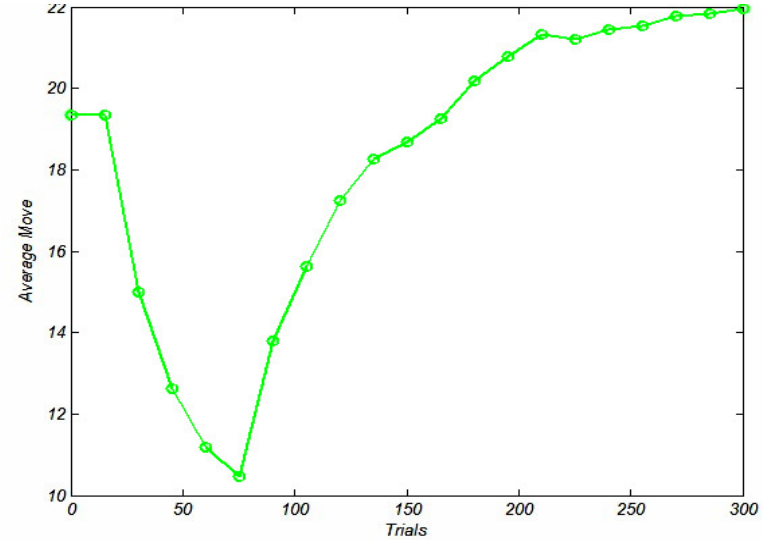
در شکل ۴-۳ روند رشد مقادیر Q در سه روش متفاوت در محیط آموزشی پلکان مارپیچ[1] نشان داده شده است. روش اول، یادگیری مستقل و بدون همکار مبتنی بر یادگیری Q است. در این حالت سه عامل بدون به اشتراک گذاری اطلاعات، انجام یک وظیفه یکسان را می‌آموزند. روش دوم یادگیری مشارکتی بر مبنای خبرگی و با استفاده از روش WSS و روش سوم یادگیری مشارکتی بر مبنای خبرگی چندمعیاره است که در آن جدول مشارکتی چندمعیاره جایگزین جدول‌های Q عامل‌ها می‌شود. در نمودارهای موجود در شکل ۴-۳ مقادیر بیشینه Q در طول ۲۰ گام همکاری رسم شده است. همان‌طور که در شکل نیز نشان داده شده است، استفاده از جدول مشارکتی چندمعیاره به صورت جایگزینی به رشد بیش از حد مقادیر Q منجر می‌شود در صورتی که در دو روش دیگر رشد مقادیر Q به صورت یکنواخت صورت می‌پذیرد. در روش اول و دوم مقادیر Q در طول ۲۰ گام همکاری تا مقداری برابر ۲.۵ رشد کرده در حالی که در روش سوم مقدار بیشینه Q تا حدود$۸*۱۰^{۱۴}$ رشد کرده است. در شکل ۴-۴ روند واگرایی روش سوم نشان داده شده است.

لازم به ذکر است که نتایج ارائه شده در شکل‌های ۴-۳ و ۴-۴ صرفا برای نشان دادن رفتار روش است و در این بخش از ذکر جزییات مربوط به پیاده‌سازی پرهیز می‌شود. در فصل بعد، جزییات پیاده‌سازی به طور دقیق معرفی خواهد شد. با توجه به دلایل مطرح شده و تفاوت نحوه رشد دو جدول مشارکتی تولید شده در روش WSS و روش پیشنهادی، استفاده از جدول مشارکتی تولید شده توسط روش پیشنهادی به صورت جایگزین‌کردن جدول مشارکتی با جدول Q عامل‌ها منجر به واگرایی روش خواهد شد.

[1] Maze



شکل ۴-۳- مقایسه روند رشد مقادیر بیشینه جدول مشارکتی Q مبتنی بر خبرگی چندمعیاره در مقایسه با سایر روش‌ها (الف) یادگیری Q مستقل بدون همکار (ب) یادگیری مشارکتی بر مبنای خبرگی (پ) یادگیری مشارکتی بر مبنای خبرگی چندمعیاره

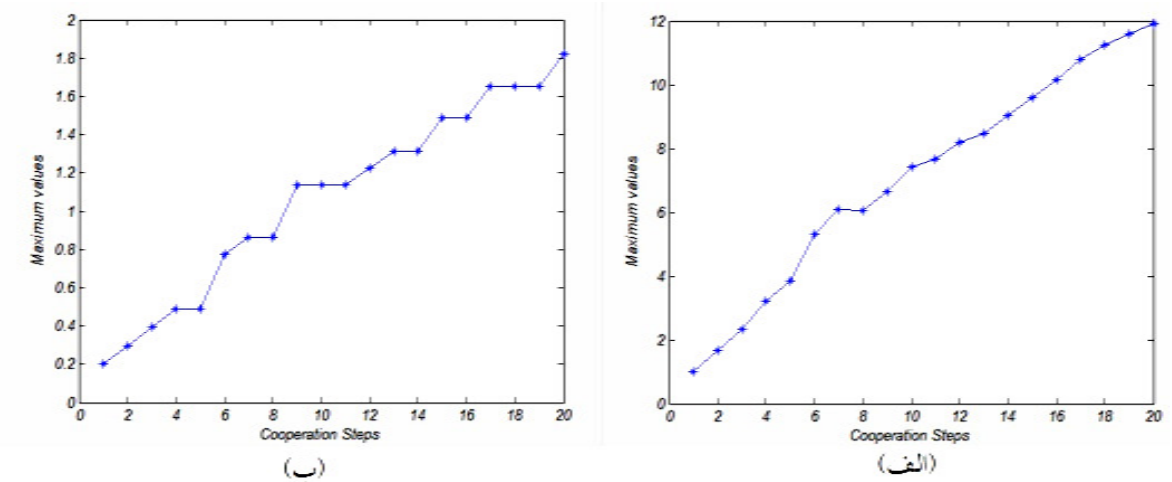


شکل ۴-۴- روند واگرایی روش پیشنهادی در حالت استفاده از جدول مشارکتی چندمعیاره به صورت جایگزین کردن آن با جدول Q عامل‌ها

ایده معرفی شده در این پایان‌نامه برای رفع مشکل بوجود آمده، از ایده موجود در [۴۲] که در فصل قبل با عنوان تسریع یادگیری Q با استفاده از مکاشفه معرفی شد، الهام گرفته شده است. همان‌طور که گفته شد، جدول مشارکتی تولید شده نشان‌دهنده دانش جمعی عامل‌ها است و لذا می‌توان از این دانش به عنوان یک راهنما برای انتخاب بهتر اعمال عامل‌ها استفاده کرد. در روش پیشنهادی از مقادیر جدول مشارکتی به عنوان مقادیر Q در تابع انتخاب عمل استفاده می‌شود. رابطه ۴-۱۰ نحوه انتخاب عمل توسط تابع بولتزمن بر اساس جدول مشارکتی چند معیاره تولید شده را نشان می‌دهد.

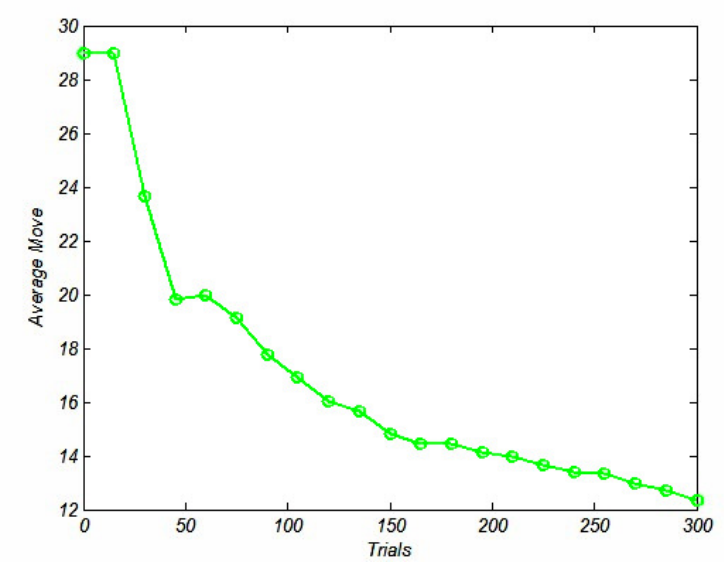
$$\pi(s_t) = \arg max_{a_t} \left(\frac{e^{CoQ_{MCE}(s_t,a_t)}}{\sum e^{CoQ_{MCE}(s_t,a_t)}}\right) \qquad ۴-۱۰$$

همان‌طور که در شبه کد الگوریتم پیشنهادی نیز بیان شده است، از جدول مشارکتی تولید شده فقط در انتخاب عمل استفاده می‌شود و عامل‌ها پس از انتخاب عمل بر مبنای جدول مشارکتی تولید شده، با استفاده از سیگنال تقویت دریافتی از محیط جدول Q قبلی خود را به‌روز رسانی می‌کنند و دوباره در چرخه همکاری بعدی، جدول مشارکتی بر اساس جدول‌های Q هر سه عامل ساخته خواهد شد. بدین ترتیب روند رشد بیش از اندازه جدول مشارکتی Q که در روش سوم موجود در شکل ۴-۳ نشان داده شد، محدود می‌شود و لذا از اثر مخرب رشد مقادیر جدول بر انتخاب عمل کاسته می‌شود. شکل ۴-۵ نحوه رشد مقادیر جدول مشارکتی Q و رشد مقادیر Q یکی از عامل‌های حاضر در سیستم چندعامله را نشان می‌دهد. همان‌طور که در شکل دیده می‌شود، محدودتر شدن رشد جدول مشارکتی Q می‌تواند انتخاب عمل را به خوبی هدایت کند و در نتیجه رشد مقادیر جدول Q عامل‌ها نیز محدود می‌شود. شکل ۵-۵ نیز روند همگرایی یادگیری را پس از استفاده جدول مشارکتی به صورت راهنما در انتخاب عمل پیشنهادی نشان می‌



شکل ۴-۵- (الف) نحوه رشد مقادیر بیشینه جدول مشارکتی Q، (ب) نحوه رشد مقادیر بیشینه جدول Q یکی از عامل‌های حاضر در سیستم

دهد.



شکل ۴-۶- روند همگرایی یادگیری پس از استفاده از جدول مشارکتی به صورت راهنما در انتخاب

۴-۴ اثبات درستی روش پیشنهادی

در این بخش درستی منطق استفاده شده در روش پیشنهادی با عنوان یک قضیه اثبات می‌شود. همان‌طور که در بخش قبل گفته شد، ایده مورد استفاده برای حل مشکل رشد جدول Q، استفاده از جدول مشارکتی به عنوان راهنما در انتخاب عمل می‌باشد. در فصل قبل الگوریتم HAQL برای تسریع یادگیری Q با استفاده از مکاشفه‌ها معرفی شد. برای اثبات درستی منطق روش پیشنهادی، می‌توان از قضایای مطرح شده در [۴۱] برای اثبات صحت الگوریتم HAQL، استفاده کرد. در مراجع مرتبط با الگوریتم HAQL [۴۲-۴۱]، بر این نکته تاکید شده است که در صورتی که مکاشفه تعریف شده کران‌دار باشد، الگوریتم با استفاده از آن واگرا نخواهد شد و مقادیر جدول Q نامتناهی نخواهند شد. برای اثبات درستی روش پیشنهادی، کافی است ثابت شود که جدول مشارکتی Q معادل مجموع جدول Q عامل و یک مکاشفه است که مکاشفه مذکور کران‌دار است و لذا جدول Q هیچ‌گاه دارای مقادیر نامتناهی نخواهد شد و الگوریتم واگرا نخواهد شد.

قضیه- اگر در یک یادگیری مشارکتی Q چندعامله که n عامل در سیستم حضور دارند و هر کدام از عامل‌ها در یک محیط مستقل و جداگانه با خصوصیات محیط مارکوف معین قرار دارند، در حالی که تعداد حالت‌ها و عمل‌های قابل انتخاب محدود است، مقادیر پاداش‌های دریافتی در محیط نیز به صورت کراندار و نرخ تخفیف γ نیز بین صفر و یک قرار دارد، انتخاب عمل بر مبنای رابطه ۴-۱۰ موجب تولید مقادیر نامتناهی در مقادیر جدول Q نخواهد شد.

اثبات – به دلیل این که n عامل به طور مستقل در حال یادگیری هستند و صورت مساله در واقع چگونگی بهبود انتقال دانش بین آن‌هاست، لذا در اثبات قضیه مباحث مربوط به انجام عمل مشترک بین عامل‌ها و تغییر محیط مطرح نمی‌شود. طبق آن‌چه در [۴۱] گفته شده است می‌توان با تعریف یک مکاشفه مناسب، به تسریع یادگیری Q تک عامله کمک شایانی کرد. طبق [۴۱] رابطه انتخاب عمل در زمان حضور یک مکاشفه به صورت رابطه ۴-۱۱ است:

$$\pi(S_t) = \arg max_{a_t} \left(\frac{e^{\widehat{Q}(S_t,a_t)+\varepsilon H_t(S_t,a_t)}}{\sum e^{\widehat{Q}(S_t,a_t)+\varepsilon H_t(S_t,a_t)}}\right) \quad ۱۱\text{-}۴$$

کافی است ثابت شود که $\widehat{Q}(S_t,a_t) + \varepsilon H_t(S_t,a_t) = CoQ_{MCE}(S_t,a_t)$ برقرار است در حالی که مکاشفه مورد استفاده کراندار است(رابطه ۴-۱۲). در ادامه این اثبات مقدار ε برابر یک در نظر گرفته شده است.

$$(\forall\, s_t, a_t)\; h_{min} \le H(s_t,a_t) \le h_{max} \quad ۱۲\text{-}۴$$

$$\widehat{Q}(S_t,a_t) + H_t(S_t,a_t) = CoQ_{MCE}(S_t,a_t)$$

$$H_t(S_t,a_t) = CoQ_{MCE}(S_t,a_t) - \widehat{Q}(S_t,a_t) \quad ۱۳\text{-}۴$$

$$\mathrm{CoQ_{MCE}} = \sum_{i=1}^{6} \mathrm{CoQ_i} \quad ۱۴\text{-}۴$$

$$CoQ_i = \sum_{j=1}^{n} w_{ij} \times Q_j(S_t,a_t) \quad ۱۵\text{-}۴$$

$$0 \le w_{ij} \le 1 \quad ۱۶\text{-}۴$$

بر طبق لم ۱ و لم۲ مطرح شده در [۴۱] مقادیر بیشینه و کمینه Q نیز به صورت رابطه ۴-۱۷ است.

$$\frac{r_{min}}{1-\gamma} \le Q_j(S_t,a_t) \le \frac{r_{max}}{1-\gamma} \quad ۱۷\text{-}۴$$

$$\min\left(0,\frac{r_{min}}{1-\gamma},\frac{r_{max}}{1-\gamma}\right) \le w_{ij} \times Q_j(S_t,a_t) \le \max\left(0,\frac{r_{min}}{1-\gamma},\frac{r_{max}}{1-\gamma}\right) \quad ۱۸\text{-}۴$$

$$n \times \min\left(0,\frac{r_{min}}{1-\gamma},\frac{r_{max}}{1-\gamma}\right) \le \sum_{j=1}^{n} w_{ij} \times Q_j(S_t,a_t) \le n \times \max\left(0,\frac{r_{min}}{1-\gamma},\frac{r_{max}}{1-\gamma}\right) \quad ۱۹\text{-}۴$$

$$n \times \min\left(0,\frac{r_{min}}{1-\gamma},\frac{r_{max}}{1-\gamma}\right) \le CoQ_i \le n \times \max\left(0,\frac{r_{min}}{1-\gamma},\frac{r_{max}}{1-\gamma}\right) \quad ۲۰\text{-}۴$$

با استفاده از روابط ۴-۱۷ و ۴-۲۱ می‌توان اثبات کرد که مکاشفه مورد استفاده کراندار است و لذا درستی قضیه اثبات می‌شود.

$$6n \times \min\left(0, \frac{r_{min}}{1-\gamma}, \frac{r_{max}}{1-\gamma}\right) \leq \sum_{i=1}^{6} CoQ_i \leq 6n \times \max\left(0, \frac{r_{min}}{1-\gamma}, \frac{r_{max}}{1-\gamma}\right) \quad ۲۱-۴$$

$$6n \times \min\left(0, \frac{r_{min}}{1-\gamma}, \frac{r_{max}}{1-\gamma}\right) - \frac{r_{min}}{1-\gamma} \leq \sum_{i=1}^{6} CoQ_i - \hat{Q}(S_t, a_t) \leq 6n \times \max\left(0, \frac{r_{min}}{1-\gamma}, \frac{r_{max}}{1-\gamma}\right) - \frac{r_{max}}{1-\gamma} \quad ۲۲-۴$$

$$6n \times \min\left(0, \frac{r_{min}}{1-\gamma}, \frac{r_{max}}{1-\gamma}\right) - \frac{r_{min}}{1-\gamma} \leq H_t(S_t, a_t) \leq 6n \times \max\left(0, \frac{r_{min}}{1-\gamma}, \frac{r_{max}}{1-\gamma}\right) - \frac{r_{max}}{1-\gamma} \quad ۲۳-۴$$

۴-۵ نتیجه گیری

در این فصل ابتدا مفهوم خبرگی چندمعیاره و لزوم استفاده از آن از دیدگاه روانشناختی و محاسباتی مورد بررسی قرار گرفت و سپس روش یادگیری مشارکتی Q مبتنی بر خبرگی چندمعیاره پیشنهاد داده شد. یکی از نورآوری‌های موجود در روش پیشنهادی استفاده از جدول مشارکتی در انتخاب موثر اعمال است. در انتهای فصل دلایل درستی این نوآوری در قالب یک قضیه و اثبات آن بیان شده است. روش پیشنهاد داده شده دارای پشتوانه خوبی از دیدگاه نظری است و در فصل بعد عملکرد آن بر روی دو محیط آزمایشی متفاوت مورد ارزیابی قرار خواهد گرفت.

فصل پنجم
شبیه‌سازی و آزمایش‌های انجام گرفته

۵-۱ مقدمه

برای ارزیابی الگوریتم پیشنهادی از شبیه‌سازی دو محیط آموزشی مطرح در یادگیری ماشین (پلکان مارپیچ و صید-صیاد) استفاده شده است. انتخاب این دو نوع مساله به دلیل ماهیت متفاوت یادگیری در آن‌ها صورت گرفته است. در مساله پلکان مارپیچ، عامل با نوعی وظیفه نقشه‌سازی مواجه است. در واقع عامل در پی رسیدن به نقطه هدف تعیین شده بدون برخورد با موانع موجود در مسیر و در کوتاه‌ترین زمان ممکن است. به دلیل ثابت بودن محیط، مساله پلکان مارپیچ ماهیتی ایستا[1] دارد و در چنین مساله‌ای جدول Q عامل یادگیرنده قادر به نشان دادن مسیر بهینه عامل در محیط خواهد بود. در مساله صید و صیاد، عامل صیاد در حال یادگیری رفتار بهینه برای شکار است. هدف کلی عامل صیاد، آموختن چگونگی تعقیب صید و به دام انداختن آن در کوتاه‌ترین زمان ممکن است. در محیط صید و صیاد به دلیل حرکت هدف، مساله ماهیتی پویا دارد و جدول Q عامل یادگیرنده نشانگر رفتار بهینه عامل با توجه به مکان صید در محیط است. تفاوت موجود در ماهیت دو مساله به خوبی قادر است تا کیفیت روش پیشنهادی را ارزیابی کند.

نکته دیگری که در ارزیابی عملکرد روش پیشنهادی مهم است، کیفیت سطح سادگی وظیفه است. اگرچه در مسائل ساده، زمان انجام آزمایش‌های یادگیری کاهش می‌یابد ولی تفاوت کارآیی الگوریتم‌های یادگیری در

[1] Static

چنین مسائلی محسوس نیست. دو محیط آموزشی مطرح شده ارائه‌دهنده دو سطح متفاوت از پیچیدگی وظیفه هستند. محیط پلکان مارپیچ به عنوان یک وظیفه ساده و در مقابل محیط صید و صیاد به عنوان یک وظیفه پیچیده محسوب می‌شوند. در این فصل ابتدا محیط‌های شبیه‌سازی پلکان مارپیچ و صید-صیاد به همراه جزییات و فرضیات در نظر گرفته شده در پیاده‌سازی معرفی می‌شوند و سپس نتایج حاصل از به‌کارگیری الگوریتم پیشنهادی در این محیط‌ها مورد بررسی قرار خواهد گرفت. به منظور ارزیابی هر چه دقیق‌تر و معنادارتر کارآیی روش‌ها، پنج نوع آزمایش مختلف طراحی شده است که در ادامه فصل به طور کامل معرفی خواهند شد.

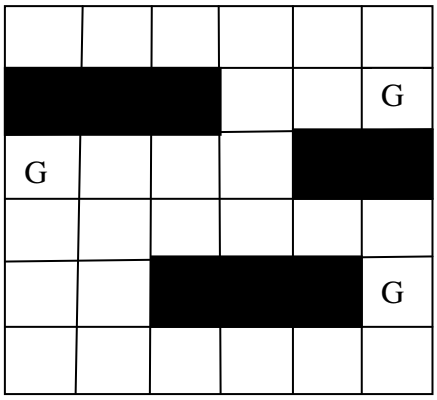
۵-۲ معرفی محیط‌های آموزشی مورد استفاده

در حوزه سیستم‌های چندعامله، چندین محیط آموزشی وجود دارد که برای سنجش میزان کارآیی روش‌های پیشنهاد شده به کار می‌رود. دو محیط آموزشی پلکان مارپیچ و صید و صیاد جزو معتبرترین محیط‌های آزمایشی برای سنجش الگوریتم‌های جدید است[۴۸]. در حوزه یادگیری ماشین نیز از این دو محیط برای مطالعه و مقایسه فرآیند-های یادگیری مختلف استفاده می‌شود. در ادامه این بخش، دو محیط ذکر شده معرفی خواهند شد.

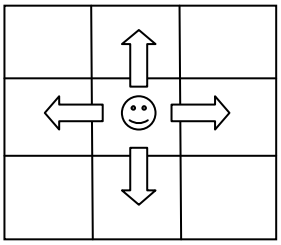
۵-۲-۱ مساله پلکان مارپیچ

محیط در این شبیه‌سازی یک پلکان مارپیچ به ابعاد ۶*۶ است که در شکل ۵-۱ نمایش داده شده است. در این محیط تعدادی مانع قرار داده شده و هر عامل وظیفه دارد که رسیدن به خانه‌های هدف که با حرف G برچسب-گذاری شده‌اند، را بیاموزد. در واقع عامل‌ها در این محیط در حال یادگیری وظیفه‌ی پیداکردن مسیر بهینه تا خانه‌های هدف موجود در محیط هستند. عامل‌ها از الگوریتم یادگیری Q که نوعی یادگیری تقویتی به شمار می‌رود، برای یادگیری وظیفه مسیریابی استفاده می‌کنند. هر تلاش یادگیری با قرار گرفتن تصادفی عامل در یکی از خانه‌های خالی محیط شروع می‌شود و هنگامی پایان می‌پذیرد که عامل به خانه هدف که دارای برچسب G است، برسد. هر عامل در هر خانه قادر است در صورت نبود مانع در مسیر به هر یک از چهار جهت مختلف که در شکل ۵-۲ نشان داده شده‌اند، حرکت کند.

در طول آموزش، در صورت برخورد عامل به مانع مقدار تنبیه ۱ و در صورت رسیدن به هدف مقدار پاداش ۱۰دریافت می‌کند. زمانی که عامل به مانع برخورد نکرده باشد و به هدف هم نرسیده باشد، متناسب با فاصله‌ای که تا نزدیک‌ترین هدف دارد و بر حسب نزدیکی‌اش به هدف، پاداش دریافت می‌کند.



شکل ۵-۱- محیط پلکان مارپیچ



تصویر ۵-۲- اعمال ممکن در محیط

پاداش در این حالت به صورت رابطه ۵-۱ محاسبه می‌شود:

$$Reward = \frac{1}{distance\ between\ the\ agent\ and\ the\ goal} \qquad ۱-۵$$

در محیط پلکان مارپیچ هر خانه از محیط (به جز موانع موجود) معادل یک حالت در جدول Q است و از این رو در یک پلکان مارپیچ همانند شکل ۵-۱، ۲۸ حالت مختلف وجود دارد و در هر حالت حداکثر ۴ عمل امکان‌پذیر است. جدول‌های Q، جدول‌هایی به ابعاد ۴*۲۸ هستند.

۵-۲-۲ مساله صید و صیاد

مساله آموزش شکار به صیاد از جمله مسائل کلاسیک در یادگیری است و بستر آزمایشی مناسبی برای روش‌های یادگیری به شمار می‌رود[۴۹]. نمونه این مساله در حیوانات و حتی در انسان نیز قابل مشاهده است. حیوانات وحشی، صید خود را تعقیب می‌کنند، صید نیز از دست آن‌ها می‌گریزد و در حین تعقیب برای فریب آنان حرکات مارپیچ انجام می‌دهد و گاه به دلیل اینرسی حرکتی صیاد، موفق به فرار می‌شود. صیاد نیز سعی می‌کند حرکات صید را پیش‌بینی کرده و راهی کوتاه برای رسیدن به او انتخاب کند. پیش‌بینی دقیق‌تر احتمال موفقیت بیشتری به همراه دارد. مسائلی همچون انهدام اهداف متحرک در جنگ نیز شبیه این مساله است[۴۰].

در این مساله چندعامله، دو نوع عامل صیاد و صید در محیط به تعامل می‌پردازند. عامل صیاد در پی یافتن، تعقیب، پیش‌بینی حرکات آینده و رسیدن به عامل صید است. خصوصیات محیطی که صید و صیادها در آن قرار گرفته‌اند و چگونگی در نظر گرفتن حالات و اعمال نقش مهمی در شبیه‌سازی مساله یاد شده دارد.

در این پایان‌نامه، از یک محیط پیوسته دوبعدی ۱۰*۱۰ که یک عامل صیاد و یک عامل صید در آن قرار دارند، استفاده شده است. سه عامل در سه محیط جداگانه و مستقل از یکدیگر قرار گرفته‌اند و پس از انجام چرخه-های یادگیری فردی، در سیکل همکاری تجربیات بدست آمده را با یکدیگر به اشتراک می‌گذارند. عامل‌های صیاد در این محیط در حال یادگیری وظیفه‌ی شکار عامل صید هستند و از الگوریتم یادگیری Q برای یادگیری وظیفه شکار استفاده می‌کنند. عامل صید هم در این شبیه‌سازی در حالت فاقد یادگیری در نظر گرفته شده است و حرکتی تصادفی دارد.

عامل‌های صیاد دارای یک میدان دید هستند و می‌توانند حضور صید را تنها در این محدوده مکان‌یابی کنند. در واقع عامل صیاد، تنها هنگامی صید را مشاهده می‌کند که صید در فاصله‌ای کمتر از حداکثر میدان دید او قرار گرفته باشد. در این پیاده‌سازی عامل‌ها برای حرکت از تعیین دو مولفه سرعت و زاویه حرکت بهره می‌برند. هر صیاد می‌تواند با سرعتی بین ۰ و ۱ حرکت کند و هر صید نیز می‌تواند با سرعتی بین ۰ و ۰.۵ حرکت کند. حداکثر سرعت صیاد می‌بایست بیش از حداکثر سرعت صید باشد تا احتمال شکار صید توسط صیاد وجود داشته باشد. هم‌چنین زاویه حرکت هر دو نوع عامل صید و صیاد می‌تواند از ۰ تا ۳۶۰ درجه تغییر یابد.

زمانی که صیاد بتواند خود را به فاصله‌ای کمتر از ۰.۵ نسبت به صید برساند، قادر به شکار صید خواهد بود و در این صورت پاداشی برابر با R خواهد گرفت و در سایر حالات تنبیه‌ای برابر P دریافت خواهد کرد. حالت صیاد بر مبنای مختصات صید در دستگاه محلی صیاد تعیین می‌شود. اگر به عنوان مثال عامل صیاد در مختصات (x_h, y_h) و عامل صید در مختصات (x_p, y_p) قرار گرفته باشند، مختصات صید در دستگاه محلی صیاد به صورت (d_{xh}, d_{yh}) خواهد بود که توسط روابط ۵-۲ و ۵-۳ تعیین می‌شود.

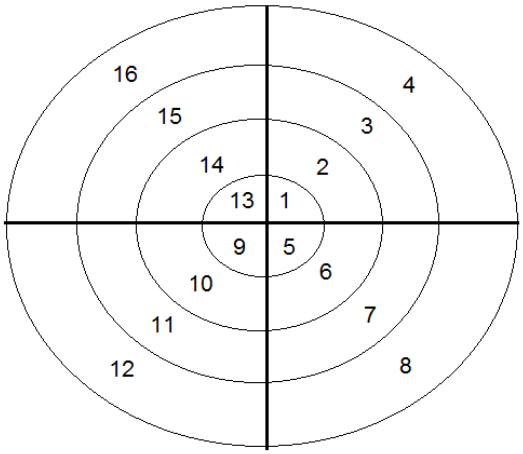
$$d_{xh} = x_p - x_h \qquad ۵\text{-}۲$$

$$d_{yh} = y_p - y_h \qquad ۵\text{-}۳$$

اگر صید در میدان دید صیاد قرار نداشته باشد، یک حالت پیش‌فرض برای صیاد در نظر گرفته می‌شود. اعمال صیاد نیز ترکیبی از سرعت و زاویه حرکت است(رابطه ۵-۴).

$$action = (v, \theta) \qquad ۵\text{-}۴$$

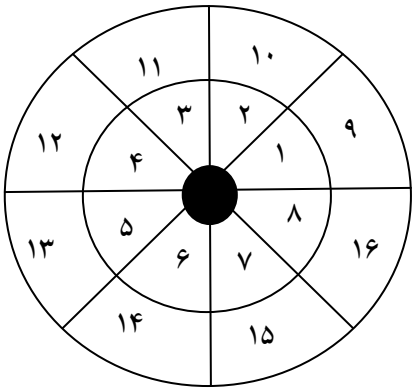
همان‌طور که گفته شد، عامل‌های صیاد قابلیت یادگیری با استفاده از الگوریتم یادگیری Q را دارند. جدول Q تنها می‌تواند تعداد محدودی حالت و عمل را پوشش دهد، از این رو می‌بایست اعمال و حالت‌های صیاد گسسته-سازی شود. هر صیاد میدان دیدی برابر ۲ دارد. از این رو اجزای مختصاتی d_{xh}, d_{yh} هر کدام در بازه‌ی [۲و۲-] تغییر می‌کنند. برای گسسته‌سازی این محدوده، می‌توان آن را به بازه‌هایی با طول ۰.۵ تقسیم‌بندی کرد. تصویر ۵-۳ حالت‌های گسسته‌سازی شده را برای یک صیاد نشان می‌دهد. تعداد حالات صیاد با در نظر گرفتن حالت پیش‌فرض – حالتی که صید در محدوده دید صیاد قرار ندارد – برابر ۱۷ خواهد بود.



تصویر ۵-۳- تقسیم‌بندی حالت صیاد: هر قسمت نشان‌دهنده مکانی است که اگر صید در آن قرار بگیرد، صیاد در حالت متناظر با شماره نوشته شده در آن قرار خواهد گرفت. حالت شماره ۱۷ حالت پیش‌فرض است.

قدم بعدی گسسته‌سازی اعمال صیاد است. برای گسسته‌سازی سرعت، می‌توان بازه [۱و۰] را به دو زیر بازه هر کدام به طول ۰.۵ تقسیم‌بندی کرد. برای گسسته‌سازی زاویه نیز بازه [۳۶۰و۰] قابل تقسیم به ۸ بازه کوچک‌تر به اندازه ۴۵ درجه است. اعمال صیاد، ترکیبی از سرعت و تغییر جهت است در نتیجه تعداد کل اعمال برابر با ۱۶ می‌باشد. صیاد از میان ۱۶ عمل ممکن یک عمل را با احتمالی متناسب با توزیع بولتزمن انتخاب می‌کند. عمل انتخاب شده، محدوده سرعت و زاویه حرکت را مشخص می‌کند. صیاد سرعت و زاویه دقیق‌تر را به طور تصادفی با توزیع یکنواخت روی محدوده تعیین شده، انتخاب می‌کند. شکل ۵-۴ همه اعمال ممکن صیاد را نشان می‌دهد. هر عمل ترکیبی از سرعت و تغییر زاویه است و با انجام هر عمل، صیاد در یکی از ۱۶ خانه قرار می‌گیرد.

همان‌گونه که پیشتر توضیح داده شد، عامل صید استفاده شده در شبیه‌سازی الگوی حرکت تصادفی دارد و اعمال خود را به تصادف و با توزیع یکنواخت از میان اعمال ممکن انتخاب می‌کند.



شکل ۵-۴- اعمال ممکن صیاد در محیط

۵-۳ معرفی حالت‌های شبیه‌سازی

در همه شبیه‌سازی‌ها سه عامل در سه محیط جداگانه حضور دارند و مساله مشارکت به طراحی مکانیزمی مناسب برای اشتراک اطلاعات بین سه عامل برای بهبود نحوه یادگیری‌شان اطلاق می‌شود. برای بررسی بیشتر و بهتر، آزمایش‌ها در دو حالت مختلف بر روی محیط‌ها انجام شده است. در حالت اول فرض بر این است که عامل‌های موجود در سیستم قبل از انجام مشارکت، تعداد تلاش‌های یادگیری یکسانی را انجام داده‌اند و در حالت دوم فرض شده که عامل‌ها قبل از انجام مشارکت تعداد تلاش‌های متفاوتی را انجام داده‌اند. زمانی که به عامل‌ها اجازه داده می‌شود که تعداد تلاش‌های متفاوتی را انجام دهند، در واقع هر کدام از آن‌ها میزان تجربیات متفاوتی را بدست خواهند آورد و هر کدام سطوح مختلفی از خبرگی خواهند داشت. بررسی‌های انجام شده در [۲۴] نشان‌دهنده این واقعیت است که در حالت تعداد تلاش‌های متفاوت، روش‌های یادگیری مشارکتی بر مبنای خبرگی کارآیی بهتری دارند. آن‌چه در این پایان‌نامه مد نظر است تاکید بر موثرتر بودن انتقال اطلاعات همه جانبه – خبرگی چندمعیاره – به جای اطلاعات محدود است. از این رو عملکرد روش پیشنهادی بر روی حالت اول نیز بررسی شده است و نتایج آن با سایر روش‌های موجود در این حالت نیز مقایسه شده است.

هر آزمایش یادگیری شامل N_s گام مشارکت[1] است. هر گام مشارکت شامل N_c چرخه یادگیری مستقل و یک چرخه همکاری است. پس از انجام چرخه‌های یادگیری مستقل، نوبت به همکاری میان عامل‌ها می‌رسد و به تعداد تلاش انجام شده در طول N_c چرخه یادگیری مستقل، بازه مشارکت[2] گفته می‌شود. هر چرخه یادگیری نیز شامل B تلاش یادگیری مستقل عامل‌های هدف‌یاب است. در هر تلاش یادگیری مستقل تنها یک عامل صیاد در

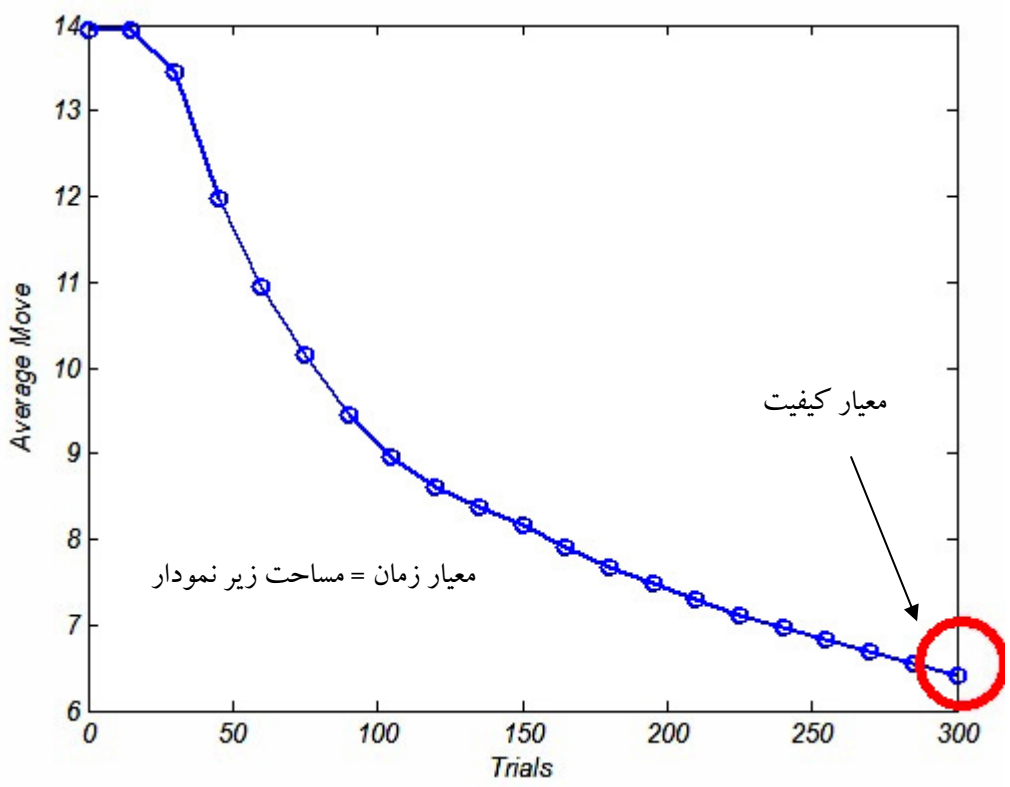
[1] Cooperation Step
[2] Cooperation Interval

محیط قرار داده می‌شود و عامل صیاد بر اساس الگوریتم یادگیری Q که در فصل ۳ توضیح داده شد، به یادگیری مستقل می‌پردازد. در ابتدای هر تلاش یادگیری مستقل عامل‌ها به مکانی تصادفی در محیط منتقل می‌شوند. تلاش یادگیری مستقل زمانی به پایان می‌رسد که عامل صیاد موفق به شکار صید شده باشد. در چرخه همکاری تمامی عامل‌های صیاد به حالت مشارکت در یادگیری تغییر حالت می‌دهند و طبق الگوریتم مطرح شده در فصل قبل به مبادله خبرگی‌ها، ساخت جدول مشارکتی مبتنی بر خبرگی چندمعیاره می‌پردازند.

۵-۴ معرفی آزمایش‌های طراحی شده و هدف آن‌ها

همان‌گونه که در فصل‌های پیشین مطرح شد، هدف کلیه الگوریتم‌های یادگیری افزایش کیفیت و سرعت یادگیری است. در این پایان‌نامه از دو معیار کیفیت و زمان برای ارزیابی روش‌ها بهره گرفته شده است. معیار کیفیت، متوسط تعداد حرکات گروهی است که پس از اتمام یادگیری، هر یک از عامل‌ها برای رسیدن به هدف نیاز دارند. این معیار میزان موثر بودن یادگیری را نشان می‌دهد و هم‌چنین می‌توان از آن برای بررسی رفتار روش‌ها در مورد همگرایی نیز استفاده کرد. به عبارت دیگر هر چه تعداد حرکات مورد نیاز برای رسیدن به هدف کم‌تر باشد، روش کیفیت بالاتری دارد. در واقع معیار کیفیت بیان‌کننده پیشرفت نهایی در رفتار عامل است که از یادگیری بدست آمده است.

معیار زمان بر اساس متوسط مجموع تعداد حرکات گروهی که عامل‌ها در طول یادگیری انجام می‌دهند، تعریف شده است. از این معیار می‌توان برای مقایسه زمان یادگیری در روش‌های مختلف استفاده کرد. بدیهی است که هر چه زمان مصرف شده برای انجام یادگیری در حالی که یادگیری کیفیت خوبی دارد، کمتر باشد روش موثرتر خواهد بود و سریع‌تر به جواب بهینه همگرا شده است. به عبارت دیگر معیار زمان به نوعی بیان کننده میزان موثر بودن رفتار عامل در طول فرآیند یادگیری است. در شکل ۵-۵ معیارهای کیفیت و زمان بر روی نمودار حاصل از یادگیری نشان داده شده است. لازم به ذکر است که در این نمودار، محور افقی معرف تعداد تلاش‌های یادگیری است و محور عمودی نیز متوسط تعداد حرکات لازم برای رسیدن به هدف را در طول تلاش‌های انجام شده، نشان می‌دهد. برای ارزیابی کارآیی الگوریتم‌های پیشنهادی، پنج نوع آزمایش متفاوت طراحی شده است که هر کدام یکی از مزیت‌های روش پیشنهادی را ملموس‌تر می‌کنند. در این بخش آزمایش‌های طراحی شده و هدف آن‌ها معرفی خواهند شد.



شکل ۵-۵- معیارهای کیفیت و زمان

- **آزمایش اول: مقایسه روش پیشنهادی با روش‌های موجود در دو حالت تجربیات یکسان و تجربیات متفاوت عامل‌ها**

همان‌گونه که گفته شد، روش‌هایی که تا کنون در حوزه یادگیری مشارکتی بر مبنای خبرگی پیشنهاد داده شده‌اند، در حالتی که تجربیات عامل‌ها متفاوت است موفق‌تر از حالتی که تجربیات یکسانی دارند، عمل کرده‌اند. هدف از انجام آزمایش اول مقایسه کلی روش پیشنهادی و سایر روش‌های موجود در این زمینه است. هم‌چنین آزمایش‌ها در دو حالت تعداد تلاش یکسان و تعداد تلاش متفاوت نیز انجام شده‌اند تا عام‌تر بودن روش پیشنهادی نیز در مقایسه با سایر روش‌ها مورد بررسی قرار بگیرد.

- **آزمایش دوم: بررسی اثر افزایش دما بر همکاری**

پارامتر دما میزان تصادفی بودن انتخاب اعمال را در طول یادگیری کنترل می‌کند. بدیهی است که تنظیم درست این پارامتر نرخ اکتشاف را در محیط تعیین خواهد کرد. هر اندازه دما در محیط بالاتر باشد، تصادفی بودن در انتخاب اعمال بالاتر خواهد رفت و در نتیجه عامل امکان انجام تجربیات مختلفی را خواهد داشت. هدف از طراحی آزمایش دوم بررسی تاثیر افزایش دما و در نتیجه آن افزایش گوناگونی تجربیات عامل‌ها بر عملکرد روش پیشنهادی است.

- **آزمایش سوم: بررسی اثر طول بازه مشارکت بر کیفیت یادگیری**

هدف از طراحی آزمایش سوم، بررسی اثر تاثیر طول بازه مشارکت تعریف‌شده بر بهبود روش پیشنهادی است. افزایش طول بازه مشارکت به معنی افزایش فرصت عامل‌ها برای یادگیری مستقل و جمع‌آوری تجربیات بیشتر است. در دنیای انسانی در یک تیم که به صورت مشارکتی انجام یک عمل را فرا می‌گیرند، تعیین صحیح بازه مشارکت نقش مهمی در کیفیت یادگیری دارد. در واقع طراحی این آزمایش تلاش برای پاسخ به این سوال است که آیا در دنیای عامل‌ها نیز روابطی همانند دنیای انسانی برقرار است یا خیر؟

- **آزمایش چهارم: بررسی اثر افزایش تعداد معیارهای خبرگی مورد استفاده**

یکی از نوآوری‌های روش پیشنهادی، نگاه همه جانبه به تجربیات عامل‌ها است. روش پیشنهادی جنبه‌های مختلف رفتاری عامل را می‌سنجد و سپس بر اساس آن عامل به تصمیم‌گیری درباره آینده‌اش می‌پردازد. هدف از طراحی آزمایش چهارم، نشان دادن تاثیر همه‌جانبه بودن اطلاعات بر عملکرد روش پیشنهادی است. در این آزمایش عملکرد روش پیشنهادی به ازای استفاده از تعداد متفاوتی از معیارهای خبرگی مورد بررسی قرار می‌گیرد.

- **آزمایش پنجم: بررسی پایایی روش نسبت به حضور اغتشاش**[1]

یکی از ویژگی‌های مورد توجه در سیستم‌های چندعامله قابلیت تحمل‌پذیری خطا در آن‌ها است[۲و ۴۵]. روش پیشنهادی بر مبنای انتقال اطلاعات بین عامل‌ها بنیاد نهاده شده است. در دنیای واقعی سیستم‌های ارتباطی مورد استفاده برای انتقال اطلاعات همواره مقداری اغتشاش نیز به محتوای اطلاعاتی مورد مبادله می‌افزایند. هدف از طراحی آزمایش پنجم بررسی عملکرد روش پیشنهادی در حضور اغتشاش است. نتایج حاصل از این آزمایش بیان‌کننده قابلیت روش پیشنهادی برای پیاده‌سازی در محیط‌های واقعی است.

۵-۵ نتایج شبیه‌سازی و آزمایش‌های انجام گرفته

در آزمایش اول روش اشتراک وزن‌دار استراتژی بر مبنای تمامی معیارهای خبرگی به علاوه روش‌های معدل‌گیری ساده[۲][۱۰] و یادگیری بدون همکار[۳] و روش پیشنهادی بررسی شده است. به دلیل ذات تصادفی یادگیری Q و برای اطمینان بیشتر، نتایج نمایش‌داده‌شده در همه آزمایش‌ها حاصل ۲۰ بار اجرای آزمایش و میانگین‌گیری بین نتایج است.

---

[1] Noise
[2] Simple Averaging
[3] Individual Learning

برای ارائه ملموس‌تر نتایج، روند تغییر متوسط تعداد حرکات گروهی عامل‌ها تا رسیدن به هدف در گام‌های همکاری در نمودارهایی رسم شده است. این نمودارها بیان کننده کیفیت پویایی رفتار روش پیشنهادی هستند. در نمودارها و جدول‌های مطرح شده در این فصل از حروف اختصاری زیر برای مشخص‌کردن روش‌های تعیین خبرگی و انواع آزمایش‌ها استفاده شده است:

IL = Individual Learning,
SS = Strategy Sharing,
Nrm = WSS (Normal),
Abs = WSS (Absolute),
Po = WSS (Positive),
No = WSS (Negative),
Gr = WSS (Gradient),
Av= WSS (Average move),
MCE= Multi Criteria Expertness based cooperative learning.

۵-۵-۱ پارامترهای یادگیری و مشارکت

برخی از پارامترهای یادگیری و مشارکت در آزمایش‌های مختلف تغییر می‌کنند و به همین دلیل در توضیح هر آزمایش مقادیر متغیر آن‌ها ذکر خواهد شد اما در جاهایی که به صراحت ذکر نشده باشد از مقادیر نرخ یادگیری (β) = ۰.۰۱، پارامتر دما (T) = ۰.۴، پارامتر تخفیف (γ)= ۰.۹ استفاده شده است. در ابتدای یادگیری سلول‌های جدول Q با صفر مقداردهی اولیه شده‌اند. ضریب تاثیرپذیری از دیگران (α) نیز برابر ۰.۹ مقداردهی شده است. پارامترهای مشارکت در محیط‌های متفاوت و در حالت‌های مختلف شبیه‌سازی متفاوت است. در زیر بخش‌های بعدی مقادیرپارامترهای مشارکت مورد استفاده در آزمایش‌ها به ازای حالت‌های مختلف معرفی می‌شوند.

- محیط آزمایشی پلکان مارپیچ

آزمایش‌های انجام شده در این محیط در دو حالت اجرا شده است. در حالت اول، عامل‌های حاضر در سیستم تعداد تلاش‌های یکسانی را انجام می‌دهند و از نظر میزان تجربه‌ای که بدست آورده‌اند، مشابه هستند. در حالت دوم، عامل‌های حاضر در سیستم تعداد تلاش‌های متفاوتی را در طول یادگیری انجام می‌دهند و از این رو میزان تجربیاتی که بدست خواهند آورد، متفاوت است. در حالت اول، هر عامل در یک چرخه یادگیری مستقل ۵ تلاش انجام می‌دهد و پارامترهای N_C و N_S با ۱ و ۲۰۰ مقداردهی شده‌اند و عامل‌ها در مجموع و به صورت تیمی ۳۰۰۰ تلاش یادگیری انجام می‌دهند. در حالت دوم در هر چرخه یادگیری مستقل عامل اول ۴ تلاش، عامل دوم ۲ تلاش و

عامل سوم ۱ تلاش انجام می‌دهند و پارامترهای N_C و N_S با ۲ و ۲۰۰ مقداردهی شده‌اند و عامل‌ها در مجموع و به صورت تیمی ۲۸۰۰ تلاش یادگیری انجام می‌دهند.

- محیط آزمایشی صید و صیاد

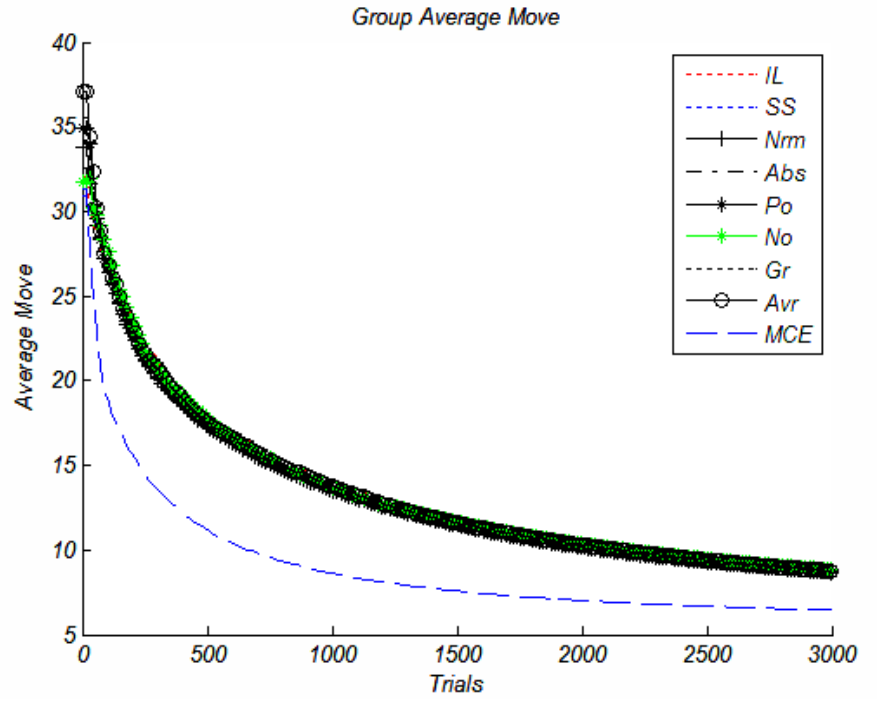
آزمایش‌های انجام شده در این محیط نیز در دو حالت اجرا شده است. در حالت اول، عامل‌های حاضر در سیستم تعداد تلاش‌های یکسانی را انجام می‌دهند و از نظر میزان تجربه‌ای که بدست آورده‌اند، مشابه هستند. در حالت دوم، عامل‌های حاضر در سیستم تعداد تلاش‌های متفاوتی را در طول یادگیری انجام می‌دهند و از این رو میزان تجربیاتی که بدست خواهند آورد، متفاوت است. در حالت اول، هر عامل در یک چرخه یادگیری مستقل ۳ تلاش انجام می‌دهد و پارامترهای N_C و N_S با ۵ و ۳۰ مقداردهی شده‌اند و عامل‌ها در مجموع و به صورت تیمی ۱۳۵۰ تلاش یادگیری انجام می‌دهند. در حالت دوم در هر چرخه یادگیری مستقل عامل اول ۶ تلاش، عامل دوم ۳ تلاش و عامل سوم ۱ تلاش انجام می‌دهند و پارامترهای N_C و N_S با ۵ و ۳۰ مقداردهی شده‌اند و عامل‌ها در مجموع و به صورت تیمی ۱۵۰۰ تلاش یادگیری انجام می‌دهند.

۵-۵-۲ آزمایش اول- مقایسه روش پیشنهادی با سایر روش‌ها

در این بخش علاوه بر سنجش کارآیی روش در مقایسه با سایر روش‌ها بر اساس دو معیار کیفیت و زمان، در شکل‌های ۵-۶ تا ۵-۹ پویایی رفتار سیستم در روش پیشنهادی و سایر روش‌ها نشان داده شده است. نتایج نشان می‌دهند که روش پیشنهادی نقش موثری در سرعت بخشیدن به همگرایی در یادگیری و همین‌طور بهبود کیفیت آن دارد.

جدول ۵-۱- مقایسه پارامترهای کیفیت و زمان در روش‌های مختلف در محیط پلکان مارپیچ - تعداد تلاش‌های یکسان

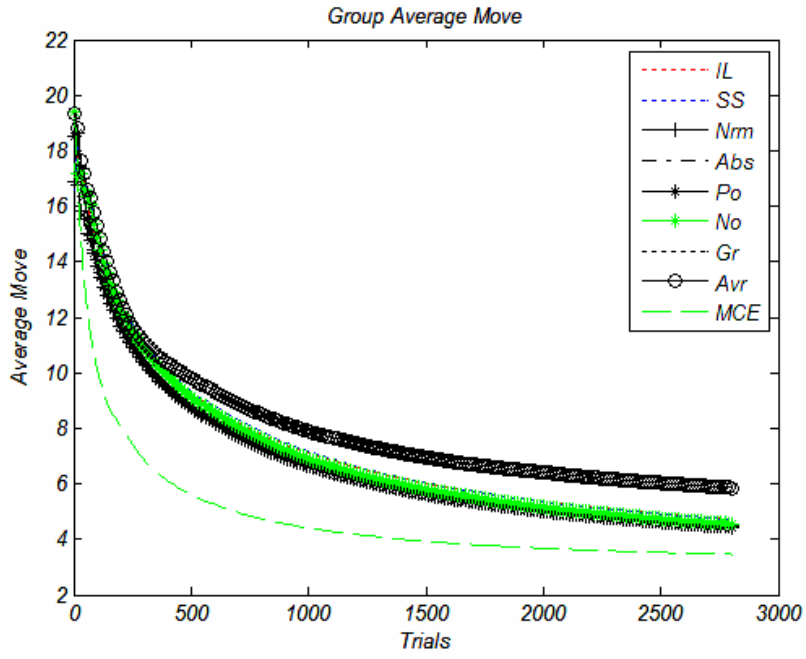
	IL	SA	Nrm	Ab	Po	Ne	Gr	Av	MCE
کیفیت	۸.۸۸۳۷	۸.۷۴۱۶	۸.۷۴۹۹	۸.۷۱۲۶	۸.۷۸۷۱	۸.۷۴۰۱	۸.۶۸۲۹	۸.۸۲۵۱	۶.۵۲۸۷
درصد بهبود		٪۱.۵	٪۱.۵	٪۱.۹	٪۱	٪۱.۶	٪۲.۲	٪۰.۶	٪۲۶
زمان	۲۷۵۱	۲۶۹۹	۲۶۷۱	۲۶۸۸	۲۷۱۷	۲۶۹۶	۲۶۹۰	۲۷۳۹	۱۸۸۱
درصد بهبود		٪۱.۸	٪۲.۹	٪۲.۲	٪۱.۲	٪۱.۹	٪۲.۲	٪۰.۴۳	٪۳۱



شکل ۵-۶- پویایی رفتار روش در محیط پلکان مارپیچ در حالت تعداد تلاش یکسان

جدول ۵-۲- مقایسه پارامترهای کیفیت و زمان در روش‌های مختلف در محیط پلکان مارپیچ - تعداد تلاش‌های متفاوت

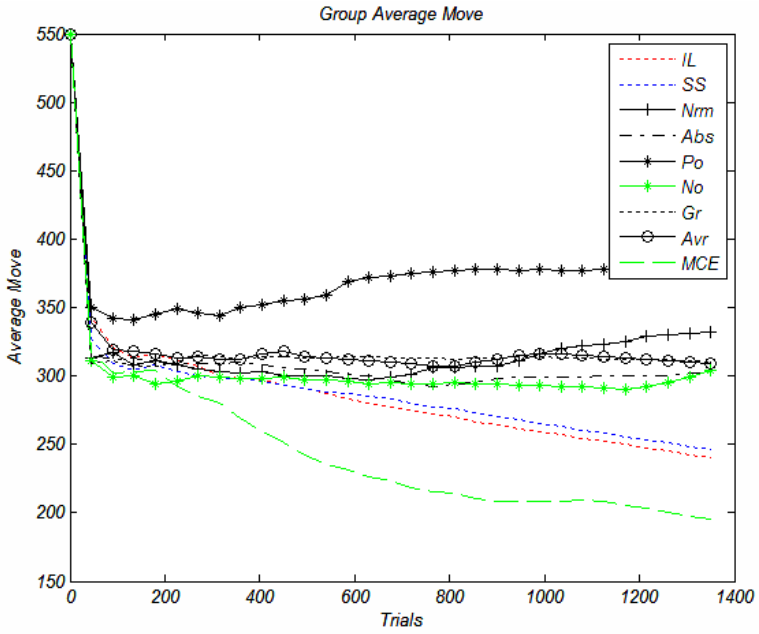
	IL	SA	Nrm	Ab	Po	Ne	Gr	Av	MCE
کیفیت	۴.۶۹۱۱	۴.۷۰۸۱	۴.۵۲۲۶	۴.۴۸۳۵	۴.۵۱۰۵	۸.۲۵۱۷	۴.۳۶۷۱	۷.۰۲۰۳	۴.۱۲۰۲
درصد بهبود		٪-۰.۳۶	٪۳.۵	٪۴.۴	٪۳.۸	٪-۷۵	٪۶.۹	٪-۴۹	٪۱۲.۱۶
زمان	۱۴۴۵.۴	۱۴۴۸.۶	۱۳۹۷.۹	۱۳۸۵.۲	۱۳۹۲.۷	۱۹۷۳.۹	۱۳۶۳.۷	۱۸۶۴.۱	۱۰۹۸.۱
درصد بهبود		٪-۰.۲۲	٪۳.۲	٪۴.۱	٪۳.۶	٪-۳۶.۵۶	٪۵.۶	٪-۲۸	٪۲۴.۰۲



شکل ۵-۷- پویایی رفتار روش در محیط پلکان مارپیچ در حالت تعداد تلاش متفاوت

جدول ۵-۳ - مقایسه پارامترهای کیفیت و زمان در روش‌های مختلف در محیط صید و صیاد - تعداد تلاش‌های یکسان

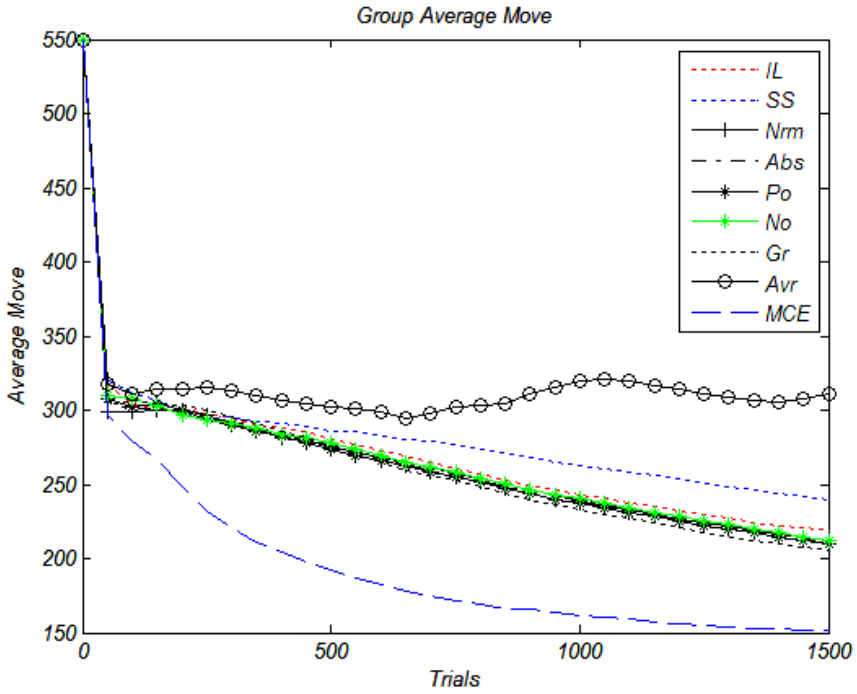
MCE	Av	Gr	Ne	Po	Ab	Nrm	SA	IL	
۱۹۵.۳۶	۳۰۸.۸۸	۳۰۹.۸	۳۰۴.۴۵	۳۷۸.۹۹	۳۰۳.۲	۳۳۲.۱۱	۲۴۶.۲۹	۲۴۰.۴۳	کیفیت
٪۱۸.۷	٪-۲۸	٪-۲۸	٪-۲۶	٪-۵۷	٪-۲۶	٪-۳۸	٪-۲.۴		درصد بهبود
۷۶۸۲.۲	۹۹۶۵	۹۹۱۷.۴	۹۴۴۰.۶	۱۱۵۱۹	۹۶۳۳.۳	۹۸۸۳.۲	۸۹۶۶.۵	۸۹۲۸	زمان
٪۱۳.۹۵	٪-۱۱	٪-۱۱	٪-۵.۷	٪-۲۹	٪-۷.۸	٪-۱۰	٪-۰.۴		درصد بهبود



شکل ۵-۸- پویایی رفتار روش در محیط صید و صیاد در حالت تعداد تلاش یکسان

جدول ۵-۴ - مقایسه پارامترهای کیفیت و زمان در روش‌های مختلف در محیط صید و صیاد - تعداد تلاش‌های متفاوت

MCE	Av	Gr	Ne	Po	Ab	Nrm	SA	IL	
۱۵۱.۴۳	۳۱۰.۵۱	۲۰۵.۲	۲۱۲.۳۷	۲۱۰.۰۱	۲۱۱.۲۶	۲۱۲.۲۵	۲۳۹.۳۸	۲۱۸.۱۶	کیفیت
٪۳۰.۵	٪-۴۲	٪۵.۹	٪۲.۶	٪۳.۷	٪۳.۱	٪۲.۷	٪-۹.۷		درصد بهبود
۶۱۷۰.۲	۹۸۲۷.۳	۸۱۶۷.۵	۸۲۸۵.۹	۸۲۱۱.۸	۸۲۴۹.۲	۸۲۴۷.۱	۸۷۷۷.۶	۸۳۹۰.۳	زمان
٪۲۶.۴۶	٪-۱۷	٪۲.۶	٪۱.۲	٪۲.۱	٪۱.۶	٪۱.۷	٪-۴.۶		درصد بهبود



شکل ۵-۹- پویایی رفتار روش در محیط صید و صیاد در حالت تعداد تلاش متفاوت

در [۴۰] آزمایش‌های زیادی بر روی عملکرد روش اشتراک وزن‌دار استراتژی (WSS) در حالتی که تعداد تلاش عامل‌ها یکسان و متفاوت است، انجام شده است. لازم به ذکر است که در [۴۰] آزمایش‌ها در محیط صید و صیاد انجام گرفته‌اند. در [۴۰] با استفاده از آزمایش‌های انجام شده بر این نکته تاکید شده است که روش WSS در حالتی که تعداد تلاش‌ها یکسان است، روش خوبی به شمار نمی‌رود و بهبود چندانی ایجاد نمی‌کند. در مقابل زمانی که تعداد تلاش‌ها متفاوت است، روش WSS عملکرد خوبی دارد. زمانی که تعداد تلاش‌ها متفاوت است، عامل‌های حاضر در سیستم سطوح خبرگی متفاوتی خواهند داشت و از این رو مشارکت بر مبنای خبرگی معنای بیشتری می‌یابد.

در تحقیق حاضر، آزمایش اول بر روی دو محیط پلکان مارپیچ و صید و صیاد و در هر دو حالت ذکر شده انجام گرفته است. پلکان مارپیچ نسبت به صید و صیاد محیط ساده‌تری به شمار می‌رود. در هر دو محیط روش WSS در حالت تعداد تلاش متفاوت بهتر از تعداد تلاش یکسان عمل می‌کنند. در محیط پلکان مارپیچ به دلیل ساده بودن محیط، عملکرد WSS در حالت تعداد تلاش یکسان نیز قابل قبول است و در مجموع روش WSS در محیط پلکان مارپیچ که محیط ساده‌تری است، عملکرد بهتری دارد.

طبق نتایج جدول ۵-۱ تا ۵-۴، روش یادگیری مشارکتی بر مبنای خبرگی چندمعیاره (MCE) در هر دو حالت و در هر دو محیط قادر است که بهبود خوبی در یادگیری بدون همکار ایجاد کند و از این رو می‌توان گفت که روش پیشنهادی نسبت به سایر روش‌های مبتنی بر خبرگی روش عام‌تری محسوب می‌شود و در حالت‌ها و محیط-

های متفاوتی قادر است یادگیری را به طور چشمگیری بهبود بخشد. در شکل‌های ۵-۶ تا ۵-۹ پویایی روش‌های مورد آزمایش نشان داده شده است. همان‌طور که در شکل‌ها نیز دیده می‌شود، MCE در مقایسه با سایر روش‌ها موجب تسریع همگرایی در یادگیری می‌شود و به این دلیل قادر است پارامتر زمان و کیفیت را به خوبی بهبود ببخشد.

۵-۵-۳ آزمایش دوم – بررسی اثر افزایش دما بر همکاری

پارامتر دما در الگوریتم یادگیری Q میزان تصادفی بودن در انتخاب اعمال را کنترل می‌کند. همانطور که در [۳۶] ذکر شده است، یکی از شرط‌های همگرایی یادگیری Q این است که در طول یادگیری همه جفت‌های حالت و عمل ممکن به اندازه کافی دیده شده باشند. به عبارت دیگر تنظیم صحیح پارامتر دما نرخ اکتشاف در محیط را تنظیم می‌کند و بر روند همگرایی یادگیری تاثیرگذار است. جداول ۵-۵ تا ۵-۱۲ نتایج حاصل از افزایش دما بر همکاری را نشان می‌دهند.

جدول ۵-۵- بررسی اثر تغییر پارامتر دما بر معیار کیفیت در روش پیشنهادی و درصد بهبود نسبت به یادگیری بدون همکار- تعداد تلاش یکسان– محیط پلکان مارپیچ

پارامتر دما	۰.۱	۰.۳	۰.۴	۰.۵	۰.۶	۰.۷	۰.۹
یادگیری Q	۶.۸۵۲۸	۸.۳۳۹۸	۸.۸۴۶۸	۹.۳۵۳۴	۹.۹۰۶۵	۱۰.۳۰۷	۱۱.۱۲۱
MCE	۶.۱۶۰۲	۶.۳۳۱۸	۶.۵۱۳۷	۶.۶۸۸۹	۶.۷۶۵۲	۶.۹۲۵۶	۷.۲۲۷۵
درصد بهبود	٪۱۰.۱	٪۲۴.۰۷	٪۲۶.۳۷	٪۲۸.۴۸	٪۳۱.۷	٪۳۲.۸	٪۳۵.۰۱

جدول ۵-۶- بررسی اثر تغییر پارامتر دما بر معیار زمان در روش پیشنهادی و درصد بهبود نسبت به یادگیری بدون همکار- تعداد تلاش یکسان– محیط پلکان مارپیچ

پارامتر دما	۰.۱	۰.۳	۰.۴	۰.۵	۰.۶	۰.۷	۰.۹
یادگیری Q	۲۰۱۸	۲۵۸۳.۲	۲۷۳۶.۱	۲۹۰۶	۳۰۸۲	۳۱۸۹.۷	۳۳۹۹.۱
MCE	۱۵۷۴.۸	۱۷۷۵.۱	۱۸۶۶.۵	۱۹۵۹.۹	۱۹۸۹.۷	۲۰۷۰.۹	۲۱۸۶.۲
بهبود	٪۲۱.۹۶	٪۳۱.۲۸	٪۳۱.۷	٪۳۲.۵۵	٪۳۵.۴۴	٪۳۵.۰۷	٪۳۵.۶۸

جدول ۵-۷- بررسی اثر تغییر پارامتر دما بر معیار کیفیت در روش پیشنهادی و درصد بهبود نسبت به یادگیری بدون همکار- تعداد تلاش متفاوت– محیط پلکان مارپیچ

پارامتر دما	۰.۱	۰.۳	۰.۴	۰.۵	۰.۶	۰.۷	۰.۹
یادگیری Q	۳.۷۳۵۷	۴.۴۲۱۱	۴.۷۱۰۵	۴.۹۷۲۷	۵.۱۹۲۵	۵.۴۰۲۴	۵.۸۶۱۸
MCE	۵.۴۷۵۳	۳.۴۴۴۶	۳.۹۸۹۶	۳.۵۸۹۵	۳.۸۹۲	۳.۶۶۹۴	۳.۷۵۴۲
درصد بهبود	٪-۴۶	٪۲۲.۰۸	٪۱۵.۳	٪۲۷.۸۱	٪۲۵.۰۴	٪۳۲.۰۷	٪۳۵.۹۵

جدول ۵-۸- بررسی اثر تغییر پارامتر دما بر معیار زمان در روش پیشنهادی و درصد بهبود نسبت به یادگیری بدون همکار- تعداد تلاش متفاوت- محیط پلکان مارپیچ

پارامتر دما	۰.۱	۰.۳	۰.۴	۰.۵	۰.۶	۰.۷	۰.۹
یادگیری Q	۱۰۷۶.۳	۱۳۵۸.۴	۱۴۴۸.۱	۱۵۳۴.۲	۱۶۲۱.۹	۱۶۷۹.۲	۱۸۰۸.۲
MCE	۱۱۲۴.۵	۹۴۷.۷۴	۱۰۸۱	۱۰۴۶.۲	۱۱۱۲.۹	۱۰۷۱.۶	۱۱۰۹.۸
درصد بهبود	٪-۴.۴	٪۳۰.۲۳	٪۲۵.۳۵	٪۳۱.۸	٪۳۱.۳۸	۳۶.۱۸	٪۳۸.۶۲

جدول ۵-۹- بررسی اثر تغییر پارامتر دما بر معیار کیفیت در روش پیشنهادی و درصد بهبود نسبت به یادگیری بدون همکار- تعداد تلاش یکسان- محیط پلکان مارپیچ

پارامتر دما	۰.۱	۰.۳	۰.۴	۰.۵	۰.۶	۰.۷	۰.۹
یادگیری Q	۱۶۰.۲۸	۲۱۸.۵	۲۳۸.۲۷	۲۵۶.۱۴	۲۶۵.۳۷	۲۷۲.۹۱	۲۸۲.۵۲
MCE	۱۸۳.۰۵	۲۰۹.۶۱	۲۱۳.۵۷	۲۲۴.۳۶	۲۲۲.۹۱	۲۳۴.۸۵	۲۴۲.۴
درصد بهبود	٪-۱۴.۲	٪۴	٪۱۰.۳	٪۱۲.۴	٪۱۶	٪۱۳.۹	٪۱۴.۲

جدول ۵-۱۰- بررسی اثر تغییر پارامتر دما بر معیار زمان در روش پیشنهادی و درصد بهبود نسبت به یادگیری بدون همکار- تعداد تلاش یکسان- محیط صید و صیاد

پارامتر دما	۰.۱	۰.۳	۰.۴	۰.۵	۰.۶	۰.۷	۰.۹
یادگیری Q	۶۶۰۹.۳	۸۵۰۶.۵	۸۷۵۵.۳	۹۱۶۱.۷	۹۲۱۶.۶	۹۴۰۵	۹۵۴۵.۲
MCE	۶۷۹۷.۳	۷۷۷۶.۴	۷۸۷۴.۲	۸۱۶۶.۲	۸۵۰۹.۴	۸۵۰۴.۴	۸۶۴۸.۸
بهبود	٪-۲.۸	٪۸.۵	٪۱۰	٪۱۰.۸	٪۷.۶	٪۹.۵	٪۹.۳

جدول ۵-۱۱- بررسی اثر تغییر پارامتر دما بر معیار کیفیت در روش پیشنهادی و درصد بهبود نسبت به یادگیری بدون همکار- تعداد تلاش متفاوت- محیط صید و صیاد

پارامتر دما	۰.۱	۰.۳	۰.۴	۰.۵	۰.۶	۰.۷	۰.۹
یادگیری Q	۱۵۴.۱۶	۲۰۱.۸۴	۲۱۶.۸۱	۲۳۲.۲۴	۲۴۸.۷۲	۲۵۴.۵۷	۲۷۰.۵۲
MCE	۱۶۶.۸۷	۱۴۵.۰۸	۱۴۷.۷۸	۱۵۰.۵۴	۱۵۵.۸۹	۱۶۲.۵۴	۱۶۷.۲۱
درصد بهبود	٪-۸.۲	٪۲۸.۱۲	٪۳۱.۸۳	٪۳۵.۱۷	٪۳۷.۳۲	٪۳۶.۱۵	٪۳۸.۱۸

جدول ۵-۱۲- بررسی اثر تغییر پارامتر دما بر معیار زمان در روش پیشنهادی و درصد بهبود نسبت به یادگیری بدون همکار- تعداد تلاش متفاوت- محیط صید و صیاد

پارامتر دما	۰.۱	۰.۳	۰.۴	۰.۵	۰.۶	۰.۷	۰.۹
یادگیری Q	۶۲۳۰.۲	۷۹۹۹.۹	۸۲۸۹.۹	۸۶۴۴.۵	۸۹۵۷.۷	۹۰۰۸.۵	۹۴۷۰.۹
MCE	۵۷۲۸.۲	۵۸۳۵.۹	۶۲۲۸.۱	۶۳۲۵.۲	۶۵۰۳.۵	۶۸۸۶.۸	۷۰۵۵.۶
درصد بهبود	٪۸	٪۲۷	٪۲۴.۸	٪۲۶.۸	٪۲۷.۳	٪۲۳.۵	٪۲۵.۵

نتایج ارائه شده در جداول ۵-۵ تا ۵-۱۲ نشان دهنده این واقعیت هستند که با افـزایش دمـا تـاثیر همکـاری بـر بهبود یادگیری افزایش می‌یابد. تاثیر مثبت افزایش دما بر همکاری با توجه به نقش پارامتر دما در تعیین میزان تصـادفی بودن انتخاب اعمال، قابل تفسیر است. هر چقدر مقدار پارامتر دما بیشتر باشد، تصادفی بودن در انتخاب عمـل افـزایش می‌یابد و در نتیجه تجربیات عامل‌ها متنوع‌تر خواهد شد. همان‌طور کـه در فصـل قبـل نیـز گفتـه شـد، مفهـوم خبرگـی چندمعیاره بر مبنای همه جانبه بودن اطلاعات استوار است. در واقع افزایش دما منجر به افـزایش تنـوع اعمـال انتخـاب شده و در نتیجه افزایش کیفیت محتوای اطلاعات مبادله شده بین عامل‌ها می‌شود.

۵-۵-٤ آزمایش سوم - بررسی اثر طول بازه مشارکت بر کیفیت یادگیری

همان‌گونه که گفته شد، در اکثر موارد قواعد حاکم بر دنیای مشارکت انسانی به دنیـای عامـل‌هـای مصـنوعی قابل تعمیم هستند. در این آزمایش انتظار می‌رود که در وظایف ساده‌ای مانند پلکـان مـارپیچ، طـول بـازه کوچـک‌تـر عملکرد بهتری را به دنبال داشته باشد. با انتخاب طول بازه کوچک در یادگیری مشارکتی، به عامل‌ها این فرصـت داده می‌شود که جزییات موجود در یادگیری خود را به سایرین انتقال دهند. اگر بازه مشارکت بیش از اندازه بزرگ باشـد، جزییاتی که در حین یادگیری بدست آمده‌اند، کمرنگ خواهند شد و عامل‌ها صرفا کلیات موضوعی که یاد گرفته‌اند را به یکدیگر انتقال می‌دهند. در برخی محیط‌های پیچیده اگر بازه مشارکت بسیار کوچـک باشـد، عامـل‌هـا فرصـت مناسبی برای یادگیری مستقل نخواهند داشت و در نتیجه با وجود یک بازه مشارکت کوچک هنگـامی کـه بـه چرخـه همکاری وارد می‌شوند، اطلاعات مفیدی برای انتقال به یکدیگر ندارند. در محیط‌های پیچیـده‌تـر نظیـر صـید و صـیاد انتخاب مناسب اندازه بازه مشارکت نقش مهمی در بهبود عملکـرد روش پیشـنهادی دارد. جـدول‌هـای ۵-۱۳ و ۵-۱۴ عملکرد روش پیشنهادی را دو محیط آزمایشی به ازای مقادیر مختلف طول بازه مشارکت نشان می‌دهند. لازم به ذکـر است که طول بازه مشارکت بر حسب تعداد تلاش یادگیری انجام شده در طول بازه تعیین می‌شود.

جدول ۵-۱۳- بررسی اثر طول بازه مشارکت بر معیار کیفیت در روش پیشنهادی در محیط پلکان مارپیچ و درصد بهبود نسبت به یادگیری بدون همکار

طول بازه مشارکت	۱۵	۳۰	۶۰	۷۵	۳۰۰	۶۰۰	۱۵۰۰
بهبود در تلاش یکسان	٪۲۶.۶۶	٪۲۶.۳۴	٪۲۴.۶۰	٪۲۴.۱۱	٪۱۶.۷۹	٪۱۱.۵۱	٪۳.۶
طول بازه مشارکت	۷	۱۴	۲۸	۵۶	۱۴۰	۲۸۰	۵۶۰
بهبود در تلاش متفاوت	٪۲۵.۹۱	٪۲۵.۶۴	٪۲۴.۸۲	٪۲۳.۶۱	٪۲۰.۰۴	٪۱۵.۴۷	٪۱۰.۷۱

جدول ۵-۱۴- بررسی اثر طول بازه مشارکت بر معیار کیفیت در روش‌های پیشنهادی در محیط صیدو صیاد و درصد بهبود نسبت به یادگیری بدون همکار

طول بازه مشارکت	۹	۱۸	۴۵	۹۰	۴۵۰
بهبود در تلاش یکسان	٪۱۷.۸	٪۱۵.۴۱	٪۱۴.۰۸	٪۱۰.۱۹	٪-۱۹
طول بازه مشارکت	۱۰	۲۰	۵۰	۱۰۰	۵۰۰
بهبود در تلاش متفاوت	٪۳۰.۶۹	٪۳۱.۶۴	٪۳۵.۲۳	٪۲۷.۶۵	٪۱۸.۱۱

نتایج ارائه شده در جدول‌های ۵-۱۳ و ۵-۱۴ انتظارات مطرح شده در بخش معرفی آزمایش را برآورده می‌کنند. در محیط پلکان مارپیچ به دلیل ساده‌تر بودن وظیفه، هر اندازه که طول بازه مشارکت کوچکتر باشد، درصد بهبود روش پیشنهادی نسبت به یادگیری بدون همکار بیشتر خواهد بود. در محیط صید و صیاد در حالت تعداد تلاش یکسان هم با طول بازه مشارکت کوچک‌تر جواب‌های بهتری بدست آمده است. در حالت تعداد تلاش متفاوت که نسبت به حالت تعداد تلاش یکسان سخت‌تر نیز به شمار می‌رود، در صورتی که طول بازه مشارکت نه خیلی بزرگ و نه خیلی کوچک تعریف شود، درصد بهبود روش بیشینه خواهد شد.

می‌توان گفت که در وظایف ساده در بازه‌های مشارکتی کوتاه‌تر، عامل‌ها اطلاعات جزئی‌تر و دقیق‌تری از محیط دارند. به عبارت دیگر عامل‌ها به زمان کمی برای درک محیط و سپس ورود به چرخه همکاری نیازمندند. در مقابل در وظایف پیچیده‌تر، عامل‌ها به زمان بیشتری برای درک محیط و پس از آن انجام همکاری نیاز دارند. طول بازه مشارکت مناسب باید به اندازه‌ای باشد که در آن خبرگی عامل در حدی است که هنوز شکست‌هایش را به خاطر دارد و قدر پیروزی‌هایش را نیز می‌داند. به عبارت دیگر طول بازه مشارکت باید به اندازه‌ای باشد که وزن تجربیات مختلف عامل تقریبا با هم برابر باشد. یعنی عامل فرصت کافی برای بدست آوردن تجربیات مختلف داشته باشد. اگر فرصت داده شده به عامل برای کسب تجربه بیش از حد مناسب باشد، آن‌گاه در نتیجه گذشت زمان وزن تجربیات خوب نسبت به تجربیات بد بیشتر خواهد شد و تنوع تجربیات عامل کاهش خواهد یافت. در واقع آن چه که مفهوم خبرگی چندمعیاره را قدرتمند می‌سازد، گوناگونی تجربیات است و از این رو تعیین بازه صحیح مشارکت نقش مهمی در بیشینه نمودن درصد بهبود روش داراست.

۵-۵-۵ آزمایش چهارم – بررسی اثر تعداد معیارهای خبرگی مورد استفاده

همان‌طور که در فصل قبل بیان شد، یکی از نوآوری‌های روش پیشنهادی، نگاه همه جانبه به تجربیات عامل‌ها است. روش پیشنهادی جنبه‌های مختلف رفتاری عامل را می‌سنجد و سپس بر اساس آن به تصمیم‌گیری درباره آینده‌اش می‌پردازد. بدیهی است هر چه تعداد جنبه‌های مختلفی که عامل رفتار خود را بر اساس آن‌ها می‌سنجد بیشتر

باشد، کیفیت تصمیم‌گیری عامل بهبود خواهد یافت. در این آزمایش برای بررسی صحت این ادعا، تعداد مختلفی معیار خبرگی در ساختن جدول مشارکتی استفاده شده‌اند. نتایج ارائه شده در جدول‌های ۵-۱۵ و ۵-۱۶ به خوبی ادعای مطرح شده در مورد غنای جدول مشارکتی حاصل از تاثیر جنبه‌های رفتاری مختلف عامل را تایید می‌کنند. با افزایش تعداد معیارهای خبرگی مورد استفاده درصد بهبود در معیارهای کیفیت و زمان به طور چشمگیری افزایش می‌یابد. لازم به ذکر است که هدف از آزمایش چهارم بررسی اثر تعداد معیارهای خبرگی مورد استفاده است و بنابراین ترتیب و چگونگی انتخاب معیارهای مورد استفاده در این آزمایش مورد سوال نبوده است.

جدول ۵-۱۵- بررسی اثر تعداد معیارهای خبرگی مورد استفاده در روش‌های پیشنهادی در محیط پلکان مارپیچ و درصد بهبود نسبت به یادگیری بدون همکار

			تعداد معیارهای خبرگی مورد استفاده	۱	۲	۳	۴	۵	۶
تلاشی یکسان	MCE	کیفیت		٪۰.۹۱	٪۱۴.۴	٪۲۰.۴۷	٪۲۳.۹۵	٪۲۵.۵۳	٪۲۷.۴۱
		زمان		٪۰.۷۱	٪۱۵.۲۱	٪۲۲.۷	٪۲۸.۳۶	٪۳۱.۰۱	٪۳۳.۵۲
تلاشی متفاوت	MCE	کیفیت		٪۴.۷	٪۱۶.۳۶	٪۲۰.۷۶	٪۲۳.۳۸	٪۱۷.۳۶	٪۱۸.۷۴
		زمان		٪۳.۵	٪۱۸.۲۴	٪۲۳.۱۷	٪۲۷.۹۷	٪۲۴.۹۱	٪۲۷.۳۹

جدول ۵-۱۶- بررسی اثر تعداد معیارهای خبرگی مورد استفاده در روش‌های پیشنهادی در محیط صید و صیاد و درصد بهبود نسبت به یادگیری بدون همکار

			تعداد معیارهای خبرگی مورد استفاده	۱	۲	۳	۴	۵	۶
تلاشی یکسان	MCE	کیفیت		٪-۱۶	٪۰.۶	٪۱.۵	٪۱۲.۱۹	٪۱۴.۹	٪۱۶.۷
		زمان		٪-۵.۲	٪۲.۴	٪۳.۱	٪۸.۹	٪۱۲.۳	٪۱۵.۴
تلاشی متفاوت	MCE	کیفیت		٪۶.۹	٪۱۵.۴	٪۲۳.۵	٪۲۹.۲۴	٪۳۱.۰۷	٪۳۰.۵۸
		زمان		٪-۴.۸	٪۹.۵	٪۱۶.۳	٪۲۲.۱۱	٪۲۴.۱۱	٪۲۳.۹۵

۵-۵-۶ آزمایش پنجم - بررسی پایایی روش نسبت به حضور اغتشاش

قابلیت تحمل‌پذیری خطا همواره به عنوان یکی از ویژگی‌های مورد توجه در مورد سیستم‌های چندعامله و سیستم‌هایی که در آن‌ها یادگیری ماشین مورد استفاده قرار گرفته، مطرح بوده است. روش پیشنهادی این پایان‌نامه در حوزه روش‌های یادگیری مشارکتی مبتنی بر انتقال اطلاعات دسته‌بندی می‌شود. یکی از مواردی که همواره در مورد روش‌های جدید پیشنهادی در این حوزه مطرح بوده، قابلیت به کارگیری آن‌ها در محیط‌های واقعی است. در محیط-های واقعی سیستم‌های ارتباطی مورد استفاده برای انتقال اطلاعات همواره مقداری اغتشاش نیز به محتوای اطلاعاتی

مورد مبادله می‌افزایند. از این رو پایایی روش طراحی شده در برابر اغتشاش یکی از ویژگی‌های مهم و مطلوب جهت استفاده از آن در کاربردهای واقعی به شمار می‌رود.

همان‌طور که در فصل قبل مطرح شد، جدول مشارکتی مبتنی بر معیار خبرگی از دید عاملی که مقدار خبرگی‌اش از سایرین کمتر است، ساخته می‌شود. از این رو در آزمایش پنجم برای شبیه‌سازی حضور اغتشاش در ارتباطات فرض شده است که به خانه‌های جدول Q عامل‌هایی که مقدار خبرگی‌شان بیشتر است، مقدار اغتشاش تصادفی M که از توزیع نرمال (0,N) پیروی می‌کند، اضافه شده است. در یک سیستم واقعی برای ساخت جدول مشارکتی از دیدگاه عاملی که خبرگی‌اش کمتر است، جدول دو عامل دیگر می‌بایست از طریق سیستم ارتباطی به عامل سوم منتقل شود و از این رو احتمال مغشوش شدن دو جدول در حین انتقال وجود دارد. هدف از طراحی آزمایش پنجم بررسی عملکرد روش پیشنهادی در حضور اغتشاش است. نتایج حاصل از انجام این آزمایش – جدول‌های ۵-۱۷ و ۵-۱۸ - بیان‌کننده توانایی روش در تحمل اغتشاش و امکان پیاده‌سازی آن در محیط‌های واقعی است.

جدول ۵-۱۷- بررسی پایایی روش پیشنهادی نسبت به اغتشاش در محیط صید و صیاد و درصد بهبود نسبت به یادگیری بدون همکار در حالت تعدادتلاش یکسان

			مقدار اغتشاش	۰	۰.۲	۰.۵	۰.۷	۲	۳
بلکمان ماریچ	MCE	کیفیت		٪۲۶.۸	٪۲۶.۲۵	٪۲۴.۵۸	٪۱۷.۶	٪-۱۵۴	٪-۱۸۳
		زمان		٪۳۴.۴۶	٪۳۰.۹۵	٪۲۸.۶۲	٪۲۰.۶۳	٪-۸۰	٪-۱۰۷
صید و صیاد	MCE	کیفیت		٪۹.۰۵	٪۲.۵	٪-۱۳.۶	٪-۲۳.۳	٪-۷۵.۱	٪-۹۳
		زمان		٪۱۱.۶۹	٪۵.۴	٪-۱۲.۳	٪-۱۹.۵	٪-۵۹	٪-۷۲

جدول ۵-۱۸- بررسی پایایی روش پیشنهادی نسبت به اغتشاش در محیط صید و صیاد و درصد بهبود نسبت به یادگیری بدون همکار در حالت تعدادتلاش متفاوت

			مقدار اغتشاش	۰	۰.۲	۰.۵	۰.۷	۲	۳
بلکمان ماریچ	MCE	کیفیت		٪۲۵.۴۶	٪۲۵.۳۷	٪۱۶.۳۱	٪۱۲.۸۵	٪-۱۳۹	٪-۱۶۴
		زمان		٪۳۲.۳۶	٪۳۱.۱۸	٪۲۲.۹۳	٪۱۸.۸۱	٪-۷۷	٪-۹۹.۶
صید و صیاد	MCE	کیفیت		٪۲۹.۹۰	٪۲۴.۶۶	٪۹.۵	٪-۳.۷	٪-۶۰	٪-۸۸.۵
		زمان		٪۲۳.۵۴	٪۱۸.۲۹	٪۴.۶	٪-۳.۵	٪-۴۸	٪-۷۰

۵-۶ نتیجه گیری

در این فصل، دو محیط آموزشی متفاوت همراه با جزییات پیاده‌سازی آن‌ها معرفی شده‌اند. برای ارزیابی روش یادگیری مشارکتی پیشنهادی هفت آزمایش متفاوت طراحی شده است که هر کدام یکی از مزیت‌های روش پیشنهادی را نشان می‌دهند. روش پیشنهادی نسبت به سایر روش‌های یادگیری مشارکتی مبتنی بر خبرگی نتایج بهتری دارد. یکی از مزیت‌های روش پیشنهادی در مقایسه با روش‌های موجود مبتنی بر انتقال اطلاعات، نقش بارز همه جانبه بودن اطلاعات در آن است. در صورتی که جدول مشارکتی بر اساس اطلاعات همه جانبه ساخته شود، مشارکت بر اساس روش پیشنهادی تاثیر بسیار بهتری خواهد داشت. نتایج آزمایش‌ها نشان می‌دهند که همه جانبه بودن اطلاعات را می‌توان با ترکیب جنبه‌های اطلاعاتی بیشتر و یا با افزایش دما بدست آورد. یکی دیگر از مزیت-های روش پیشنهادی پایا بودن آن نسبت به حضور اغتشاش در ارتباطات بین عامل‌ها است.

فصل ششم
نتیجه گیری

۶-۱ مقدمه

در این پایان‌نامه روشی موثر برای مشارکت در یادگیری بین عامل‌های حاضر در یک سیستم چندعامله به منظور بهبود یادگیری عامل‌ها ارائه شده است. دو دیدگاه کلی موجود در حوزه یادگیری مشارکتی در سیستم‌های چندعامله به صورت زیر است:

- مشارکت به منظور یادگیری هماهنگی برای رسیدن به یک هدف مشترک، در این حالت همه عامل‌ها در یک محیط قرار دارند و مشارکت در یادگیری به منظور آموختن شیوه‌های هماهنگی عامل‌ها برای رسیدن به یک هدف مشترک تعریف شده است.
- مشارکت به منظور بهبود یادگیری هر یک از عامل‌ها در آموختن یک کار یکسان، در این حالت هر یک از عامل‌ها در محیطی جداگانه و به طور مستقل فرآیند یادگیری یک کار یکسان را انجام می‌دهند و با استفاده از انتقال اطلاعات بین یکدیگر یادگیری اعضای گروه بهبود داده می‌شود.

در دسته دوم عامل‌ها بوسیله ارتباطات مستقیم و یا غیر مستقیم، اطلاعات حاصل از فرآیند یادگیری خود را با دیگران به اشتراک می‌گذارند. در صورتی که انتقال اطلاعات و به کارگیری آن‌ها در طول یادگیری به خوبی تعریف شده باشد، مشارکت قادر است که یادگیری اعضای گروه را بهبود بخشد. در این دسته از روش‌ها عملکرد روش‌های یادگیری مشارکتی به اطلاعاتی که بین عامل‌ها مبادله می‌شود و روشی که بر مبنای آن اطلاعات مبادله

شده در طول یادگیری مورد استفاده قرار می‌گیرند، بستگی دارد. تا کنون روش‌های متفاوتی برای انتقال اطلاعات بین عامل‌های یادگیرنده پیشنهاد داده شده است که هر کدام برگرفته از ایده‌هایی ساده از زندگی دسته‌جمعی انسان‌ها و حیوانات هستند. در روش‌های مختلف اطلاعات متفاوتی بین عامل‌ها به اشتراک گذاشته می‌شود. اطلاعاتی نظیر پارامترهای یادگیری، جدول‌های Q عامل‌ها، سیگنال تقویتی دریافتی از محیط، عمل انجام شده تا کنون مورد استفاده قرار گرفته‌اند.

۶-۲ نوآوری‌ها و نتایج کلی پایان‌نامه

در این پایان‌نامه مفهوم جدیدی به نام خبرگی چندمعیاره معرفی شده است که قادر است به خوبی اطلاعات همه‌جانبه‌ای را در مورد عامل‌ها و آن چه که تا کنون یاد گرفته‌اند، ارائه دهد. در روش پیشنهادی جدول مشارکتی که بر اساس خبرگی چندمعیاره ساخته شده است، به عنوان اطلاعات بین عامل‌ها مبادله می‌شود. جدول مشارکتی حاصل روش نوینی برای ارائه اطلاعات جمعی عامل‌های حاضر در سیستم است. هم‌چنین در این پایان‌نامه اطلاعات مبادله شده به صورت راهنمایی عامل‌های حاضر در سیستم در حین انتخاب عمل مورد استفاده قرار گرفته‌اند. این نوع استفاده از دانش مبادله‌شده به تسریع روند همگرایی کمک شایانی می‌کند و از واگرایی یادگیری به دلیل رشد مقادیر جدول مشارکتی جلوگیری خواهد کرد.

با استناد به نتایج آزمایش‌های انجام گرفته در فصل پنجم، روش پیشنهادی ارائه شده در این پایان‌نامه نسبت به سایر روش‌های یادگیری مشارکتی مبتنی بر خبرگی عملکرد بسیار بهتری را در هر دو حالت تعداد تلاش یکسان و تعداد تلاش متفاوت دارد. روش پیشنهادی علاوه بر تسریع روند همگرایی در یادگیری، کیفیت جواب نهایی را نیز بهبود بخشیده است. همانند اکثر روش‌های مبتنی بر انتقال اطلاعات و استفاده از تجربیات، روش پیشنهادی نیز در نرخ‌های یادگیری پایین عملکرد قابل توجهی دارد و با افزایش نرخ یادگیری مزیت‌های خود را از دست می‌دهد.

با افزایش دما در محیط، عملکرد روش پیشنهادی به طور محسوسی افزایش می‌یابد و این امر با توجه به استفاده از جدول مشارکتی تولید شده که بر اساس تجربیات مختلف عامل‌ها شکل گرفته، دور از انتظار نیست چرا که با افزایش دما در سیستم عامل‌ها توانایی کسب تجربیات غنی‌تر و متفاوت‌تری را خواهند داشت و از این رو جدول مشارکتی حاوی اطلاعات غنی‌تری خواهد بود.

یکی دیگر از مزایای روش پیشنهادی پایا بودن آن نسبت به اغتشاش موجود در ارتباطات بین عامل‌ها است. از این رو روش پیشنهادی قابل استفاده در سیستم‌های چند رباته (چندعامله) خواهد بود. در روش پیشنهادی تعریف درست بازه مشارکت نقش مهمی در عملکرد بهتر روش دارد.

۶-۳ راهکارهای آینده و پیشنهادها

در این پایان‌نامه با بهبود کیفیت اطلاعات مبادله شده بین عامل‌ها و همین‌طور تغییر موثر نحوه استفاده از اطلاعات، روش یادگیری مشارکتی جدیدی ارائه شده است. روش پیشنهادی از نظر تئوری و تجربی هنوز در مراحل اولیه رشد خود به سر می‌برد. به نظر می‌رسد تحقیقات آینده در این زمینه می‌تواند مطابق پیشنهادهایی که در ادامه مطرح می‌شود، انجام پذیرد.

- به نظر می‌رسد یکی از راهکارهای آینده برای بهبود روش پیشنهادی یافتن مکانیزمی خودکار برای کنترل بهینه دما به منظور غنی‌تر شدن تجربیات عامل‌ها است.
- به نظر می‌رسد بررسی کارآیی روش در محیط‌هایی که ماهیت‌های متفاوتی دارند و هم‌چنین محیط‌های با ابعاد بزرگتر و ساختار پیچیده‌تر بتواند راهکارهای جدیدی را برای ادامه تحقیق در اختیار بگذارد.
- آن چه در این پایان‌نامه بررسی شد، در حوزه مشارکت به منظور بهبود یادگیری هر یک از عامل‌ها در آموختن یک کار یکسان دسته‌بندی می‌شود. پیشنهاد می‌شود که روش پیشنهادی بر روی حوزه تحقیقاتی مشارکت به منظور یادگیری هماهنگی برای رسیدن به یک هدف مشترک نیز بررسی شود.

مراجع

[1] Russell, S. and Norving, P., *Artificial Intelligence: A Modern Approach, Second Edition*, Prentice Hall, 2006.

[2] Wooldrige, M. *An Introduction to Multi-Agent Systems,* John Wiley & Sons, Second Edition, 2009.

[3] Vidal, J., *Fundamentals of Multi-Agent Systems,* Copy Right By J. Vidal, 2007.

[4] Panait, L., Luke, S., "Cooperative Multi-Agent learning: The State of the Art", *Journal of Autonomous Agents and Multi-Agent Systems*, vol. 11, Issue. 3, November 2005, pp.387-434.

[5] Smith, E., "Human Cooperation: Perspectives from Behavioral Ecology", *Proceedings of Dahlem Conference on Genetic and Cultural Evolution of Cooperation*, Berlin, June 2002, pp.401-427.

[6] Mitchell, T., *an Introduction to Machine Learning* , Mc-GrowHill, 1997.

[7] Nunes, L., Oliveira, E., "Advice-Exchange Amongst Hetergeneous Learning Agents: Expriments in the Pursuit Domain", *Proceedings of the Second International Joint Conference on Autonomous Agents & Multi-agent Systems, AAMAS 2003*, Melbourne, Victoria, Australia, 2003, pp. 1084-1085.

[8] Whitehead, S. D., "A Complexity Analysis of Cooperative Mechanisms in Reinforcement Learning", *Proceedings of the Ninth National Conference on Artificial Intelligence, 1991*, pp. 607-613.

[9] Whitehead, S., Ballard, D., "A study of Cooperative Mechanisms for faster Reinforcement learning", *Technical Report 365, Computer Science Dept., University of Rochester*, February 1991.

[10] Tan, M ., "Multi-Agent Reinforcement Learning: Independent vs. Cooperative Agents" , *Proceedings of Tenth. Int. Conf. Machine Learning, Amherst, MA, 1993,* pp. 487-494.

[11] Berenji. H. R., Vengerov. D., "Cooperation and coordination between fuzzy reinforcement learning agents in continuous state partially observable Markov decision processes" *Proceedings of the 8th IEEE International Conference on Fuzzy Systems*, pp. 621-627.

[12] Berenji, H.R., Vengero, D., "Advantages of Cooperation between Reinforcement Learning Agents in Difficult Stochastic Problems*", In Proceedings of the 9th IEEE International Conference on Fuzzy Systems*, 2000, pp. 871-876.

[13] Trevarthen, C., "Learning About Ourselves, From Children: Why A Growing Human brain needs interesting companion companions", *Journal of Research and Clinical Center for Child Development*, vol. 26, Feb. 2004, pp. 9-44.

[14] kuniyoshi, y., "Learning by Watching: Extracting Reuseable Task Knowledge from Visual Observation of Human Performance", *IEEE Transaction Robot. Automat*,vol. 10, issue. 6, 1994, pp 799-822.

[15] Yamaguchi, T., Tanaka, Y., Yachida, M., "Speed up Reinforcement Leraning between Two Agents with Adaptive Mimetism", *Proceedings of IEEE Conf. Intell. Robot. Syst. (IROS), 1997,* pp. 594-600.

[16] Yamaguchi, T., Miura, M., Yachida, M., " Multi-Agent Reinforcement Learning Adaptive Mimetism", *Proc. Fifth IEEE Int. Conf. Emerging Technol. Factory Automat. (ETFA), pp.288-294*.

[17] Garland, A., Alterman, R., "Multi-Agent Learning through Collective Memory", In *Adaptation, Co-evolution and Learning in Multi-agent Systems: Papers from the 1996 AAAI Spring Symposium*, Menlo Park, CA, March 1996, pp. 33-38.

[18] Salomon, G., *Distributed Cognition: Psychological and Educational Consideration,* New York: Cambridge University Press, 1993.

[19] Garland, A., Alterman, R., "Preparation of Multi-Agent Knowledge for Reuse*", Technical Report, Waltham: AAAI Fall Sumposium on Adaptation of Knowledge for Reuse* , 1995.

[20] Nunes, L., Oliveira, E., "On Learning by Exchanging Advic", *in Proceedings of the Artificial Intelligence and the Simulation of Behavior Convention, Second Symposium on Adaptive Agents and Multi-Agent Systems (AISB/AAMAS-II),* Imperial College, London, April 2002.

[21] Nunes, L., Oliveira, E., "Cooperative Learning using Advice-Exchange", *in Adaptive Agents and Multi-Agent Systems, LNCS*, vol. 2636, Jan 2003, pp. 33-48.

[22] Nunes, L., Oliveira, E., "Advice-Exchange between Evolutionary Algorithms and Reinforcement Learning Agents: Expriments in the Pursuit Domain", *Proceedings of the Artificial Intelligence and the Simulation of Behavior Convention, Third Symposium on Adaptive Agents and Multi-Agent Systems (AISB03/AAMAS02),* Aberystwyth, Wales, 2003.

[23] Nunes, L., Oliveira, E., "Exchanging Advice and Learning to Trust", *Seventh Int. Conf. on Cooperative Information Agents (CIA-03),* Helsinki, Finland, 2003.

[24] Nili Ahmadabadi, M., Asadpour, M., Khodaabakhsh, Seyyed H., Nakano, E., "Expertness Measuring in Cooperative Learning", *Proceedings of the 2000 IEEE/RSJ Inter. Conf. on Intelligent Robots and Systems, 2000, pp. 2261-2267*.

[25] Akbarzadeh, M. R., Rezaei, S., Naghibi, M. B., "A Fuzzy Adaptive Algorithm for Expertness Based Cooperative Learning Application to Herding Problem", *Proceedings of 22th International Conference of the North American Fuzzy Information Processing Society*, 2003, pp. 317-322.

[26] Mastour Eshgh, S., Nili Ahmadabadi, M., " Extension of Weighted Strategy Sharing in Cooperative Q-Learning for Specialized Agents", *Proceedings of the 9th Inter. Conf. on Neural Information Processing, 2002, pp. 106-110*.

[27] Nili Ahmadabadi, M., Imanipour, A., Araabi, B., Asadpour, M., Siegwart, R., "Knowledge-based Extraction of Area of Expertise for Cooperation in Learning", *Proceedings of the 2006 IEEE/RSJ Inter. Conf. on Intelligent Robots and Systems,* Beijing, China, 2006, pp. 3700-3705.

[28] Nadjar Araabi, B., Mastoureshgh, S., Nili Ahmadabadi, M., "A Study on Expertise of Agents and Its Effects on Cooperative Q-Learning", *IEEE Transactions on systems, man, and cybernetics* , vol. 37, no.2, April 2007, pp. 398-409.

[29] Ritthipravat, P., Maneewarn, T., Wyatt, J., Laowattana, D., "Comparison and Analysis of Expertness Measure in Knowledge Sharing Among Robots", *Springer-Verlag Berlin Heidelberg, LNAI* 4031, 2006, pp. 60-69.

[30] Carver, N., Lesser, V. “The Evolution of Blackboard Control Architectures”. *Expert Systems with Applications Special Issue on the Blackboard Paradigm and Its Applications*, 1992, 7(1): pp.1-30.

[31] McManus, J.W., Bynum, W. L. “Design and Analysis Techniques for Concurrent Blackboard Systems”. *IEEE Transactions on Systems, Man and Cybernetics*, 1996, 26(6): pp. 669-680.

[32] Yang, Y., Tian, Y., Mei, H., “Cooperative Q Learning Based on Blackboard" Architecture”, *Proceedings of 2007 International Conference on Computational Intelligence and Security Workshops*, pp.224-227.

[33] Yang, M., Tian, Y., Liu, Xiaomei, "Cooperative Q-Learning Based on Maturity of the Policy", *Proceedings of the 2009 IEEE International Conference on Mechatronics and Automation*, August 9-12, Changchun, China.

[34] Sutton, R. S., Barto, A. G., *Reinforcement Learning: An Introduction Adaptive Computation and Machine Learning*, MIT Press, 1998.

[35] Watkins, C. J. C. H., *Learning from Delayed Rewards*, PhD dissertation, king's College, Cambridge, U.K., May 1989.

[36] Watkins, C. J. C. H., Dayan, P., "Q-Learning (technical note)", *in Machine Learning, Special Issue on Reinforcement learning*, Cambridge, MA: MIT Press, 1998, pp. 55-68.

[37] Dorigo, M., Gambardella, M., "Ant Colony System: a Cooperative Learning Approach to the Traveling Salesman Problem*", IEEE Transactions Evolutionary Computation*, Vol. 1, 1997, pp. 53-66.

[38] McDonald, D. W., Ackerman, M. S., “Just Talk to Me: A Field of Expertise Location", *Proceedings of ACM Conf. Comput.-Supported Cooperative Work,* Seattle, WA, 1998, pp. 315-324.

[39] Nili Ahmadabadi, M., Asadpour, M., Nakano, E., "Cooperative Q'learning : The Knowledge Sharing Issue", *Journal of Advanced Robotics*, vol. 15, no. 8, 2002, pp. 815-832.

[۴۰] م. اسدپور، «بررسی همکاری در یادگیری در ربات‌های جابجا کننده اجسام»، پایان‌نامه کارشناسی ارشد، دانشکده فنی، دانشگاه تهران، ۱۳۷۸.

[41] BIANCHI, R. A. C.; COSTA, A. H. R., "The use of heuristics to speedup Reinforcement Learning", *Boletim Interno, No. BT/PCS/0409. Escola Politécnica da USP*, São Paulo, 2004.

[42] Bianchi, R. A. C., Ribeiro, C. H. C., Costa, A. H. R., "Heuristically Accelerated Q-Learning: a New Approach to Speed Up Reinforcement Learning", *Lecture notes in Artificial Intelligence*, 3171, 2004, pp. 245-254.

[43] Shavelson, R. J., Hubner, J. J., Stanton, G. C., "Self-concept: Validation of Construct Interpretations", *Review of Educational Research*, Vol. 46, 1976, pp.407-441.

[44] Shavelson, R. J., Bolus, R., "Self-concept: The Interplay of Theory and Methods ", *Journal of Educational Psychology*, Vol. 74, 1982, pp. 3-17.

[45] Weiss, G., *Multi agent Systems: A Modern Approach to Distributed Artificial Intelligence*, MIT Press, 2000.

[46] Lesser, V. R., “Multi-Agent Systems: An Emerging Sub discipline Of AI” *ACM Computing Surveys*, Vol. 27, No 3, September 1995.

[47] Doran, J., Palmer, M., *The EOS Project: integrating two models of Paleolithic social change*, UCL Press, London, 1995.

[48] Benda, M., Jagannathan, V., Dodhiawala, R., "on Optimal Cooperation of Knowledge Sources – an Empirical Investigation", *Technical Report, BCS-G2010-28, Boeing Advanced Technology Center, Boeing Computing Services*, Seattle, Washington, 1986.

Multi-Criteria Expertness based Cooperative Learning in Multi-Agent Systems

Esmat Pakizeh Hajyyar
e.pakizehhajyyar@ec.iut.ac.ir

Date of Submission: 2010/10/16

Department of Electrical and Computer Engineering

Isfahan University of Technology, Isfahan 84156-83111, Iran

Degree: M.Sc. Language: Farsi

Supervisor: Maziyar Palhang, palhang@cc.iut.ac.ir

Abstract

The aim of this thesis is to introduce a new algorithm for cooperative learning in multi-agent systems. Since cooperation is the key to success in most biological and artificial communities, the capability of cooperation in multi-agent systems is critical in achieving better solutions. Due to having more knowledge and information resources, multi-agent cooperative learning is expected to result in higher efficiency and faster learning compared to individual learning.

Better cooperative strategies may speed up and improve learning. Nowadays, most of researches in multi-agent cooperative learning field focus on Reinforcement Learning (RL) as their basic learning method. RL is one of the more prominent machine learning technologies because of its unsupervised learning structure and continuous learning ability, even in a dynamic operating environment. Applying this learning to cooperative multi-agent systems not only allows each individual agent to learn from its own experience, but also offers the opportunity for the individual agents to learn from other agents in the system so that the speed of learning can be accelerated.

During the life cycle, human learns through different experiences over different time periods of his life. Sometimes the experience is quite successful and sometimes it completely fails. Individual's character is formed based on all of the gained experiences whether they are good or bad. In human societies, people who want advice about completing a task, get help from someone who has more experience. In other words a person will be evaluated based on his total personality which is composed of his different expertness measures. This fact could be generalized into world of agents. Therefore the main contribution of thesis is relied on this fact. In this thesis a new Multi-Criteria Expertness based cooperative learning method is proposed that benefits from all of expertness measures and attempts to cooperate more efficiently. Experimental results performed on maze world and hunter-prey domain show the high potential of proposed method in producing better cooperative learning.

Keywords:

Multi-Agent Systems, Cooperative Learning, Multi-Criteria Expertness, Reinforcement Learning