
بسم الله الرحمن الرحيم

دانشگاه صنعتی اصفهان

دانشکده مهندسی برق و کامپیوتر

بهبود کیفیت و سرعت یادگیری در سیستم‌های چندعامله با استفاده از معیار جدید خبرگی و انتگرال فازی

پایان‌نامه کارشناسی ارشد مهندسی کامپیوتر – هوش‌مصنوعی و رباتیک

داریوش حسن‌پور آده

استاد راهنما

دکتر مازیار پالهنگ

پاییز ۱۳۹۵

دانشگاه صنعتی اصفهان

دانشکده مهندسی برق و کامپیوتر

پایان‌نامه کارشناسی ارشد رشته مهندسی کامپیوتر – هوش‌مصنوعی و رباتیک آقای

داریوش حسن‌پور آده

تحت عنوان

بهبود کیفیت و سرعت یادگیری در سیستم‌های چندعامله با استفاده از معیار جدید خبرگی و انتگرال فازی

در تاریخ ۱۳۹۵/۱۰/۲۰ توسط کمیته تخصصی زیر مورد بررسی و تصویب نهایی قرار گرفت:

۱ـ استاد راهنمای پایان‌نامه	دکتر مازیار پالهنگ
۳ـاستاد داور	دکتر عبدالرضا میرزایی
۴ـاستاد داور	دکتر محمد حسین منشئی
سرپرست تحصیلات تکمیلی دانشکده	دکتر محمد رضا تابان

تشکر و قدردانی

پروردگار منّان را سپاسگزارم .....

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

دلتنگی‌های آدمی را باد ترانه‌ای می‌خواند
رویاهایش را آسمان پر ستاره نادیده می‌گیرد
و هر دانه‌ی برفی به اشکی نریخته می‌ماند.
سکوت سرشار از سخنان ناگفته است؛
از حرکات ناکرده،
اعتراف به عشق‌های نهان،
و شگفتی‌های به زبان نیامده،
دراین سکوت حقیقت ما نهفته است؛
حقیقت تو و من.

برای تو و خویش
چشمانی آرزو می‌کنم،
که چراغ‌ها و نشانه‌ها را در ظلمات‌مان ببیند.
گوشی،
که صداها و شناسه‌ها را در بیهوشی‌مان بشنود.
برای تو و خویش،
روحی،
که این‌همه را در خود گیرد و بپذیرد.
و زبانی
که در صداقت خود ما را از خاموشی خویش بیرون کشد،
و بگذارد از آن‌چیزها که در بندمان کشیده است، سخن بگوییم.

پنجه درافکنده‌ایم با دست‌هایمان
به جای رها شدن
سنگین سنگین بر دوش می‌کشیم
بار دیگران را
به جای همراهی کردن‌شان!
عشق ما نیازمند رهایی است نه تصاحب
در راه خویش ایثار باید نه انجام وظیفه...

بی‌اعتمادی دری است
خودستایی، چفت و بست غرور است
و تهی دستی، دیوار است و لولاست
زندانی را که در آن محبوس رآی خویش‌ایم
دلتنگی‌مان را برای آزادی و دلخواه دیگران بودن
از رخنه‌هایش تنفس می‌کنیم...

ـ مارگوت بیگل

فهرست مطالب

فهرست تصاویر

فهرست جداول

چکیده

معمولا در دنیایی واقعی هنگامی که افراد برای انتقال دانش گرد هم می‌آیند و از تجربیات خوب و بد گذشته خود سخن می‌گویند هرکسی متناسب با جایگاهی که دارد دارای دانشی می‌باشد و در این انتقال دانش‌ها تجربیات هیچ کسی را نمی‌توان نادیده گرفت ولی گاها پیش می‌آید که تجربیات و دانش فردی دارای بار محتویاتی بیشتری نسبت به اطرافیان خود می‌باشد، مردم معمولا از دانش فرد خبره‌تر بیشتر بهره می‌برند تا افراد دیگر. دستاوردهای این پژوهش بر مبنای همین فلسفه بنا شده است که سخن و دانش هرکسی باید شنیده شود. انتگرال فازی یکی از قوی‌ترین و منعطف‌ترین ابزارهای ریاضی برای ترکیب اطلاعات می‌باشد، لذا در این پژوهش از انتگرال فازی برای شنیدن بازتاب ندای دانش هر عامل در دانش جمعی و مدل کردن اطلاعات(دانش‌های) غیرافزایشی استفاده شده است. ولی در این راه مشکلاتی نیز وجود داشت و آن این بود که چگونه منصفانه بفهمیم که کدام عامل خبره‌تر از دیگری می‌باشد؟ در گذشته روش‌های متنوعی برای تخمین این معیار ارائه شد است که از شمارش میزان پاداش‌های مثبت و منفی عامل‌ها گرفته تا محاسبات پیچیده‌ای چون معیارهای شوک و کوتاه‌ترین مسیر تجربه شده. در طی پژوهش که منجر به نگارش این پایان‌نامه گردید احساس شد که تمامی روش‌های قبلی در یک چیز مشترکند: بسیار پیچیده و غیر منعطف!

وجود این فصل مشترک ناکارا انگیزه‌ای شد که در صدد ارائه‌ای معیاری برآیم که نه تنها ساده باشد بلکه در زندگی روزمره ما انسان‌ها هم تجلی داشته باشد. در پی این هدف ما به ارائه‌ی تئوری جامعی برای خبرگی پرداختیم که می‌تواند منشع بسیاری از تعاریف خبرگی، در آینده گردد؛ نهایتا با استفاده از تئوری خبرگی معرفی شده تعریفی برای یک معیار خبرگی جدید ارائه دادیم و نشان دادیم که تئوری و تعریف خبرگی جدید نسبت به تعاریف قبلی بسیار کارآمد بوده است.

واژه‌های کلیدی: ۱ـ سیستم‌های چندعامله، ۲ـ یادگیری مشارکتی، ۳ـ یادگیری تقویتی، ۴ـ دانش غیرافزایشی، ۵ـ انتگرال فازی.

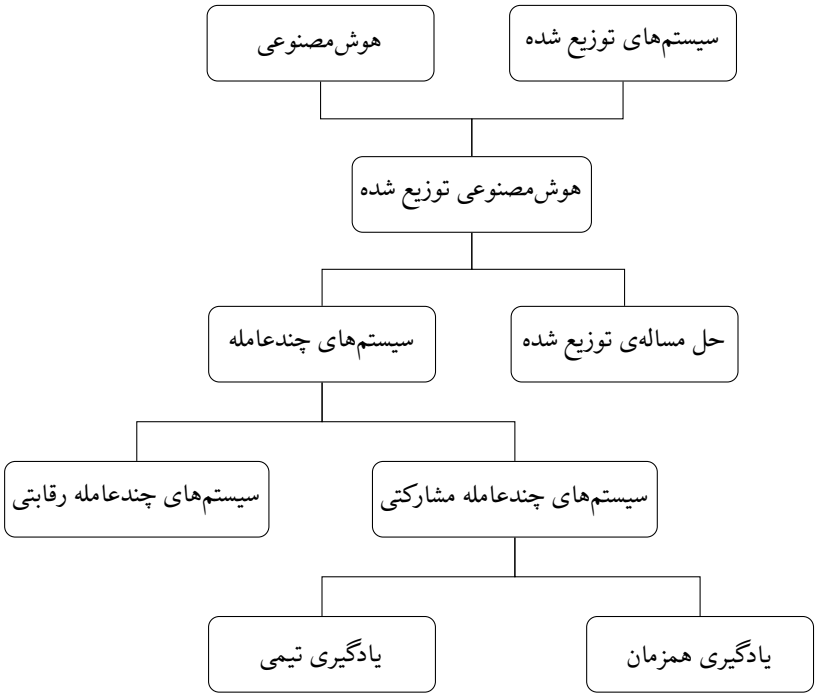
فصل اول

مقدمه

رایانه‌ها از قدرت بالایی در محاسبات برخوردار هستند تا جایی که محاسباتی که عامل انسانی ممکن است در چند سال انجام دهد را می‌توانند در کسری از ثانیه انجام دهند. در سال‌هایی که از عمر کامپیوتر گذشته پیشرفت‌های فراوانی صورت گرفته است که بر سرعت اجرای محاسبات کامپیوتر افزوده است. اما باوجود تمام این پیشرفت‌ها هنوز هم مسائل زیادی هستند که کامپیوتر نمی‌تواند در زمان قابل‌قبول آن‌ها را حل کند. یک روش که در راستای افزایش سرعت حل مسائل و با بهره‌گیری از زندگی جمعی عامل‌های انسانی پیشنهاد شد تقسیم مسئله به چندین زیر مسئله و حل هر یک توسط یک رایانه بوده است این موضوع شاخه‌ای تحت عنوان سیستم‌های توزیع‌شده را به وجود آورده است.

از طرف دیگر یکی از مسائل مهم در دنیای رایانه‌ها که ازنظر محاسباتی به زمان زیاد نیاز دارد موضوعات هوش مصنوعی هستند که در ترکیب با سیستم‌های توزیع‌شده به هوش مصنوعی توزیع‌شده بدل شده‌اند. هوش مصنوعی توزیع‌شده نیز در قالب حل مسائل هوش مصنوعی و یادگیری فعالیت می‌کند. در حل مسائل توزیع‌شده یک مسئله که قابلیت اجرای موازی داشته باشد را به بخش‌هایی تقسیم کرده و هر بخش در رایانه‌ای حل‌شده و نهایتاً نتایج ترکیب می‌شوند این ترکیب می‌تواند یک یا چند مرتبه انجام شود.

دسته دیگر از مسائل هوش مصنوعی توزیع شده، مسائل یادگیری هستند که از زیر مجموعه‌ی سیستم‌های چندعاملی شناخته می‌شوند؛ این سیستم‌ها در راستای یادگیری از چند عامل بهره می‌برند. بر اساس روابطی که



شکل ۱ ـ ۱: جایگاه پژوهش انجام شده [۱، ۲]

می‌توان بین این عامل‌ها تعریف کرد سیستم‌های چندعاملی می‌توانند رقابتی یا مشارکتی باشند در سیستم‌های چند عامله رقابتی عامل‌ها به دنبال افزایش سود شخصی خود هستند که در زمان‌های زیادی به قیمت کاهش سود دیگر عامل‌ها خواهد بود؛ در سیستم‌های چندعاملی مشارکتی که موضوع پژوهش پیش رو نیز هست عامل‌ها به دنبال افزایش سود گروهی هستند. در شکل ۱ ـ ۱ جایگاه یادگیری مشارکتی در سیستم‌های چند عامله آورده شده است.

سیستم‌های چند عامله مشارکتی در دو دیدگاه موردبررسی قرار می‌گیرند در دیدگاه اول عامل‌ها در یک محیط قرارگرفته و سعی در یادگیری نحوه‌ی تعامل با محیط و یکدیگر را دارند و این یادگیری در این راستا است که عامل‌ها بتوانند در محیط با همکاری هم به اهداف مشترک و مشخص‌شده برسند. در دیدگاه دوم عامل‌ها تعاملی ندارند یعنی که قرار نیست با همکاری یکدیگر به یک هدف مشخص دست‌یابند بلکه فقط سعی دارند با همکاری یکدیگر به یادگیری برسند این عامل‌ها در محیط‌های جداگانه و مشابهی قرار می‌گیرند و در طول فرایند یادگیری با هم ارتباط دارند. در این ارتباط داده‌های به دست آمده را به یکدیگر منتقل می‌نمایند تا زمانی که عامل‌ها تمام محیط را به‌خوبی مورد شناسایی قرار دهند.

۱ ـ ۱ یادگیری مشارکتی در سیستم‌های چند عامله

انسان در طول حیات خود یادگیری زیادی انجام می‌دهد اما اگر قرار بود هر انسان از صفر شروع به جمع‌آوری اطلاعات کند و از عامل‌های دیگر یادگیری نداشته باشد بدون شک هنوز انسان‌ها همانند انسان‌های اولیه زندگی

می‌کردند. این رشدی که امروزه در زندگی انسانی دیده می‌شود مدیون انتقال اطلاعات و دانش بین عامل‌های انسانی است. بر همین اساس در [۳] یادگیری مشارکتی را سیستمی می‌داند که عامل‌ها در ان با همکاری یکدیگر به یادگیری یک وظیفه مشترک می‌پردازند. می‌توان آثار مثبت یادگیری مشارکتی در سیستم‌های چند عامله را چنین برشمرد.

- افزایش سرعت و دقت یادگیری.
- آزادسازی عامل از بهینگی محلی.
- کمک به تنظیم پارامترهای محلی عامل‌ها.

نکته مهمی که باید در نظر داشت عامل‌های موجود در یادگیری مشارکتی است. در بعضی از روش‌ها عامل‌های یک سیستم از توانایی‌های متفاوتی برخوردار هستند. به مجموعه عامل‌های موجود در این محیط‌ها تیم گفته می‌شود [۱، ۲] و در سیستم‌های دیگر مشابه پژوهش پیش رو عامل‌ها با توانایی‌های یکسان در نظر گرفته می‌شوند. گذشته از تفاوت بین عامل‌ها، یادگیری مشارکتی نیز همانند دیگر شاخه‌های هوش مصنوعی از چالش‌های فراوانی برخوردار است.فارغ از چالش‌های مشترکی که بیت یادگیری مشارکتی و روش یادگیری استفاده در ان وجود دارد چالش‌های جدیدی در یادگیری مشارکتی وجود دارد که می‌توان با سؤالاتی تعدادی از این چالش‌ها را نشان داد.

- چه اطلاعاتی باید بین عامل‌ها ردوبدل شود؟
- چه زمان باید اطلاعات منتقل شود؟
- ترکیب داده‌های دریافتی باید به چه صورت باشد؟
- بر اساس چه معیاری می‌توان عامل‌ها را مقایسه کرد؟

در رابطه با هریک از این چالش‌ها کارهای فراوانی چون پنددهی، تقلید و خبرگی انجام شدم است که تعدادی از آن‌ها در فصل دوم آورده شده است. تمام این روش‌های تشریح شده در فصل دوم از یادگیری تقویتی به عنوان الگوریتم اصلی عامل برای یادگیری نحوه‌ی تعامل با محیط بهره برده‌اند و سعی در ارائه‌ی روش جهت ترکیب مناسب داده‌ها نموده‌اند. در ترکیب داده‌ها همیشه نیازی به معیاری جهت سنجش میزان درستی داده‌ها (دانش هر عامل) وجود دارد اما از آنجایی که عامل‌ها محیط را نمی‌شناسند در نتیجه دستیابی به معیاری صحیح برای این منظور کار دشواری می‌باشد.

در روش‌هایی چون خبرگی سعی شده تا معیارهایی جهت سنجش داده‌ها ارائه شود اما در کار پنددهی عامل‌ها زمانی که از داده‌ی خود مطمئن باشند به عامل دیگر بازخورد می‌دهند و این بازخورد در درک عامل از محیط موثر خواهد بود. معمولا معیارهای معرفی شده در پژوهش‌های صورت گرفته بروی خبرگی یا با دیدگاه خیلی جزئی به بررسی خبرگی عامل‌ها می‌پردازند یا بصورت خیلی کلی؛ در حالت خلاصه کار اصلی که در این پژوهش انجام دادیم ارائه‌ی چهارچوب کلی برای تولید انواع معیارها و سپس ارائه‌ی معیاری که خیلی کلی یا جزئی نباشد و در عین حال بتواند عملکرد بهتری نسبت به روش‌های قبلی ارائه دهد.

۱ـ۲ اهداف و نوآوری‌های پایان‌نامه

بطور خلاصه هدف این پژوهش ارائه‌ی معیاری نرم (ساده در عین موثر بودن) بجهت محاسبه‌ی خبرگی عامل‌ها با در نظر گرفتن خاصیت غیرافزایشی [۴] خبرگی و دانش عامل‌ها می‌باشد – خاصیت غیرافزایشی دانش (خبرگی) عامل‌ها می‌گوید که ارزش دانش (خبرگی) چند عامل باهم لزوما برابر با مجموع ارزش دانش (خبرگی) تک‌تک آن‌ها نمی‌باشد. در طی دست‌یابی به هدف تعیین شده در این پژوهش ابتدا چهارچوبی به نام «فرضیه‌ی خبرگی» معرفی شد که توانایی استخراج معیارهای زیادی برای محاسبه‌ی خبرگی از طریق این فرضیه میسر باشد؛ سپس با استفاده از فرضیه‌ی خبرگی معرفی شده و با در نظر داشتن هدف تعیین شده برای این پژوهش در مورد ارائه‌ی معیار نرم، معیار خبرگی جدیدی به نام «میزان ارجاع» تعریف شد.

در طی این پژوهش الگوریتمی برای ترکیب دانش عامل‌ها با در نظر داشتن میزان خبرگی معرفی شده هر عامل ارائه شد. در این الگوریتم از انتگرال فازی به عنوان عملگر ترکیب کننده دانش عامل‌ها استفاده کردیم و طبق آزمایش‌ها نشان دادیم که انتگرال فازی چوکت می‌تواند نتایج بهتری نسبت به روش‌های سنتی چون میانگیری وزن‌دار تولید کند زیرا انتگرال فازی چوکت می‌تواند خاصیت غیرافزایشی مساله را برخلاف میانگین وزنی مدل کند. دستآوردهای این پژوهش به صورت خلاصه به شرح زیر می‌باشد:

- معرفی چهارچوبی به‌نام «فرضیه‌ی خبرگی» برای تعریف معیارهای خبرگی جدید.
- تعریف معیار خبرگی جدید به‌نام «میزان ارجاع» در چهارچوب معرفی شده توسط «فرضیه‌ی خبرگی».
- استفاده از «انتگرال فازی چوکت» در ترکیب دانش‌های عامل‌ها با توجه به میزان خبرگی عامل‌ها.
- تعریف معیاری جدید به‌نام «میزان باروری» به جهت سنجش سرعت یادگیری الگوریتم‌ها.
- بررسی تاثیر سیاست‌های انتخاب عمل εـحریصانه[1] در یادگیری مشارکتی – پژوهش‌های قبلی این موضوع را مورد بررسی قرار نداده‌اند.

[1] ε-greedy

- اثبات صحت فرضیه و معیار خبرگی معرفی شده در این پژوهش با توجه نتایج آزمایش‌ها.

۱ـ۳ ساختار پایان نامه

در ادامه‌ی گزارش در فصل دوم سعی شده کارهای انجام‌شده در این زمینه تشریح شود؛ در فصل سوم موضوعاتی که برای درک روش پیشنهادی در این پژوهش لازم است بیان شده است؛ سپس در فصل چهارم روش پیشنهادی تشریح و در فصل پنجم آزمایش‌های موردنیاز جهت نمایش عملکرد روش پیشنهادی آورده شده است. نهایتاً در فصل ششم جمع‌بندی از مطالب ارائه شده در این پایان‌نامه صورت گرفته است.

فصل دوم

مرور کارهای پیشین

۲ـ۱ مقدمه

در سال‌های گذشته پژوهش‌های فراوانی در سیستم‌های چند عامله انجام‌شده است که در این پژوهش‌ها محققان سعی داشته‌اند مزایای کار گروهی در انسان را در رایانه نیز ایجاد نمایند. یکی از قابلیت‌های عامل‌های هوشمند که می‌تواند باکار گروه سریع‌تر و بهتر شود موضوع یادگیری است که در این زمینه هم کارهایی انجام‌شده که معمولاً الگوبرداری از عامل‌های انسانی بوده است. همان‌طور که می‌دانیم انسان تنها از یک مکانیزم دررسیدن به یادگیری بهره نمی‌برد، عامل‌های انسانی با تقلید از عامل‌هایی که دارای اطلاعات بیشتری هستند توانسته‌اند یادگیری خود را بهبود دهند؛ عامل‌های انسانی در شرایط بحرانی زندگی از عامل‌های باتجربه‌تر پند می‌گیرند، عامل‌های انسانی در مراتبی از خبرگی قرار دارند؛ همه این موارد الگوهایی مناسب بوده که توانسته یادگیری در سیستم‌های چندعاملی را بهبود بخشد. اما می‌توان کارهایی که در یادگیری مشارکتی انجام می‌شود را به دسته‌هایی تقسیم کرد، هر دسته از پژوهش‌های انجام‌شده در این رشته سعی در رفع یک یا چند چالش از چالش‌های این رشته داشته‌اند. پژوهش پیش رو را می‌توان از دسته پژوهش‌های یادگیری مشارکتی دانست که سعی در حل مشکل ترکیب داده‌های عامل‌ها دارند که روش‌های ارائه‌شده در این فصل نیز روش‌هایی هستند که در تقسیم داده‌های یادگیری مشارکتی فعالیت کرده‌اند. پیچیدگی ترکیب داده‌های عامل به این دلیل است که معیار مناسبی جهت مشخص کردن داده‌ی درست وجود ندارد. در بسیاری از کارهایی که در این فصل ارائه خواهد شد در ترکیب

داده‌ها معیار جایگزینی معرفی‌شده و آن معیاری جهت نمایش برتری عامل است. ایده این جایگذاری ازآنجاست که عاملی که از برتری برخوردار باشد داده‌های بهتری نیز نسبت به عامل‌های دیگر خواهد داشت.

۲ـ۲ اشتراک‌گذاری اطلاعات

برای اولین بار در [۵] اشتراک‌گذاری داده‌ها در سیستم‌های چند عامله مورد ارزیابی قرار گرفت. هدف این بررسی نمایش اثر اشتراک‌گذاری داده‌ها در مقابل سیستم‌های تک عاملی بود. نتیجه این پژوهش نشان داد که اگر اشتراک‌گذاری به خوبی انجام شود می‌تواند سرعت و کیفیت یادگیری را به‌صورت چشم‌گیری افزایش دهد. در این پژوهش سه نوع اشتراک‌گذاری موردبررسی قرار گرفت در نوع اول که اشتراک‌گذاری ادراک نام گرفت عامل‌ها تنها نتایج مشاهدات خود را به اشتراک می‌گذاشتند، در نوع دوم اشتراک‌گذاری سه‌تایی حالت، عمل، کیفیت اشتراک‌گذاری شده و اشتراک‌گذاری واقعیت نامیده شد و نهایتاً در نوع سوم اشتراک‌گذاری که اشتراک‌گذاری سیاست خوانده می‌شود اطلاعات داخلی عامل‌ها که منبع استخراج سیاست آنهاست به اشتراک گذاشته شده است. اشتراک‌گذاری در این پژوهش با یک میانگین‌گیری ساده بین اطلاعات عامل‌ها انجام می‌شد. در این پژوهش که SA[1] نامیده شد است ثابت‌شده ممکن است اشتراک‌گذاری سربارهایی در ترکیب داده‌ها به سیستم بیفزاید یا در شروع یادگیری از سرعت یادگیری بکاهد اما در طول یادگیری این سربارها جبران شده و اشتراک داده‌ها می‌تواند به‌صورت چشم‌گیری در افزایش سرعت سیستم‌های چند عامله مؤثر باشد.

۲ـ۳ یادگیری مشترک

برنجی و همکاران در سال ۱۳۷۸ (۱۹۹۹ م.) روشی تحت عنوان یادگیری مشترک مطرح کردند [۶]. در این روش اشتراک‌گذاری با در نظر گرفتن تنها یک سیاست برای تمام عامل‌ها انجام شد .نتایج این پژوهش نشان می‌دهد که در دسته بزرگی از مسائل روش‌های یادگیری مشترک می‌تواند مفیدتر از روش‌های یادگیری مستقل باشد. فرآیند یادگیری در این روش به این صورت است که عامل ها در محیط اقداماتی انجام می‌دهند و بعد ازدریافت پاداش عمل بروزرسانی را در یک داده مشترک انجام می‌دهند و در انتخاب عمل نیز از همان داده مشترک بهره می‌برند. این به این معنی است که عامل ها دیگر برای خود داده مستقلی ندارند. در این پژوهش حتی یادگیری با منطق فازی ادغام‌شده است و نویسندگان سعی کردند اثر فازی کردن داده‌ها در یادگیری مشارکتی را نمایش دهند.

[1] Simple Averaging

۲ـ۴ تقلید

انسان در طول زندگی برای رسیدن به یادگیری روش‌های متفاوتی دارد. گاهی برای رسیدن به یادگیری باید آزمایش کرد گاهی تحلیل کرد و گاهی تجربه اما یک روش که انسان از آن مخصوصاً در مراحل رشد بسیار بهره می‌برد تقلید است. همین موضوع باعث شده که در یادگیری مشارکتی نیز به تقلید عامل‌ها از هم توجه شود. بر همین اساس نونس و همکاران با ایده برداری از تقلید در انسان پیشنهاد کردند که رابطه عامل‌ها از طریق تقلید از یکدیگر باشد [۷].

موضوع دیگری که در مورد تقلید عامل‌های انسانی باید در نظر گرفته می‌شد این است که عامل‌های انسانی از عامل‌های انسانی تقلید می‌کنند که اطلاعات بیشتری دارند. در پیاده‌سازی انجام‌شده نیز بر همین اساس سه نوع تقلید پیشنهاد می‌شود. تقلید می‌تواند به‌صورت ساده باشد. پیشنهاد داده‌شده است که عامل‌ها همیشه از عامل‌های همسایه (همسایگی در این روش بر اساس همسایگی محلی است چراکه عامل‌هایی که در منطقه یکسانی قرار دارند کمک بیشتری می‌توانند به هم کنند) خود تقلید نمایند. این موضوع یک دور در عامل‌ها ایجاد می‌کند که هر عامل منتظر می‌ماند تا عامل دیگر حرکتی انجام دهد. برای رفع این موضوع نوع دیگری از تقلید به نام تقلید شرطی مطرح می‌شود در تقلید شرطی عامل از کسانی تقلید می‌کند که عملکرد بهتری نسبت به او داشته‌اند در این حالت موضوع دور و انتظار عامل‌ها برطرف شده است. اما درروش سوم که تقلید انطباقی نام دارد عامل همیشه تقلید نکرده و تقلید بر اساس یک احتمال انجام می‌شود.

۲ـ۵ حافظه جمعی

گارلند و همکاران در سال ۱۳۷۵ (۱۹۹۶ م.) ایده جدید خود را با عنوان یادگیری حافظه جمعی مطرح کردند [۸، ۹]. در یادگیری حافظه جمعی که برگرفته از شناخت توزیع‌شده در علوم اجتماعی می‌باشد عامل‌ها تجارب خود را در یک حافظه مشترک نگهداری می‌کنند. هر عامل در زمان برخورد با مشکلات می‌تواند با بهره‌گیری از این تجارب راه درست را پیدا کند. این روش در دو دیدگاه مورد ارزیابی قرارگرفته است. در دیدگاه اول عامل‌ها الگوهای موفق خود در طول یادگیری را در حافظه مشترک نگهداری می‌کنند تا در زمان نیاز تمام عامل‌ها با استفاده از این الگوها بتوانند راه‌حل مشکلات خود را پیدا کنند. در دیدگاه دیگر احتمال موفقیت عامل‌ها نگه‌داری می‌شود که با بهره‌گیری از این داده می‌توان میزان موفقیت عامل‌ها در اعمال مختلف را ارزیابی کرده و در جهت بهبود طراحی سیستم مورد ارزیابی قرارداد. لازم به ذکر است که در این پژوهش‌ها حافظه جمعی را در دو حالت حافظه مرکزی و حافظه توزیع‌شده بین عامل‌ها مورد ارزیابی قرار داده است.

۲ـ۶ پند

در سال ۱۳۸۱ (۲۰۰۲ م.) نونس و همکاران باردیگر روشی جدید با عنوان پند دهی مطرح کردند [۱۰]. در جوامع انسانی پند دادن بسیار رواج داشته و در زمان مشکلات بسیار کارا می‌باشد. یک عامل انسانی در زمان برخورد با مشکلات از عامل‌هایی که اطلاعات بیشتری دارند پند گرفته و مشکلات خود را حل می‌کند. عاملی انسانی که دارای اطلاعاتی است هم اطلاعات خود را باتجربه کردن و یا گرفتن پند در زمان‌های دیگر به‌دست می‌آورد. مشخصاً یادگیری تقویتی در حالت معمول با تجارب به یادگیری می‌رسد. اگر هر تجربه را بازخوردی از محیط در نظر بگیریم هر پند را نیز می‌توان بازخوردی از عامل‌های دیگر دانست. با این ایده دیگر حتی نیازی نیست که عامل‌ها از روش‌های یکسانی در یادگیری بهره ببرند زیرا پند دادن به عامل‌ها را می‌توان فارغ از روش یادگیری پیاده‌سازی کرد. ایده پردازان پند در [۱۱] کار قبل خود را کامل‌تر کرده و این ایده را به‌صورتی که عامل‌ها در یک محیط به تعامل می‌پرداختند پیاده‌سازی کردند. هر عامل بعد از رسیدن به هر حالت موقعیت خود را به عامل‌های دیگر ارسال می‌نماید. عامل‌هایی که تجربه مشابهی داشته‌اند در پاسخ مقداری را به عنوان میزان ارزش عمل انجام شده برای عامل ارسال می‌کنند و عامل از این مقادیر همانند پاداش دریافتی از محیط بهره می‌برد.

۲ـ۷ یادگیری مشارکتی بر مبنای خبرگی

تشریح یادگیری مشارکتی بر مبنای خبرگی را با یک سؤال می‌توان آغاز کرد. آیا عامل‌ها در شناخت محیط از خبرگی یکسانی برخوردار هستند؟ مسلماً چنین نیست، در [۱۲] ایده یادگیری مشارکتی بر مبنای خبرگی با عنوان WSS[1] مطرح می‌شود. همان‌طور که در تشریح روش SA مطرح شد در این روش با میانگین‌گیری از اطلاعات عامل‌ها ترکیب انجام می‌شود. در این میانگین‌گیری تمام عامل‌ها به یک اندازه سهیم هستند. ایده پردازان WSS با طرح این موضوع که میزان خبرگی عامل‌ها یکسان نیست سعی کردند هر عامل در ترکیب داده‌ها به میزان توانایی و خبرگی خودش مؤثر باشد.

نویسندگان با ارائه معیارهایی میزان خبرگی عامل‌ها را سنجیده و بر همین اساس داده‌ها باهم ترکیب می‌شوند. در WSS روال یادگیری به دو فاز یادگیری مستقل و یادگیری مشارکتی شکسته شده است. در یادگیری مستقل هر عامل به‌طور مستقل به یادگیری می‌پردازد این یادگیری منجر به کسب اطلاعاتی می‌شود که در فاز یادگیری مشارکتی باهم ترکیب می‌شوند. یادگیر در فاز یادگیری مستقل چندین چرخه یادگیری را تجربه می‌کند. تعداد این چرخه‌ها می‌تواند در بین عامل‌ها یکسان و یا متفاوت باشد. اما باید در انتخاب تعداد چرخه‌های یادگیری هر فاز یادگیری مستقل دقت کرد چراکه اگر این تعداد کم در نظر گرفته شود عامل اطلاعات کافی را جمع آوری

[1] Weighted Strategy Sharing

نکرده است و اگر زیاد در نظر گرفته شود از تأثیر یادگیری مشارکتی خواهد کاست.

در فاز دوم یادگیری عامل‌ها باید به یادگیری مشارکتی بپردازند. در آغاز این فاز میزان خبرگی عامل‌ها سنجیده می‌شود و پس از ان داده‌ها ترکیب‌شده و جداول Q عامل‌ها بروز رسانی می‌شود. در [۱۲] روش‌هایی جهت ترکیب دادها ارائه شده است. در یکی از روش‌ها جدول تمام عامل‌ها با بهره‌گیری از میزان خبرگی میانگین‌گیری شده و جدول تولیدشده به تمام عامل‌ها داده شود که در صورت انجام این کار بعد از فاز یادگیری مشارکتی تمام عامل‌ها جدول Q یکسانی خواهند داشت. درروش دیگری پیشنهادشده که هر عامل جدول جدید خود را با ترکیب جدول خود با جدول عامل‌های خبره‌تر از خودش تولید کند. در این ترکیب نیز هر عامل به میزان خبرگی خودش در ترکیب داده‌ها سهم خواهد داشت.

در WSS با در نظر گرفتن خبرگی عامل‌ها تأثیر زیادی در بهبود یادگیری مشارکتی داشته است اما نکته‌ای که در نظر گرفته نشده است اینجاست که میزان خبرگی عامل‌ها در دامنه‌های مختلف بسیار متفاوت بوده و بهتر است که در ترکیب داده‌ها این دامنه‌ها هم در نظر گرفته شود. در [۱۳] با در نظر گرفتن دامنه خبرگی عامل‌ها سعی شده تا نقصان WSS برطرف شود. بعدازآن در [۱۴] سعی شده تا استفاده از جدول Q یک عامل در ترکیب داده‌ها قطعی نباشد. در این راستا در فاز ترکیب برای اطلاعات هر عامل احتمالی در نظر گرفته‌شده است که نشان‌دهنده احتمال حضور اطلاعات آن عامل در ترکیب داده‌ها است. میزان این احتمال نیز بر اساس تفاوت میزان خبرگی عامل‌ها محاسبه‌شده است. در ادامه تعدادی از معیارهای خبرگی معرفی‌شده در [۱۲] خواهد آمد.

- **معیار خبرگی معمولی:** در این معیار میزان خبرگی عامل‌ها بر اساس مجموع پاداش‌های دریافتی آنها در نظر گرفته‌شده است. درنتیجه عاملی که میزان پاداش منفی کمتر و میزان پاداش مثبت بیشتری گرفته است را عامل خبره‌تر می‌داند.

- **معیار خبرگی مثبت:** در این معیار سعی شده با شمارش پاداش‌های مثبت عامل‌ها میزان خبرگی اندازه‌گیری شود. ایده انتخاب این معیار این بوده که عاملی که پاداش مثبت بیشتری گرفته است از خبرگی بالاتری برخوردار است.

- **معیار خبرگی منفی:** این معیار برعکس معیار خبرگی مثبت با این ایده که عاملی که پاداش منفی بیشتری دارد نقاط بحرانی بیشتری را می‌شناسد عمل شده و تعداد پاداش‌های منفی عامل یادگیری را شمارش می‌نماید.

- **معیار خبرگی قدر مطلق:** در معیار خبرگی قدر مطلق میزان خبرگی عامل با محاسبه مجموع قدر مطلق پاداش‌ها دریافتی او انجام می‌شود. درنتیجه به پاداش‌های منفی و مثبت ارزش یکسانی داده‌شده است.

- **معیار خبرگی گرادیان:** در این معیار مانند معیار اول عمل می‌شود با این تفاوت که میزان افزایش سیگنال دریافتی نسبت به آخرین دوره‌ی یادگیری مشارکتی را معیار خبرگی قرار داده است، هرچقدر این اختلاف بیشتر و مثبت باشد نشان می‌دهد که عامل نسبت به دوره‌ی قبل خبره‌تر شده است.
- **معیار خبرگی میانگین تعداد قدم‌ها:** این معیار برعکس پنج معیار دیگر به‌جای تأکید بر روی پاداش‌ها میانگین تعداد قدم‌های عامل در چرخه‌های یادگیری را معیار می‌داند. این انتخاب با این ایده انجام‌شده که عامل‌های خبره‌تر با تعداد قدم‌های کمتر چرخه‌های یادگیری را به اتمام می‌رسانند.

۲ـ۸ یادگیری مشارکتی بر مبنای تخته‌سیاه

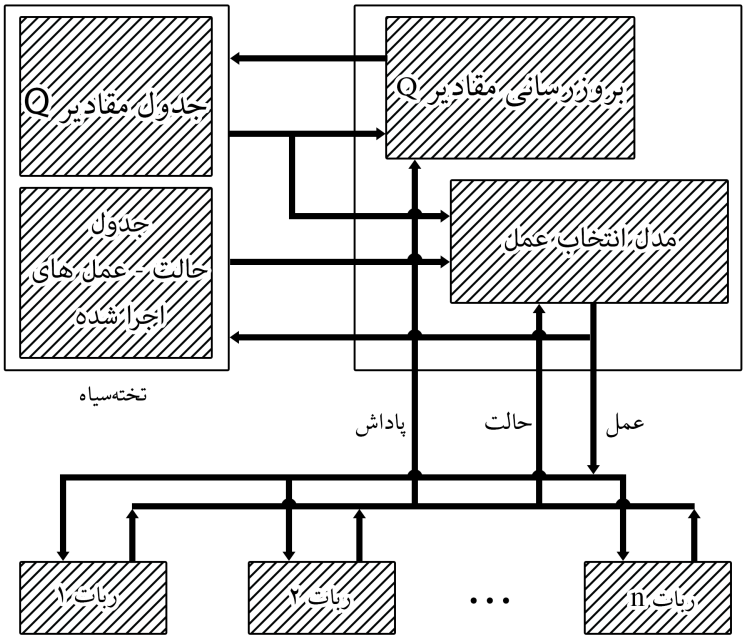
در [۱۵] سازوکار تخته‌سیاه مطرح شد. تخته‌سیاه یک حافظه مرکزی است که تمام عامل‌ها به آن دسترسی دارند. در این روش عامل‌ها به‌طور مستقیم باهم ارتباط نداشته و ارتباطات از طریق همین تخته‌سیاه انجام می‌شود. هر عامل می‌تواند بر روی تخته نوشته و یا از آن بخواند. درروش پیشنهادشده در [۱۵] به این شکل است که عامل بعد از رسیدن به هر موقعیت حالت خود را به تخته‌سیاه اعلام می‌کند و تخته‌سیاه عملی را بر اساس حالت جاری به عامل برمی‌گرداند. عامل بعد از انجام آن عمل و دریافت بازخورد از محیط این بازخورد را به تخته‌سیاه برمی‌گرداند.

تخته‌سیاه دودسته از داده‌ها را نگه‌داری می‌کند. دسته اول داده‌ها همان جدول Q عامل‌ها است و دسته دوم از داده‌ها عمل‌های انجام‌شده توسط هر عامل است. همان‌طور که مشخص است در این روش بروز رسانی جدول Q و انتخاب عمل از عامل به تخته‌سیاه منتقل‌شده و مشخصاً جدول Q باید در تخته‌سیاه پردازش شود. اما دسته دوم اطلاعات صرفاً جهت کمک به انتخاب عمل عامل‌ها انجام می‌شود. به‌عنوان‌مثال اگر عامل در حالتی قرار گیرد و عملی تجربه نشده باشد آن عمل پیشنهاد می‌شود. پس ذخیره‌سازی دسته دوم اطلاعات در جهت مدیریت اکتشاف و بهره‌برداری عامل‌ها از اطلاعات است. در شکل ۲ـ۱ مکانیسم تخته‌سیاه نمایش داده‌شده است.

۲ـ۹ یادگیری تقویتی تعاملی

در سال ۱۳۸۵ (۲۰۰۶ م.) لیما و همکاران سه الگوریتم به نام‌های BEST-Q, AVG-Q و PSO-Q را ارائه دادند که این دسته از الگوریتم‌ها مساله‌ی یادگیری مشارکتی تقویتی را به صورت جستجوی جدول Q توسط عامل‌ها تعریف کرده است [۱۶]. لیما و همکاران ابتدا معیاری ارائه دادند برای سنجش میزان خوب بودن مقادیر جداول Q که در واقع متشکل از مقادیر پاداش‌های دریافتی عامل‌ها می‌باشد، که این معیار شامل جمع تخفیف یافته[1] پاداش‌ها در هر چرخه‌ی یادگیری می‌باشد به‌صورتی که به پاداش‌های نهایی ارزش بیشتری می‌دهد. سپس با

[1] Discounted Sum



شکل ۲ـ۱: شماتیک مکانیزم روش تخته‌سیاه برای یادگیری تقویتی مشارکتی [۲]

استفاده از این معیار به جستجوی بهترین جدول Q از طریق عامل‌ها می‌پردازد؛ که این مساله وجه تمایز کار لیما با روش پیشنهادی در این پژوهش می‌باشد. الگوریتم BEST-Q در مرحله‌ی به اشتراک‌گذاری دانش، بهترین دانش (با توجه به معیار معرفی شده) را به عنوان دانش جمعی در نظر می‌گیرد. الگوریتم AVG-Q بهترین دانش را با دانش هر عامل میانگین‌گیری می‌کند و به عنوان دانش همان عامل در نظر می‌گیرد و در الگوریتم PSO-Q الگوریتم PSO [۱۷] را به عنوان مدل جستجو کننده دانش جمعی در نظر گرفته است.

یکی از معایب این روش‌ها این است که یادگیری تقویتی را به صورت یک مساله‌ی جستجوی مقادیر جداول Q در نظر گرفته است و از آنجایی که ماهیت الگوریتم‌ها به صورت جستجو می‌باشد، اثباتی جهت اینکه این جستجوهای مقادیر جداول Q به مقادیر بهینه یعنی جدول Q^* همگرا خواهند شد، وجود ندارد. از دیگر معایب این روش‌ها نحوه‌ی محاسبه‌ی میزان بهینگی مقادیر جداول عامل‌ها می‌باشد، بطوری که این روش‌ها به پاداش‌های دریافتی در اواخر چرخه‌ی یادگیری عامل‌ها ارزش بیشتری می‌دهند، این ممکن است در ابتدای یادگیری که عامل‌ها در اوایل چرخه‌ی یادگیری خود حرکت‌های بیهوده‌ی زیادی انجام دهند منطقی به نظر بیاید، ولی بعد از آنکه تعدادی چرخه‌ی یادگیری سپری شد و عامل به دانش نسبی خوبی از محیط خود دست یافت حرکت‌های ابتدایی به اندازه‌ی حرکت‌های نهایی ارزش دارند (و البته شاید هم ارزش بیشتری داشته باشند) زیرا که عامل برای دست یافتن به هدف، نسبت به محیط اطراف اهداف شناخت بهتری دارد (به علت خاصیت شوک یادگیری تقویتی [۲]) که در این شرایط در حالت کلی میزان بهینگی عامل‌ها را بیشتر میزان بهینگی اعمال در اوایل چرخه‌ی یادگیری عامل تعیین می‌کند. معیاری که روش‌های فوق‌الذکر از آن، جهت سنجش خبرگی عامل‌ها در

نظر گرفته‌اند این مساله را نادیده می‌گیرد که باعث می‌شود خبرگی عامل‌ها را نتوان به درستی تعیین کرد.

۲ـ۱۰ یادگیری مشارکتی بر مبنای پختگی سیاست

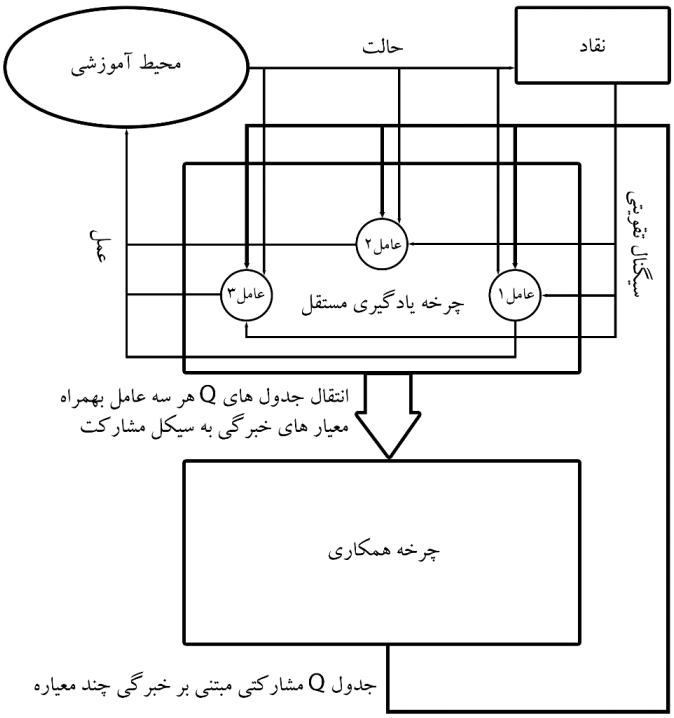
یانگ و همکاران در سال ۱۳۸۸ (۲۰۰۹ م.) روشی با عنوان یادگیری مشارکتی بر مبنای خبرگی چند معیاری ارائه دادند [۱۸]. این روش تا حدودی ترکیب روش تخته‌سیاه با WSS هست. در این روش عامل‌ها حافظه مرکزی خود یا تخته‌سیاه را دارند که وجود تخته‌سیاه عامل‌ها را از شکستن بازه یادگیری به دو فاز بی‌نیاز می‌سازد. در روشی چون WSS یادگیری به دو فاز یادگیری مستقل و یادگیری مشارکتی شکسته می‌شد تا عامل‌ها داده‌های خود را به اشتراک بگذارند اما زمانی که عامل‌ها دائما می‌توانند داده‌های خود را بر روی تخته‌سیاه نوشته و بخوانند ارتباط از طریق همین تخته‌سیاه انجام خواهد شد.

اما عامل‌ها بر خلاف روش تخته‌سیاه از ارتباط مستقیم هم در تصمیم‌گیری‌ها و انتخاب اعمال بهره می‌برند. عاملی که در وضعیت انتخاب عمل قرار گرفته می‌تواند از عامل‌های دیگر بیاموزد. در این روش جهت شناخت عامل‌هایی که اطلاعات خوبی دارند و می‌توانند آموزگار باشند از معیارهای خبرگی ارائه‌شده در WSS استفاده‌شده است. با این کار عامل از عامل‌هایی می‌آموزد که واقعاً از خبرگی بالاتری برخوردار هستند. این کار باعث می‌شود که در شروع یادگیری که عامل‌ها داده کمی دارند و نیز عامل آموزگاری نداشته با کمک اطلاعات و دستورات تخته‌سیاه عمل کند و بعد طی مراحلی از یادگیری که عامل‌ها داده‌های زیادی کسب کردند با بهره بردن از نظرات دیگر عامل‌ها انتخاب‌های بهتری داشته باشند.

۲ـ۱۱ یادگیری مشارکتی بر مبنای خبرگی چند معیاری

پاکیزه و همکاران در سال ۱۳۹۲ (۲۰۱۳ م.) با نقد روش WSS روشی جدید ارائه کردند [۱۹]. ایشان با اشاره به این موضوع که خبرگی در یک‌رشته نبوده در کار خود از ترکیب ۶ معیار خبرگی WSS در کنار هم بهره برده‌اند. ایشان تأکیددارند که عامل‌های انسانی درزمینه‌های مختلف خبرگی‌های متفاوتی دارند و این موضوع در عامل‌های هوشمند نیز وجود دارد. این پژوهش هر یک از معیارهای ارائه‌شده در WSS را مانند یک زمینه در عامل انسانی دانسته و درروش خود از تمام این معیارها در کنار هم بهره برده‌اند.

پاکیزه و همکاران مانند WSS یادگیری را در دو فاز یادگیری مستقل و یادگیری مشارکتی تقسیم می‌نمایند عامل‌ها در فاز یادگیری مشترک از هر معیار برای ترکیب داده‌های جدول Q بهره می‌برند و بعد از ترکیب جدول به‌وسیله هر معیار ۶ جدول مشارکتی تولید می‌شود که هر یک بر اساس یک معیار خبرگی است. آن‌ها برای ترکیب این جداول آنها را باهم جمع می‌کنند. اما موضوعی که وجود دارد این است که جدول تولیدشده به‌وسیله جمع چندین جدول دیگر خواص جدول Q را ندارد. برای رفع این مشکل این جدول را نه در جایگزینی با جدول Q



شکل ۲ـ۲: شمایی از یادگیری مشارکتی برمبنای خبرگی عامل‌ها [۱]

عامل‌ها بلکه در کنار جدول Q عامل نگه‌داری می‌نمایند. به عبارت دیگر هر عامل دو جدول دارد یک جدول Q که بر اساس یادگیری تقویتی است و جدول دیگر که جدول مشارکتی عامل‌ها است. پاکیزه و همکاران پیشنهاد کردند که از جدول مشارکتی که خواص جدول Q عامل‌ها را ندارد صرفاً برای انتخاب عمل استفاده شود و عامل بر اساس این جدول عمل را انتخاب کرده انجام دهد سپس جدول Q خود را بروز رسانی نماید. جهت درک بهتر این روش شمای کلی آن در شکل ۲ـ۲ آورده شده است.

۲ـ۱۲ تسریع یادگیری مشارکتی با بهره‌گیری از کوتاه‌ترین فاصله تجربه‌شده

میرزایی در سال ۱۳۹۵ (۲۰۱۶ م.) جهت تسریع در یادگیری مشارکتی دو معیار جدید را ارائه کرد [۲]. معیار اول یک معیار مکاشفه است که کوتاه‌ترین فاصله تجربه‌شده توسط عامل از هر حالت و عمل را شمارش می‌کند. ایشان نام این معیار را SEP[1] گذاشته است. معیار دیگر که شوک نام‌گذاری شده است میزان شناخت عامل از هر حالت و عمل را محاسبه می‌نماید.

میرزایی برخلاف دیگران فقط در فاز ترکیب داده‌های یادگیری مشارکتی تغییر ایجاد نکرده است. وی در فاز انتخاب عمل توسط عامل‌های مشارکتی نیز از جدول SEP در کنار جدول Q استفاده کرده است. استدلال ایشان در انجام این کار چنین بوده که عامل‌های یادگیری تقویتی در فازهای اول یادگیری داده زیادی ندارند و ازآنجایی‌که جدول SEP با سرعت بیشتری به‌روزرسانی می‌شود بهتر است انتخاب اعمال در فازهای اولیه یادگیری

[1] Shortest Experienced Path

بیشتر بر اساس SEP انجام شود. ایشان با استفاده از شوک که نمایشی از میزان شناخت عامل از هر حالت و عمل است تعادلی بین بهره‌برداری از جدول SEP و جدول Q برقرار کرده است. در شروع یادگیری که شناخت عامل کمتر است بیشتر انتخاب بر اساس SEP انجام می‌شود و در طول یادگیری با افزایش میزان شناخت عامل از محیط انتخاب عمل بر اساس جدول Q افزایش می‌یابد.

همچنین میرزایی در پژوهش خود در فاز ترکیب داده‌ها نیز روش جدیدی ارائه داد. ازآنجایی‌که وی یک جدول جدید به سیستم افزوده است در فاز ترکیب داده‌ها جدول SEP عامل‌ها را نیز ترکیب می‌نماید. همچنین جداول SEP عامل‌ها را تنها با یک حداقل‌گیری باهم ترکیب کرده و به عامل‌ها برمی‌گرداند. سپس ترکیب جداول Q عامل‌ها به‌صورت محلی انجام می‌گیرد به این صورت که هر سطر از جدول که نمایش یک حالت از محیط است به‌صورت جداگانه بروز رسانی می‌شود. ایشان در ترکیب داده‌های هر سطر عامل‌ها را به دو گروه تقسیم نموده و داده‌های هر گروه را جداگانه ترکیب می‌نماید. این تقسیم‌بندی بر اساس رابطه بین سیاست‌های استخراج‌شده از جدول Q و SEP عامل در یک حالت هست. وی عامل‌هایی که سیاست استخراج‌شده از جدول Q و SEP در آنها همخوانی داشته باشد در یک گروه و عامل‌هایی که سیاست استخراج‌شده آنها عمل‌های متفاوتی را پیشنهاد می‌کنند را در گروه دیگر قرار داده است. ترکیب داده‌های هر گروه با استفاده از میزان شناخت عامل از آن حالت (شوک) انجام می‌شود به این صورت که داده‌های عملی که شناخت بیشتری دارند بیشتر مورداستفاده قرار می‌گیرند. در فصل بعد روش محاسبه ارائه شده توسط میرزایی تشریح شده است.

۲ـ۱۳ نتیجه‌گیری

روش‌های ارائه‌شده در این فصل در ترکیب داده‌ها در یادگیری مشارکتی عمل می‌نمایند. در بعضی از این روش‌ها سعی شده رابطه عامل‌ها به‌صورت غیرمستقیم باشد و در روش‌هایی چون WSS سعی شده تا زمانی برای جمع‌آوری داده‌ها به عامل داده شود و بعد از آن‌یک فاز ترکیب داده‌ها وجود داشته باشد. در این روش‌ها زمان ترکیب و ارتباط عامل‌ها مشخص و ثابت است و در روش‌هایی دیگر چون پند دهی عامل‌ها هر زمان که نیاز به کمک داشته باشند می‌توانند از عامل‌های دیگر بهره ببرند. باید تأکید کرد که روش‌های ارائه‌شده در این پژوهش همانند روش WSS زمان ترکیب داده‌های ثابتی در نظر گرفته‌شده است.

فصل سوم

مفاهیم علمی پیش‌نیاز پایان‌نامه

۳ـ۱ مقدمه

در این فصل سعی شده تا موضوعاتی که درروش پیشنهادی به‌کاررفته‌اند و به درک بهتر موضوع کمک می کنند تشریح می‌شوند در این جهت ابتدا در مورد روش های یادگیری و روش یادگیری Q که معمولاً در کارهای یادگیری مشارکتی استفاده میشود توضیح داده شده است. مطمئناً شناخت یادگیری تقویتی حتی به صورت جزئی می‌تواند در درک یادگیری مشارکتی بسیار مؤثر باشد. ازآنجایی‌که بررسی عملکرد روش پیشنهادی با روش یادگیری مشارکتی بر مبنای کوتاه‌ترین فاصله تجربه‌شده انجام می‌شود در ادامه به تشریح معیارهای SEP و شک پرداخته خواهد شد. بعد از آن محیط‌های آزمایشی و معیارهای ارزیابی استفاده شده در آزمایش‌های این پژوهش تشریح خواهد شد. همچنین از آنجایی که در این پژوهش از انتگرال فازی چوکت استفاده شده است گذری خلاصه بر اندازه‌گیری‌های فازی و غیرافزایشی شده است و سپس از بین انتگرال‌های فازی دو انتگرال همه کاره سوگنو و چوکت که می‌توان به‌روی هر نوع داده‌ای اعمال کرد را معرفی کردیم و نشان داده شده است چرا در کاربرد مورد استفاده در این پژوهش فقط از انتگرال چوکت بهره برده شده است.

الگوریتم ۳-۱ الگوریتم یادگیری Q [۲۰]

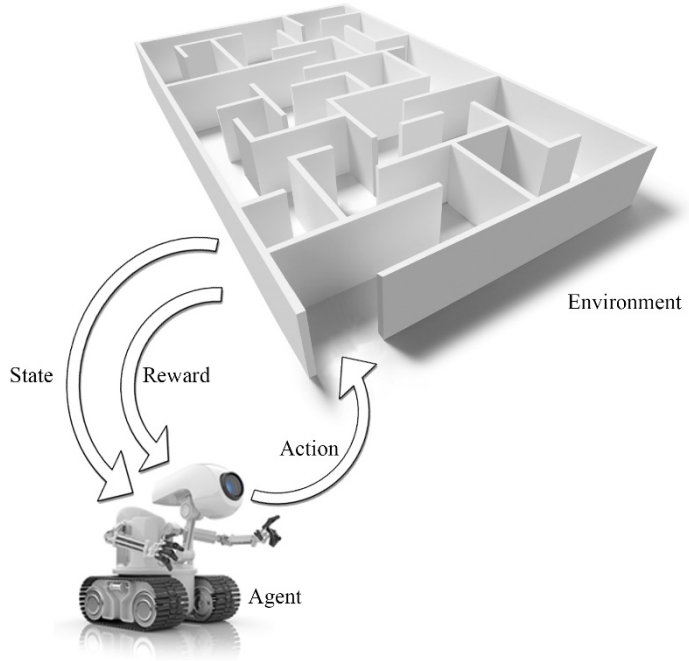
```
procedure Q-LEARNING
Ensure: Intialize the Q matrix;
    while not End Of Learning do
        Visit the state s;
        Select an action a based on an action selection policy;
        Carry out the a and observe a reward r at the new state s';
        Q[s,a] ← Q[s,a] + α(r + λ max_a'(Q[s',a']) − Q[s,a]);
        s ← s';
    end while
    return Q;
end procedure
```

۳ـ۲ یادگیری تقویتی

یادگیری تقویتی معمولا به آن سری از روش‌های یادگیری گفته میشود که عامل به دنبال رسیدن به یادگیری از طرق ارتباط با محیط است. در این دسته از روش‌ها به عامل ارزش اعمال گفته نمی‌شود و عامل با تعاملی که با محیط دارد باید ارزش اعمال را کشف نماید. در این روش فرایند یادگیری به بخش‌هایی با عنوان چرخه یادگیری شکسته می‌شود، که الگوریتم ۳-۱ به ازای هر چرخه‌ی یادگیری به اجرا در می‌آید. هر چرخه یادگیری از قرار دادن عامل در یک حالت تصادفی شروع و تا رسیدن به یک حالت پایانی ادامه دارد. در طول هر چرخه تا زمان رسیدن به حالت پایانی عامل وظیفه دارد عمل را انتخاب نماید و پس از دریافت پاداش عمل انجام‌شده به بروز رسانی اطلاعات بپردازد. بروز رسانی داده‌ها بر اساس فرایند تصادفی مارکف و روش برنامه نویسی پویا می‌تواند انجام شود. بر اساس تئوری ارائه‌شده در فرایند تصادفی مارکف باید انتخاب عمل به صورتی باشد که انجام هر عمل در هر حالت ضمانت شود. معمولا در یادگیری تقویتی این انتخاب عمل به‌وسیله روش‌هایی چون بولتزمن انجام میشود. در شکل ۳ـ۱ می‌توان فرایند یادگیری را مشاهده کرد.

۳ـ۳ روش‌های انتخاب عمل

همان‌طور که گفته شد یکی از وظایف عاملی که از یادگیری تقویتی استفاده می‌کند انتخاب عمل در طول یادگیری است. عامل بعد از رسیدن به هر حالت تا زمان رسیدن به یک حالت پایانی باید اعمالی انتخاب نماید. این انتخاب می‌تواند کاملاً بر اساس داده‌های جمع‌آوری شده در چرخه‌های یادگیری قبلی باشد که اصطلاحاً بهره‌برداری نامیده می‌شود. برای این کار کافی است در هر حرکت عملی انتخاب شود که در جدول Q ارزش بالاتری دارد اما این کار باعث می‌شود تا عامل مسیرهایی را تکرار نماید و شرط بررسی تمام‌مسیرها که از فرایند تصادفی مارکو به یادگیری تقویتی رسیده است را ارضا نمی‌کند. پس عامل نیاز است در طول یادگیری گاهی



شکل ۳ـ۱: شمایی از فرایند یادگیری تقویتی در تعامل با محیط [۲]

فارغ از بهترین عمل اعمال دیگر را نیز بررسی نماید. که این بررسی اعمال دیگر را اکتشاف گویند. در بدترین حالت می‌توان گفت که در طول یادگیری عامل به‌صورت تصادفی اعمال را انتخاب نماید اما این موضوع فرایند یادگیری را بسیار طولانی می‌کند. پس عامل باید برای رسیدن به یک تعادل در اکتشاف و بهره‌برداری یک روش مناسب در انتخاب اعمال داشته باشد. در ادامه دو روش انتخاب عمل که در یادگیری مشارکتی مورداستفاده قرار می‌گیرد خواهد آمد.

۳ـ۳ـ۱ εـحریصانه

در εـحریصانه جهت رسیدن به یک تعادل در اکتشاف و بهره‌برداری یک پارامتر ε که مقداری بین ۰ و ۱ است در نظر گرفته میشود. سپس به‌احتمال ε اکتشاف و به‌احتمال $(1-\varepsilon)$ بهره‌برداری از داده‌های جدول Q انجام می‌شود. جهت پیاده‌سازی این روش یک مقدار تصادفی ξ بین ۰ و ۱ تولیدشده و درصورتی‌که این ξ کوچک‌تر از ε باشد حرکت تصادفی و در غیر این صورت بهترین حرکت بر اساس داده‌های جدول Q انجام میشود.

۳ـ۳ـ۲ بولتزمن

روش εـحریصانه توانسته تا حدودی بین اکتشاف و بهره‌برداری تعادل ایجاد نماید اما احتمال انتخاب اعمال در εـحریصانه ثابت بوده و ارتباطی باارزش اعمال در جداول Q ندارد. روش بولتزمن با بهره‌گیری از رابطه ۳ـ۱ سعی دارد تا احتمال انتخاب عمل iام در یک موقعیت s را بر اساس ارزش اعمال در همان موقعیت در جدول Q محاسبه شود [۲۰]. همچنین در این رابطه مشخص است این روش دارای پارامتری به‌عنوان τ برای

جدول ۳ـ۱: ساختار جدول CP [۲]

...	آخرین عمل انجام شده در حالت i	...
...	حالت بعدی که مشاهده شده	...
...	شماره آخرین گام حرکت مشاهده حالت i	...

کنترل حساست به اختلاف ارزش بین اعمال در نظر گرفته‌شده است هرچقدر این میزان بزرگ‌تر باشد به اختلاف ارزش‌ها اهمیت کمتری می‌شود. یعنی در صورتی که $\tau \to \infty$ تمامی اعمال ممکن در موقعیت s به احتمال یکسانی انتخاب می‌شود و در صورتی که $\tau \to 0$ به صورت حریصانه عملی انتخاب می‌شود که مقدار Q بیشتری دارد.

$$p_Q(a|s) = \frac{e^{\frac{Q[s,i]}{\tau}}}{\sum_b e^{\frac{Q[s,b]}{\tau}}} \qquad (۱ـ۳)$$

۳ـ۴ الگوریتم مورد مقایسه با روش پیشنهادی

در فصل دوم به تشریح روش یادگیری مشارکتی بر مبنای «کوتاه‌ترین فاصله تجربه‌شده» پرداخته شد اما از آنجایی که بررسی عملکرد روش پیشنهادی در این پژوهش با کار میرزایی [۲] انجام‌شده است که مدرن‌ترین روش ارائه‌ی شده در این زمینه می‌باشد، لذا لازم است شناخت بیشتری از معیارهای به‌کاررفته در این کار صورت گیرد.

۳ـ۴ـ۱ معیار کوتاه‌ترین فاصله تجربه‌شده

میرزایی معیاری با عنوان «کوتاه‌ترین فاصله تجربه‌شده» تعریف کرد [۲]، اما بهتر بود «کوتاه‌ترین فاصله استدلال شده» نامیده شود چراکه در روش محاسبه ارائه توسط ایشان کوتاه‌ترین مسیرهایی را که بتوان از مسیرهای تجربه‌شده استدلال کرد پیدا می‌نماید.

این معیار برای هر عمل از هر حالت مقداری در نظر می‌گیرد که در پایان هر چرخه یادگیری بروز رسانی می‌شود. جهت به‌روزرسانی این جدول یک ماتریس با نام CP[1] در نظر گرفته شده است که وظیفه نگهداری مسیر طی شده در هر چرخه را بر عهده دارد. این ماتریس به تعداد حالت‌های محیط ستون دارد و سه سطر دارد. مطابق با جدول ۳ـ۱ در سلول $(1,i)$ آخرین عمل انجام‌شده در حالت i انجام می‌شود، در سلول $(2,i)$ حالت مقصدی که عامل بعدازاین حالت مشاهده کرده و در سلول $(3,i)$ شماره آخرین گام حرکت در چرخه یادگیری فعلی که عامل در حالت i بوده نگه‌داری می‌شود.

[1] Current Path

بعد از اتمام هر چرخه یادگیری زمان بروز رسانی جدول SEP با بهره‌گیری از جدول CP است. در الگوریتم ارائه شده توسط میرزایی [۲] ابتدای سلول مربوط به عملی که در آخرین چرخه یادگیری انجام‌شده مقدار ۱ می‌گیرد چراکه انجام این حرکت باعث رسیدن به نقطه هدف شده است پس فاصله این حرکت تا مقصد برابر با یک خواهد بود.بعد از آن به ترتیب معکوس گام‌های یادگیری سلول مربوط به اعمال انجام‌شده بروزرسانی میشود.جهت بروزرسانی مقدار یک خانه برابر حداقل مقدار مربوط به اعمال مقصد در همین جدول SEP به‌علاوه ۱ خواهد بود. چرا که فاصله حالت جاری تا حالت بعدی ۱ بوده و جمع این فاصله با کوتاه‌ترین فاصله در حالت مجاور کوتاه‌ترین فاصله حالت جاری را خواهد ساخت.

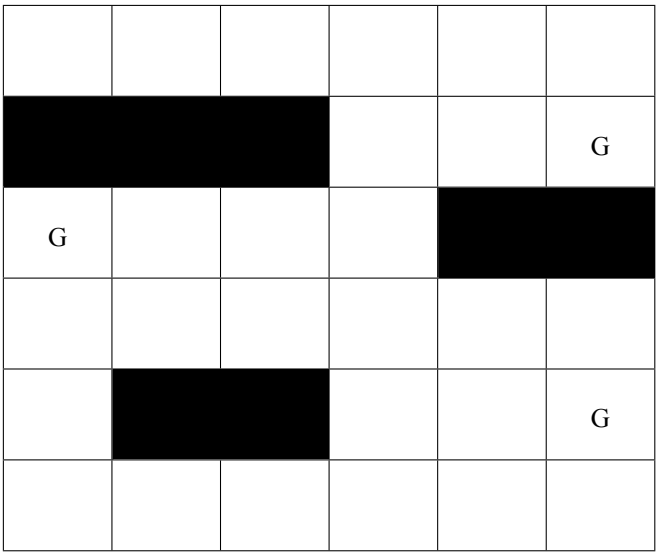
۳ـ۴ـ۲ شوک

بزرگ‌ترین دلیل رسیدن به یادگیری در یادگیری تقویتی پاداش‌های دریافتی برای انجام اعمال از محیط است. اما هرچه این اعمال به حالت پایانی نزدیک‌تر باشند باارزش‌تر هستند. میرزایی در [۲] معتقد است که پاداش دریافتی برای رسیدن به حالت‌های نهایی دارای ارزش بالاتری است و عامل حالتی که اثر بیشتری از این پاداش گرفته باشد بیشتر می‌شناسد. بر همین اساس در پارامتر شوک تعداد دفعاتی که اثر پاداش دریافتی از حالت‌های نهایی به هر خانه از جدول Q رسیده است شمارش میشود. برای شمارش این کار کافی است در شروع یادگیری یک ماتریس در ابعاد ماتریس Q با مقادیر اولیه ۰ تولید کرده و پس از انجام هر عمل در صورتی که حالت بعدی حالت پایانی باشد یا حالتی باشد که قبلاً از پاداش حالت پایانی اثر گرفته باشد مقدار مربوط به آن عمل در جدول شوک یک واحد افزایش می‌یابد.

۳ـ۵ محیط‌های آزمایش

راسل در [۲۱] محیط‌ها را با پنج دیدگاه دسته‌بندی می‌نماید.

- **مشاهده پذیر و نیمه مشاهده پذیر:** اگر عامل به کمک حسگر های خود توانایی تشخیص حالت محیط را داشته باشد محیط را مشاهده پذیر گویند. در غیر این صورت محیط نیمه مشاهده پذیر یا تا حدی قابل‌مشاهده نامیده می‌شود.
- **قطعی و غیرقطعی:** اگر بتوان حالت بعدی محیط را بر اساس سابقه اعمال و حالت فعلی مشخص کرد محیط قطعی است و در غیر این صورت محیط غیرقطعی است.
- **واقعه‌ای و غیرواقعه‌ای:** درصورتی‌که هر مرحله از مراحل دیگر مستقل باشد محیط را دوره‌ای می‌نامیم.
- **ایستا و پویا:** اگر محیط در مدت‌زمان بین درک و انتخاب عمل تغییر کند پویا و در غیر این صورت ایستا



شکل ۳ـ۲: محیط پلکان مارپیچ [۲]

است.

- **گسسته و پیوسته:** اگر مشاهدات و اعمال به شکل جداگانه تعریف شوند محیط را گسسته گویند.

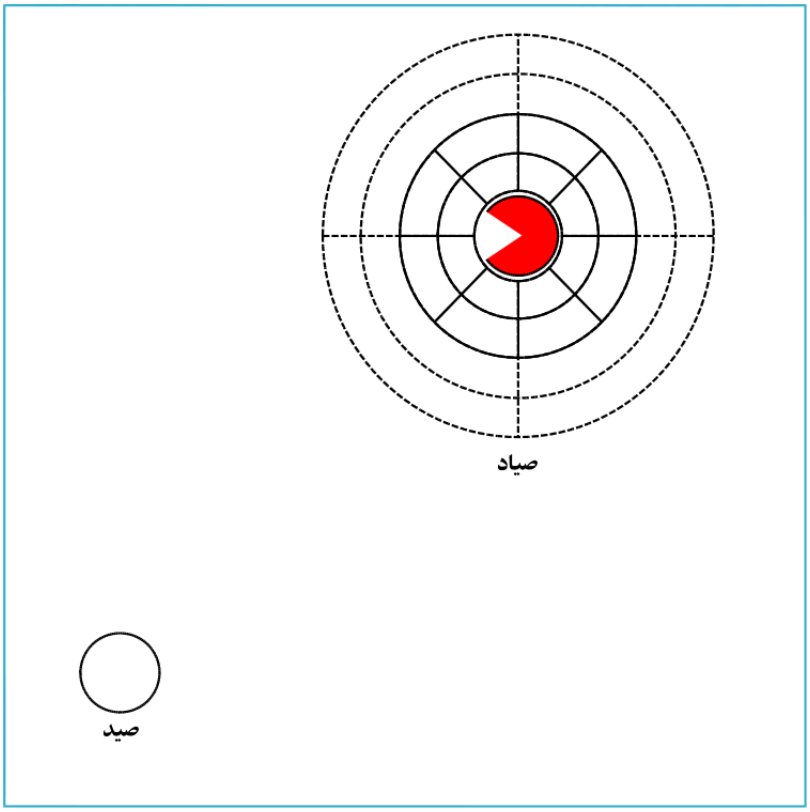
معمولا بعد از ارائه‌ی هر روش یادگیری نیاز است تا عملکرد آن روش در محیط‌های مختلف مورد ارزیابی قرار گیرد. در این پژوهش نیز دو محیط جهت این امر مورد استفاده قرار گرفته است که در ادامه تشریح میشوند.

۳ـ۵ـ۱ محیط پلکان مارپیچ

پلکان مارپیچ [۱، ۲] همان‌طور که در شکل ۳ـ۲ مشخص است یک محیط ایستا است که از یک مربع 6×6 شامل ۳ خانه هدف، تعدادی دیوار و تعدادی خانه آزاد تشکیل‌شده است. عامل در این محیط باید سیاست رسیدن از هر حالت به یک حالت هدف را با استفاده از چهار عمل اصلی بالا، پایین، چپ، راست را یادگیری نماید. درصورتی که انجام عمل توسط عامل باعث انتقال عامل به خانه هدف شود به عامل پاداش ۱۰، در صورتی که عمل انتخابی عامل را به دیوار بزند پاداش ۱ـ، در غیر این صورت معکوس فاصله حالت جاری تا نزدیک‌ترین هدف را به عنوان پاداش دریافت می‌نماید.

۳ـ۵ـ۲ محیط صید و صیاد

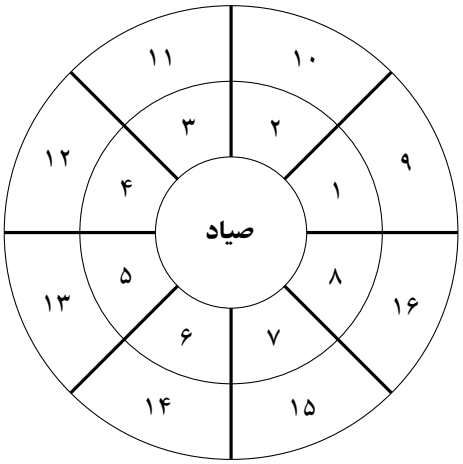
همانطور که در شکل ۳ـ۳ محیط صید و صیاد آورده شده است، محیطی است شامل یک مربع 10×10 شامل دو عامل صید و صیاد که در پیاده‌سازی‌ها معمولا عامل صید به‌صورت تصادفی حرکت کرده عامل صیاد روش شکار را یادگیری می‌نماید [۱، ۲]. این محیط برخلاف محیط پلکان مارپیچ یک محیط پیوسته است که عامل‌ها می‌توانند در هر نقطه از آن قرار گیرند. حرکت عامل‌ها در این محیط هم به‌صورت پیوسته است به این صورت که عامل صیاد می‌تواند به هر نقطه با شعاع یک و عامل صید به هر نقطه با شعاع ۰٫۵ در اطرافشان حرکت



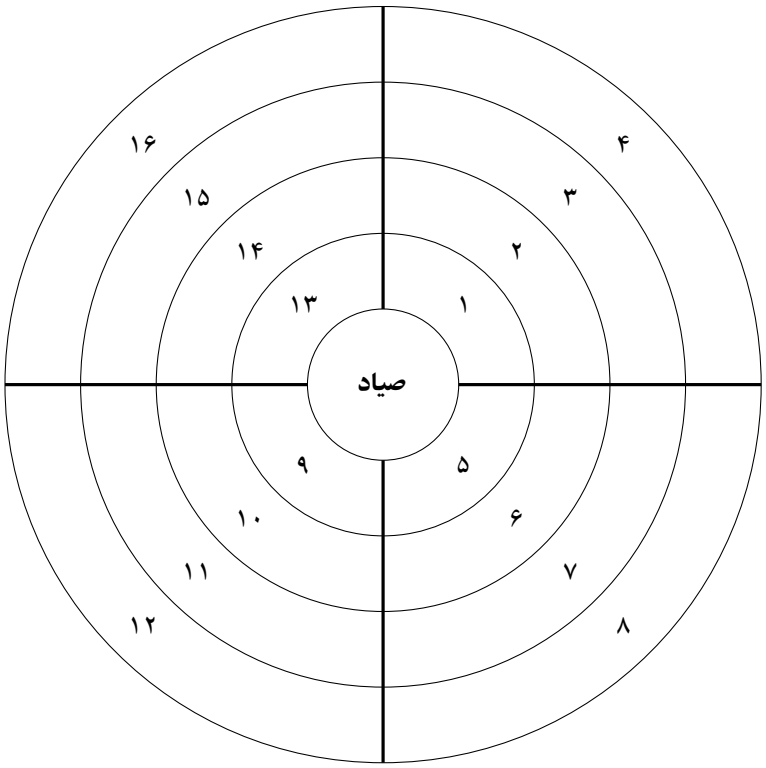
شکل ۳ـ۳: محیط صید و صیاد

نماید. اما از آنجایی که روش‌های یادگیری تقویتی نیاز به تعداد اعمال مشخص دارد باید این پیوستگی گسسته سازی شود. در اینجا برای گسسته سازی مطابق شکل ۳ـ۴(آ) زاویه حرکت به فاصله‌های ۴۵ درجه‌ای و فاصله حرکت به نیم و یک تقسیم‌شده است که جمعاً تولید ۱۶ عمل برای عامل می‌نماید.

موضوع بعد حالت‌های قرارگیری عامل است .در محیط صید و صیاد برای صیاد یک دامنه دید در نظر گرفته می‌شود که درصورتی‌که عامل صید در فاصله کمتر مساوی از دامنه دید صیاد باشد صیاد قادر به دیدن صید خواهد بود. از آنجایی که این محل قرارگیری عامل صید نسبت به صیاد یک مقدار پیوسته است نیاز به گسسته سازی دارد. در کار پیش رو دامنه دید عامل صیاد برابر ۲ در نظر گرفته‌شده است و برای گسسته سازی این دامنه دید را مطابق شکل ۳ـ۴(ب) به زوایای ۹۰ درجه و فاصله‌ی به اندازه‌های ۰/۵ تقسیم می‌نماییم. پس تعداد حالت‌ها زمانی که صید در دامنه دید قرارگرفته باشد برابر با ۱۶ حالت خواهد بود این شانزده به‌علاوه یک حالت که عامل در دامنه دید نباشد ۱۷ حالت را برای سیستم به وجود میاورد. همان طور که گفته شد حرکت‌های صید نیز در این سیستم به‌صورت تصادفی انجام میشود. عامل صیاد در صورتی که با انجام یک عمل صید را شکار کند پاداش ۱۰ و در غیر این صورت پاداش ۰/۱ ـ دریافت می‌نماید.

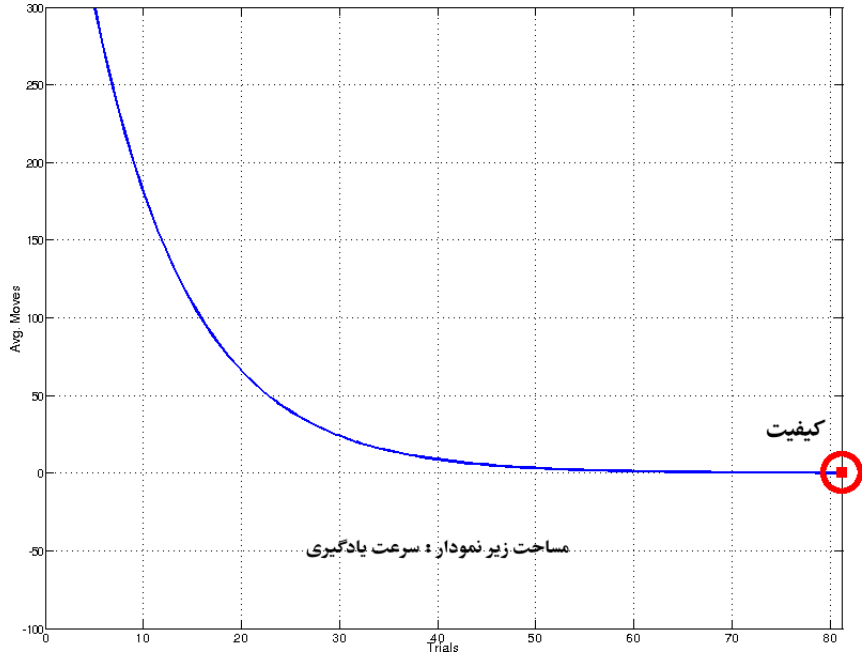


(آ) ۱۶ عمل تعریف شده برای صیاد



(ب) دامنه دید تعریف شده برای صیاد

شکل ۳ـ۴: دامنه‌ی دید و حالت تعریف شده برای عامل صیاد در محیط صید و صیاد [۲].



شکل ۳ـ۵: سرعت و کیفیت یادگیری از معیارهای ارزیابی و مقایسه‌ی عملکرد الگوریتم‌های یادگیری تقویتی می‌باشد [۲].

۳ـ۶ معیارهای ارزیابی

ارزیابی هر سیستمی نیاز به معیارهایی دارد، که در یادگیری مشارکتی نیز دو معیار «سرعت و دقت یادگیری» وجود دارد که بر اساس میانگین تعداد قدم‌ها در هر چرخه یادگیری محاسبه می‌شوند [۱، ۲]. در صورتی که تعداد قدم‌های چرخه‌های یادگیری نگه‌داری و به صورتی نموداری که در محور عمودی میانگین تعداد قدم‌ها و در محور افقی چرخه یادگیری باشد رسم شود، می‌توان آخرین نقطه از محور را که نشانه میانگین تعداد چرخه‌ها در آخرین چرخه یادگیری است را به‌عنوان کیفیت و مساحت زیر این نمودار را به‌عنوان سرعت یادگیری در نظر گرفت که هدف کمینه کردن این معیارها است؛ در شکل ۳ـ۵ این معیارها نمایش داده‌شده است.

۳ـ۷ اندازه‌گیری و انتگرال فازی

برای درک روش پیشنهادی نیاز به داشتن اطلاعات پایه در مورد اندازه‌گیری‌های فازی[1] و انتگرال فازی که با هدف جمع‌آوری اطلاعات[2] ارائه شده‌اند، داریم. اندازه‌گیری‌های فازی پیش‌زمینه‌ای بر انتگرال‌های فازی هستند که قبل از آشنایی با انتگرال‌های فازی نیاز به معرفی اندازه‌گیری‌های فازی داریم. اگر فرض کنیم که تعدای منبع اطلاعاتی $X = \{x_1, x_2, \cdots, x_n\}$ که این منابع اطلاعاتی اطلاعات دریافتی از حسگرها، پاسخ‌های داده شده به یک پرسشنامه و غیره وجود داشته باشند. اندازه‌گیری فازی میزان ارزش اطلاعاتی این منابع را در اختیار ما می‌گذارد. معمولا اندازه‌گیری فازی توسط تابع $g : 2^{|X|} \to [0, 1]$ تعریف می‌شود که ورودی آن یک

[1] Fuzzy measures

[2] Aggregate Information

زیرمجموعه‌ای از منابع اطلاعاتی می‌باشد و خروجی آن یک مقدار مابین صفر و یک که میزان ارزش اطلاعاتی که آن زیرمجموعه از منابع اطلاعاتی ورودی تابع را مشخص می‌کند. این تابع باید دارای شرایط مرزی تعریف شده و یکنواختی باشد که در ادامه به معرفی شرایط می‌پردازیم [۲۲]:

۱. **شرایط مرزی:** اگر اطلاعاتی در دست نداریم ارزش صفر را دارد و کلیه اطلاعات ارزش ۱ را دارد.

$$g(\emptyset) = 0, \quad g(X) = 1 \tag{۲-۳}$$

۲. **یکنواختی ـ غیر کاهشی:** اگر اطلاعات بیشتری به دست آمد ارزش کلیه اطلاعات که شامل اطلاعات جدید می‌باشد حداقل به اندازه زمانی است که آن اطلاعات جدید بدست نیامده است.

$$A \subseteq B \subseteq X \Rightarrow g(A) \leq g(B) \leq 1 \tag{۳-۳}$$

مقادیر تابع g یا توسط کارشناس ارائه می‌شود یا توسط یک تابعی مدل می‌شود، یکی از توابع معروف برای تخمین مقادیر تابع g تابع اندازه‌گیری‌ـλ سوگنو[1] می‌باشد که به صورت زیر تعریف می‌شود [۲۳].

$$g(\{x_1, \cdots, x_l\}) = \frac{1}{\lambda}\left[\prod_{i=1}^{l}(1+\lambda g_i) - 1\right] \tag{۴-۳}$$

که در معادله ۳ـ۴ مقدار g_iها مقادیر ارزش هریک از منابع اطلاعاتی است و λ بگونه‌ای تعیین می‌گردد که $g(X) = 1$ شود که این مقدار برابر با جواب معادله‌ی زیر باشد.

$$\lambda + 1 = \prod_{i=1}^{n}(1+\lambda g_i), \quad \lambda \in (-1, \infty) \tag{۵-۳}$$

نکته‌ای که در رابطه با تابع اندازه‌گیری‌ـλ سوگنو باید توجه کرد این است که به ازای مقادیر n مختلف باید ریشه‌یابی بروی متغیر λ صورت گیرد؛ این ویژگی باعث می‌شود که این تابع در بعضی از کاربردها کارایی نداشته باشد. به عنوان مثال در کار این پژوهش از آنجایی که تعداد عامل‌ها متفاوت می‌باشد، در صورت استفاده از تابع اندازه‌گیری‌ـλ باید به ازای هر دفعه تغییر در تعداد عامل‌ها (از آنجایی که مقادیر n تغییر می‌کند) یکبار روی متغیر λ ریشه‌یابی صورت بگیرد که امکان همچین ریشه‌یابی‌ای بدون سربار محاسباتی سنگین میسر نیست. لذا در کاربردهایی که تعداد n متغیر می‌باشد استفاده از تابع اندازه‌گیری‌ـλ عاقلانه نیست.

[1] Sugeno λ-Measure

انتگرال فازی در واقع یک تعمیمی به روش میانگین وزنی[1] می‌باشد بطوری که نه تنها مشخصه‌های مهم تک تک ویژگی‌ها را در نظر می‌گیرد بلکه اطلاعات تعاملات بین ویژگی‌ها را نیز در نظر می‌گیرید [۲۴]. از میان انتگرال‌های فازی دو انتگرال سوگنو[2] و چوکت[3] از الگوریتم‌هایی هستند که می‌توانند بر روی هر اندازه‌گیری فازی مورد استفاده واقع شوند [۲۵]. فرض کنیم که تابعی چون $h : X \to [0,1]$ وجود دارد که مقادیر منابع اطلاعاتی را به بازه‌ی $[0,1]$ نگاشت می‌کند. در واقع h تابع پشتیبان[4] منابع اطلاعاتی می‌باشد. انتگرال فازی سوگنو به صورت ۳ـ۶ تا ۳ـ۸ تعریف می‌شود [۲۵، ۲۶]:

$$\int_s h \circ g = \mathcal{S}_g(h) = \bigvee_{i=1}^{n} h(x_{\pi_i^s}) \wedge g(A_i^s) \qquad (۳ـ۶)$$

$$h \xrightarrow{\pi^s} h(x_{\pi_1^s}) \leq h(x_{\pi_2^s}) \leq \cdots \leq h(x_{\pi_n^s}) \qquad (۳ـ۷)$$

$$A_i^s = \{x_{\pi_i}^s, x_{\pi_{i+1}}^s, \cdots, x_{\pi_n}^s\} \qquad (۳ـ۸)$$

در روابط ۳ـ۶ تا ۳ـ۸ نماد s بیان‌گر «سوگنو» می‌باشد. در انتگرال سوگنو لازم است که مقادیر منابع اطلاعاتی را مرتب کنیم که π^s عملگر جایگشت انتگرال فازی سوگنو می‌باشد که خروجی مقادیر تابع h را به ترتیب صعودی مرتب می‌کند. نمادهای $\vee$ و $\wedge$ به ترتیب عمگرهای $\max$ و $\min$ می‌باشد. در این انتگرال ابتدا مقادیر دریافتی از منابع اطلاعاتی به تابع پشتیبان h ارسال می‌شود و سپس مقادیر خروجی تابع پشتیبان به ازای همه‌ی اطلاعات دریافتی را به صورت صعودی توسط عملگر جایگشت π^s مرتب می‌شود. مجموعه‌ی A_i^s اندیس عناصر مرتب شده مقادیر تابع پشتیبان از اندیس iام تا اندیس n می‌باشد. سپس طبق آنچه که در ۳ـ۶ آمده است از کوچکترین مقدار $h(x_{\pi_1^s})$ شروع می‌کنیم با $g(A_1^s)$ کمینه‌گیری می‌کنیم و سپس می‌رویم به دومین کوچک‌ترین عنصر $h(x_{\pi_2^s})$ و همین کار را تا آخرین (بزرگترین) عنصر تکرار می‌کنیم و سپس یک بیشینه‌گیری روی این مقادیر انجام می‌دهیم که خروجی انتگرال سوگنو می‌شود.

انتگرال فازی چوکت به صورت ۳ـ۹ تعریف می‌شود [۲۵، ۲۷]. در این رابطه $f : X \to \mathbb{R}$ می‌باشد که از وجه تمایز انتگرال فازی چوکت با سوگنو می‌باشد و π^c عملگر جایگشت انتگرال فازی چوکت می‌باشد.

[1] Weighted Arithmetic Mean
[2] Sugeno
[3] Choquet
[4] Support

$$\int_c f \circ g = \mathcal{C}_g(f) = \sum_{i=1}^{n} \left(f(x_{\pi^c_{(i)}}) - f(x_{\pi^c_{(i-1)}}) \right) \cdot g(A^c_i) \qquad (۹\text{ـ}۳)$$

$$f \xrightarrow{\pi^c} f(x_{\pi^c_1}) \leq f(x_{\pi^c_2}) \leq \cdots \leq f(x_{\pi^c_n}) \qquad (۱۰\text{ـ}۳)$$

$$A^c_i = \{x^c_{\pi_i}, x^c_{\pi_{i+1}}, \cdots, x^c_{\pi_n}\} \qquad (۱۱\text{ـ}۳)$$

$$f(x_{\pi^c_0}) = 0 \qquad (۱۲\text{ـ}۳)$$

در روابط ۳ـ۹ تا ۳ـ۱۱ نماد c بیان‌گر «چوکت» می‌باشد. عمکرد انتگرال چوکت شباهت نزدیکی با انتگرال سوگنو دارد به این صورت که در انتگرال چوکت مقادیر دریافتی از منابع اطلاعاتی را به تابع پشتیبان $f(\cdot)$ ارسال می‌شود و خروجی این تابع را به ازای تمامی ورودی‌ها توسط عملگر جایگشت π^c به صورت صعودی مرتب می‌کنیم. تعریف A^c_i در ۳ـ۱۱ مشابه انتگرال سوگنو می‌باشد. طبق تعریف انجام شده در ۳ـ۱۲ مقدار ۰ را به اول مقادیر مرتب شده $f(\cdot)$ در ۳ـ۱۰ اضافه می‌کنیم. سپس توسط رابطه‌ی ۳ـ۹ مجموع ضرب اختلاف دو عنصر متوالی مرتب شده $f(\cdot)$ در $g(A^c_i)$ را به عنوان خروجی انتگرال چوکت حساب می‌کنیم.

انتگرال‌های فازی سوگنو و چوکت در حالت کلی دارای تفاوت‌هایی هستند که از جمله‌ی مهم‌ترین این ویژگی‌ها تفاوت تعریف توابع h و f در این انتگرال‌ها می‌باشد که باعث می‌شود انتگرال چوکت برای مسائلی که مقادیر اعداد حائز اهمیت است، مناسب باشد و از طرف دیگر انتگرال سوگنو زمانی مطلوب است که تنها ترتیب اعداد مد نظر باشد [۲۸]. به همین علت در این پژوهش انتگرال فازی چوکت مورد استفاده قرار گرفته است زیرا که ورودی انتگرال اعداد کاملا معنی‌دار می‌باشد و اعمال تابع h بروی مقادیر منابع اطلاعاتی، معانی آن‌ها را تغییر داده و اطلاعات نامطلوبی تولید خواهد کرد.

۳ـ۸ نتیجه‌گیری

در این فصل مطالبی که برای درک بهتر روش پیشنهادی نیاز است شرح داده شده است. در روش پیشنهادی از یادگیری Q با سیاست‌های انتخاب عمل بولترمن، εـحریصانه و انتگرال فازی چوکت استفاده شده است. در ادامه، در فصل بعدی روش پیشنهادی توضیح داده شده است.

فصل چهارم

روش پیشنهادی

۴ـ۱ مقدمه

در این فصل جزییات روش پیشنهادی به طور مفصل معرفی خواهد شد، روش ارائه شده در حالت کلی از دو قسمت تشکیل شده است؛ اولین و مهم‌ترین قسمت ارائه یک معیار خبرگی جدید به نام معیار خبرگی «ارجاع» که برای هر عامل در هر چرخه یادگیری محاسبه و در یک «ماتریس ارجاع» نگه‌داری می‌شود. دومین قسمت مربوط به ترکیب دانش‌های عامل‌ها هستند که با استفاده از یک مدل انتگرال فازی، صورت می‌گیرد. همانطور که در فصل بعدی نیز نشان داده خواهده شد استفاده از مدل انتگرال فازی به دلیل خواصی مهمی که این مدل دارد باعث می‌شود سرعت و کیفیت یادگیری به طرز چشم‌گیری افزایش یابد. در این فصل ابتدا به معرفی معیار «ارجاع» و دلیل استفاده از آن می‌پردازیم سپس یادگیری مشارکتی چندعامله با استفاده از ماتریس ارجاع و انتگرال فازی معرفی خواهد شد و در نهایت نشان داده خواهد شد که استفاده از انتگرال فازی می‌تواند نتایج بهتری را نسبت به مدل های سنتی چون مدل مجموع وزنی[1] ارائه دهد.

[1] Weighted Sum

۴ـ۲ معیار خبرگی ـ ماتریس ارجاع و خاطره

در دنیای واقعی «خبرگی» تعاریف متعددی به خود گرفته است، در روانشناسی خبرگی به معنی عملکرد برتر عامل تلقی می‌شود. در جامعه شناسی خبره به فردی گفته می‌شود که برچسب خبرگی توسط یک گروهی به فرد زده شده است و آن گروه به توانایی که آن فرد در اختیار دارد علاقه‌مند[1] است. در فلسفه خبره به فردی گفته می‌شود که دانشی که فرد تازه‌کار در اختیار ندارد را دارا می‌باشد [۲۹]. اگر تعاریف مختلف «خبرگی» را بررسی کنیم می‌بینیم که همه‌ی تعاریف در واقع تعبیری از میزان کیفیت عمکرد عامل نسبت به دیگر عامل‌ها می‌باشد. این تعبیر کلی از «خبرگی» انگیزه‌ای شد که درصدد معرفی معیاری برآییم که در حالت کلی بتوان به کلیه‌ی تعاریف «خبرگی» قابل تعمیم باشد.

فرضیه ۴ـ۱ (خبرگی)**.** *فرض می‌کنیم عامل $\mathcal{A}$ در محیط $\mathcal{E}$ در پی رسیدن به یک مجموعه اهداف $\mathcal{G} \subseteq \{g_1, g_2, \cdots, g_n\}$ می‌باشد. میزان خبرگی عامل رابطه‌ی معکوسی با میزان تلاش عامل برای رسیدن به اهداف تعریف شده خود دارد.*

طبق آنچه که در فرضیه بالا آورده شده است از بین چند عاملی که در یک محیط و یک مجموعه از اهداف فعالیت می‌کنند، عاملی خبره‌تر است که تلاش کمتری برای رسیدن به آن مجموعه اهداف می‌کند. شاید این مساله در نگاه اول نامتعارف به ذهن برسد ولی در فعالیت‌های روزمره ما انسان‌ها نیز به کررات شاهد این امر می‌باشیم. به عنوان مثال رانندگی دو فرد مبتدی و حرفه‌ای را در نظر بگیریم؛ فرد مبتدی هنگام رانندگی تمام حواس خود را معطوف به رانندگی می‌کند تلاش بسیار زیادی برای کنترل نسبت میزان کلاچ و گاز می‌کند و هنگام رانندگی به طور طبیعی رانندگی نمی‌کند و ... ولی فرد خبره کلیه موارد ذکر شده را بطور خودکار و طبیعی انجام می‌دهد بطوری که انگار رانندگی مانند دیگر رفتارهای طبیعی وی چون نفس کشیدن می‌باشد، که بصورت خودکار صورت می‌پذیرد. از این گونه مثال‌ها از کاربرد فرضیه ۴ـ۱ در زندگی روزمره ما زیاد می‌توان یافت.

توجه شود که در فرضیه ۴ـ۱ عبارت «میزان تلاش» عامل می‌تواند در کاربردهای مختلف تعابیر مختلفی به خود بگیرد، مثلا در مثال راننده‌ی مبتدی و خبره میزان نسبت مسافت طی شده بر زمان رانندگی را می‌توان به عنوان «میزان تلاش» عامل در نظر گرفت که در شرایط یکسان راننده‌ی خبره‌تر به طور نسبی در زمان کوتاه‌تری یک مسافت مشخصی را طی خواهد کرد (در رد کردن پیچ و خم‌های ترافیک و مدت زمان ترمز و ... زمان کمتری را تلف می‌کند). یا به عنوان مثال دیگر، دانشجوی قوی و دانشجوی ضعیف را مورد بررسی قرار دهیم، دانشجویی خبره هست که زمان کمتری را صرف حل صحیح یک مساله خاص کند (با فرض اینکه دانشجوها حتما باید مساله را حل کنند). همانطور که دیدیم کمیت «میزان تلاش» عامل برای مسائل مختلف معیار متفاوتی را دربر می‌گیرد ولی همگی از همان اصل معرفی شده در فرضیه ۴ـ۱ تبعیت می‌کنند.

در یادگیری مشارکتی با استفاده از فرضیه ۴ـ۱ می‌توان با تعریف ۴ـ۱ یک معیار خبرگی جدید را معرفی

[1] Interested

کرد که مبنا و پایه‌ی دستآوردهای این پژوهش می‌باشد.

تعریف ۴ـ۱ (معیار خبرگی «میزان ارجاع»)**.** فرض می‌کنیم مجموعه‌ای از عامل‌ها $\mathbb{A} = \{\mathcal{A}_1, \mathcal{A}_2, \cdots, \mathcal{A}_m\}$ در محیط $\mathcal{E}$ در پی رسیدن به یک مجموعه اهداف $\mathcal{G} \subseteq \{g_1, g_2, \cdots, g_n\}$ می‌باشند. اگر ما به طور مجازی و دلخواه محیط $\mathcal{E}$ را به k ناحیه مانند e_i افراز کنیم بطوری که $\mathcal{E} = \{e_i \mid \cup_{i=1}^{k} e_i = \mathcal{E} \wedge \forall i, j \in \{1, 2, \cdots, k\} \wedge i \neq j : e_i \cap e_j = \varnothing\}$ میزان ارجاع هر عامل در هر ناحیه را میزان حضور آن عامل را در آن ناحیه تعریف می‌کنیم.

در تشریح آنچه که در تعریف ۴ـ۱ آمده است می‌توان گفت که در سیستم‌های چندعاملی که همگی عوامل در یک محیط به صورت مستقل در حال فعالیت هستند؛ محیط را به چند ناحیه دلخواه افراز می‌کنیم که اجتماع نواحی باهم کل محیط $\mathcal{E}$ را تشکیل دهند و هیچ دو ناحیه‌ای اشتراکی باهم نداشته باشند [۳۰]. در این چنین افرازی از محیط، در هرناحیه عاملی که نسبت به بقیه خبره‌تر است، نسبت به بقیه عوامل در همان ناحیه میزان تمایل حضور کمتری را از خود نشان می‌دهند. به عبارت دیگر عاملی که خبره‌تر است تمایل دارد کوتاه‌ترین مسیر رسیدن به اهداف خود را طی کند که نهایتا منجر خواهد شد که میزان حضور عامل در هریک از نواحی محیط کمینه شود.

آنچه که در فرضیه ۴ـ۱ در مورد «میزان تلاش» عامل آمده است در تعریف ۴ـ۱ در به صورت «میزان حضور عامل در هر ناحیه» تعریف شده است. بطوری که طبق فرضیه مطرح شده میزان خبرگی عامل در هر ناحیه رابطه‌ی معکوسی با میزان حضور عامل در همان ناحیه را دارد. زیرا اگر عامل نسبت به محیط خود شناخت کامل‌تری داشته در هنگام تلاش برای رسیدن به اهداف خود به علت شناخت خوبی که از محیط دارد کمتر در محیط پرسه می‌زند (کمتر تلاش می‌کند) و با تعداد گام کمتری به سمت اهداف خود حرکت می‌کند – در واقع مسیر بهتری/کوتاه‌تری برای رسیدن به هدف را می‌شناسد. این موضوع در نهایت منجر می‌شود که عاملی که در هر ناحیه خبره‌تر است در همان ناحیه میزان پرسه زدن (حضور/تلاش) کمتری نسبت به دیگر عامل‌ها که از خبرگی نسبی کمتری برخوردار است را داشته باشد.

معیار تعریف شده در تعریف ۴ـ۱ قبلا به صورت جزیی توسط احمدآبادی و همکاران [۱۲] ارائه شده است ولی معیار تعریف شده در این پژوهش تفاوت‌هایی با معیار احمدآبادی و همکاران دارد که به شرح زیر است:

۱. **میانگین تعداد قدم‌ها:** احمدآبادی و همکاران میانگین تعداد قدم‌های رسیدن به هدف(یا طبق تعریف ۴ـ۱ میانگین میزان ارجاع عامل در کل محیط – در زمانی که کل محیط را یک ناحیه در نظر بگیریم) را به عنوان معیار خبرگی در نظر گرفته‌اند در حالی که در تعریف ۴ـ۱ حرفی از میانگین آورده نشده است. ایرادی که معیار احمدآبادی و همکاران دارد این است که هنگامی که می‌خواهیم خبرگی عامل‌ها را بسنجیم صحیح نیست میانگین تعداد گام‌ها در نظر بگیرم زیرا ممکن است عامل در ابتدا بسیار نادان

بوده ولی بعد از طی مدتی به وسیله‌ی تجاربی خاص به عاملی بسیار دانا تبدیل شود و اگر میانگین‌گیری صورت گیرد آنگاه نادانی گذشته به میزان خبرگی کنونی تاثیر گذاشته و خبرگی عامل کمتر از میزان واقعی تخمین زده شود. در تعریف ۴ـ۱ خبرگی کنونی عامل مورد نظر است و کاری با مسیری که عامل برای کسب خبرگی کنونی‌اش طی کرده است نداریم.

۲. **انعطاف:** معیار احمدآبادی و همکاران از انعطاف برخوردار نیست و در خبرگی عامل‌ها را بصورت میانگین خبرگی در کل محیط محاسبه می‌کند در حالی که طبق تعریف ۴ـ۱ خبرگی عامل در نواحی مختلف از محیط قابل محاسبه است و همانطور که بعدها خواهیم دید خبرگی عامل‌ها در هر ناحیه به عنوان معیاری برای ترکیب دانش عامل‌ها نسبت به آن ناحیه مورد استفاده واقع خواهد؛ زیرا که عاملی ممکن است در حالت کلی محیط را آنچنان نشناخته باشد ولی در یک یا چند ناحیه بخصوص این عامل شناخت کامل‌تری از آن نواحی داشته باشد که معیار احمدآبادی و همکاران نمی‌تواند این مساله را در نظر بگیرد.

تا به اینجا گفته شد که عاملی که از خبرگی بیشتری برخوردار است لزوما کمتر در محیط پرسه می‌زند و با طی کردن مسیر کوتاه‌تر به سمت اهداف خود، تلاش کمتری می‌کند ولی چند سوال در اینجا مطرح می‌شود که برای حل مساله نیازمند پاسخ به آن‌ها هستیم.

۱. میزان حضور عامل در نواحی مختلف، در محیطی که از dـبعد تشکیل شده است چگونه مدل شود؟

۲. اگر عاملی که در هر چرخه یادگیری به یکی از نواحی کلا وارد نشد و میزان پرسه زدن عامل در آن ناحیه صفر شود؛ آیا این مقدار کمینه پرسه زدن، نشان دهنده‌ی خبرگی عامل در آن ناحیه است؟

۳. چگونه در معیار خبرگی ارائه شده باید مساله عدم حضور عامل در یکی از نواحی را مدل کرد، بگونه‌ای که اثر سوئی بر تجربه‌ی دیگر عامل‌ها در آن نواحی، در هنگام ترکیب دانش عامل‌ها نداشته باشد؟

پاسخ به این سوالات برای حل مساله با استفاده از معیار خبرگی پیشنهادی (تعریف ۴ـ۱) ضروری است. در پاسخ به سوال اول، ما به ازای کلیه‌ی نواحی یک ماتریسی به نام «ماتریس ارجاع» (یا به اختصار REFMAT[1]) در نظر میگیریم که در ابتدا صفر مقداردهی شده‌اند و هر دفعه که عامل از حالتی به حالت دیگر می‌رود مقدار آن ناحیه‌ای که حالت جدید در آن واقع است یک واحد افزایش می‌یابد بدین وسیله میزان حضور عامل در نواحی مختلف را می‌شماریم. همانطور که در قسمت آزمایش‌های این پایان‌نامه نشان داده شده است میزان ریز یا درشت بودن این نواحی در کیفیت نتیجه تاثیرگذار نیست! یعنی عملا چه ما در حالت کلی، کل محیط را به عنوان یک ناحیه در نظر بگیریم و میزان حضور عامل در این ناحیه را بشماریم (که معادل می‌شود با تعداد گام‌های عامل در طی رسیدن به هدف) یا در حالت جزئی به ازای هر حالت موجود را یک ناحیه در نظر بگیریم (که معادل می‌شود

[1] Reference Matrix

با تعداد ملاقات هر یکی از موقعیت‌ها توسط عامل) به یک نتیجه می‌رسیم. با توجه به دلیل فوق‌الذکر در پاسخ به سوال دوم، اگر تعداد نواحی زیاد باشد (مثلا هر موقعیت یک ناحیه باشد – حداکثر تعداد نواحی) ممکن است عامل در طی رسیدن به هدف برخی از نواحی را کلا ملاقات نکند و مقدار ارجاع به آن نواحی صفر شود و از طرفی طبق تعریف ۴ـ۱ عاملی که تعداد حضور کمتری در نواحی مختلف داشته باشد از خبرگی بیشتری در آن نواحی برخوردار است و در این شرایط که مقدار ارجاع عامل به ناحیه‌ای صفر باشد را نمی‌توان به خبرگی عامل در آن ناحیه نسبت داد زیرا که آن عامل در کل، آن ناحیه را ملاقات نکرده است که بخواهد تجربه‌ای را در تعامل با آن ناحیه کسب کند تا بتواند خبرگی خود را در آن ناحیه افزایش دهد. برای حل این مشکل و پاسخ به سوال سوم، ماتریسی جدیدی به نام ماتریس خاطره (یا به اختصار RCMAT[1]) را معرفی می‌کنیم. این ماتریس وظیفه‌ی نگه‌داری آخرین تعداد ارجاعات عامل به هرکدام از نواحی تعریف شده را دارد و در زمان‌هایی که مقدار یک ناحیه در ماتریس REFMAT صفر باشد مقدار آن ناحیه از ماتریس RCMAT بروز رسانی می‌شود که میزان پرسه زدن عامل در آن ناحیه در آخرین باری عامل آن ناحیه را ملاقات کرده است را نشان می‌دهد؛ در صورتی که مقدار پرسه زدن یک ناحیه در ماتریس REFMAT مقداری غیر صفر باشد مقدار ماتریس RCMAT با مقدار کنونی REFMAT آن ناحیه بروز رسانی می‌شود.

دلیل استفاده از ماتریس RCMAT این است که در یادگیری تقویتی عامل زمانی می‌توان دانش (سیاست/خبرگی) خود را نسبت به نحوه‌ی عمل در یک موقعیت بهبود ببخشد که آن موقعیت را ملاقات کند. حال اگر عامل موقعیتی را ملاقات نکند دانش وی در آن موقعیت ثابت خواهد ماند به همین دلیل اگر عامل ناحیه‌ای را ملاقات نکند و مقدار REFMAT آن ناحیه صفر باشد می‌دانیم که دانش (خبرگی) عامل در آن ناحیه در این چرخه‌ی یادگیری ثابت مانده است و در صورتی که دوباره در آن ناحیه قرار می‌گرفت، حدودا به همان میزان آخرین ملاقات در آن محیط پرسه خواهد زد. به عبارت دیگر در یک چرخه یادگیری ناحیه‌ی ملاقات نشده، تقریبا به میزان آخرین تعداد ارجاع شده برای آن ناحیه، مورد ارجاع واقع می‌شد.

۴ـ۳ یادگیری مشارکتی Q با استفاده از ماتریس ارجاع و انتگرال فازی

آنچه که تا به اکنون در مورد روش پیشنهادی این پژوهش آورده شده، معرفی یک معیار خبرگی است، که در تمامی موقعیت‌های دنیای واقعی به وفور مشاهده می‌شود و آن ارائه این فرضیه است که عامل خبره‌تر برای رسیدن به یک مجموعه از اهداف تلاش نسبی کمتری نسبت به دیگر عامل‌ها با خبرگی کمتر در شرایط یکسان می‌کند. حال که معیاری برای میزان خبرگی عامل‌ها در اختیار داریم چالش بعدی برای بهبود کیفیت و سرعت یادگیری مشارکتی ارائه‌ی روشی برای ترکیب دانش‌های عامل‌ها از محیط (جداول Q آن‌ها) با استفاده از معیار ارائه شده

[1] Recall Matrix

می‌باشد. روش ترکیب باید بگونه‌ای باشد که کیفیت و سرعت یادگیری مشارکتی عامل‌ها را در طی زمان نسبت به زمانی که عامل‌ها بدون مشارکت یاد می‌گیرند بهتر کند. همچنین کیفیت و سرعت یادگیری همبستگی مستقیمی داشته باشند با تعداد عامل‌هایی که درحال اشتراک گذاری هستند؛ به عبارت دیگر در صورت افزایش تعداد عامل‌هایی که دانش‌های خود را به اشتراک می‌گذارند مدل ترکیب کننده‌ی دانش‌های آن عامل‌ها باید بتواند دانش بهتری تولید کند که نهایتا منجر به بهتر شدن کیفیت و سرعت کلی یادگیری عامل‌ها شود.

در این پژوهش ما انتگرال فازی را به عنوان مدل ترکیب کننده‌ی دانش‌های عامل‌ها پیشنهاد می‌دهیم. دلیل انتخاب این مدل ویژگی‌های منحصر به فردی است که این مدل کننده در اختیار دارد که مدل را کاملا مناسب برای ترکیب دانش عامل‌ها می‌کند. که در بخش‌های آتی فصل این ویژگی‌ها و دلایل مناسب بودن آن‌ها برای ترکیب دانش عامل‌ها آورده شده است. لازم به یادآوری است که همانطور که در قسمت ۳ـ۷ این پایان‌نامه آورده شده است ما به دلایل فنی از انتگرال فازی چوکت استفاده می‌کنیم که در بخش های بعدی این دلایل نیز بطور مفصل شرح داده می‌شود.

۴ـ۳ـ۱ الگوریتم پیشنهادی

در این قسمت به معرفی الگوریتم پیشنهادی می‌پردازیم. آنچه که در الگوریتم ۴-۱ آمده است از دو قسمت تشکیل شده است، یک قسمت مربوط به یادگیری مستقل (خطوط ۸ تا ۱۳) و قسمت دیگر مربوط به یادگیری مشارکتی (خطوط ۱۵ تا ۲۳) می‌باشد. ورودی الگوریتم تعداد عامل‌ها می‌باشد و در ابتدا ماتریس‌های Q و REFMAT و RCMAT مقداردهی می‌شود. سپس تا زمانی که یادگیری پایان نیافته است ابتدا عامل‌ها در قسمت یادگیری مستقل به صورت جدا گانه در محیط فعالیت می‌کنند که رویه‌های آورده شده در خطوط ۸ تا ۱۲ همان الگوریتم یادگیری Q متعارف می‌باشد [۳۱]. در قسمت یادگیری مستقل تنها خط ۱۳ می‌باشد که در روش پیشنهادی به شبه‌کد اضافه شده است و این تنها یک وظیفه‌ی بسیار ساده را انجام می‌دهد و آن شمارش میزان حضور عامل در هر کدام از نواحی از پیش تعیین شده است؛ $\phi(\cdot)$ یک تابع نگاشت از یک موقعیت به یک ناحیه از محیط می‌باشد.

بعد از طی یادگیری مستقل عامل‌ها به قسمت اشتراک گذاری دانش‌های خود(جداول Q) می‌رسند (خطوط ۱۵ تا ۲۳). در قسمت یادگیری مشترک ابتدا طبق آنچه که در در بخش قبلی آورده شده است جداول REFMAT و RCMAT به صورت مشترک بروزرسانی می‌شود و سپس جداول Q و REFMAT تمامی عامل‌ها به مدل ترکیب کننده فازی معرفی شده در این پژوهش فرستاده می‌شود و مدل ترکیب کننده فازی وظیفه‌ی استخراج یک دانش جدید با در نظر گرفتن ورودی‌های آن برای جایگزینی دانش قابلی عامل‌ها را دارد. لازم به ذکر است که الگوریتم ۴-۱ به صورت یک الگوریتم غیرـمتمرکز[1] بروی هرعامل می‌باشد [۲، ۱۲، ۱۹].

[1] Decentralized

الگوریتم ۴-۱ الگوریتم پیشنهادی یادگیری مشارکتی برمبنای ماتریس REFMAT و انتگرال فازی

```
1: procedure REFMAT-COOPERATIVE-LEARNING(m)
2:     Require: m > 1                                   ▷ The number of agents.
3:     Ensure: Intialize the Q matrix;
4:     Ensure: Intialize the RCMAT ← 0;
5:     Ensure: Intialize the REFMAT ← 0;
6:     while not End Of Learning do
7:         if In individual learning mode then
8:             Visit the state s;
9:             Select an action a based on an action selection policy;
10:            Carry out the a and observe a reward r at the new state s';
11:            Q[s,a] ← Q[s,a] + α(r + λ max_{a'}(Q[s',a']) − Q[s,a]);
12:            s ← s';
13:            Increment REFMAT(φ(s)) by one;
14:        else if In cooperative learning mode then
15:            K⃗ ← {};
16:            R⃗ ← {};
17:            for each agent i ← 1, m do
18:                REFMAT_i, RCMAT_i ← Conditional_Swap(REFMAT_i, RCMAT_i);
19:                K⃗.add(Q_i);
20:                R⃗.add(REFMAT_i);
21:            end for
22:            Q ← FCI_Combiner(K⃗, R⃗);
23:            REFMAT ← 0;
24:        end if
25:    end while
26: end procedure
```

الگوریتم تابع Conditional_Swap(·) بسیار ساده می‌باشد و مقادیر غیر صفر ماتریس ارجاع را در ماتریس خاطره کپی می‌کند و مقادیر صفر ماتریس ارجاع را از ماتریس خاطره جایگزین می‌کند. این تابع در الگوریتم ۴-۲ آمده است.

در این پژوهش در دوقسمت نوآوری صورت گرفته است، قسمت اول ارائه‌ی معیاری جدید برای سنجش معیار خبرگی که طبق تعریف ۴ـ۱ این معیار در خط ۱۳ الگوریتم ۴-۱ پیاده‌سازی شده است؛ نوآوری دوم نحوه‌ی ترکیب اطلاعات دانش عامل‌ها با استفاده از انتگرال فازی که در خط ۲۲ الگوریتم ۴-۱ و شرح جزییات پیاده‌سازی آن در الگوریتم ۴-۳ آمده است.

ورودی‌های الگوریتم ۴-۳ به ترتیب مجموعه‌ای از جداول Q و ماتریس‌های ارجاع(REFMAT) تمامی عامل‌ها می‌باشد بطوری که در ازای هر جدول Q یک ماتریس REFMAT متناظر وجود دارد. خروجی این الگوریتم یک جدول Q می‌باشد که از ترکیب جداول Q ورودی با درنظر گرفتن میزان خبرگی هرکدام از عامل‌ها که توسط ماتریس‌های REFMAT آن‌ها تعیین می‌شود، بدست آمده است. الگوریتم ۴-۳ به ازای کلیه‌ی

الگوریتم ۴-۲ تابع Conditional_Swap معرفی شده در الگوریتم ۴-۱

```
1: function Conditional_Swap(REFMAT, RCMAT)
2:     Require: size(REFMAT) = size(RCMAT)
3:     for each element r in REFMAT and its corresponding element c in RCMAT do
4:         if r = 0 then
5:             r = c;
6:         else
7:             c = r;
8:         end if
9:     end for
10:    return REFMAT, RCMAT
11: end function
```

موقعیت‌ها(sها در خط ۴) ابتدا مقادیر REFMAT کلیه‌ی عامل‌ها در ناحیه‌ای که آن موقعیت در آن واقع است (که توسط تابع نگاشت $\phi(\cdot)$ بدست می‌آید) را استخراج می‌کند و در برداری بنام $\vec{a}$ ذخیره می‌کند (خطوط ۶ و ۷) که در واقع میزان ارجاعات هرکدام از عامل‌ها در ناحیه‌ی $\phi(s)$ می‌باشد.

بردار $\vec{a}$ معیاری برای سنجش میزان خبرگی کلی عامل‌ها در موقعیت s است، طبق آنچه که در تعریف ۴ـ۱ آمده است در هر ناحیه عاملی خبره‌تر است که مقدار REFMAT مربوط به آن ناحیه از دیگر عامل‌ها کمتر باشد. از آنجایی که دامنه‌ی خروجی توابع $g(\cdot)$ بازه‌ی $[0,1]$ می‌باشد و طبق آنچه که در الگوریتم‌های ۴-۴ تا ۴-۷ آورده شده است خروجی تابع $g(\cdot)$ بر اساس ورودی‌های آن تخمین زده می‌شود و از طرفی نیاز به مکمل‌سازی میزان خبرگی عامل‌ها (عاملی که بیشترین ارجاع به ناحیه‌ای را دارد، کمترین میزان خبرگی را در آن ناحیه دارد و برعکس) نیاز به عادی‌سازی مقادیر ارجاعات داریم، در نتیجه در خط ۹ بعد از عادی‌سازی[1] مقادیر REFMAT عامل‌ها در ناحیه‌ی $\phi(s)$ یک مکمل‌گیری صورت می‌گیرد تا عاملی که مقدار REFMAT کمتری دارد دارای بیشترین مقدار بعد از عادی‌سازی شود. در خط ۱۰ به ازای کلیه‌ی عمل‌های ممکن در موقعیت s ابتدا مقادیر Q تک‌تک عامل‌ها را در موقعیت s و عمل a در خطوط ۱۲ و ۱۳ در بردار $\vec{x}$ ذخیره می‌کنیم و در نهایت در خط ۱۵ با استفاده از انتگرال فازی چوکت معرفی شده در ۳ـ۹ مقدار Q مشارکتی حاصل از میزان خبرگی بردار $\vec{A}$ و مقادیر Qهای تک‌تک عامل‌ها در بردار $\vec{x}$ در موقعیت s و عمل a بدست محاسبه می‌شود.

۴ـ۳ـ۲ تعیین توابع $f(\cdot)$ و $g(\cdot)$ در انتگرال فازی چوکت

بطور خلاصه در الگوریتم ۴-۳ دو بخش عمده دارد بخش اول مربوط استخراج میزان خبرگی عامل‌ها بگونه‌ای که عاملی که خبره‌تر از دارای مقدار خبرگی بیشتری باشد که این بخش در خطوط ۶ تا ۹ صورت می‌گیرد؛ بخش دیگر محاسبه‌ی مقادیر Q مشارکتی کلیه‌ی عمل‌های ممکن در یک موقعیت با درنظر گرفتن میزان خبرگی عامل‌ها و مقادیر Q آن‌ها با استفاده از انتگرال فازی چوکت که در خطوط ۱۰ تا ۱۵ صورت می‌پذیرد.

[1] Normalize

الگوریتم ۴-۳ تابع FCI_Combiner معرفی شده در الگوریتم ۴-۱

```
1: function FCI_Combiner(K⃗, R⃗)
2:     Require: length(K⃗) = length(R⃗) = m
3:     Ensure: Initialize CoQ_FCI                          ▷ The cooperative Q table.
4:     for each state s do
5:         a⃗ ← {};          ▷ Contains the normalized values of REFMATs for state s for all agents.
6:         for each REFMAT_i in R⃗ do
7:             a⃗.add(REFMAT_i(φ(s)));
8:         end for
9:         A⃗ ← 1− normalize(a⃗);
10:        for each possible action a in state s do
11:            x⃗ ← {};          ▷ Contains the Q values of action a in state s for all agents.
12:            for each Q_i in K⃗ do
13:                x⃗.add(Q_i[s, a]);
14:            end for
15:            CoQ_FCI[s, a] ← Σ_{i=1}^{m} (f(x_{π(i)}) − f(x_{π(i−1)})) · g(A⃗_i)   ▷ The Choquet Integral.
16:        end for
17:    end for
18:    return CoQ_FCI;
19: end function
```

آنچه که در خط ۱۵ الگوریتم ۴-۳ مورد توجه واقع شود این است که توابع $f(\cdot)$ و $g(\cdot)$ چگونه باید تعریف شوند؟ برای تعیین تابع $f(\cdot)$ منطقی که در این پژوهش استفاده کردیم بدین صورت است که از آنجایی که خروجی تابع $g(\cdot)$ یک مقدار عددی[1] بدون واحد می‌باشد و همچنین برای اینکه خروجی انتگرال فازی خط ۱۵ را بتوان به عنوان مقادیر جدول Q مشارکتی جدید در نظر گرفت تا بتوانیم در خطوط ۲۲ الگوریتم ۴-۱ به عنوان جدول Q تک‌تک عامل‌ها جایگذاری کنیم باید خروجی انتگرال فازی خط ۱۵ الگوریتم ۴-۳ از جنس جدول‌های Q عامل‌ها باشد در نتیجه تابع $f(\cdot)$ باید یک تابع خطی بصورت ۴ـ۱ باشد تا خروجی انتگرال فازی همجنس مقادیر $\vec{x}$ باشد.

$$f(\omega) = a\omega + b \qquad (۴ـ۱)$$

متغیرهای a و b در ۴ـ۱ می‌تواند به عنوان پارامترهای وفقی[2] در میزان کیفیت جدول Q مشارکتی خروجی الگوریتم ۴-۳ موثر واقع شود ولی با این‌حال در این پژوهش مقادیر a و b هردو به ترتیب مقادیر ثابت ۱ و صفر در نظر گرفته شده‌اند که یعنی از تابع همانی به عنوان تابع $f(\cdot)$ استفاده شده است؛ دلیل انتخاب مقادیر ۱ و صفر ثابت برای متغیرهای a و b این است که از بین تمام مقادیری که a و b می‌تواند بگیرند، استفاده از تابع همانی به

[1] Scalar

[2] Addaptive Parameters

عنوان تابع $f(\omega)$ به نظر منطقی می‌رسد زیرا هیچ انتقال[1] و درشت‌نمایی[2] بروی مقادیر ω (و نهایتا مقادیر جدول Q) اعمال نمی‌کند و به داده‌ها را به‌همان صورت که هستند می‌بیند – البته در صورت ارائه‌ی روشی که بتواند این مقادیر را به صورت وفقی تغییر دهد، شهود[3] این را می‌گوید که می‌تواند در بهبود نتایج حاصله کمک کند؛ مثلا در بازه‌های زمانی مشخص با تغییر مقدار درشت‌نمایی بتواند به افزایش سرعت یادگیری کمک کند.

تابع $g(\cdot)$ یک ورودی مرتب شده طبق آنچه که در ۳ـ۱۱ آمده است می‌گیرد و در الگوریتم ۴-۳ تعیین این تابع تاثیر زیادی بروی کیفیت خروجی الگوریتم خواهد داشت ولی چالش‌هایی برای تعیین این تابع داریم؛ تابع $g(\cdot)$ باید دارای ویژگی‌های زیر باشد:

۱. **پویا[4] باشد:** از آنجایی که تابع $g(\cdot)$ میزان اندازه‌گیری غیرافزایشی[5] منابع اطلاعاتی را در اختیار می‌گذارد [۴]، نیاز داریم تعیین کنیم که کدام منابع اطلاعاتی(در اینجا خبرگی عامل‌ها) در کنار هم چه ارزش افزوده‌ای دارد؛ ولی از آنجایی که در حین یادگیری مشترک روشی برای تعیین این ارزش افزوده نداریم بنابراین باید تابع $g(\cdot)$ بصورت پویا بتواند مقادیر این ارزش افزوده را تخمین بزند.

۲. **قابل گسترش[6] باشد:** چون تعداد عامل‌ها در محیط متغیر است لذا باید تابع $g(\cdot)$ بگونه‌ای باشد که به ازای تغییر تعداد عامل‌ها (که تغییر در تعداد اعضای بردار $\vec{A}$ را در پی دارد) قابل گسترش باشد.

یکی از روش‌های تخمین $g(\cdot)$ که دو ویژگی بالا را داشته باشد، تابع اندازه‌گیری‌ـλ سوگنو می‌باشد ولی این تابع نیاز به ریشه‌یابی روی متغیر λ دارد که طبق آنچه که در ۳ـ۵ آمده است به ازای تعداد عامل‌های مختلف نیاز به ریشه‌یابی معادلات غیرخطی دارد. بدلیل محاسبات سنگین و وقت‌گیر این ریشه‌یابی و همچنین نتایج حاصل از دستاوردهای این پژوهش (که در فصل نتیجه‌گیری آورده شده است)، در آزمایش‌ها صورت گرفته در این پژوهش از تابع اندازه‌گیری‌ـλ سوگنو به عنوان تابع $g(\cdot)$ استفاده نشده است. یک سری توابع در این پژوهش بجهت استفاده، آزمایش و نتیجه‌گیری به عنوان $g(\cdot)$ معرفی شده است که این توابع در الگوریتم‌های ۴-۴ تا ۴-۷ آمده‌اند.

در الگوریتم ۴-۴ به ازای هر ورودی دلخواه مقدار ثابت ۱ به عنوان خروجی برگشت داده می‌شود، این بدین معنی است که ارزش افزوده‌ی هرنوع ترکیب اطلاعاتی(خبرگی) برای ما دارای حداکثر ارزش می‌باشد و این مساله باعث می‌شود که نتیجه‌ی انتگرال فازی خط ۱۵ الگوریتم ۴-۳ مقداری معادل با مقدار خبره‌ترین عامل(عاملی که کمترین پرسه را در محیط مربوطه داشته) را به عنوان مقدار جدید جدول Q مشارکتی تولید کند.

[1] Shift
[2] Magnification
[3] Intuition
[4] Dynamic
[5] Non-additive
[6] Expandable

الگوریتم ۴-۴ الگوریتم Const-One برای تخمین تابع $g(\cdot)$ در الگوریتم ۴-۳

```
1: function Const-One(A_i)
2:     if length(A_i) ≥ m then
3:         return 1;
4:     else if length(A_i) = 0 then
5:         return 0;
6:     else
7:         return 1;
8:     end if
9: end function
```

الگوریتم ۴-۵ الگوریتم Max برای تخمین تابع $g(\cdot)$ در الگوریتم ۴-۳

```
1: function Max(A_i)
2:     if length(A_i) ≥ m then
3:         return 1;
4:     else if length(A_i) = 0 then
5:         return 0;
6:     else
7:         return max_i(A_i);
8:     end if
9: end function
```

در الگوریتم ۴-۵ میزان خبرگی خبره‌ترین عامل به عنوان خروجی تابع $g(\cdot)$ برگشت داده می‌شود. در الگوریتم ۴-۶ خروجی، میانگین خبرگی عامل‌ها در نظر گرفته شده است و در الگوریتم ۴-۷ طبق رابطه‌ی نوشته شده میانگین kام میزان خبرگی‌ها به عنوان خروجی برمی‌گردد به طوری که بزرگترین خبرگی در عدد k و کوچکترین خبرگی در عدد ۱ و هر آنچه که مابین این دو خبرگی وجود دارد در اندیس ترتیب مرتب شده آن‌ها ضرب می‌شود و میانگین این مجموع محاسبه و برگشت داده می‌شود.

نکته‌ای که در مورد الگوریتم ۴-۶ باید توجه کرد این است که با اینکه این الگورتم شرط «یکنوایی» انتگرال چوکت را ارضا نمی‌کند ولی همانطور که در فصل آزمایش‌ها خواهیم دید باعث واگرایی الگوریتم پیشنهادی نمی‌شود که در این مساله جای تحقیق بیشتری دارد که شرایط لازم توابع $g(\cdot)$ کاربرد مورد استفاده این پژوهش چگونه باید باشد، زیرا در الگوریتم ۴-۶ با نقض شدن شرط یکنوایی به همگرایی الگوریتم لطمه‌ای وارد نشد!

۴-۴ علت کارکرد انتگرال فازی چوکت در انتقال دانش

در این قسمت به بررسی شهودی اینکه چرا انتگرال فازی چوکت برای انتقال(ترکیب) دانش‌های عامل‌ها می‌تواند موثر واقع باشد می‌پردازیم. این شهود بعدها در آزمایش‌ها نشان داده خواهد شد که صحت دارد. انتگرال فازی چوکت یک سری ویژگی‌ها دارد که برای مدل کردن انتقال دانش آن را کاندیدای مناسبی می‌کند. از مهم‌ترین

الگوریتم ۴-۶ الگوریتم Mean برای تخمین تابع $g(\cdot)$ در الگوریتم ۴-۳

function Mean($\vec{A}_i$)
    if length($\vec{A}_i$) $\geq m$ then
        return 1;
    else if length($\vec{A}_i$) = 0 then
        return 0;
    else
        return $\frac{\sum_{j=1}^{\text{length}(\vec{A}_i)} \vec{A}_i(j)}{\text{length}(\vec{A}_i)}$;
    end if
end function

الگوریتم ۴-۷ الگوریتم K-Mean برای تخمین تابع $g(\cdot)$ در الگوریتم ۴-۳

function K-Mean($\vec{A}_i$)
    if length($\vec{A}_i$) $\geq m$ then
        return 1;
    else if length($\vec{A}_i$) = 0 then
        return 0;
    else if length($\vec{A}_i$) = 1 then
        return $\vec{A}_i(1)$;
    else
        $\vec{B}_i$ = Sort-Ascending($\vec{A}_i$);
        return $\min\{\frac{\sum_{k=1}^{\text{length}(\vec{B}_i)} k \cdot \vec{B}_i(k)}{(\sum_{j=1}^{\text{length}(\vec{B}_i)} j)-1}, 1\}$;
    end if
end function

ویژگی‌ها می‌توان به موارد زیر اشاره کرد [۲۶].

۱. **محدود است:** اگر شرایط مرزی تابع $g(\cdot)$ برقرار باشد انتگرال فازی هیچ‌گاه بیشتر از حداکثر مقدار $f(x_{\pi_i})$ها و کمتر از حداقل مقدار آن‌ها خروجی نمی‌دهد [۲۷]. یعنی دانش تولیدی خارج از محدوده‌ی دانش فعلی عامل‌ها نمی‌باشد فقط ترکیب مناسبی از این دانش‌ها به عنوان خروجی برگشت داده می‌شود که این در کاربرد یادگیری تقویتی به این معنی است که هیچ‌گاه مقادیر جدول Q بیشتر یا کمتر از آنچه که تجربه شده نمی‌شود؛ در نتیجه در صورت کران‌دار بودن پاداش‌های دریافتی از محیط جدول Q خروجی انتگرال فازی نیز کران‌دار است که نتیجه می‌دهد الگوریتم پیشنهادی حتما همگرا خواهد شد [۱].

۲. **می‌تواند اندازه‌گیری‌های غیرافزایشی را مدل کند:** معمولا روش‌هایی که تا کنون در این زمینه ارائه شده است از میانگین وزنی خبرگی عامل‌ها برای بدست آوردن جدول Q مشترک استفاده کرده‌اند [۲، ۱۲، ۱۹]. این در حالی هست که میانگین وزن‌دار قسمتی از مدل اندازه‌گیری غیرافزایشی می‌باشد. بنابراین با درنظر گرفتن مدل‌های غیرافزایشی که در ماهیت مساله هست قدرت و انعطاف بیشتری نسبت به روش‌هایی که

فقط از میانگین وزنی استفاده کرده‌اند، در اختیار داریم.

۴-۴-۱ اثبات همگرایی روش پیشنهادی

در مورد همگرایی روش پیشنهادی می‌توان گفت که از آنجایی که در خطوط ۸ تا ۱۲ الگوریتم پیشنهادی ۴-۱ که یادگیری تقویتی Q بدون دخل و تصرفی آورده شده است، اثبات همگرایی یادگیری تقویتی عامل‌ها به قوت خود باقی است [۳۲]. حال باید نشان دهیم که ترکیب یادگیری تقویتی با انتگرال فازی همگرایی یادگیری تقویتی را برهم نمی‌زند.

اثبات. فرض می‌کنیم که به تعداد l عدد عامل وجود دارد که در هر مرحله‌ی یادگیری مشارکتی در زمان t، مقادیر $Q(s,a)$ عامل‌ها به ترتیب میزان بهینگی مقادیر آن‌ها به صورت زیر می‌باشند.

$$Q^t_{\pi_1}(s,a) \preceq Q^t_{\pi_2}(s,a) \preceq \cdots \preceq Q^t_{\pi_l}(s,a) \preceq Q^*(s,a), \quad \forall s,a,t \tag{۴-۲}$$

که π_1 اندیس عاملی است که نسبت به دیگر عامل‌ها دارای مقدار $Q(s,a)$ با کمترین بهینگی می‌باشد و π_l اندیس بهینه‌ترین مقدار $Q(s,a)$ است و عملگر $\preceq$ به معنی «کمتر یا مساوی بودن از دیدگاه بهینگی» می‌باشد و $Q^*(s,a)$ مقدار سیاست بهینه در (s,a) می‌باشد. طبق اثبات همگرایی یادگیری تقویتی (بدون در نظر گرفتن انتگرال فازی) زمانی که پاداش‌های دریافتی محیطی محدود و نرخ یادگیری محدود به بازه‌ی $[0,1)$ باشد آنگاه داریم [۳۲]:

$$Q^t_{\pi_i}(s,a) \to Q^*(s,a) \quad \text{as} \quad t \to \infty, \quad \forall s,a,i \in [1 \cdots l] \tag{۴-۳}$$

با توجه به اینکه که انتگرال فازی به ازای هر (s,a) مقداری مابین حداکثر و حداقل مقادیر جداول Q عامل‌ها در آن (s,a) را تولید می‌کند، طبق خاصیت محدود بودن انتگرال فازی چوکت می‌توانیم بگوییم که دانش تولیدی حاصل از ترکیب جداول $Q_i, \forall i \in [1 \cdots l]$ عامل‌ها در زمان t به جدول Q^t_{REFMAT} می‌رسیم که دارای خاصیت زیر است.

$$Q^t_{\pi_1}(s,a) \preceq Q^t_{\text{REFMAT}}(s,a) \preceq Q^t_{\pi_l}(s,a), \quad \forall s,a,t \tag{۴-۴}$$

از ۴-۲ تا ۴-۴ می‌توان به این نتیجه رسید از آنجایی که $Q^t_{\pi_1}(s,a)$ به ازای هر t دلخواه و با توجه به

شرایط ذکر شده در اثبات همگرایی [۳۲] نهایتا همگرا می‌شود و همچنین $Q^t_{\pi_l}(s,a)$ نیز در $t \to \infty$ همگرا می‌شود و کلیه‌ی جداول مابین این دو یعنی $Q^t_{\pi_i}(s,a), \forall s,a,t,i \in [1 \cdots l]$ نیز همگرا می‌شوند (۴ـ۲ و ۴ـ۳) ؛ در نتیجه جدول تولیدی توسط انتگرال فازی چوکت که خاصیت ۴ـ۴ در آن برقرار می‌باشد نیز در $t \to \infty$ همگرا خواهد شد. □

۴ـ۵ نتیجه‌گیری

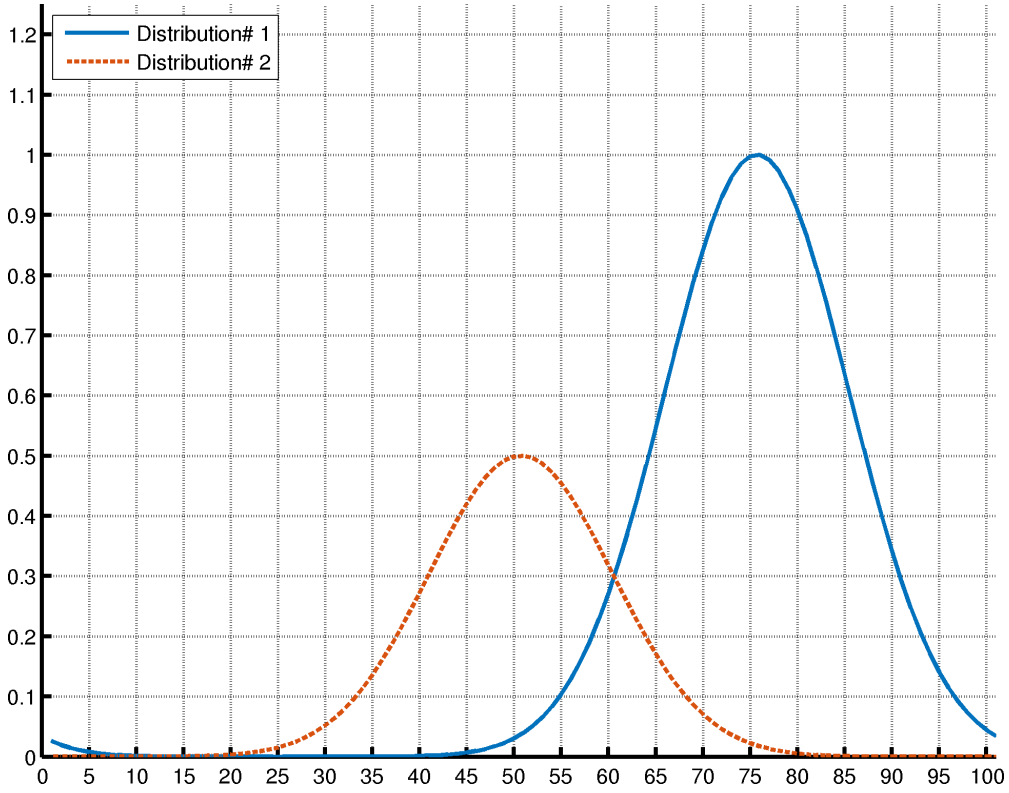
در این فصل به‌طور مفصل به معرفی روش پیشنهادی پرداختیم، ابتدا مفاهیم و تعاریف بنیادی روش پیشنهادی را ارائه دادیم، بعد از معرفی الگوریتم پیشنهادی به توضیح قسمت‌های مختلف این الگوریتم به‌صورت جز پرداختیم و در نهایت با اثبات همگرایی روش پیشنهادی به مبحث این فصل خاتمه بخشیدیم. در فصل بعدی آزمایش‌های انجام شده به‌جهت تائید صحت روش پیشنهادی آورده شده است.

فصل پنجم

نتایج شبیه‌سازی و آزمایش‌ها

۵ـ۱ مقدمه

در این فصل به ارائه‌ی آزمایش‌های صورت گرفته بروی روش پیشنهادی می‌پردازیم و در طی این آزمایش‌ها روش پیشنهادی را با روش کوتاه‌ترین مسیر تجربه شده(یا به اختصار SEP) مقایسه می‌کنیم که آخرین و مدرن‌ترین روش ارائه شده در جهت بهبود یادگیری مشارکتی می‌باشد [۲]. آزمایش‌ها بروی دو محیط «پلکان مارپیچ» و «صید و صیاد» صورت گرفته است. آزمایش‌ها به دو دسته تقسیم بندی شده است؛ دسته اول آزمایش‌هایی که روش پیشنهادی را در مقابل روش SEP قرار می‌دهد و عملکرد روش پیشنهادی را مورد سنجش قرار می‌دهد. دسته دوم آزمایش‌ها مربوط به آزمون رفتار روش پیشنهادی در صورت تغییر در پارامترهای مختلف آن می‌باشد. همچنین اثر استفاده از سیاست‌های انتخاب عمل مختلف در الگوریتم ۴-۱ نیز بررسی شده است. در روش‌های مرتبط قبلی [۲، ۱۹] که این پژوهش ادامه‌ی کار آن‌ها می‌باشد فقط از سیاست انتخاب عمل بولتزمن استفاده کرده‌اند؛ در این پژوهش علاوه بر بولتزمن تاثیر استفاده از روش εـحریصانه بروی هردو روش پیشنهادی و SEP نیز مورد بررسی واقع گردیده است.



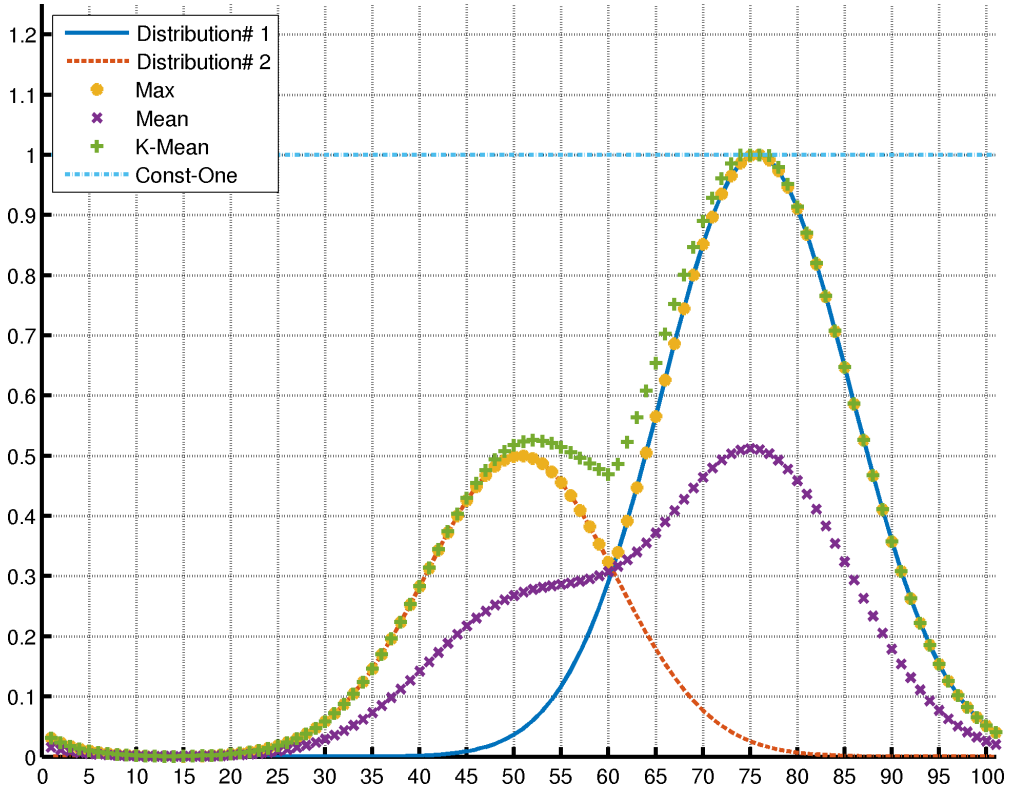
شکل ۵ـ۱: دو توزیع فرضی بجهت نمایش نحوه‌ی رفتار الگوریتم‌های ۴-۴ تا ۴-۷ بروی آن‌ها.

۵ـ۲ رفتار الگوریتم‌های معرفی شده برای $g(\cdot)$

در این قسمت به بررسی رفتار الگوریتم‌های ۴-۴ تا ۴-۷ معرفی شده برای $g(\cdot)$ بروی دو توزیع فرضی خواهیم پرداخت، زیرا که در طی اجرای آزمایش‌های مختلف نتایج تاثیر این توابع بر اجرای الگوریتم پیشنهادی ۴-۱ آورده شده است، لذا به‌جهت درکت علت تاثیرهای مختلف هرکدام از این توابع بروی نتیجه‌ی الگوریتم پیشنهادی در آزمایش‌ها، درک نحوه‌ی رفتار الگوریتم‌های ۴-۴ تا ۴-۷ ضروری است.

برای نمایش نحوه‌ی رفتار هرکدام از الگوریتم‌ها دو توزیع فرضی که در شکل ۵ـ۱ آورده شده است، را در نظر می‌گیریم. در صورت اعمال الگوریتم‌های ۴-۴ تا ۴-۷ بروی دو توزیع آورده شده در شکل ۵ـ۱ توزیع‌های جدیدی بصورت آنچه که در شکل ۵ـ۲ آمده است بدست می‌آیند. همانطور که در شکل ۵ـ۲ می‌بینیم اعمال الگوریتم Const-One بروی دو توزیع مقدار ثابت ۱ را برمی‌گرداند. اعمال الگوریتم Max در هر نقطه حداکثر مقدار هر دو توزیع را برمی‌گرداند. الگوریتم Mean میانگین دو توزیع را در هر نقطه حساب می‌کند و در نهایت الگوریتم K-Mean هردو توزیع را محاسبه میکند که همانطور که می‌بینیم الگوریتم K-Mean به سبب ماهیت الگوریتم به سمت بیشترین مقدار پیش‌قدر[1] می‌باشد.

[1] Bias



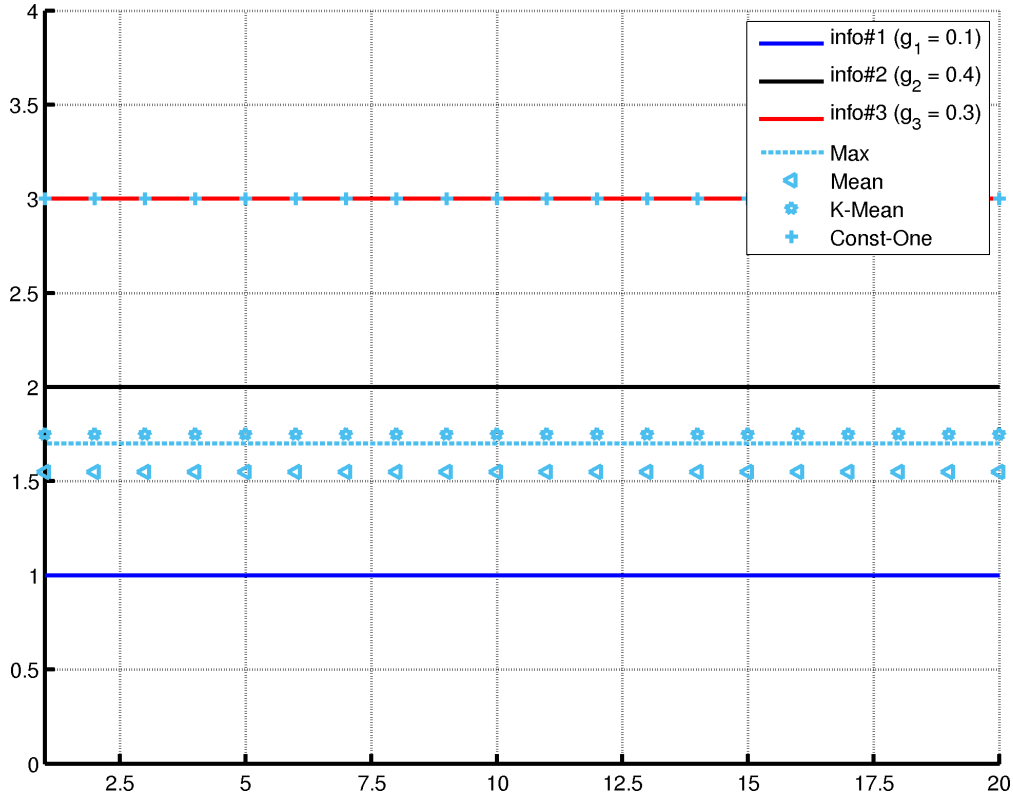
شکل ۵ـ۲: نمایش توزیع‌های جدید بدست آمده بعد از اعمال الگوریتم‌های ۴-۴ تا ۴-۷ بروی دو توزیع فرضی شکل ۵ـ۱

۵ـ۲ـ۱ تعابیر مختلف انتگرال فازی چوکت از داده‌ها برمبنای $g(\cdot)$

الگوریتم‌های ۴-۴ تا ۴-۷ به تنهایی فقط در نقش یک عملگر بازی می‌کند ولی در هنگام ترکیب دانش با انتگرال فازی چوکت به دانش خروجی الگوریتم از دیدگاه‌های متفاوتی نگاه می‌کنند. از آنجایی که در فصل‌های قبلی نیز آورده شد انتگرال فازی در واقع یک تعمیم الگوریتم دهنده‌ی میانگین وزنی می‌باشد که علاوه بر ویژگی‌هایی که روش میانگین وزنی ارائه می‌دهد می‌تواند اندازه‌گیری‌های غیرافزایشی را نیز مدل کند. لذا با تغییر تابع $g(\cdot)$ می‌توان باعث شد که انتگرال فازی چوکت تعابیر مختلفی از داده‌های ورودی خود ارائه دهد. از بین الگوریتم‌ها فقط الگوریتم Const-One دارای تعبیر صریح ریاضی می‌باشد که در رابطه‌ی ۵ـ۱ آمده است، بقیه‌ی الگوریتم‌ها دارای تعابیر صریح نیستند و فقط می‌توانیم بر اساسی نمایشی که در شکل ۵ـ۲ آمده است شهودی از نحوه‌ی تغییر رفتار انتگرال فازی به ازای هریک از الگوریتم‌ها ارائه داد.

$$g = \text{Const-One}(\cdot) \equiv \begin{cases} g(X) & = 1 \\ g(\emptyset) & = 0 \\ g_{A \subseteq X}(A) & = 1 \end{cases} \Rightarrow \mathcal{C}_g(f) \equiv \max\{f(x_{\pi^c_{(1)}}), \cdots, f(x_{\pi^c_{(n)}})\} \quad (۱ـ۵)$$

برای نمایش شهودی نحوه‌ی تغییر رفتار انتگرال فازی چوکت در شکل ۵ـ۳ سه منبع اطلاعاتی با مقادیر $y = 1$ و $y = 2$ و $y = 3$ در نظر گرفته شده است و مقدار ارزش هرکدام از این‌ها به ترتیب $g = [0.1 \quad 0.4 \quad 0.3]^T$



شکل ۵ـ۳: نمایش رفتار انتگرال فازی بروی منابع اطلاعاتی $y = 1$ و $y = 2$ و $y = 3$ به ازای توابع $g(\cdot)$های مختلف.

در نظر گرفته شده است. سپس انتگرال فازی چوکت را با در نظر گرفتن تابع همانی به عنوان تابع $f(\cdot)$ بر روی این ۳ منبع اطلاعاتی اعمال کردیم و همانطور که می‌بینیم مقداری که انتگرال فازی چوکت به ازای $g = \text{Const-One}(\cdot)$ تولید می‌کند برابر با حداکثر مقدار منابع اطلاعاتی دریافتی می‌باشد. در حالت کلی هرچقدر میانگین تابع $g_{A\subseteq X}(A)$ به سمت مقدار ۱ متمایل باشد خروجی انتگرال فازی چوکت به سمت بیشینه مقدار منابع اطلاعاتی پیش‌قدر می‌شود و در صورتی که این میانگین به سمت صفر متمایل باشد خروجی به کمینه مقدار پیش‌قدر می‌شود.

۵ـ۳ مقایسه‌ی روش پیشنهادی با روش کوتاه‌ترین مسیر تجربه شده

در این قسمت به مقایسه‌ی روش پیشنهادی با روش «کوتاه‌ترین مسیر تجربه شده» که بروزترین تکنیک ارائه شده در این شاخه از یادگیری مشارکتی می‌باشد می‌پردازیم [۲]. کلیه‌ی این آزمایش‌ها در دو محیط «پلکان مارپیچ» و «صید و صیاد» صورت گرفته است. نتیجه‌ی هر آزمایش حاصل میانگین ۲۰ اجرای مستقل تمامی الگوریتم‌ها می‌باشد. همچنین به غیر از مواردی که صراحتا قید شده است تعداد عامل‌ها ۳ عدد می‌باشد – البته بدیهی است که یادگیری مستقل تک عامله(یا به اختصار IL[1]) شامل این قاعده نمی‌باشد. همچنین در کلیه‌ی

[1] Individual Learning

جدول ۵ـ۱: لیست اختصارهای استفاده شده در این فصل

اختصار	معنی
REFMAT	روش پیشنهادی
IL	یادگیری مستقل تک عامله
SEP	روش کوتاه‌ترین مسیر تجربه شده
wsum	میانگین وزنی
fci-max	الگوریتم Max به عنوان مدل کننده‌ی تابع $g(\cdot)$
fci-mean	الگوریتم Mean به عنوان مدل کننده‌ی تابع $g(\cdot)$
fci-k-mean	الگوریتم K-Mean به عنوان مدل کننده‌ی تابع $g(\cdot)$
fci-const-one	الگوریتم Const-One به عنوان مدل کننده‌ی تابع $g(\cdot)$
Rand-Walk	جستجوی کاملا تصادفی محیط

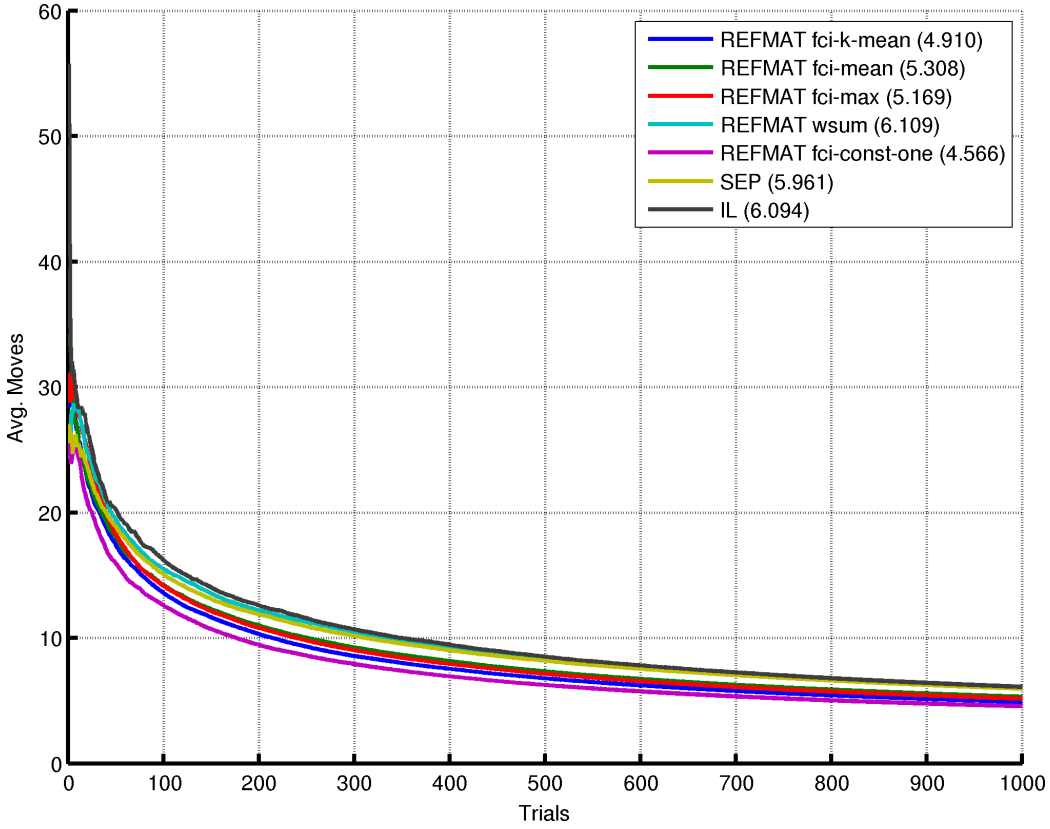
آزمایش‌ها عامل‌ها از ۲۰۰ چرخه یادگیری بهره می‌برند و در هر چرخه عامل ۵ بار تلاش[1] مستقل می‌کند که در مجموع ۱۰۰۰ تلاش صورت می‌گیرد. کلیه‌ی پارامترهای مربوط قسمت یادگیری مستقل الگوریتم ۴-۱ اعمال شده در آزمایش‌ها این فصل منطبق بر پارامترهای تعریف شده کار میرزایی می‌باشد که نتایج قابل قیاس باشند [۲]. در ضمن در این فصل اختصارهای جدول ۵ـ۱ را نیز داریم.

در این فصل در حالت کلی ما در دو بخش سیاست انتخاب عمل «بولتزمن» و «εـحریصانه» (که از این به بعد، به اختصار «تابع بولتزمن» و «تابع εـحریصانه» خطاب خواهیم کرد.) به مقایسه‌ی نتایج می‌پردازیم. طبق آنچه که در ادامه مشاهده خواهیم کرد چه در صورت استفاده از تابع بولتزمن و چه تابع حریصانه روش پیشنهادی چه در سرعت یادگیری و چه در کیفیت یادگیری بهتر از روش SEP می‌باشد.

برای اینکه نشان دهیم که استفاده از انتگرال فازی در بهبود نتیجه تاثیر بسزایی دارد از تابع میانگین وزنی (یا به اختصار wsum[2]) نیز استفاده کرده‌ایم. بدین صورت که بجای اینکه بعد از استخراج میزان خبرگی هر عامل، جدول Q هر عامل را متناسب با میزان خبرگی‌ای که دارد در دانش جمعی دخیل می‌کنیم تا جدول Q مشارکتی تولید شود. تابع میانگین وزنی روشی است که در پژوهش‌های اخیر به کررات از آن استفاده کرده‌اند [۲، ۱۲، ۱۹]. یکی از اهداف ما در این پژوهش نمایش قدرت انتگرال‌های فازی می‌باشد.

[1] Trial

[2] Weighted Sum



شکل ۵ـ۴: مقایسه در سرعت و کیفیت یادگیری با تابع بولتزمن در محیط پلکان مارپیچ

۵ـ۳ـ۱ مقایسه در محیط پلکان مارپیچ

آزمایش‌های مربوط به این قسمت در ۴ بخش صورت گرفته است؛ ۱. مقایسه در سرعت و کیفیت یادگیری، ۲. مقایسه در پیچیدگی زمانی، ۳. مقایسه در میزان باروری، ۴. مقایسه تاثیر تعداد عامل‌ها بر میزان کیفیت و سرعت یادگیری.

سیاست انتخاب عمل «بولتزمن»

مقایسه در سرعت و کیفیت یادگیری: نتایج حاصل از اجرای الگوریتم‌ها در محیط پلکان مارپیچ در شکل ۵ـ۴ آمده است. در این شکل محور افقی تعداد تلاش‌های یادگیری عامل را نشان می‌دهد که در تلاش اول عامل بدون دانش اولیه شروع به تعامل با محیط می‌کند و در تلاش ۱۰۰۰ام عامل به اجرای خود پایان می‌دهد. محور عمودی نمودار میانگین تعداد قدم‌های عامل را نشان می‌دهد. اعداد کناری برچسب‌ها (گوشه بالا سمت راست) متوسط تعداد قدم در آخرین تلاش عامل می‌باشد که انتظار می‌رود عامل آگاهی نسبی کاملی از محیط دارد را نشان می‌دهد که این عدد هرچقدر کمتر باشد نشان می‌دهد که عامل در طی رسیدن به هدف تعداد گام کمتری برداشته است و در نتیجه دانش و شناخت بهتری از محیط دارد.

همانطور که مشاهده می‌شود روش SEP دارای ۲٪ بهبود نسبت به IL می‌باشد در حالی که روش پیشنهادی

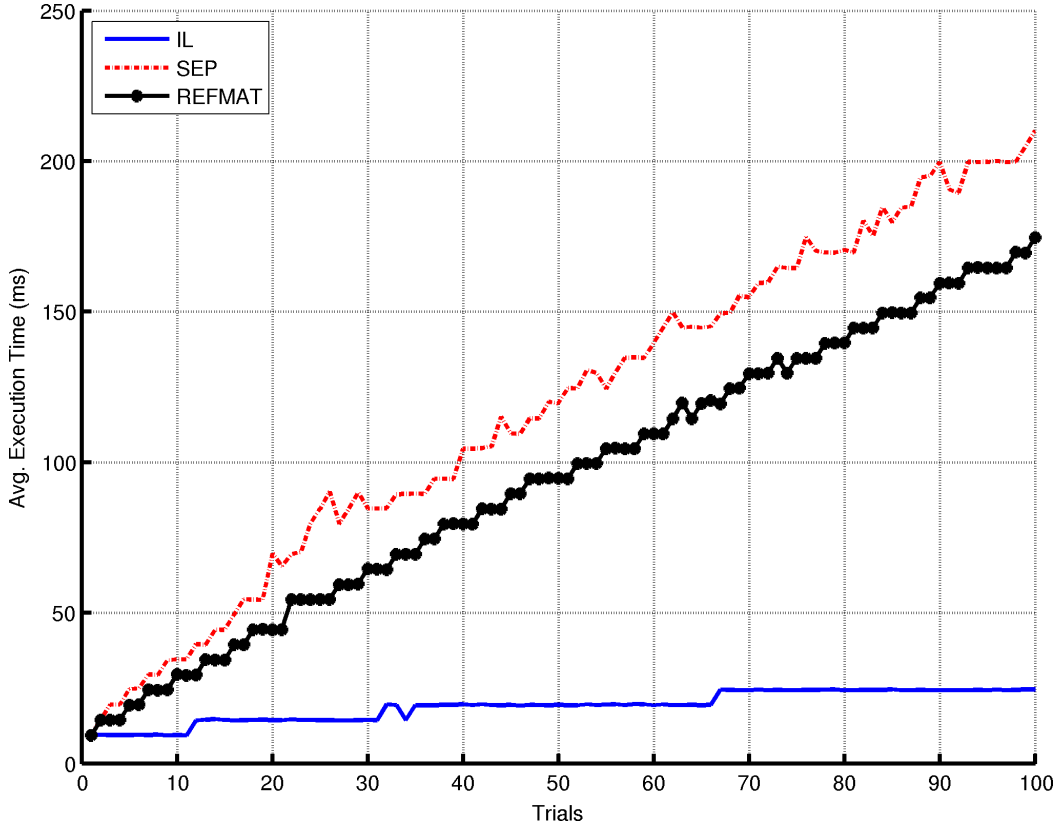
جدول ۵ـ۲: مقایسه در میزان درصد بهبود کیفیت یادگیری در محیط پلکان مارپیچ با تابع بولتزمن

			REFMAT				
	IL	SEP	wsum	fci-mean	fci-max	fci-k-mean	fci-const-one
IL	0.0						
SEP	2.2	0.0					
wsum	-0.2	-2.3	0.0				
fci-mean	14.9	12.5	15.1	0.0			
fci-max	18.0	15.5	18.2	2.7	0.0		
fci-k-mean	24.0	21.4	24.2	7.9	5.1	0.0	
fci-const-one	33.6	30.7	33.8	16.2	13.2	7.7	0.0

در زمانی که از انتگرال فازی استفاده می‌کند در بدترین حالت دارای ٪۱۸ بهبود و در بهترین حالات دارای ٪۳۳ بهبود می‌باشد که نسبت به روش SEP تقریبا ۹ الی ۱۶ برابر نتیجه را بهبود داده است. در صورتی که از میانگین وزنی بجای انتگرال فازی استفاده شود نتایج با اختلاف اندکی (کمتر از ۱-٪) بدتر از یادگیری IL بوده است که نشان می‌دهد که استفاده از انتگرال فازی چقدر می‌تواند نسبت به روش‌های سنتی و معمولی چون میانگین وزنی موثر واقع شود. نتایج این قسمت را می‌توان در جدول ۵ـ۲ خلاصه کرد. همچنین طبق آنچه در فصل ۳ آورده شده سرعت یادگیری به صورت مساحت زیرنمودار در شکل ۵ـ۴ محاسبه می‌شود، به صورت دیداری[1] می‌توانیم ببینیم روش پیشنهادی دارای کمترین مساحت زیر منحنی می‌باشد که نشان می‌دهد دارای بیشترین سرعت یادگیری می‌باشد.

مقایسه در سرعت اجرا: در این قسمت به مقایسه‌ی پیچیدگی زمانی روش پیشنهادی با روش SEP مورد بررسی قرار می‌گیرد، برای محاسبه‌ی پیچیدگی زمانی به روش ریاضی کار بسیار دشوار و پرخطایی می‌باشد – زیرا محاسبه‌ی پیشچیدگی تک‌تک بخش‌های الگوریتم که خودشان از زیربخش‌های مختلف و پیچیده‌ای تشکیل شده است کاری پر خطا می‌باشد؛ در اینجا ما بجای محاسبه‌ی پیچیدگی زمانی ریاضی دو الگوریتم از مدت زمانی که طول می‌کشد برنامه در سیستم اجرا و خاتمه یابد استفاده می‌کنیم. در شکل ۵ـ۵ میانگین زمانی ۲۰ اجرای مستقل برحسب میلی‌ثانیه به ازای هریک از تعداد تلاش‌ها آورده شده است. همان‌طور که در شکل ۵ـ۵ مشاهده‌ی می‌شود الگوریتم IL دارای حداکثر سرعت اجرا می‌باشد زیرا که هیچ سربار محاسباتی یادگیری مشترک را ندارد؛ هدف یادگیری اشتراکی این است که می‌خواهد در ازای یک سری سربار محاسباتی کیفیت و سرعت «یادگیری»

[1] Visual



شکل ۵ـ۵: مقایسه در سرعت اجرای روش‌ها به ازای تعداد تلاش‌های متفاوت برحسب میلی‌ثانیه با تابع بولتزمن در محیط پلکان مارپیچ

عامل‌ها را افزایش دهد. با در نظر داشتن این موضوع همانطور که قبلا دیدیم روش پیشنهادی سرعت و کیفیت یادگیری را بیشتر از روش SEP افزایش می‌دهد و در اینجا نیز می‌بینیم که دارای سرعت اجرای بیشتری نسبت به روش SEP می‌باشد که نشان از بهینه‌گی روش پیشنهادی نسبت به روش SEP می‌دهد.

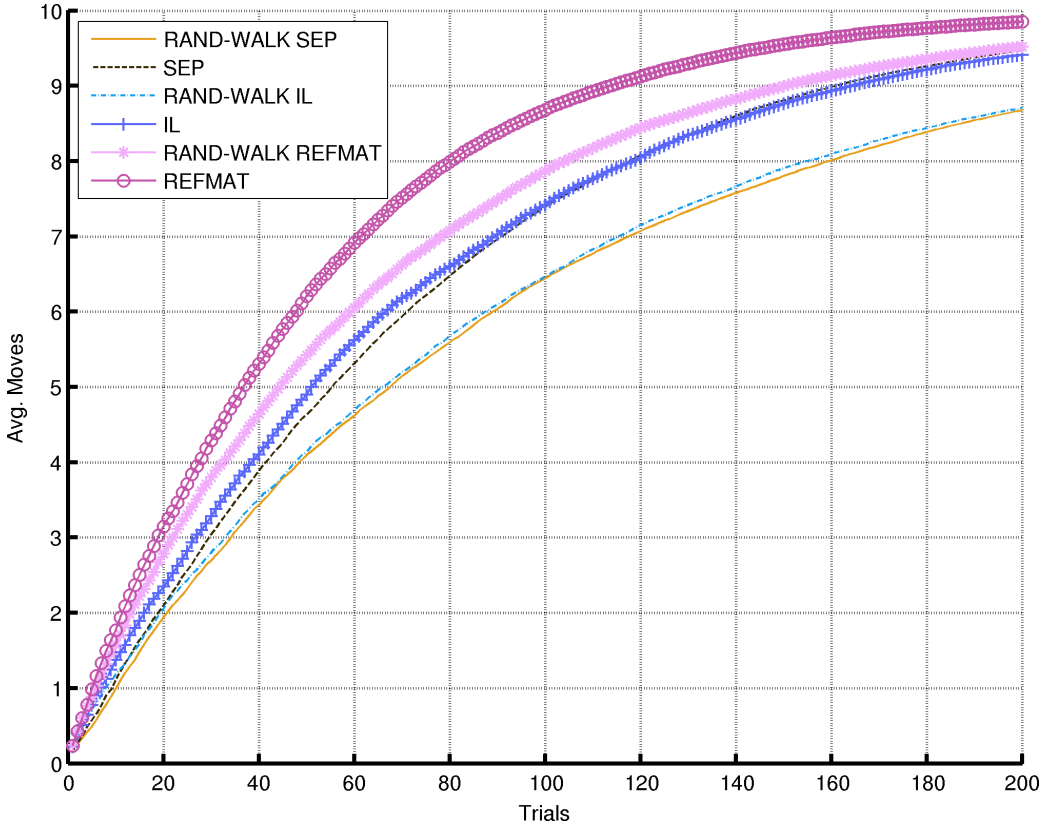
مقایسه در میزان باروری:

تعریف ۵ـ۱ (باروری)**.** اگر فرض کنیم الگوریتم یادگیری تقویتی $\psi_Q(\mathcal{E})$ وجود دارد که در محیط $\mathcal{E}$ فعالیت می کند و دانش خود را در جدولی مانند Q ذخیره می‌کند، باروری الگوریتم $\psi_Q(\mathcal{E})$ در هر لحظه را حداکثر مقدار جدول Q در آن لحظه تعریف می‌کنیم.

تعریف ۵ـ۲ (سرعت باروری)**.** اگر فرض کنیم الگوریتم یادگیری تقویتی $\psi_Q(\mathcal{E})$ وجود دارد که در محیط $\mathcal{E}$ فعالیت می کند و دانش خود را در جدولی مانند Q ذخیره می‌کند، سرعت باروری الگوریتم $\psi_Q(\mathcal{E})$ در هرلحظه را سرعت همگرایی حداکثر مقدار جدول Q به سمت حداکثر پاداش محیط قابل دریافت تعریف می‌کنیم.

تعریف ۵ـ۳ (میزان باروری)**.** انتگرال سرعت باوری را میزان باروری الگوریتم $\psi_Q(\mathcal{E})$ که در محیط $\mathcal{E}$ فعالیت می کند و دانش خود را در جدولی مانند Q ذخیره می‌کند، تعریف می‌کنیم.

فرضیه ۵ـ۱ (معیاری جدید برای سرعت یادگیری)**.** *طبق تعاریف ۵ـ۲ و ۵ـ۳ الگوریتمی میزان باروری بیشتری دارد که*



شکل ۵ـ۶: نمودار باروری الگوریتم‌ها مختلف با تابع بولتزمن در محیط پلکان مارپیچ

سریع‌تر مقادیر جدول Q خود را به سمت بیشنه مقداری که می‌توانند داشته باشد سوق دهد. معمولا در الگوریتم‌های یادگیری تقویتی Q این کار با تنظیم مقدار سرعت یادگیری α صورت می‌گیرد که باعث می‌شود الگوریتم‌ها با سرعت بیشتری به یادگیری نحوه‌ی تعامل با محیط بپردازند. لذا در شرایط یکسان می‌توان گفت الگوریتمی بهتر عمل می‌کند که نحوه‌ی تعامل با محیط را سریع‌تر نسبت به دیگر الگوریتم‌ها یاد می‌گیرید و میزان باروری بیشتری داشته باشد.

در شکل ۵ـ۶ آورده شده است حداکثر میزان جدول Q روش‌ها در هر تلاش آورده شده است. همانطور که قبلا در تعریف محیط پلکان مارپیچ آورده شده است حداکثر مقدار پاداش این محیط مقدار ۱۰ می‌باشد لذا همان‌طور که مشاهده می‌شود الگوریتم‌ها با مساحت‌های زیر نمودار متفاوتی مقادیر جدول Q خود را به سمت حداکثر مقدار پاداش قابل دریافت از محیط سوق می‌دهد که نشان از میزان باروری الگوریتم‌ها می‌دهد.

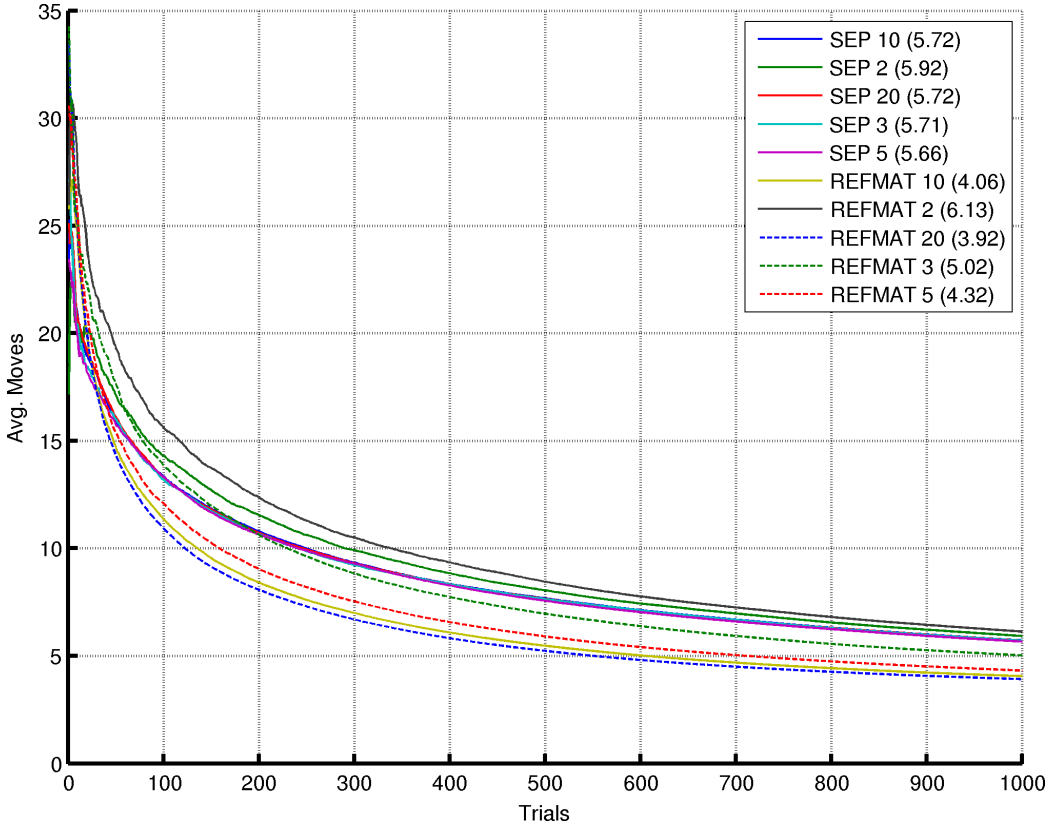
در شکل ۵ـ۶ منظور از RAND-WALK حرکت کاملا تصادفی می‌باشد، به این صورت که عامل بعد از هر حرکت جدول Q خود را بروز رسانی می‌کند ولی هنگام انتخاب عمل در تابع بولتزمن مقدار $\tau \to +\infty$ در نظر گرفته می‌شود تا میزان احتمال تمامی حرکت‌ها یکسان شود و در نتیجه حرکتی به صورت تصادفی انتخاب شود. همانطور که در قسمت‌های قبل دیدیم روش پیشنهادی هم در کیفیت و هم در سرعت یادگیری بهبود چشم‌گیری دارد و از طرفی هم در نمودار ۵ـ۶ دارای بیشترین میزان باروری (مساحت زیرنمودار) حداکثر مقدار جدول Q می‌باشد که این مساله تایید کننده‌ی فرضیه ۵ـ۱ می‌باشد.

دلیل وجود نتایج آزمایش اجرای RAND-WALK در این قسمت این است که بررسی کنیم در صورتی که عامل بصورت کورکورانه حرکت کند روش معرفی شده و SEP چقدر در میزان بارور شدن جدول Q عامل‌ها موثرند؟ به عبارت دیگر، در صورتی که استراتژی خاصی جهت انتخاب عمل وجود نداشته باشد، روش‌ها چقدر قدرت باروری دارند؟ همانطور که در شکل ۵ـ۶ مشاهده می‌کنیم روش معرفی شده در زمانی که به صورت تصادفی اقدام به انتخاب عمل می‌کند بیشتر از زمانی که IL با استفاده از تابع بولتزمن اقدام به انتخاب عمل می‌کند جدول Q را بارور می‌کند که از قدرت روش ارائه شده خبر می‌دهد. همچنین در مورد روش SEP می‌بینیم که در زمانی که بصورت تصادفی اقدام به عمل می‌کند باروری کمتری نسبت به روش پیشنهادی و IL دارد؛ یعنی میزان باروری روش SEP وابستگی زیادی به سیاست انتخاب عمل دارد و در صورت نداشتن سیاست انتخاب عمل خاصی بشدت عملکردش کاسته می‌شود ولی در روش پیشنهادی میزان این وابستگی از شدت کمتری برخوردار است که از دیگر امتیازها مثبت روش پیشنهادی می‌باشد.

مقایسه تاثیر تعداد عامل‌ها بر میزان کیفیت و سرعت یادگیری: در این مقایسه سعی شده است که تاثیر یک فاکتور بنیادی سیستم‌های چندعامله مشارکتی را مورد بررسی قرار دهیم، و آن میزان تاثیر پذیری روش‌های مورد مقایسه با افزایش تعداد عامل‌ها می‌باشد. در تئوری سیستم‌های چندعامله مشارکتی دیدگاه معقول براین است که اثر تعداد عامل‌ها در کیفیت و سرعت یادگیری مشارکتی باید مثبت باشد. درغیر این صورت سیستم‌های چندعامله‌ای که تعداد عامل‌ها تاثیری در خروجی سیستم نداشته باشد، دیگر ماهیت سیستم‌های چند عامله را ندارد.

همان‌طور که در شکل ۵ـ۷ آمده است، روش پیشنهادی و روش SEP به ازای تعداد عامل‌های ۲، ۳، ۵، ۱۰ و ۲۰ عدد به تعداد ۲۰ بار اجرا درآمده و میانگین اجراها به نمودار کشیده شده است. همانطور که می‌بینیم روش SEP در زمانی که ۲۰ عامل در حال یادگیری و اشتراک گذاری دانش‌های خود هستند نسبت به زمانی که فقط ۲ عامل در حال تعامل مشارکتی با محیط هستند فقط ٪۳ در خروجی الگوریتم تاثیر مثبت داشته است. این در حالی است که در همین شرایط میزان بهبود نتیجه‌ی روش پیشنهادی ٪۵۶ می‌باشد. که نشان می‌دهد روش SEP نسبت به افزایش تعداد عامل‌ها رفتاری تقریبا خنثی از خود نشان می‌دهد درحالی که روش پیشنهادی در ازای افزایش تعداد عامل‌ها به دلیل اینکه دانش جمعی نیز افزایش می‌یابد کیفیت خروجی آن نیز بهتر می‌شود.

نتیجه‌گیری: نتیجه‌ای که از مقایسه‌ی روش پیشنهادی در هر چهار مقایسه‌ی بالا می‌توان گرفت این است که روش پیشنهادی بهبود چشم‌گیری به روش SEP در محیط پلکان مارپیچ و سیاست انتخاب عمل بولتزمن داده است.



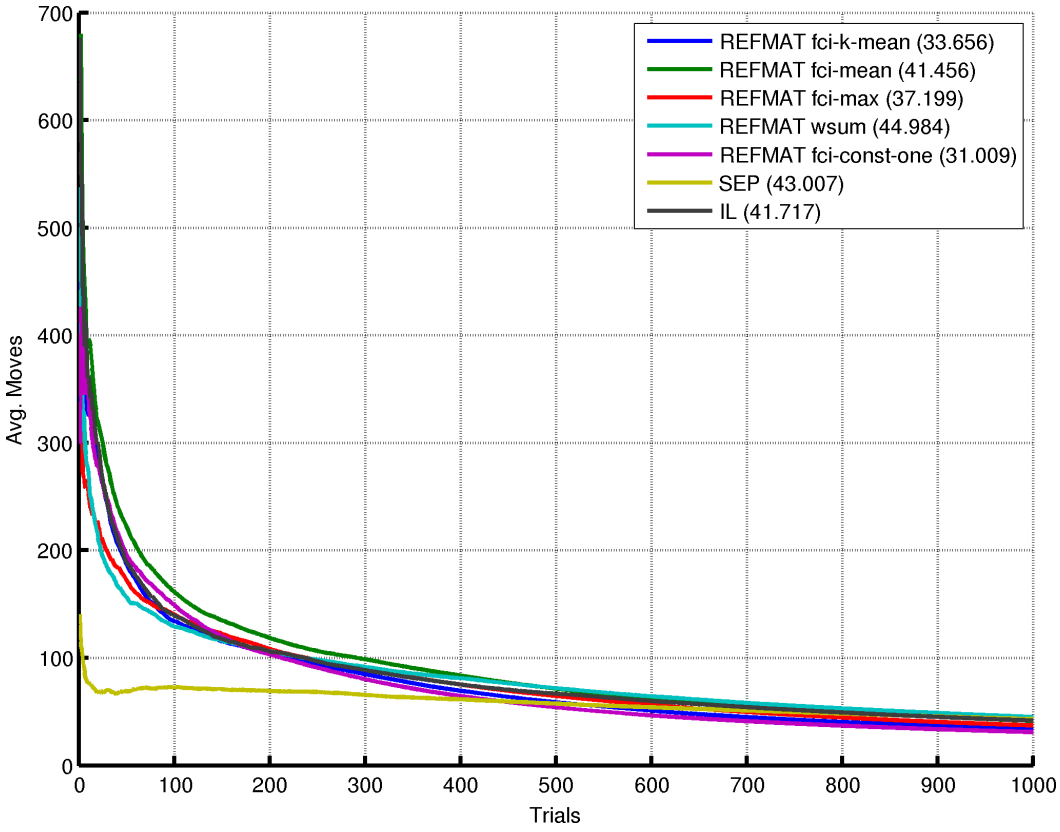
شکل ۵ـ۷: مقایسه تاثیر تعداد عامل‌ها میزان کیفیت و سرعت یادگیری با تابع بولتزمن در محیط پلکان مارپیچ

سیاست انتخاب عمل «εـحریصانه»

مقایسه در سرعت و کیفیت یادگیری: نتایج حاصل از اجرای الگوریتم‌ها در محیط پلکان مارپیچ در شکل ۵ـ۸ آمده است. شرایط این آزمایش مشابه با شرایط آزمایش با تابع بولتزمن می‌باشد.

همانطور که مشاهده می‌شود روش SEP دارای ۳-٪ بهبود نسبت به IL می‌باشد در حالی که روش پیشنهادی در زمانی که از انتگرال فازی استفاده می‌کند در بدترین حالت دارای ۰/۶٪ بهبود و در بهترین حالات دارای ۳۴٪ بهبود می‌باشد که نسبت به روش SEP تقریبا ۴ الی ۳۸ برابر نتیجه را بهبود داده است. در صورتی که از میانگین وزنی بجای انتگرال فازی استفاده شود نتایج با اختلافی حدود ۷-٪ بدتر از یادگیری IL بوده است که نشان می‌دهد که استفاده از انتگرال فازی چقدر می‌تواند نسبت به روش‌های سنتی و معمولی چون میانگین وزنی موثر واقع شود. البته در شکل ۵ـ۸ باید توجه کرد که روش SEP در همان ابتدای کار خود به شدت میانگین حرکت عامل‌ها را کاهش داده ولی به دلیل ماهیت الگوریتم SEP اشباع جداول الگوریتم توانایی ادامه‌ی سرشکن کردن بیشتر میانگین حرکت عامل‌ها را ندارد. میانگین نتایج این قسمت را می‌توان در جدول ۵ـ۳ خلاصه کرد.

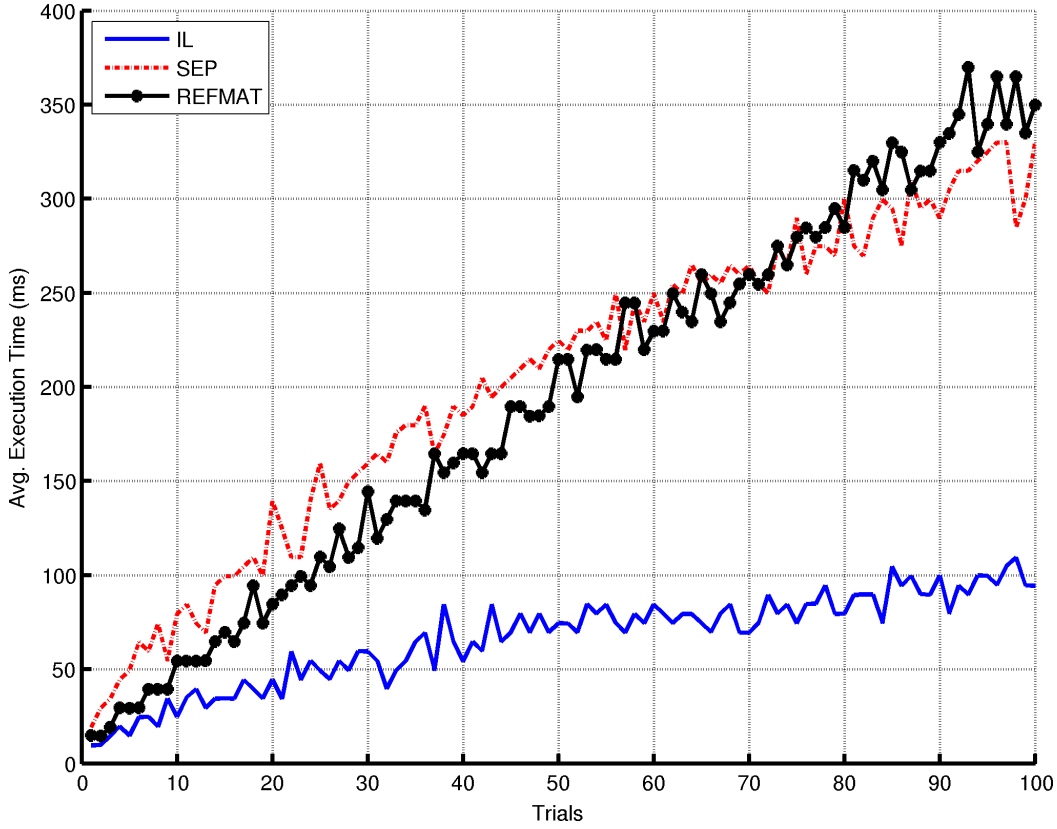
مقایسه در سرعت اجرا: در شکل ۵ـ۹ میانگین زمانی ۲۰ اجرای مستقل برحسب میلی‌ثانیه به ازای هریک از تعداد تلاش‌ها آورده شده است. همان‌طور که در این شکل مشاهده‌ی می‌شود الگوریتم IL دارای حداکثر سرعت



شکل ۵ـ۸: مقایسه در سرعت و کیفیت یادگیری با تابع εـحریصانه در محیط پلکان مارپیچ

جدول ۵ـ۳: مقایسه در میزان درصد بهبود کیفیت یادگیری در محیط پلکان مارپیچ با تابع εـحریصانه

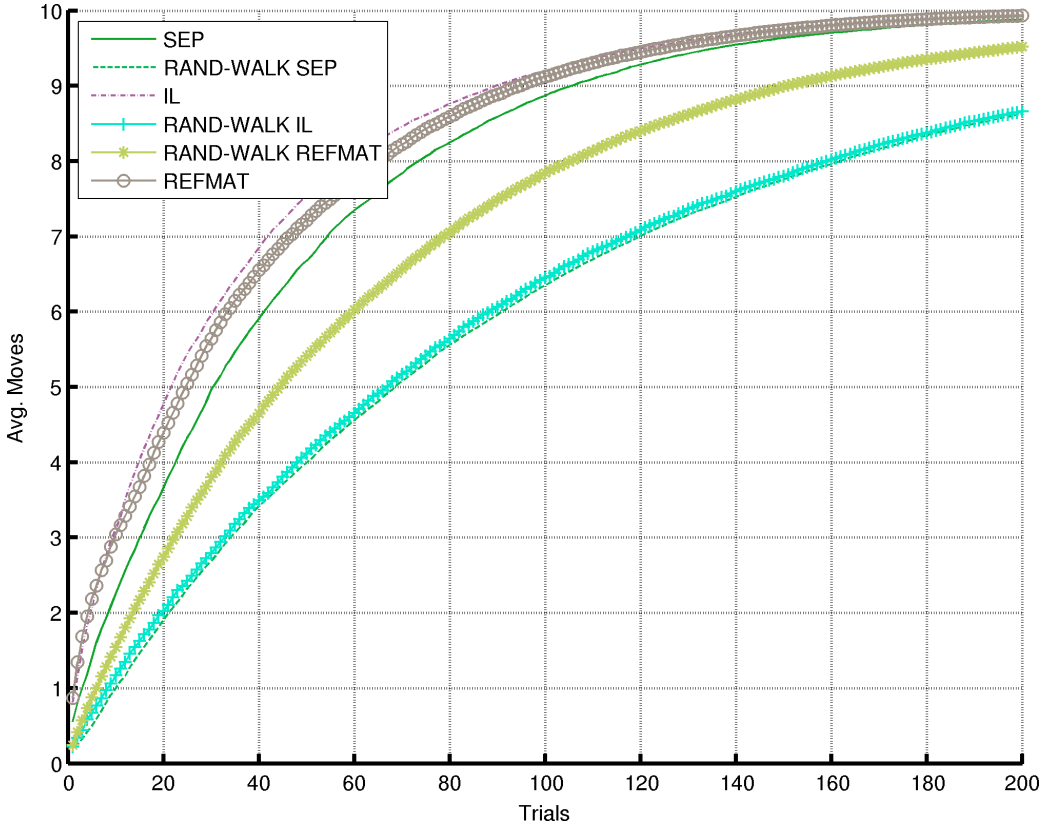
			REFMAT				
	IL	SEP	wsum	fci-mean	fci-max	fci-k-mean	fci-const-one
IL	0.0						
SEP	-3.0	0.0					
wsum	-7.3	-4.4	0.0				
fci-mean	0.6	3.7	8.5	0.0			
fci-max	12.2	15.6	20.9	11.5	0.0		
fci-k-mean	24.0	27.8	33.7	23.2	10.5	0.0	
fci-const-one	34.5	38.7	45.1	33.7	20.0	8.5	0.0



شکل ۵ـ۹: مقایسه در سرعت اجرای روش‌ها به ازای تعداد تلاش‌های متفاوت برحسب میلی‌ثانیه با تابع εـحریصانه در محیط پلکان مارپیچ

اجرا می‌باشد زیرا که هیچ سربار محاسباتی یادگیری مشترک را ندارد؛ هدف یادگیری اشتراکی این است که می‌خواهد در ازای یک سری سربار محاسباتی کیفیت و سرعت «یادگیری» عامل‌ها را افزایش دهد. با در نظر داشتن این موضوع همانطور که قبلا دیدیم روش پیشنهادی سرعت و کیفیت یادگیری را بیشتر از روش SEP افزایش می‌دهد و در اینجا نیز می‌بینیم که دارای سرعت اجرای بیشتری نسبت به روش SEP می‌باشد که نشان از بهینه‌گی روش پیشنهادی نسبت به روش SEP می‌دهد.

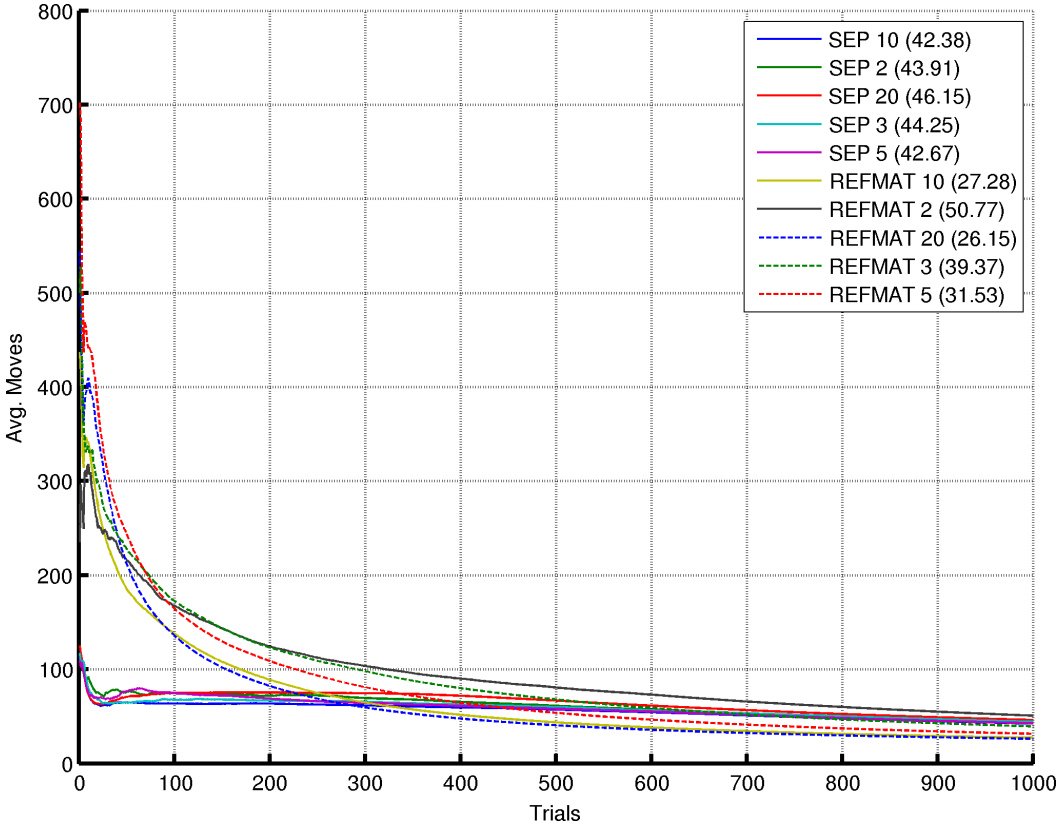
مقایسه در میزان باروری: در شکل ۵ـ۱۰ میزان باروری IL از کلیه‌ی روش‌ها بهتر بوده (با اندک اختلاف نسبت روش پیشنهادی) ولی همچنان باروری روش پیشنهادی از روش SEP بیشتر بوده است و همچون آزمایش مشابه با تابع بولتزمن در اینجا نیز نشان داده شده است که روش SEP کاملا وابسته است به این‌که در هنگام انتخاب عمل بر اساس دانش عامل عمل شود و اگر عامل بدون درنظر گرفتن دانش عامل حرکتی اتخاذ کند میزان باروری عامل بشدت تحت تاثیر قرار می‌گیرد در حالی که در روش پیشنهادی در شرایط یکسان از کلیه الگوریتم‌ها میزان باروری بیشتری دارد.



شکل ۵ـ۱۰: نمودار باروری الگوریتم‌ها مختلف با تابع εـحریصانه در محیط پلکان مارپیچ

مقایسه تاثیر تعداد عامل‌ها میزان کیفیت و سرعت یادگیری: همان‌طور که در شکل ۵ـ۱۱ آمده است، روش پیشنهادی و روش SEP به ازای تعداد عامل‌های ۲، ۳، ۵، ۱۰ و ۲۰ عدد به تعداد ۲۰ بار اجرا درآمده و میانگین اجراها به نمودار کشیده شده است. همانطور که می‌بینیم روش SEP در زمانی ۲۰ عامل در حال یادگیری و اشتراک گذاری دانش‌های خود هستند نسبت به زمانی که فقط ۲ عامل در حال تعامل مشارکتی با محیط هستند ۸-٪ در خروجی الگوریتم تاثیر منفی داشته است؛ بدین معنی که در زمانی که از تابع εـحریصانه استفاده شود روش SEP به افزایش تعداد عامل فقط منجر به بدتر شدن عملکرد عامل‌ها در یادگیری مشارکتی می‌شود. این در حالی است که در همین شرایط میزان بهبود نتیجه‌ی روش پیشنهادی ۹۲٪ می‌باشد که نشان می‌دهد روش پیشنهادی در ازای افزایش تعداد عامل‌ها به دلیل اینکه دانش جمعی نیز افزایش می‌یابد کیفیت خروجی آن نیز بطور چشم‌گیری بهتر می‌شود. در حالی که در روش SEP اگر کار نتایج بدتر نشود بهتر نمی‌شود که از ضعف بزرگ روش SEP خبر می‌دهد.

نتیجه‌گیری: نتیجه‌ای که از مقایسه‌ی روش پیشنهادی در هر چهار مقایسه‌ی بالا می‌توان گرفت همچون نتیجه‌ای که از نتایج تابع بولتزمن، روش پیشنهادی بهبود چشم‌گیری به روش SEP در محیط پلکان مارپیچ و سیاست انتخاب عمل εـحریصانه داده است.



شکل ۵ـ۱۱: مقایسه تاثیر تعداد عامل‌ها میزان کیفیت و سرعت یادگیری با تابع εـحریصانه در محیط پلکان مارپیچ

جدول ۵ـ۴: مقایسه در سرعت و کیفیت یادگیری نسبت کیفیت نتیجه‌ی حاصل از تابع εـحریصانه نسبت به تابع بولتزمن

		Boltzmann	
		SEP	REFMAT
ε-greedy	SEP	7.27	9.42
	REFMAT	5.20	6.79

مقایسه‌ی بین نتایج حاصل از سیاست انتخاب عمل بولتزمن و εـحریصانه

در حالت کلی در محیط پلکان مارپیچ تابع بولتزمن نتایج یکنواتر و پایدارتری[1] نسبت به تابع εـحریصانه از خود نشان داد و در هر دوی این توابع روش پیشنهادی نتیجه‌ی بهتری نسبت به روش SEP ارائه داد. در این قسمت به مقایسه‌ی نتایج بدست آمده توسط هر دو روش در هر دو سیاست انتخاب عمل می‌پردازیم.

مقایسه در سرعت و کیفیت یادگیری: مقایسه‌ی این قسمت را بطور خلاصه می‌توان در جدول ۵ـ۴ مشاهده کرد. که نسبت کیفیت نتیجه‌ی حاصل از تابع εـحریصانه نسبت به تابع بولتزمن همگی بزرگتر از ۱ می‌باشد، که نشان می‌دهد که استفاده از تابع εـحریصانه در کیفیت خروجی تاثیری منفی دارد.

[1] Stable

جدول ۵ـ۵: مقایسه در نسبت میانگین سرعت اجرای حاصل از استفاده تابع εـحریصانه نسبت به تابع بولتزمن

		Boltzmann		
		SEP	REFMAT	IL
ε-greedy	SEP	1.64	2.05	10.23
	REFMAT	1.72	2.15	10.73
	IL	0.56	0.70	3.49

جدول ۵ـ۶: مقایسه در نسبت میزان باروری حاصل از استفاده تابع εـحریصانه نسبت به تابع بولتزمن

		Boltzmann		
		SEP	REFMAT	IL
ε-greedy	SEP	1.08	1.25	1.23
	REFMAT	1.03	1.20	1.18
	IL	1.09	1.27	1.25

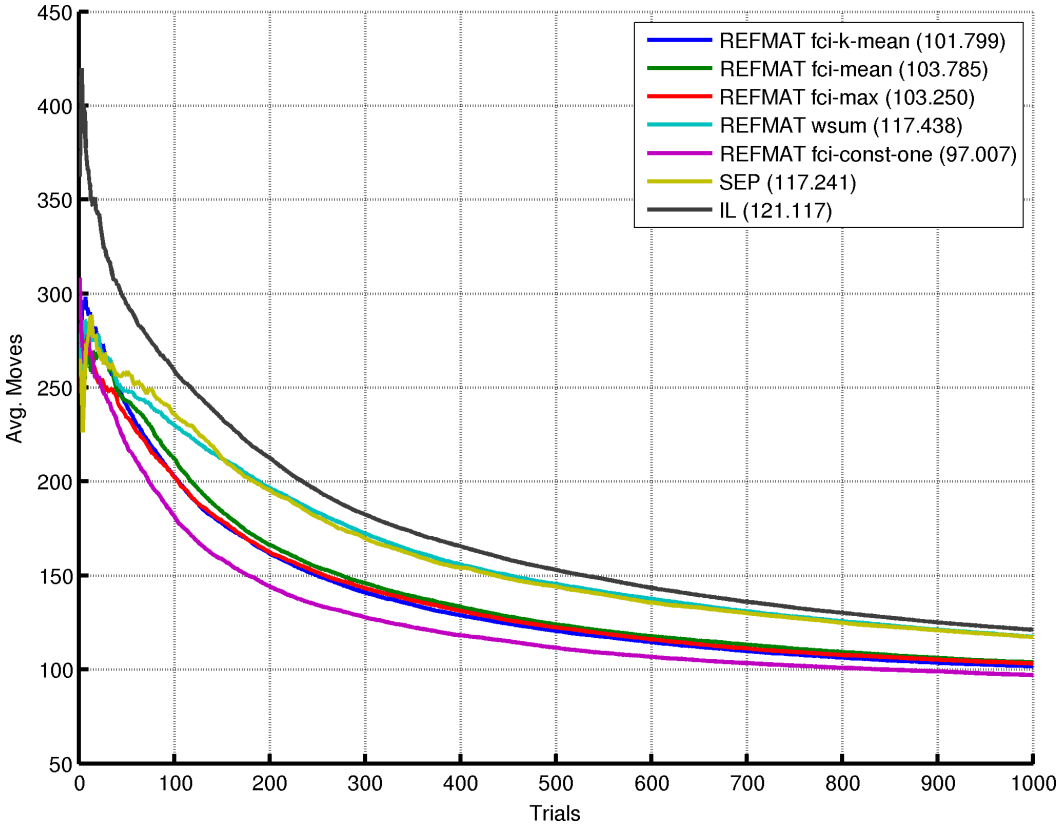
مقایسه در سرعت اجرا: در جدول ۵ـ۵ نسبت میانگین سرعت اجرای روش‌ها آمده است، قطر اصلی این جدول همگی مقادیر بزرگتر از ۱ دارد که نشان می‌دهد هر روش در زمانی که از تابع εـحریصانه استفاده می‌کند زمان بیشتری را تلف می‌کند(صرف جستجوی بی‌مورد محیط می‌کند) نسبت به زمانی که از تابع بولتزمن استفاده می‌کند. این مساله نشان می‌دهد که تابع بولتزمن سریع‌تر عامل را به سمت اهداف هدایت می‌کند – که این نکته در قسمت «مقایسه‌ی سرعت و کیفیت یادگیری» نیز قابل استنتاج است.

مقایسه در میزان باروری: همانطور که در جدول ۵ـ۶ آمده است همه‌ی مقادیر نسبت‌ها بیشتر از ۱ می‌باشد که بدین معنی است که استفاده از تابع εـحریصانه با این حال که کیفیت و سرعت یادگیری کمتری نسبت به تابع بولتزمن دارد و عامل‌ها در حالت کلی زمان زیادی صرف گشت و گذار در محیط می‌کند به نسبت باعث باروری بیشتر جدول Q می‌شود.

مقایسه تاثیر تعداد عامل‌ها بر میزان کیفیت و سرعت یادگیری: در جدول ۵ـ۷ نسبت شیب تاثیر تعداد عامل‌ها میزان کیفیت نتیجه‌ی حاصل از تابع εـحریصانه نسبت به تابع بولتزمن آمده است؛ همانطور که مشاهده می‌شود در زمانی که از تابع εـحریصانه استفاده می‌شود در روش پیشنهادی تاثیر تعداد عامل‌ها به مراتب بیشتر از زمانی است که از تابع بولتزمن استفاده می‌کنیم. این در حالی می‌باشد که در روش SEP اضافه کردن عامل‌ها به محیط تفاوت زیادی در دانش خروجی الگوریتم در هر دو تابع ایجاد نمی‌کند.

جدول ۵ـ۷: مقایسه نسبت شیب تاثیر تعداد عامل‌ها میزان کیفیت نتیجه‌ی حاصل از تابع εـحریصانه نسبت به تابع بولتزمن

		Boltzmann	
		SEP	REFMAT
ε-greedy	SEP	0.59	0.09
	REFMAT	73.02	10.67



شکل ۵ـ۱۲: مقایسه در سرعت و کیفیت یادگیری در محیط صید و صیاد با تابع بولتزمن در محیط صید و صیاد

۵ـ۳ـ۲ مقایسه در محیط صید و صیاد

همانند مقایسه در محیط پلکان مارپیچ، آزمایش‌های مربوط به این قسمت در ۴ بخش صورت گرفته است؛ ۱. مقایسه در سرعت و کیفیت یادگیری، ۲. مقایسه در پیچیدگی زمانی، ۳. مقایسه در میزان باروری، ۴. مقایسه تاثیر تعداد عامل‌ها بر میزان کیفیت و سرعت یادگیری.

سیاست انتخاب عمل «بولتزمن»

مقایسه در سرعت و کیفیت یادگیری: نتایج حاصل از اجرای الگوریتم‌ها در محیط صید و صیاد در شکل ۵ـ ۱۲ آمده است. همانطور که مشاهده می‌شود روش SEP دارای ۳٪ بهبود نسبت به IL می‌باشد ولی همانطور که مشاهده می‌کنیم روش پیشنهادی بهبود چشم‌گیری نسبت به روش SEP دارد و در بدترین حالت در صورت

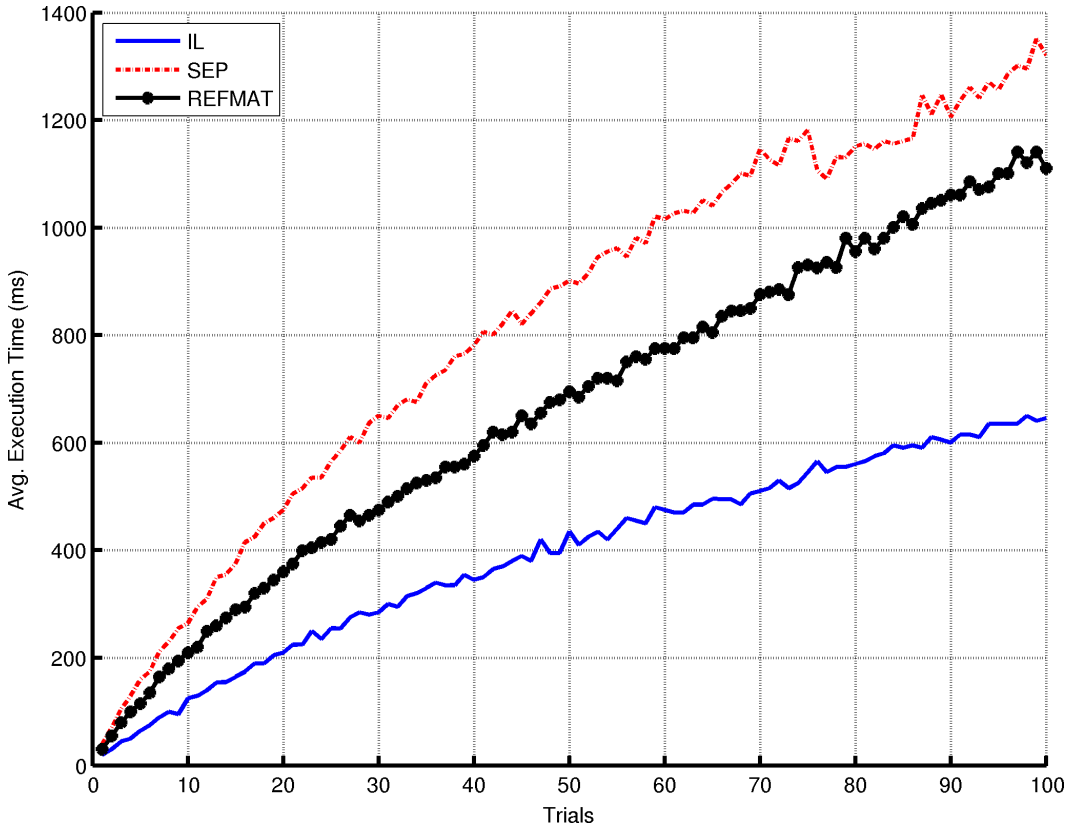
جدول ۵ـ۸: مقایسه در میزان درصد بهبود کیفیت یادگیری در محیط صید و صیاد با تابع بولتزمن

			REFMAT				
	IL	SEP	wsum	fci-mean	fci-max	fci-k-mean	fci-const-one
IL	0.0						
SEP	3.3	0.0					
wsum	3.1	-0.2	0.0				
fci-mean	16.7	13.0	13.2	0.0			
fci-max	17.3	13.5	13.7	0.5	0.0		
fci-k-mean	19.0	15.2	15.4	2.0	1.4	0.0	
fci-const-one	24.9	20.9	21.1	7.0	6.4	4.9	0.0

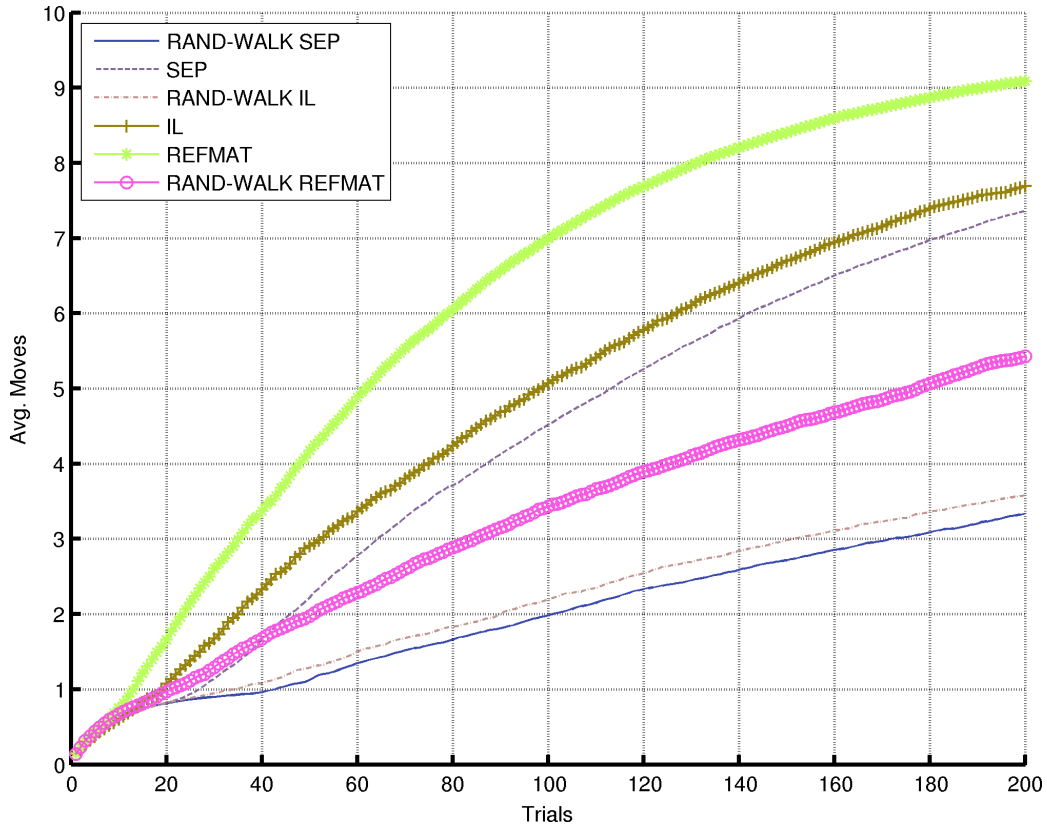
استفاده از تابع Mean به عنوان مدل‌کننده‌ی تابع $g(\cdot)$ حدود ۱۷٪ نتایج بهبود می‌یابد و در صورتی‌که بخواهیم با استفاده از تابع Const-One به صورت حریصانه عمل کنیم (یعنی در هنگام ترکیب دانش عامل‌ها، فقط دانش عاملی را درنظر بگیریم که با توجه به معیار خبرگی دارای بیشترین خبرگی می‌باشد) مطابق نتایج بدست آمده در محیط قبلی بهترین نتیجه‌ی ممکن یعنی حدود ۲۵٪ بهبود را بدست می‌آوریم؛ نتایج این قسمت را می‌توان در جدول ۵ـ۸ خلاصه کرد.

مقایسه در سرعت اجرا: در شکل ۵ـ۱۳ همانند شرایط طرح شده در محیط پلکان مارپیچ، سرعت اجرای الگوریتم‌ها آورده شده است. نتایج بدست آمده در محیط صید و صیاد همانند محیط پلکان مارپیچ نشان می‌دهد که روش پیشنهادی نسبت به روش SEP از سرعت اجرای بیشتری برخوردار است.

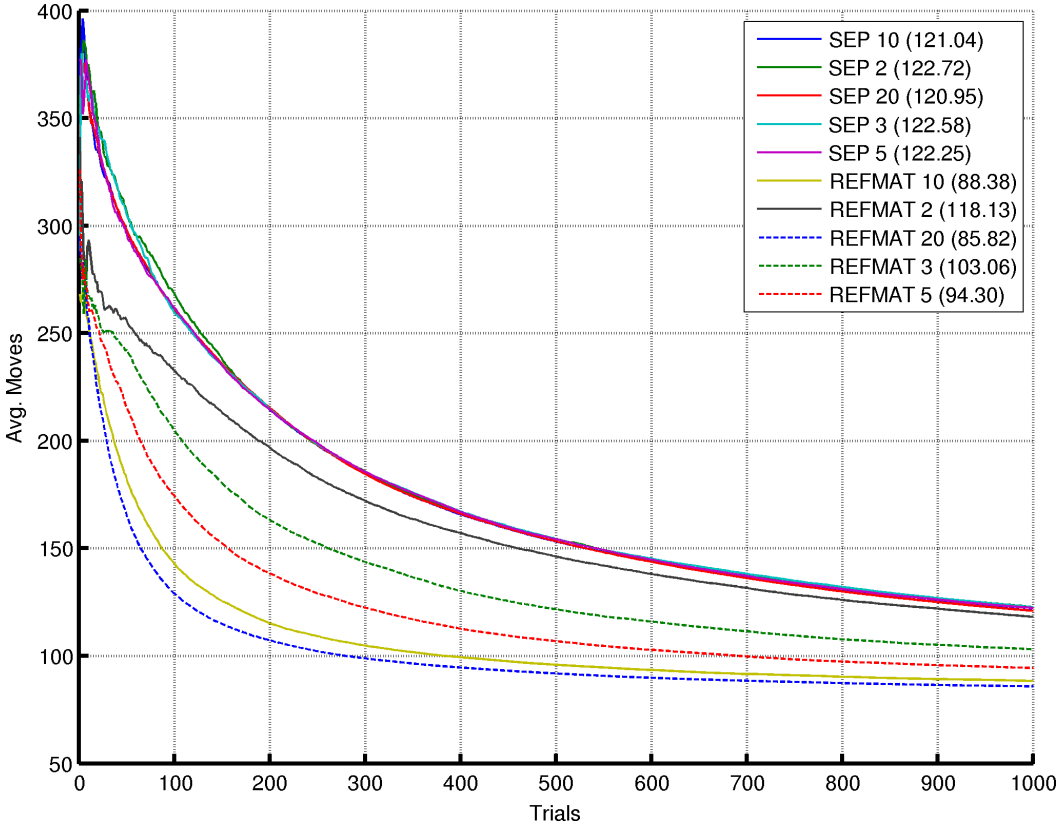
مقایسه در میزان باروری: همانطور که در شکل ۵ـ۱۴ مشاهده می‌کنیم روش معرفی شده در زمانی که به صورت تصادفی اقدام به انتخاب عمل می‌کند بیشتر از زمانی که IL و SEP با بصورت تصادفی اقدام به انتخاب عمل می‌کند جدول Q را بارور می‌کند که از قدرت روش ارائه شده خبر می‌دهد. همچنین در مورد روش SEP می‌بینیم که در زمانی که بصورت تصادفی اقدام به عمل می‌کند باروری کمتری نسبت به روش پیشنهادی و IL دارد؛ یعنی میزان باروری روش SEP وابستگی زیادی به سیاست انتخاب عمل دارد و در صورت نداشتن سیاست انتخاب عمل خاصی بشدت عملکردش کاسته می‌شود ولی در روش پیشنهادی میزان این وابستگی از شدت کمتری برخوردار است که از دیگر امتیازات مثبت روش پیشنهادی می‌باشد – همانند نتایج حاصله در محیط پلکان مارپیچ.



شکل ۵ـ۱۳: مقایسه در سرعت اجرای روش‌ها به ازای تعداد تلاش‌های متفاوت برحسب میلی‌ثانیه با تابع بولتزمن در محیط صید و صیاد



شکل ۵ـ۱۴: نمودار باروری الگوریتم‌ها مختلف با تابع بولتزمن در محیط صید و صیاد



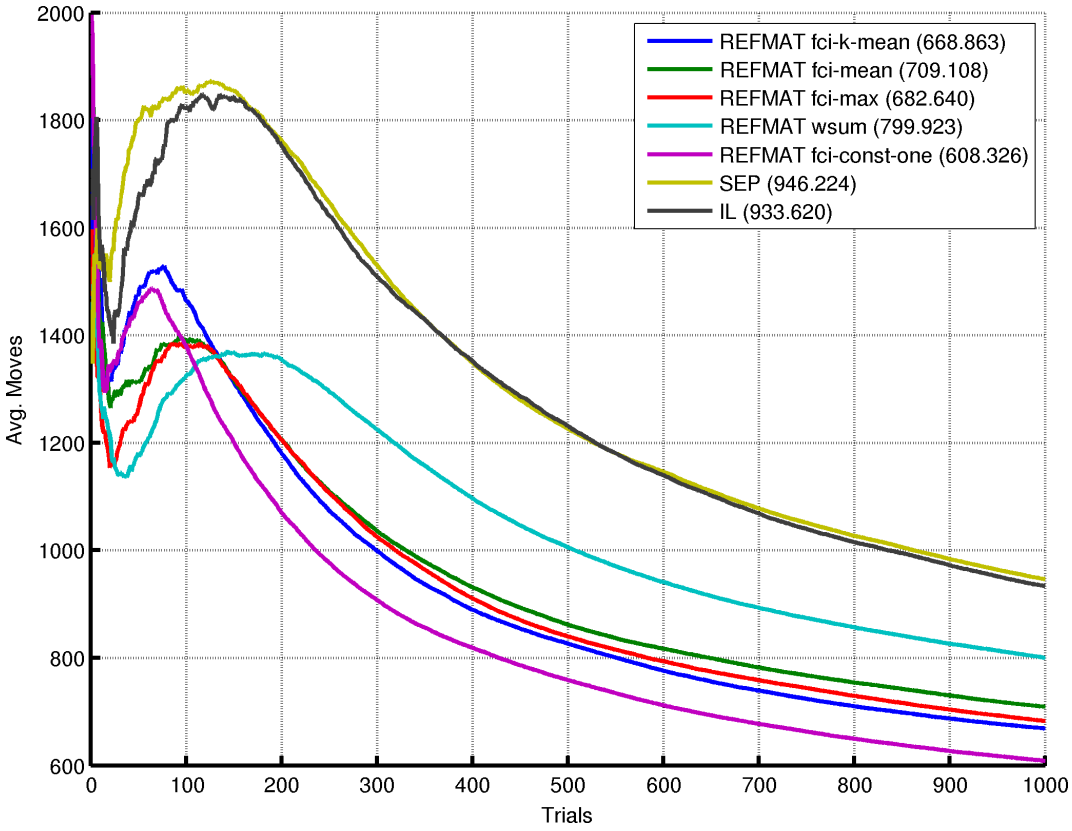
شکل ۵ـ۱۵: مقایسه تاثیر تعداد عامل‌ها میزان کیفیت و سرعت یادگیری با تابع بولتزمن در محیط صید و صیاد

مقایسه تاثیر تعداد عامل‌ها میزان کیفیت و سرعت یادگیری: همان‌طور که در شکل ۵ـ۱۵ آمده است، روش SEP در زمانی ۲۰ عامل در حال یادگیری و اشتراک گذاری دانش‌های خود هستند نسبت به زمانی که فقط ۲ عامل در حال تعامل مشارکتی با محیط هستند فقط ۲٪ در خروجی الگوریتم تاثیر مثبت داشته است. این در حالی است که در همین شرایط میزان بهبود نتیجه‌ی روش پیشنهادی ۳۸٪ می‌باشد. که نشان می‌دهد روش SEP نسبت به افزایش تعداد عامل‌ها رفتاری تقریبا خنثی از خود نشان می‌دهد درحالی که روش پیشنهادی در ازای افزایش تعداد عامل‌ها به دلیل اینکه دانش جمعی نیز افزایش می‌یابد کیفیت خروجی آن نیز بهتر می‌شود – دلایل و شهود این مساله همانند شهود مطرح شده در محیط صید و صیاد می‌باشد.

نتیجه‌گیری: نتیجه‌ای که از مقایسه‌ی روش پیشنهادی در هر چهار مقایسه‌ی بالا می‌توان گرفت این است که روش پیشنهادی بهبود چشم‌گیری به روش SEP در محیط صید و صیاد و سیاست انتخاب عمل بولتزمن داده است.

سیاست انتخاب عمل «εـ حریصانه»

مقایسه در سرعت و کیفیت یادگیری: نتایج حاصل از اجرای الگوریتم‌ها در محیط صید و صیاد در شکل ۵ـ۱۶ آمده است. شرایط این آزمایش به مشابه شرایط آزمایش با تابع بولتزمن می‌باشد.

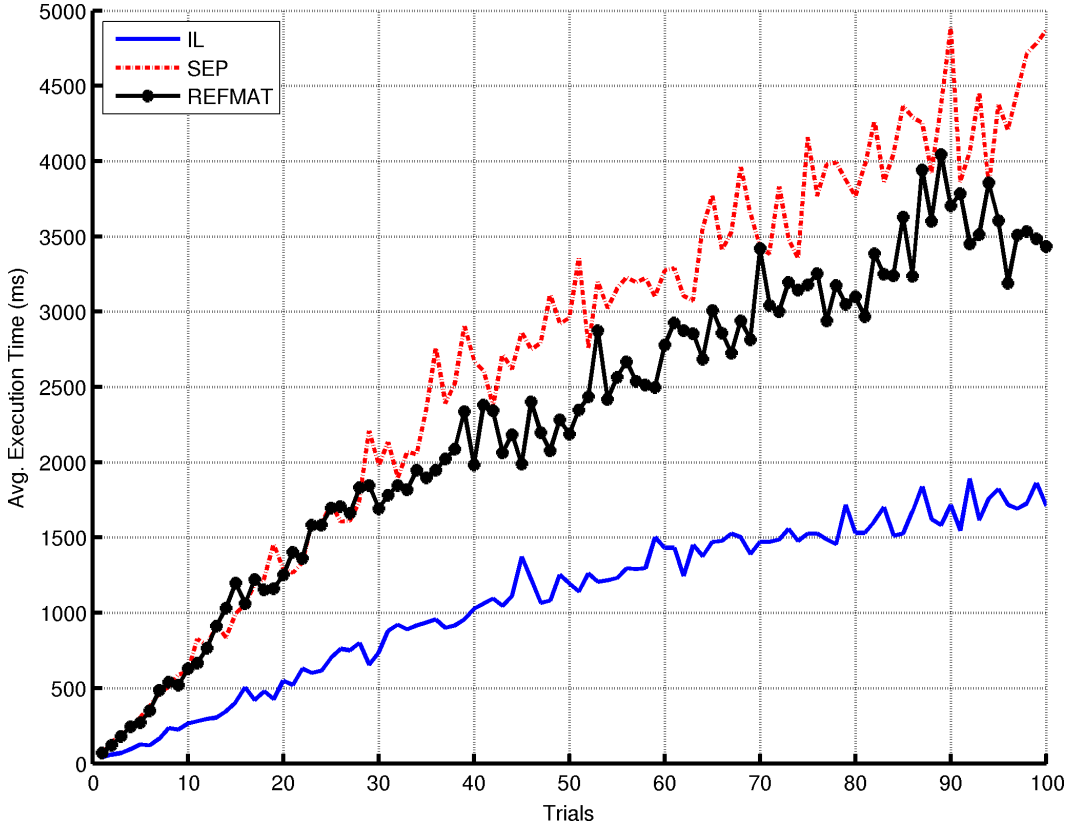


شکل ۵ـ۱۶: مقایسه در سرعت و کیفیت یادگیری با تابع εـحریصانه در محیط صید و صیاد

جدول ۵ـ۹: مقایسه در میزان درصد بهبود کیفیت یادگیری در محیط صید و صیاد با تابع εـحریصانه

			REFMAT				
	IL	SEP	wsum	fci-mean	fci-max	fci-k-mean	fci-const-one
IL	0.0						
SEP	-1.3	0.0					
wsum	16.7	18.3	0.0				
fci-mean	31.7	33.4	12.8	0.0			
fci-max	36.8	38.6	17.2	3.9	0.0		
fci-k-mean	39.6	41.5	19.6	6.0	2.1	0.0	
fci-const-one	53.5	55.5	31.5	16.6	12.2	10.0	0.0

همانطور که خلاصه‌ی این نتایج را در جدول ۵ـ۹ مشاهده می‌کنیم می‌بینیم که همانند نتایج بدست آمده در آزمایش‌های قبلی روش پیشنهادی با حدود ۵۳٪ بهبود نسبت به IL داشته است درحالی که روش SEP با تابع εـحریصانه نه تنها نتایج بهبود داده نشده است بلکه حدود ۱٪ـ بدتر شده است!

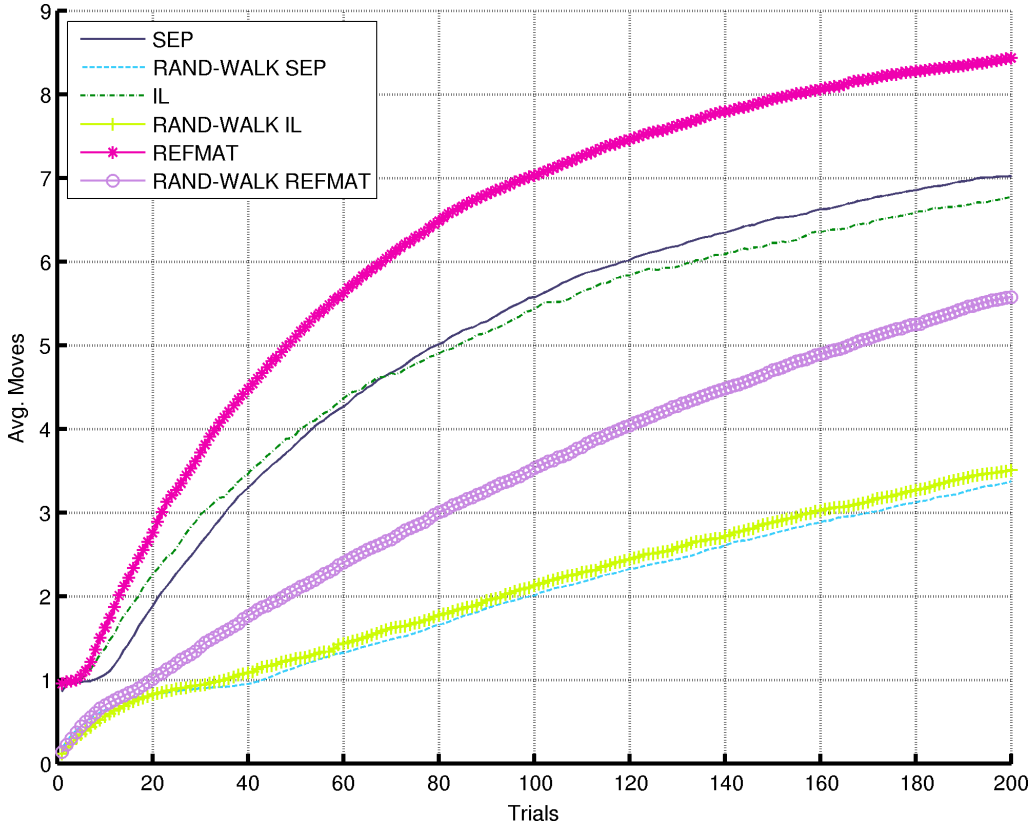


شکل ۵ـ۱۷: مقایسه در سرعت اجرای روش‌ها به ازای تعداد تلاش‌های متفاوت برحسب میلی‌ثانیه با تابع εـحریصانه در محیط صید و صیاد

مقایسه در سرعت اجرا: در شکل ۵ـ۱۷ نیز می‌بینیم که در محیط صید و صیاد نیز روش پیشنهادی دارای سرعت اجرای بیشتری نسبت به روش SEP می‌باشد که نشان از بهینه‌گی روش پیشنهادی نسبت به روش SEP می‌دهد.

مقایسه در میزان باروری: در شکل ۵ـ۱۸ میزان باروری روش پیشنهادی از دیگر روش‌ها بیشتر بوده و همانند آزمایش‌های قبلی در اینجا نیز نشان داده شده است که روش SEP در زمانی که به صورت تصادفی محیط را کاوش کند کمترین باروری را دارد که مطابق دلایل ذکر شده در مقایسه‌ی میزان باروری در آزمایش‌های گذشته این مساله نشان از ضعف بزرگ روش SEP می‌دهد.

مقایسه تاثیر تعداد عامل‌ها میزان کیفیت و سرعت یادگیری: همان‌طور که در شکل ۵ـ۱۹ آمده است، روش SEP در زمانی ۲۰ عامل در حال یادگیری و اشتراک گذاری دانش‌های خود هستند نسبت به زمانی که فحقط ۲ عامل در حال تعامل مشارکتی با محیط هستند ۹۰٪ در خروجی الگوریتم تاثیر منفی داشته است؛ بدین معنی که در زمانی که از تابع εـحریصانه استفاده شود روش SEP به افزایش تعداد عامل فقط منجر به بدتر شدن عملکرد عامل‌ها در یادگیری مشارکتی می‌شود. این در حالی است که در همین شرایط میزان بهبود نتیجه‌ی روش پیشنهادی ۵۵٪ می‌باشد. نشان می‌دهد روش پیشنهادی در ازای افزایش تعداد عامل‌ها به دلیل اینکه دانش جمعی



شکل ۵ـ۱۸ : نمودار باروری الگوریتم‌ها مختلف با تابع εـحریصانه در محیط صید و صیاد

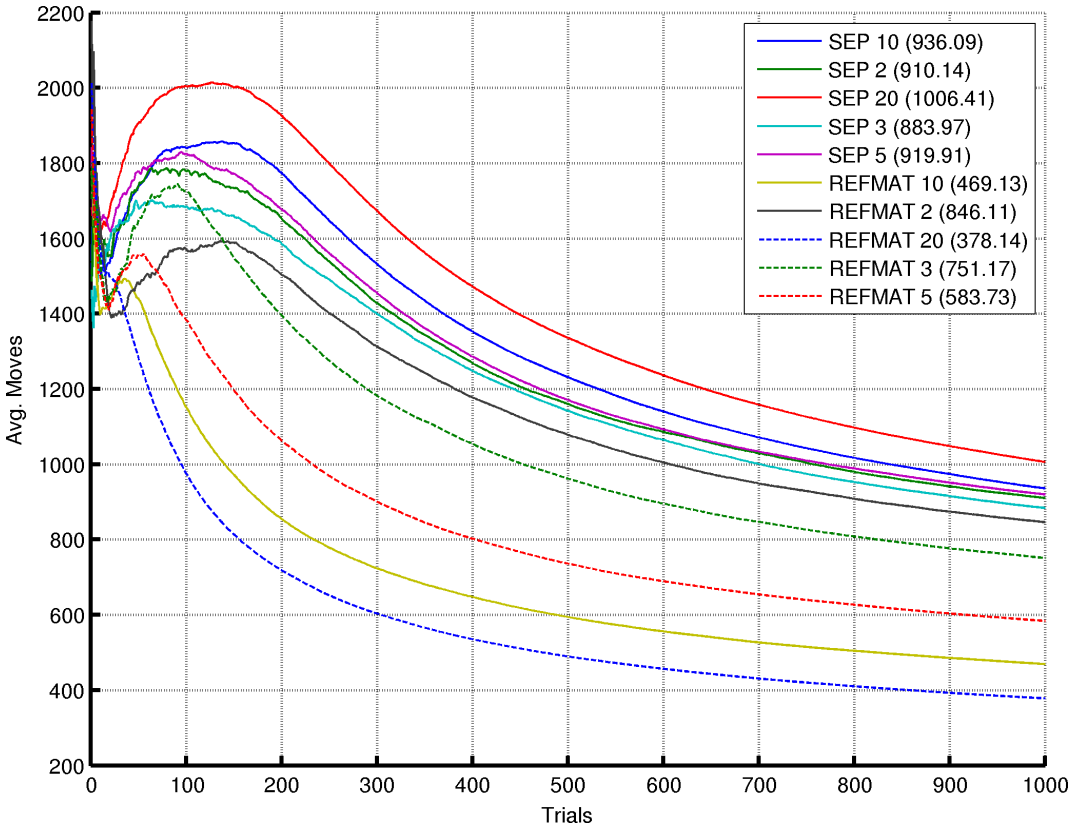
نیز افزایش می‌یابد کیفیت خروجی آن نیز بطور چشم‌گیری بهتر می‌شود. در حالی که در روش SEP اگر کار نتایج بدتر نشود بهتر نمی‌شود که از ضعف بزرگ روش SEP خبر می‌دهد.

نتیجه‌گیری: نتیجه‌ای که از مقایسه‌ی روش پیشنهادی در هر چهار مقایسه‌ی بالا می‌توان گرفت همچون نتیجه‌ای که از نتایج تابع بولتزمن، روش پیشنهادی بهبود چشم‌گیری به روش SEP در محیط صید و صیاد و سیاست انتخاب عمل حریصانه داده است.

مقایسه‌ی بین نتایج حاصل از سیاست انتخاب عمل بولتزمن و εـحریصانه

در حالت کلی همانند محیط پلکان مارپیچ در محیط صید و صیاد تابع بولتزمن نتایج یکنواتر و پایدارتری نسبت به تابع εـحریصانه از خود نشان داد و در هر دوی این توابع روش پیشنهادی نتیجه‌ی بهتری نسبت به روش SEP ارائه داد. در این قسمت به مقایسه‌ی نتایج بدست آمده توسط هر دو روش در هر دو سیاست انتخاب عمل می‌پردازیم.

مقایسه در سرعت و کیفیت یادگیری: مقایسه‌ی این قسمت را بطور خلاصه می‌توان در جدول ۵ـ۱۰ دید. که نسبت کیفیت نتیجه‌ی حاصل از تابع εـحریصانه نسبت به تابع بولتزمن همگی بزرگتر از ۱ می‌باشد، که نشان



شکل ۵ـ۱۹: مقایسه تاثیر تعداد عامل‌ها میزان کیفیت و سرعت یادگیری با تابع εـحریصانه در محیط صید و صیاد

جدول ۵ـ۱۰: مقایسه در سرعت و کیفیت یادگیری نسبت کیفیت نتیجه‌ی حاصل از تابع εـحریصانه نسبت به تابع بولتزمن

		Boltzmann	
		SEP	REFMAT
ε-greedy	SEP	8.07	9.75
	REFMAT	5.19	6.27

می‌دهد که استفاده از تابع εـحریصانه در کیفیت خروجی تاثیری منفی دارد.

مقایسه در سرعت اجرا: در جدول ۵ـ۱۱ نسبت میانگین سرعت اجرای روش‌ها آمده است، که نشان می‌دهد هر روش در زمانی که از تابع εـحریصانه استفاده می‌کند زمان بیشتری را تلف می‌کند نسبت به زمانی که از تابع بولتزمن استفاده می‌کند. همانند آنچه که در محیط پلکان مارپیچ مشاهده کردیم در اینجا نیز تابع بولزمن سریع‌تر عامل را به سمت اهداف هدایت می‌کند.

مقایسه در میزان باروری: همانطور که در جدول ۵ـ۱۲ آمده است استفاده از تابع εـحریصانه به نسبت تابع بولتزمن باعث باروری بیشتر جدول Q می‌شود.

جدول ۵ـ۱۱: مقایسه در نسبت میانگین سرعت اجرا حاصل از استفاده تابع εـحریصانه نسبت به تابع بولتزمن

		Boltzmann		
		SEP	REFMAT	IL
ε-greedy	SEP	3.27	4.10	6.95
	REFMAT	2.74	3.44	5.83
	IL	1.31	1.65	2.79

جدول ۵ـ۱۲: مقایسه در نسبت میزان باروری حاصل از استفاده تابع εـحریصانه نسبت به تابع بولتزمن

		Boltzmann		
		SEP	REFMAT	IL
ε-greedy	SEP	1.19	0.82	1.07
	REFMAT	1.49	1.03	1.35
	IL	1.17	0.80	1.05

مقایسه تاثیر تعداد عامل‌ها بر میزان کیفیت و سرعت یادگیری: همانطور که در جدول ۵ـ۱۳ مشاهده می‌شود در زمانی که از تابع εـحریصانه استفاده می‌شود در روش پیشنهادی تاثیر تعداد عامل‌ها به مراتب بیشتر از زمانی است که از تابع بولتزمن استفاده می‌کنیم. این در حالی می‌باشد که در روش SEP اضافه کردن عامل‌ها به محیط نه تنها به بهبود دانش خروجی الگوریتم کمکی نمی‌کند بلکه نتایج را بدتر نیز می‌کند!

نتیجه‌گیری: با مقایسه‌ی بین تاثیر توابع حریصانه و بولتزمن در خروجی الگوریتم‌ها به این نتیجه می‌توان رسید که تابع بولتزمن رفتاری مطمئن‌تر دارد و باعث می‌شود که روش‌ها سریع‌تر همگرا شود.

۵ـ۴ بررسی تاثیر تعداد نواحی محیط در کیفیت و سرعت یادگیری عامل‌ها در روش پیشنهادی

همانطور که در تعریف ۴ـ۱ آورده شده است، بنا به معیار خبرگی معرفی شده در این پژوهش باید محیط به تعدادی ناحیه افراز شود و سپس میزان حضور عامل در هر ناحیه را سنجیده و خبرگی عامل معکوسی از میزان حضور عامل در این نواحی می‌باشد. لذا ضروری است که در این قسمت به بررسی تاثیر تعداد نواحی محیط در کیفیت و سرعت یادگیری عامل‌ها در روش پیشنهادی بپردازیم.

جدول ۵ـ۱۳: مقایسه در نسبت شیب تاثیر تعداد عامل‌ها میزان کیفیت نتیجه‌ی حاصل از تابع εـحریصانه نسبت به تابع بولتزمن

Boltzmann

		SEP	REFMAT
ε-greedy	SEP	-2.52	-0.07
	REFMAT	379.32	10.65

۵ـ۴ـ۱ محیط پلکان مارپیچ

ما محیط پلکان مارپیچ را به ۶ ناحیه‌ی مختلف با اندازه‌های ۱×۱، ۲×۲، ... ۶×۶ (کل محیط) تقسیم‌بندی کرده‌ایم و همان‌طور که در شکل ۵ـ۲۰ آمده است اندازه‌ی این نواحی در کیفیت و سرعت یادگیری روش پیشنهادی تفاوتی ایجاد نمی‌کند و می‌توان برای کل محیط را یک ناحیه فرض کرد و میزان خبرگی کلی عامل برابر می‌شود با تعداد گام‌هایی که عامل برای رسیدن به هدف طی می‌کند.

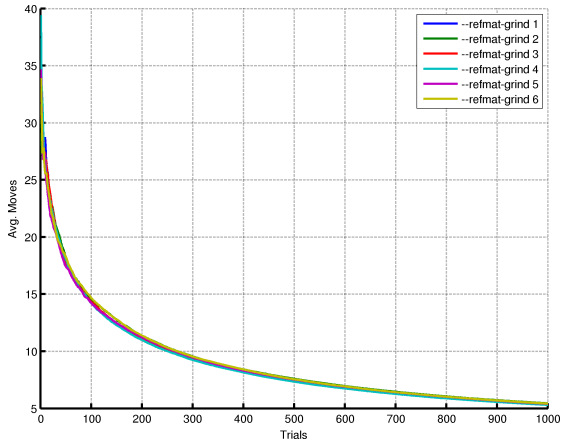
۵ـ۴ـ۲ محیط پلکان صید و صیاد

همانند محیط پلکان مارپیچ را به چند ناحیه‌ی مختلف با اندازه‌های ۱×۱، ۱×۲، ... ۱×۱۷ (کل محیط) تقسیم‌بندی کرده‌ایم و همان‌طور که در شکل ۵ـ۲۱ آمده است همچون محیط پلکان مارپیچ اندازه‌ی این نواحی در کیفیت و سرعت یادگیری روش پیشنهادی تفاوتی ایجاد نمی‌کند.

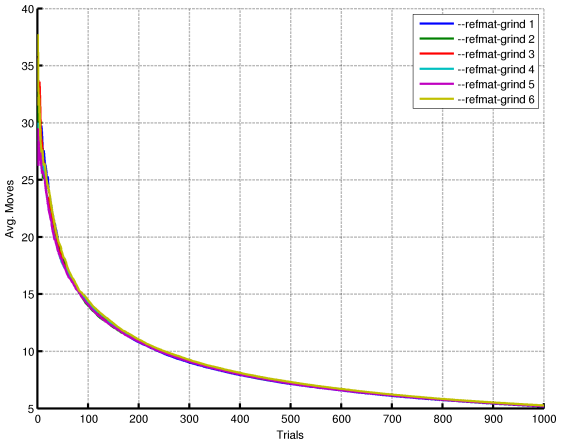
۵ـ۵ بررسی تاثیر استفاده از انتگرال فازی در بهبود دانش جمعی

همانطور که در بخش‌های قبلی دیدیم، استفاده از توابع $g(\cdot)$ مختلف در انتگرال فازی نتایج روش پیشنهادی موثر واقع شده‌اند و در نهایت با استفاده از تابع Const-One که باعث می‌شود انتگرال فازی به یک عملگر بیشینه‌گیری تبدیل شود و بطور حریصانه در هر مرحله‌ی به اشتراک‌گذاری دانش، فقط دانش عاملی را که از دیگر عامل‌ها خبره‌تر است به عنوان دانش جمعی خروجی الگوریتم در نظر بگیرد؛ در این بخش می‌خواهیم این موضوع را بررسی کنیم که تاثیر انتگرال فازی در بهبود دانش جمعی چگونه می‌باشد؟ به عنوان مثال آیا همیشه استفاده از انتگرال فازی در بهبود دانش جمعی موثر واقع است؟

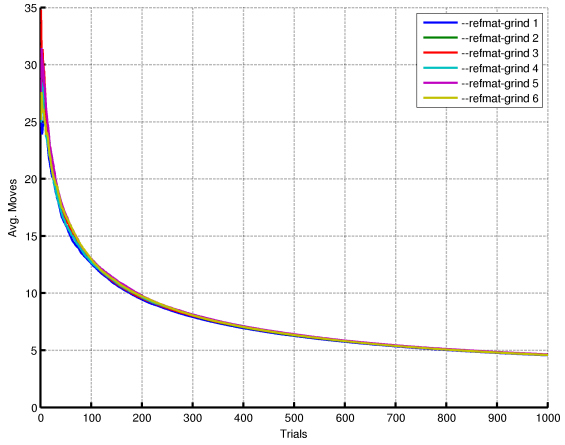
در پاسخ به این سوالات، ما روش SEP را با این‌بار با استفاده از انتگرال فازی مورد آزمون قرار دادیم که همانطور که در شکل ۵ـ۲۲ مشاهده می‌کنیم وقتی در روش SEP بجای میانگین وزنی (روش پیشنهادی SEP) از انتگرال فازی با توابع معرفی شده در الگوریتم‌های ۴-۵ تا ۴-۷ استفاده کنیم خروجی انتگرال فازی کیفیت و سرعت یادگیری را بهبود نمی‌خشید – گرچه ممکن است در صورت تعریف توابع $g(\cdot)$ دیگر نتایج بهتری تولید کند؛ ولی در صورتی که از الگوریتم ۴-۴ استفاده کنیم بهترین نتیجه‌ی ممکن را می‌گیریم.



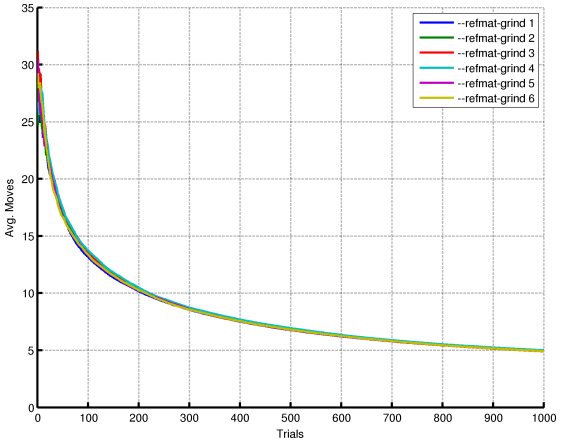
ب ـ تابع Mean به عنوان مدل‌کننده‌ی تابع $g(\cdot)$



الف ـ تابع Max به عنوان مدل‌کننده‌ی تابع $g(\cdot)$

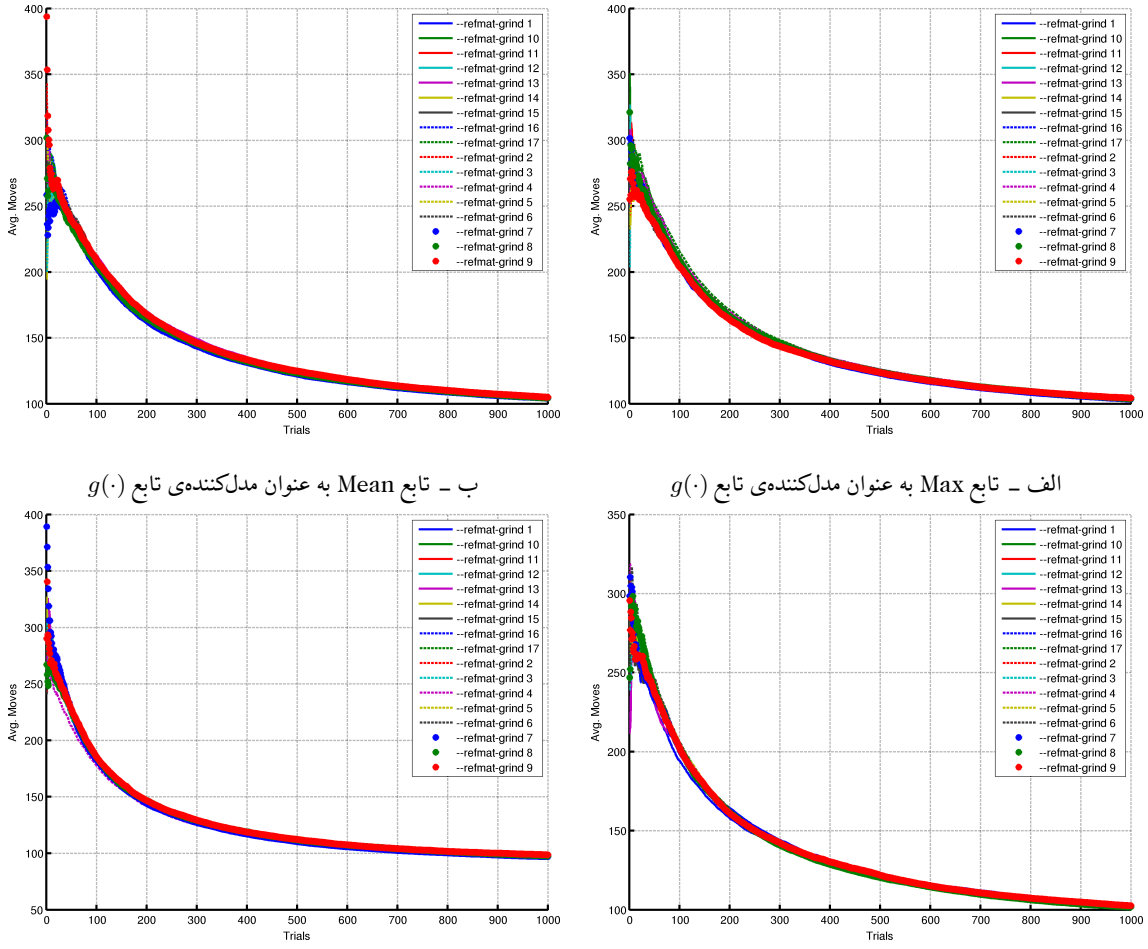


د ـ تابع Const-One به عنوان مدل‌کننده‌ی تابع $g(\cdot)$



ج ـ تابع K-Mean به عنوان مدل‌کننده‌ی تابع $g(\cdot)$

شکل ۵ـ۲۰: تاثیر ناحیه‌بندی مختلف بروی کیفیت و سرعت یادگیری در محیط پلکان مارپیچ



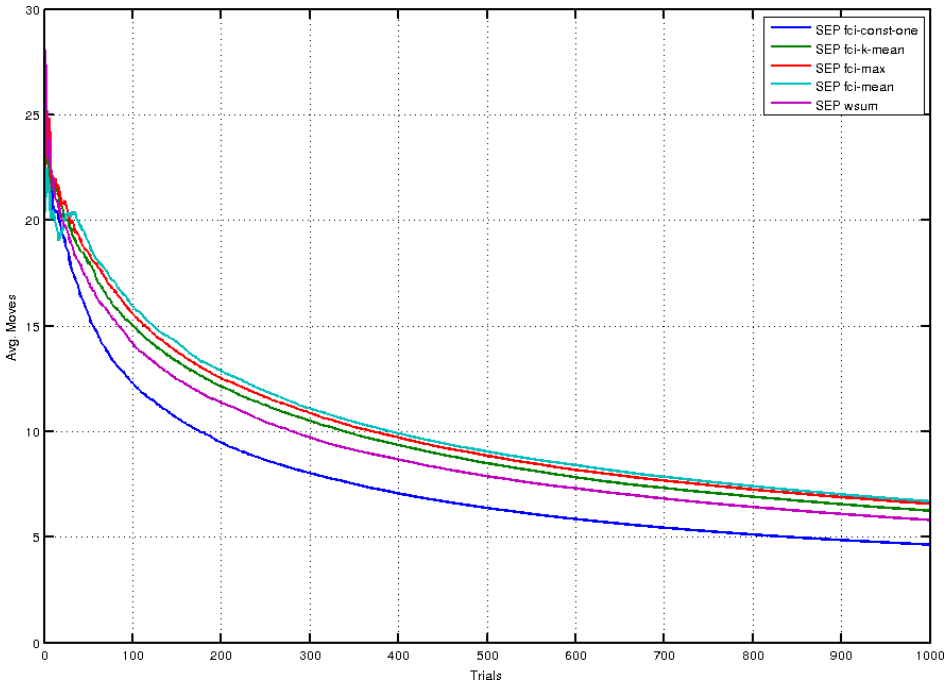
الف ـ تابع Max به عنوان مدل‌کننده‌ی تابع $g(\cdot)$

ب ـ تابع Mean به عنوان مدل‌کننده‌ی تابع $g(\cdot)$

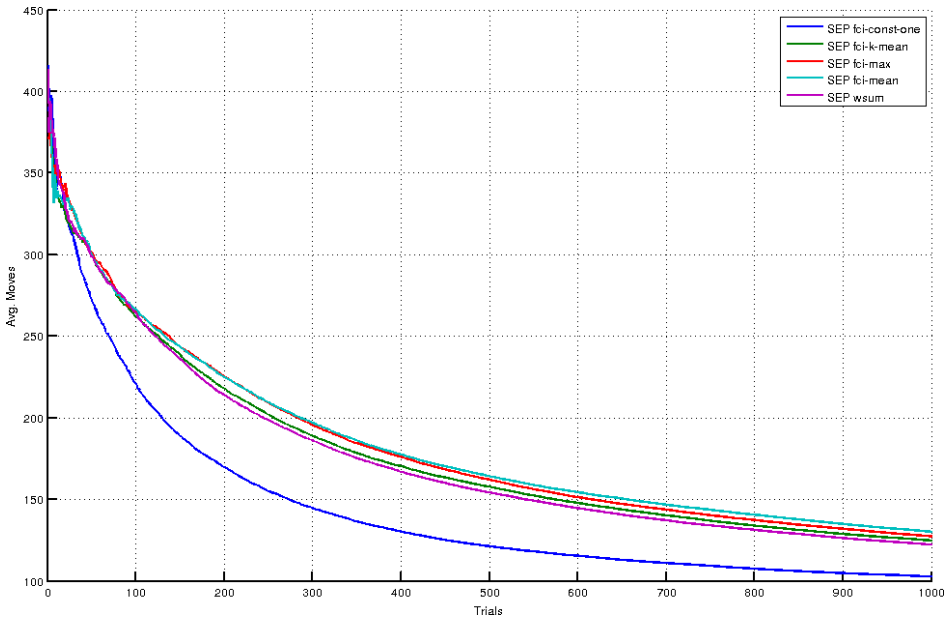
ج ـ تابع K-Mean به عنوان مدل‌کننده‌ی تابع $g(\cdot)$

د ـ تابع Const-One به عنوان مدل‌کننده‌ی تابع $g(\cdot)$

شکل ۵ـ۲۱: تاثیر ناحیه‌بندی مختلف بروی کیفیت و سرعت یادگیری در محیط صید و صیاد



الف



ب

شکل ۵ـ۲۲: تاثیر استفاده از انتگرال فازی در روش SEP بروی کیفیت و سرعت یادگیری در محیط پلکان مارپیچ. الف ـ استفاده از انتگرال فازی در روش SEP در محیط پلکان مارپیچ. ب ـ استفاده از انتگرال فازی در روش SEP در محیط صید و صیاد.

در مورد روش MCE نیز تاثیر استفاده از انتگرال فازی را نیز مورد بررسی قرار دادیم که در شکل ۵ـ۲۳ نتایج آمده است. همانطور که مشاهده می‌شود انتگرال فازی در روش MCE با توابع $g(\cdot)$ معرفی شده در این پژوهش (یعنی الگوریتم‌های ۴-۴ تا ۴-۷) باعث بهبود نتایج شده است. این بهبود در محیط پلکان مارپیچ به صورت بهبود در سرعت یادگیری و در محیط صید و صیاد در سرعت و کیفیت یادگیری مشاهده می‌شود. همانطور که در بخش‌های قبلی نیز آورده شده است، انتگرال فازی روشی جامع‌تر از میانگین‌گیری وزن‌دار می‌باشد و همانطور که از نتایج آزمایش‌های این فصل می‌توان برداشت کرد با انتخاب و معرفی مناسب تابع $g(\cdot)$ انتگرال فازی می‌تواند نتایج بهتری نسبت به روش میانگین‌گیری وزندار تولید نماید – این نتیجه‌گیری را می‌توان در بهبود نتایج توسط کلیه‌ی توابع $g(\cdot)$ معرفی شده در این پژوهش بر روی روش پیشنهادی و MCE و همچنین تابع ۴-۴ بروی روش SEP مشاهده کرد، علت بهبود نتایج این روش‌ها با استفاده از انتگرال فازی در قسمت ۴ـ۴ آورده شده است. همچنین با توجه به نتایج حاصل در شکل ۵ـ۲۳ باید به این نکته توجه کرد که انتخاب حریصانه‌ی دانش عامل‌ها به عنوان دانش جمعی (یعنی انتخاب توابع Const-One – الگوریتم ۴-۴ به عنوان تابع $g(\cdot)$ همیشه موثر نمی‌باشد.

۵ـ۶ تحلیل نتایج

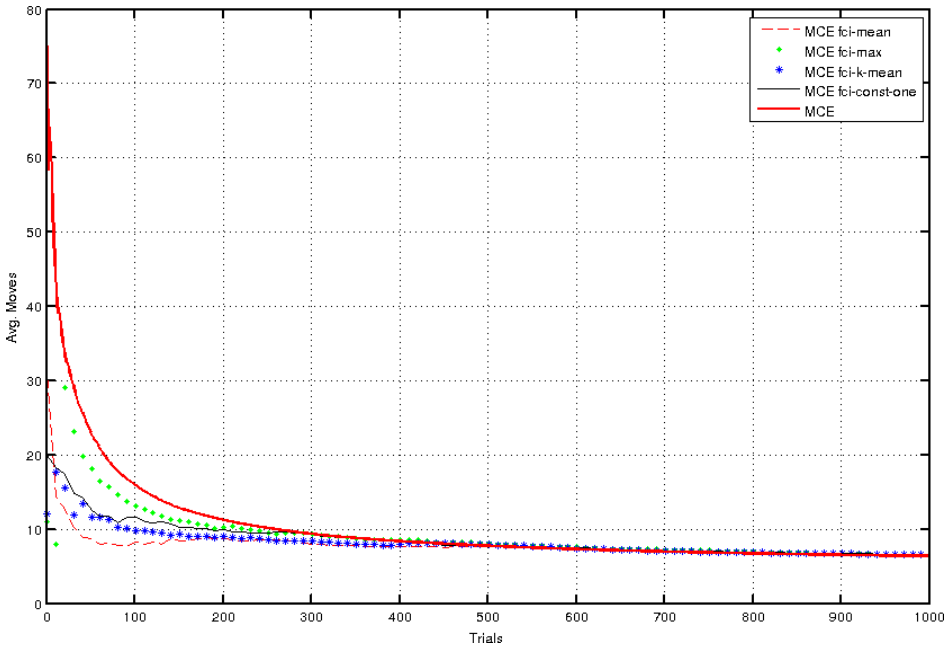
در این قسمت به بررسی و تحلیل نتایج عددی که در بخش‌های قبلی این فصل آورده شده است می‌پردازیم. در تمامی آزمایش‌های انجام شده همانطور که مشاهده می‌شود که روش پیشنهادی از هر لحاظ از روش SEP نتایج بهتری بدست آورده است لذا در این بخش به بررسی علت‌های این برتری می‌پردازیم.

۵ـ۶ـ۱ مقایسه‌ی روش SEP با روش پیشنهادی

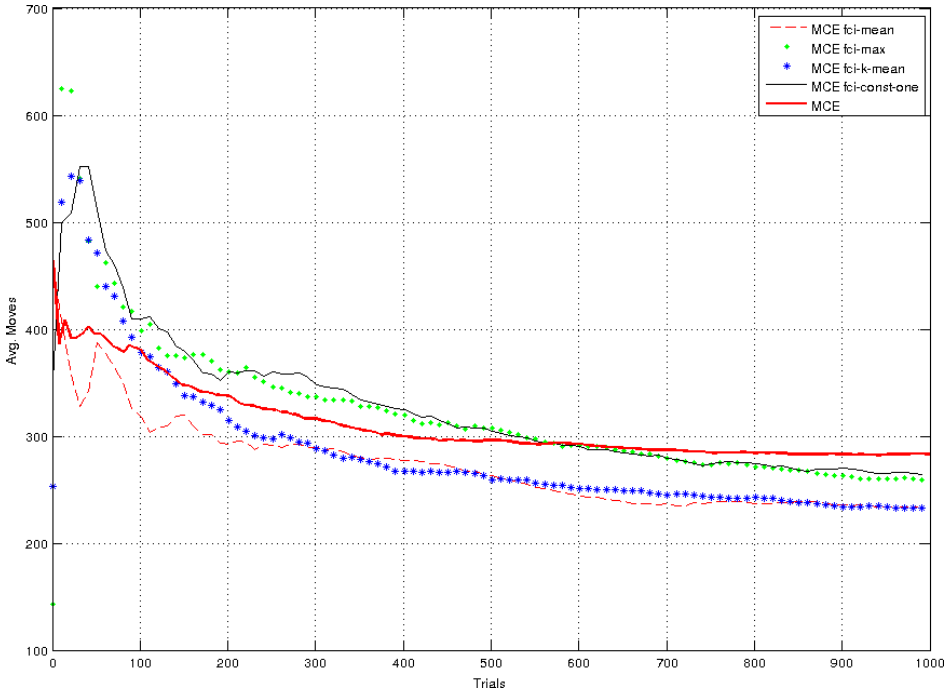
جواب این سوال که «چرا روش پیشنهادی نسبت به روش SEP با اختلاف چشم‌گیری، نتایج بهتری ارائه داده است؟» دلیل زیر می‌باشد:

۱. **فقدان وجود پراش[1] دانشی بین عامل‌ها:** برای اینکه ترکیب دانش عامل‌ها موثر واقع شود باید یک پراشی بین دانش عامل‌ها وجود داشته باشد (یعنی عامل‌ها با دانش‌های مختلف از محیط وجود داشته باشد)، اگر همه‌ی عامل‌ها یک دانش را داشته باشند (پراش دانشی کمتری داشته باشند) نباید انتظار داشت که ترکیب دانش عامل‌ها دانشی فراتر از آنچه که در حالت کلی جمع دارند، نتیجه دهد. لذا نمودار شکل ۵ـ۷ این نتیجه را می‌دهد که روش میرزایی در بین عامل‌ها با تجربیات متفاوت دانش‌های متفاوتی ایجاد نمی‌کند!

[1] Variance



الف



ب

شکل ۵ـ۲۳: تاثیر استفاده از انتگرال فازی در روش MCE بروی کیفیت و سرعت یادگیری در محیط پلکان مارپیچ. الف ـ استفاده از انتگرال فازی در روش MCE در محیط پلکان مارپیچ. ب ـ استفاده از انتگرال فازی در روش MCE در محیط صید و صیاد.

۲. **کیفیت معیارها:** معیارهای معرفی شده توسط میرزایی از کیفیت کافی برخوردار نیستند که نهایتا در کیفیت نتایج روش ارائه شده توسط ایشان تاثیر می‌گذارند.

۵ـ۶ـ۲ مقایسه‌ی تابع بولتزمن و εـحریصانه

همانطور که در آزمایش‌ها مشاهده کردیم تابع بولتزمن از هر لحاظ نسبت به تابع εـحریصانه نتایج بهتری تولید می‌کند و شهودی که به نظر می‌رسد دلیل این مساله می‌باشد این است که تابع εـحریصانه در ε٪ مواقع بصورت کاملا تصادفی عمل می‌کند. حال فرض کنید عامل بعد از طی چند دوره به دانش نسبتا خوب نسبت به محیط خود رسیده است، در این صورت در هر حالت بخوبی می‌داند که کدام عمل به صلاح‌تر است، در صورتی که از تابع بولتزمن جهت انتخاب عمل استفاده کند این تابع احتمال انتخاب عملی که در این حالت به‌صلاح است را بیشتر می‌کند و عملی که تاثیر مخربی بر دست‌یابی عامل به اهداف خود دارد را احتمال انتخاب کمتری تخصیص می‌دهد؛ به همین دلیل در هر حالت احتمال انتخاب عملی که بهینه است بیشتر است و به مرور که عامل دانش خود را نسبت به محیط کامل‌تر می‌کند این احتمال بیشتر تقویت می‌شود. در حالی که تابع εـحریصانه فارغ از میزان دانش عامل از محیط در هر حالت ε٪ احتمال دارد که کاملا تصادفی انتخاب عمل کند و در این انتخاب تصادفی محتمل است که بدترین عمل انتخاب شود و این احتمال‌ها در طور یادگیری ثابت است لذا ممکن است در یک قدمی رسیدن به هدف عامل به صورت تصادفی یک عملی را انتخاب که کند که منجر عدم دست‌یابی به هدف یا دورشدن از هدف شود، به همین دلیل تابع εـحریصانه نسبت به تابع بولتزمن باعث می‌شود عامل دیرتر به هدف برسد که همین مساله در معیارهای سنجش از قبیل سرعت و کیفیت یادگیری، سرعت اجرا تاثیر منفی می‌گذارد.

۵ـ۶ـ۳ بررسی تاثیر تعداد نواحی در کیفیت و سرعت یادگیری در روش پیشنهادی

دلیل این مساله که «با توجه به شکل‌های ۵ـ۲۰ و ۵ـ۲۱ تعداد نواحی معیار تعریف شده در تعریف ۴ـ۱ در سرعت و کیفیت یادگیری در روش پیشنهادی تاثیرگذار نیست.» این است که با توجه به تعریف ۴ـ۱ یک همبستگی[1] مثبتی میان میزان ارجاع در نواحی ریز و نواحی درشت وجود دارد، زیرا که هرچه میزان ارجاع به نواحی ریز زیاد باشد همان میزان ارجاع به نواحی درشت (که شامل آن نواحی ریز می‌باشند) نیز زیاد می‌شود و برعکس هرچه میزان ارجاع به نواحی درشت کمتر باشد میزان ارجاع به نواحی زیرمجموعه‌ی آن نواحی درشت نیز کمتر است. بنابرین با توجه نحوه‌ی تعریف معیار مورد استفاده در تعریف ۴ـ۱ تعداد نواحی در کیفیت و سرعت یادگیری تاثیرگذار نمی‌باشد.

[1] Correlation

۷ـ۵ نتیجه‌گیری

در فصل به بررسی نتایج آزمایش‌های گوناگونی که برای مقایسه‌ی روش پیشنهادی با روش SEP آورده شده بودند پرداختیم؛ مقایسه‌ها بر اساس ۴ معیار صورت گرفت: ۱. مقایسه در سرعت و کیفیت یادگیری، ۲. مقایسه در پیچیدگی زمانی، ۳. مقایسه در میزان باروری، ۴. مقایسه تاثیر تعداد عامل‌ها بر میزان کیفیت و سرعت یادگیری. همانطور که دیدیم این روش پیشنهادی در هر ۴ معیار نسبت به روش SEP برتری قابل توجهی داشت. همچنین به مقایسه‌ی تاثیر استفاده از سیاست‌های انتخاب عمل بولتزمن و εـحریصانه پرداختیم و نشان دادیم که استفاده از بولتزمن نسبت به εـحریصانه کیفیت و سرعت یادگیری بسیار بهتری بدست می‌دهد. همچنین نشان دادیم که معیار خبرگی معرفی شده در تعریف ۴ـ۱ مستقل از تعداد و اندازه‌ی افرازهای محیط می‌باشد، سپس با تحلیل نتایج آزمایش‌ها به مطالب این بخش خاتمه دادیم. در فصل بعدی به جمع‌بندی مطالب و دست‌آوردهای این پژوهش می‌پردازیم.

فصل ششم

نتیجه‌گیری و جمع‌بندی

۶ـ۱ مقدمه

اگر اندکی به مسائلی که افراد انجام می‌دهند و ما آن‌ها را در آن خبره می‌بینیم توجه کنیم، متوجه خواهیم شد که زمانی که فردی در موردی خبره می‌شود بطور طبیعی انرژی نسبتا کمتری در انجام آن مصرف می‌کند. این معیار همان معیاری است که می‌گوید عاملی در انجام وظیفه‌ای خبره‌تر است که در طی انجام آن انرژی کمتری مصرف کند. این معیار که از فلسفه‌ی بسیار ساده‌ای برخوردار است برخلاف معیارهای گذشته بسیار منعطف می‌باشد زیرا که در تعریف این معیار عبارت «میزان انرژی» می‌تواند تعابیر مختلفی به خود بگیرد و در هر مورد قابل استفاده باشد.

وجود این فلسفه انگیزه‌ای شد که در صدد ارائه‌ای معیاری برآییم که نه تنها ساده باشد بلکه در زندگی روزمره ما انسان‌ها هم تجلی داشته باشد. بعد از اندکی تفکر و تفحص در نهایت این معیار چیزی جز معیار «تنبلی» نبود! معیار تنبلی که در این پایان‌نامه با اصطلاح علمی «میزان ارجاع» ارائه شد می‌گوید که «عاملی که برای به نتیجه رساندن فعالیت‌هایش انرژی کمتری صرف کند خبره‌تر است».

در این قسمت به مروری خلاصه بر هرآنچه که در این پژوهش صورت گرفته و ارائه‌ی یک نتیجه‌گیری نهایی حاصل از این پژوهش و همچنین ارائه‌ی مسیر پژوهشی پیشنهادی برای آیندگان این زمینه از یادگیری مشارکتی خواهیم پرداخت.

۶ـ۲ نوآوری‌ها و نتایج کلی پایان‌نامه

در طی این پایان‌نامه معیار جدیدی به نام معیار «میزان ارجاع» ارائه شد که می‌گوید عاملی که کمتر در محیط مورد تعاملش پرسه بزند از خبرگی بیشتری برخوردار است و سپس با استفاده از این معیار خبرگی به سنجش عامل‌های فعال در محیط در هنگام مشارکت در دانش جمعی پرداختیم. در هنگام ترکیب دانش عامل‌ها از انتگرال فازی چوکت استفاده شد.

در طی آزمایش‌ها از میانگین وزنی نیز به جای انتگرال فازی استفاده شد و نشان داده شد که در معیار ارائه شده توسط این پژوهش انتگرال فازی توانایی بهتری نسبت به میانگین وزنی برای بهبود کیفیت و سرعت یادگیری مشارکتی دارد. همچنین از ۴ تابع به عنوان مدل کننده‌ی تابع $g(\cdot)$ استفاده شد، که هرکدام یک دیدگاهی نسبت به نحوه‌ی ترکیب دانش‌های ورودی ارائه می‌دهد. از بین این ۴ تابع، تابع Const-One در کلیه‌ی آزمایش‌ها نسبت به دیگر توابع برتری قابل توجهی از خود نشان داد؛ طبق آنچه که فصول قبلی این پایان‌نامه آورده شده این تابع معادل با حداکثرگیری بروی دانش عامل‌ها بر اساس معیار خبرگی آن‌ها می‌باشد. یعنی اینکه این تابع در واقع در هر ناحیه فقط دانش عاملی را در نظر می‌گیرد از همه خبره‌تر (تنبل‌تر) است که این امر تاییدی بر فرضیه ۴ـ۱ و متعاقبا تعریف ۴ـ۱ می‌باشد. همچنین تاثیر استفاده از انتگرال فازی در روش SEP را نیز مورد بررسی قرار دادیم و مشاهده کردیم که انتگرال فازی **با توابع $g(\cdot)$ معرفی شده در این پژوهش** در روش SEP تاثیر مثبتی ندارد ولی در این بررسی نیز تابع Const-One بیشترین بهبود را در پی‌داشت که معادل می‌شود با انتخاب حریصانه‌ی بین دانش عامل‌ها به عنوان دانش جمعی، که یک نقطه‌ی مشترک بین نتایج روش پیشنهادی و این بررسی می‌باشد؛ از طرفی بررسی‌های انجام شده در رابطه با تاثیر انتگرال فازی **با توابع $g(\cdot)$ معرفی شده در این پژوهش** در روش MCE نشان می‌دهد که انتگرال فازی موثر واقع شده است، که نشان می‌دهد در صورت انتخاب مناسب تابع $g(\cdot)$ انتگرال فازی می‌تواند جایگزین بهتری بجای میانگین‌گیری وزنی باشد. همچنین لازم به ذکر است که بررسی‌های انجام شده در استفاده از انتگرال فازی بروی هر سه روش REFMAT و SEP و MCE نشان می‌دهد که انتخاب حریصانه دانش عامل‌ها به عنوان دانش جمعی (انتخاب Const-One به عنوان تابع $g(\cdot)$) همیشه باعث حداکثر شدن سرعت و کیفیت یادگیری نمی‌شود که این نتیجه‌گیری به اهمیت استفاده از انتگرال فازی به عنوان تابع ادغام‌کننده دانش عامل‌ها می‌افزاید.

همچنین در نهایت، در انتهای فصل آزمایش‌ها نشان داده شد که می‌توان معیار خبرگی ارائه شده در تعریف ۴ـ۱ را به کل محیط خلاصه کرد؛ یعنی عاملی خبره‌تر است که میزان حضور آن در کل محیط کمتر باشد – یعنی با تعداد گام کمتری به اهداف خود برسد. همین نتیجه‌گیری باعث می‌شود که آزمودن دیگر توابع برای مدل کردن $g(\cdot)$ (مثلا تابع اندازه‌گیری‌ـλ سوگنو) نیازی نباشد.

در این پژوهش تعادلی بین کلی و جزئی نگری به عملکرد عامل‌ها در هنگام ادغام دانش‌های آن‌ها برقرار شد. همچنین تاثیر دیگر روش‌های انتخاب عمل را در ترکیب با معیارهای ارائه شده را مورد بررسی قرار گرفته است و به این نتیجه رسیدیم که تابع بولتزمن نتیجه‌ی با کیفیت‌تری را تولید می‌کند. همچنین دستاوردهای این پژوهش را با در نظر گرفتن ماهیت غیرافزایشی بودن ذات مساله ارائه دادیم.

یکی از مزایای روش پیشنهادی این است که در عین کارایی و قدرت روشی ساده در مفهومی و پیاده‌سازی می‌باشد که این سادگی طبق آنچه که در آزمایش‌ها آمده است نهایتا منجر شد که روش پیشنهادی از پیچیدگی کمتری برخوردار باشد. از دیگر مزیت روش پیشنهادی کلی بودن فرضیه خبرگی‌ای که این پژوهش برمبنای آن ارائه شد، می‌باشد که می‌توان آن را به تمامی مسائل یادگیری مشارکتی به راحتی اعمال کرد.

۶ـ۳ راهکارهای آینده و پیشنهادها

همانطور که آزمایش‌ها نشان دادند با توجه به معیار خبرگی ارائه شده در قسمت یادگیری مشارکتی اگر فقط دانش عامل خبره را در نظر بگیریم حداکثر نتیجه‌ی ممکن (در قالب روش پیشنهادی) را خواهیم گرفت. در طی این پژوهش دو مفهوم مهم ارائه شد: ۱. انتگرال فازی چوکت می‌تواند عملگر بسیار قوی‌ای نسبت به روش‌ها سنتی چون میانگین‌گیری وزنی باشد. ۲. فرضیه خبرگی معرفی شده بخوبی می‌تواند هر نوع معیار خبرگی را توجیه کند.

در این پژوهش سعی شده است که حداکثر نتیجه‌ی ممکن حاصل از استفاده از این دو مفهوم باهم را استخراج کنیم ولی پیشنهادها زیر می‌تواند شروع خوبی برای پژوهش‌های آینده در این زمینه باشد.

۱. ارائه‌ی معیار خبرگی جدیدی مبتنی بر فرضیه خبرگی(فرضیه‌ی ۴ـ۱) معرفی شده در این پژوهش که چهارچوبی کلی جهت تعریف معیارهای خبرگی را تعریف می‌کند؛ سپس آزمایش معیار خبرگی تعریف شده بجهت آزمودن فرضیه خبرگی ارائه شده.

۲. بررسی تاثیر استفاده از انتگرال فازی چوکت در پژوهش‌های گذشته.

۳. بررسی این موضوع که «آیا انتخاب دانش عاملی که از همه خبره‌تر است همیشه باعث حداکثر شدن کیفیت و سرعت یادگیری جمع می‌شود یا خیر؟».

۴. ارائه‌ی روشی جهت وفقی شدن پارامترهای معرفی شده در بخش ۴ـ۳ـ۲ با هدف بهبود کیفیت و سرعت یادگیری هرچه بیشتر روش پیشنهادی.

۵. بررسی شرایط و ویژگی‌های توابع $g(\cdot)$ در حالت کلی.

مراجع

[1] E. P. Hajyyar, "Multi-criteria expertness based cooperative learning in multi-agent systems," Master's thesis, Department of Electrical and Computer Engineering, Isfahan University of Technology, Isfahan University of Technology, Isfahan 84156-83111, Iran, 10 2010.

[2] M. ali mirzaei badizi, "Speed-up cooperative learning in multi-agent systems using shortest experimented path," Master's thesis, Department of Electrical and Computer Engineering, Isfahan University of Technology, Isfahan University of Technology, Isfahan 84156-83111, Iran, 3 2016.

[3] L. Panait and S. Luke, "Cooperative multi-agent learning: The state of the art," Autonomous agents and multi-agent systems, vol. 11, no. 3, pp. 387–434, 2005.

[4] V. Torra, Y. Narukawa, and M. Sugeno, Non-Additive Measures, pp. 3–7. Springer, 2014.

[5] M. Tan, "Multi-agent reinforcement learning: Independent vs. cooperative agents," in Proceedings of the tenth international conference on machine learning, pp. 330–337, 1993.

[6] H. R. Berenji and D. Vengerov, "Cooperation and coordination between fuzzy reinforcement learning agents in continuous state partially observable markov decision processes," in Fuzzy Systems Conference Proceedings, 1999. FUZZ-IEEE'99. 1999 IEEE International, vol. 2, pp. 621–627, IEEE, 1999.

[7] Y. Kuniyoshi, M. Inaba, and H. Inoue, "Learning by watching: Extracting reusable task knowledge from visual observation of human performance," IEEE transactions on robotics and automation, vol. 10, no. 6, pp. 799–822, 1994.

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

[9] A. Garland and R. Alterman, "Preparation of multi-agent knowledge for reuse," in Proceedings of the Fall Symposium on Adaptation of Knowledge for Reuse, vol. 26, p. 33, 1995.

[10] L. Nunes and E. Oliveira, "On learning by exchanging advice," arXiv preprint cs/0203010, 2002.

[11] L. Nunes and E. Oliveira, "Advice-exchange between evolutionary algorithms and reinforcement learning agents: Experiments in the pursuit domain," in Adaptive Agents and Multi-Agent Systems II, pp. 185–204, Springer, 2005.

[12] M. N. Ahmadabadi, M. Asadpur, S. H. Khodanbakhsh, and E. Nakano, "Expertness measuring in cooperative learning," in Intelligent Robots and Systems, 2000.(IROS 2000). Proceedings. 2000 IEEE/RSJ International Conference on, vol. 3, pp. 2261–2267, IEEE, 2000.

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

[14] P. Ritthipravat, T. Maneewarn, J. Wyatt, and D. Laowattana, "Comparison and analysis of expertness measure in knowledge sharing among robots," in International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, pp. 60–69, Springer, 2006.

[15] Y. Yang, Y. Tian, and H. Mei, "Cooperative q learning based on blackboard architecture," in Computational Intelligence and Security Workshops, 2007. CISW 2007. International Conference on, pp. 224–227, IEEE, 2007.

[16] H. Iima and Y. Kuroe, "Reinforcement learning through interaction among multiple agents," in 2006 SICE-ICASE International Joint Conference, pp. 2457–2462, IEEE, 2006.

[17] J. Kennedy, "Particle swarm optimization," in Encyclopedia of machine learning, pp. 760–766, Springer, 2011.

[18] M. Yang, Y. Tian, and X. Liu, "Cooperative q-learning based on maturity of the policy," in 2009 International Conference on Mechatronics and Automation, pp. 1352–1356, IEEE, 2009.

[19] E. Pakizeh, M. Palhang, and M. M. Pedram, "Multi-criteria expertness based cooperative q-learning," Applied intelligence, vol. 39, no. 1, pp. 28–40, 2013.

[20] D. L. Poole and A. K. Mackworth, Artificial Intelligence: foundations of computational agents, ch. 11. Cambridge University Press, 2010.

[21] S. J. Russell, P. Norvig, J. F. Canny, J. M. Malik, and D. D. Edwards, Artificial intelligence: a modern approach, vol. 2. Prentice hall Upper Saddle River, 2003.

[22] V. Torra and Y. Narukawa, "The interpretation of fuzzy integrals and their application to fuzzy systems," International journal of approximate reasoning, vol. 41, no. 1, pp. 43–58, 2006.

[23] K. Leszczy□ski, P. Penczek, and W. Grochulski, "Sugeno's fuzzy measure and fuzzy clustering," Fuzzy Sets and Systems, vol. 15, no. 2, pp. 147–158, 1985.

[24] A. F. Tehrani, W. Cheng, and E. Hullermeier, "Preference learning using the choquet integral: The case of multipartite ranking," IEEE Transactions on Fuzzy Systems, vol. 20, no. 6, pp. 1102–1113, 2012.

[25] L. M. De Campos and M. Jorge, "Characterization and comparison of sugeno and choquet integrals," Fuzzy Sets and Systems, vol. 52, no. 1, pp. 61–67, 1992.

[26] M. Grabisch, "Fuzzy integral in multicriteria decision making," Fuzzy sets and Systems, vol. 69, no. 3, pp. 279–298, 1995.

[27] T. Murofushi, M. Sugeno, and M. Machida, "Non-monotonic fuzzy measures and the choquet integral," Fuzzy sets and Systems, vol. 64, no. 1, pp. 73–86, 1994.

[28] M. Grabisch, "The application of fuzzy integrals in multicriteria decision making," European journal of operational research, vol. 89, no. 3, pp. 445–456, 1996.

[29] "Expert - wikipedia." https://en.wikipedia.org/wiki/Expert. (Accessed on 11/12/2016).

[30] E. Schechter, Handbook of Analysis and its Foundations, ch. 1, p. 16. Academic Press, 1996.

[31] R. S. Sutton and A. G. Barto, Reinforcement learning: An introduction, vol. 1. MIT press Cambridge, 1998.

[32] C. J. Watkins and P. Dayan, "Q-learning," Machine learning, vol. 8, no. 3-4, pp. 279–292, 1992.

Improvments in speed and quality of learning in multi-agent systems using a novel expertness criteria and fuzzy integral

Dariush Hasanpour Adeh

d.hasanpoor@ec.iut.ac.ir

Department of Electrical and Computer Engineering

Isfahan University of Technology, Isfahan 84156-83111, Iran

Degree: M.Sc. Language: Farsi

Supervisor: Assoc. Prof. Maziar Palhang (palhang@cc.iut.ac.ir)

Abstract

In the real world, usually, peoples are coming together for sharing their knowledge and talking from their good and bad experiences and more or less everybody has something to say. Although we cannot ignore anybody's knowledge but it's common sense to assign more weight on the most experienced person's knowledge when we are going to decide what we need to do based on consultation from people. The achievements of this research have the same philosophy, that everybody needs to be heard. Fuzzy integrals are one of the most powerful and flexible methods for hearing everybody's knowledge and extract knowledge which is useful for everybody.

One of the challenges is that how to fairly answer the "what is the agents' expertise and how to determine the most and least expert agent?" question. To answer this question, in this thesis, we have proposed «the hypothesis of expertness» which defines a framework for "expertness criteria" definitions, and based on this framework we have introduced a new expertness criteria and showed that the defined framework and criteria are much more efficient than the state of the art criteria "Shortest Experienced Path" criteria. Also, the power of using fuzzy integrals for intelligence aggregation and non-additive measuring/knowledge is demonstrated.

Key Words:

Multi-agent Systems, Cooperative Learning, Reinforcement Learning, Fuzzy Integral, Non-additive Knowledge

Isfahan University of Technology

Department of Electrical and Computer Engineering

Improvments in speed and quality of learning in multi-agent systems using a novel expertness criteria and fuzzy integral

A Thesis

Submitted in partial fulfillment of the requirements
for the degree of Master of Science

by
Dariush Hasanpour Adeh

Evaluated and Approved by the Thesis Committee, on Jan. 9. 2016

1. Maziar Palhang, Assoc. Prof. (Supervisor)
2. Abdolreza mirzaie, Asst. Prof. (Examiner)
3. Mohamad Hosein Manshaie, Asst. Prof. (Examiner)

Mohamad Reza Taban, Department Graduate Coordinator