علوم و مهندسی کامپیوتر
نشریه علمی پژوهشی انجمن کامپیوتر ایران
مجلد ۳، شماره ۳ (الف)، پائیز ۱۳۸۴
صفحات ۷۱-۷۹
مقاله عادی

بررسی نقش کشف در یادگیری تقویتی چندعامله با توجه به نوع وظیفه

امیرحسین الهی‌بخش مجید نیلی احمدآبادی بابک نجاراعرابی

دانشکده مهندسی برق و کامپیوتر، دانشگاه تهران، تهران، ایران
پژوهشکدهٔ علوم شناختی، مرکز تحقیقات فیزیک نظری و ریاضیات، تهران، ایران

چکیده

در این مقاله تنظیم دما در یادگیری تقویتی با توجه به نوع وظیفه محیط مورد بررسی قرار می‌گیرد. دسته‌بندی بر اساس نوع وظیفه محیط شامل مدل‌های عطفی، فصلی و حالت ترکیبی عطفی و فصلی خواهد بود. بر مبنای نتایج بدست آمده مقدار دمای اولیه در محیط فصلی باید بالا انتخاب شده و سرعت کاهش آن نیز کم باشد. در محیط عطفی عکس این موضوع صادق است. محیط ترکیبی عطفی و فصلی نیز با توجه به نسبت تعداد عامل‌های افزونه به تعداد عامل‌های گروه عطفی، حالت بینابینی دارد. تحلیل‌های انجام شده بر اساس تعدیل رفتار انتفاع‌گرایانه یا جستجوگرایانه بنا شده است. در این تحقیق نتایج یادگیری در قالب یادگیری تیمی و یادگیری فردی بررسی می‌شود. در یادگیری تیمی سرعت و کیفیت همگرایی با توجه به نمودار طول رویداد بر حسب شماره آن و در یادگیری فردی شباهت جدول دانش عامل‌ها به جدول دانش بهینه مد نظر بوده است.

کلمات کلیدی: یادگیری تقویتی، تنظیم دما، نوع وظیفه محیط، وظیفه عطفی، وظیفه فصلی، وظیفه ترکیبی عطفی و فصلی، عامل‌های افزونه، سیستم‌های چندعامله، رفتار جستجوگرانه، رفتار انتفاع‌گرایانه

۱- مقدمه

استفاده از سیستم‌های چندعامله گذشته از تطابق با طبیعت مسائل، مزایای گوناگونی نظیر استفاده از افزونگی در عامل‌ها را به همراه دارد، اما به لحاظ گستردگی تصمیم‌سازی و تعدد عامل‌ها، طراحی ذهن عامل در یک سیستم چندعامله امری پیچیده و مشکل بوده و در بسیاری از موارد راه‌حل سیستماتیکی برای آن وجود ندارد. استفاده مناسب از یادگیری و بهره‌گیری از همکاری در یادگیری علاوه بر آنکه طراحی سیستم‌های چندعامله را ساده‌تر می‌نماید، امکان پویایی در محیط پویا را برای سیستم فراهم می‌آورد.

۱-۱ طبقه‌بندی سیستم از نظر نوع وظیفه

یکی از نکات مهم مطرح در انتخاب و تنظیم الگوریتم یادگیری، نوع وظیفه عامل‌ها در به هدف رساندن سیستم است. در وظایف عطفی[1]، همه عامل‌ها باید با هم به صورت هم‌زمان وظیفه خود را درست انجام دهند تا سیستم به هدف برسد، اما در وظایف فصلی[2] هر کدام از عامل‌ها که وظیفه خود را درست انجام دهند، سیستم به هدف می‌رسد. تقسیم‌بندی وظیفه محیط از این منظر ابتدا در [۵] مطرح شده و در تنظیم الگوریتم یادگیری مورد توجه قرار گرفت. در تحقیق مذکور تقسیم امتیاز بین عامل‌ها بر مبنای خبرگی با توجه به نوع وظیفه محیط مد نظر قرار گرفت.

استفاده از مفهوم خبرگی عامل‌ها با توجه به جدول دانش آنها برای اولین بار در [۹و۱۰] مطرح شد. پس از آن، این مفهوم در [۵] کمی توسعه پیدا کرده و در الگوریتم یادگیری تقویتی برای تقسیم امتیاز در مدل‌های عطفی و فصلی به طور مجزا در [۴و۶] مورد استفاده قرار گرفت. در پژوهش‌های مذکور بر مبنای تحلیل‌ها و نتایج بدست آمده از [۸] روشی برای تقسیم امتیاز مبتنی بر خبرگی عامل‌ها ارائه شده است. بررسی دیگر مباحث مطرح در یادگیری تقویتی در این محیط‌ها و همچنین ترکیب دو مدل عطفی و فصلی در کنار هم (مفهوم یادگیری در وظیفه عطفی با حضور عامل‌های افزونه[3])، موضوع مهمی بود که هدف پژوهش جاری ما را تشکیل می‌دهد. ما در این مقاله تنها نتایج بدست آمده مرتبط با تنظیم دمای یادگیری که در این تحقیق به آن دست یافته‌ایم را مطرح و بررسی می‌کنیم.

گفتنی است که بر اساس بررسی‌های ما پژوهش مشابهی که از این زاویه این مطلب را مورد مطالعه قرار داده باشد وجود ندارد.

۲-۱ وظیفه ترکیبی عطفی و فصلی

وقتی ذات بسیاری از سیستم‌های چندعامله را بررسی می‌کنیم درمی‌یابیم که سیستم دارای وظیفه‌ای است که همکاری و هماهنگی همه اعضا در کنار هم، آن را تحقق می‌بخشند. به این ترتیب از این مفهوم به وجود وظیفه‌ای عطفی پی می‌بریم که با همکاری همه انجام می‌شود. از طرف دیگر وجود عامل‌های افزونه، از دیگر خصوصیات این گونه سیستم‌هاست که وظیفه فصلی را تداعی می‌کند. با نگاهی کلی می‌توان دریافت که رفتار سیستم در بسیاری از جوامع چندعامله قابل مدل شدن با وظیفه ترکیبی عطفی و فصلی (مدل عطفی با حضور عامل‌های افزونه) است. در حین فعالیت عامل‌ها در این سیستم‌ها دسته‌ای از عامل‌ها هسته انجام دهنده وظیفه را تشکیل می‌دهند و دسته دیگر نقش فعالی ندارند. تشکیل دسته فعال به صورت کاملاً پویا بوده، لزوماً تصریح خاصی ندارد و تنها انجام وظیفه مهم است. نمونه مناسب بکارگیری این مفهوم در فوتبال ربات‌ها مطرح است. در [۱۱] با بکارگیری و دسته‌بندی مجموعه عامل‌های فعال و غیر فعال در هر لحظه، به ارائه الگوریتمی برای فوتبال ربات‌ها پرداخته شده است. با این توصیف با ارائه راه‌حل مناسب برای روند یادگیری در سیستم‌های دارای وظیفه ترکیبی عطفی و فصلی، مبحث یادگیری در طیف وسیعی از سیستم‌ها به طرز مناسب‌تری استفاده خواهد شد.

در محیط مذکور برای تحقق هدف زیرمجموعه‌ای کافی (با بیش از یک عضو) از عامل‌ها باید کار صحیح را هماهنگ انجام دهند. با الهام از طبیعت درمی‌یابیم که این نوع وظیفه‌مندی در بسیاری از جوامع جانداران نیز مرسوم است، بنابراین گام برداشتن در بسط و توسعه این روش ضروری و مفید می‌نماید. ما خود از نتایج این پژوهش برای بستر مهار اجسام در [۳و۲و۱] استفاده کرده‌ایم.

۳-۱ یادگیری تقویتی نوع Q

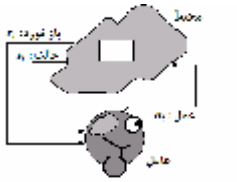
در این پژوهش مسأله تنظیم دما در یادگیری تقویتی نوع Q[۴] با توجه به وظیفه محیط مورد بررسی قرار می‌گیرد. این نوع یادگیری که ابتدا در[۱۴] al. ارائه شد، روشی مبتنی بر برنامه‌ریزی پویا است که با در نظر گرفتن محیط به صورت یک زنجیره مارکوف با حالت‌های متناهی و زمان گسسته، تقریبی از ارزش (پاداش مورد انتظار) هر حالت-عمل را در جدولی به نام Q نگهداری و به‌روزرسانی می‌کند. شکل ۱، شمایی از این مدل را نشان می‌دهد. در هر مرحله عامل با درک s_t بعنوان حالت جاری محیط، عمل a را از مجموعه اعمال ممکن(A_t) با احتمال $\Pr(a_i \mid s = s_t)$ انتخاب و اجرا می‌کند که این احتمال عموماً از رابطه (۱) که به ماشین بولتزمن موسوم است بدست می‌آید:

$$\Pr(a_i \mid s = s_t) = \frac{e^{Q(s_t,a_i)/T}}{\sum_{a_j \in A_t} e^{Q(s_t,a_j)/T}} \qquad (۱)$$

در این رابطه پارامتر T بیانگر میزان اتفاقی بودن تصمیم عامل است و برای کنترل کشف راه‌حل‌های جدید بکار می‌رود. پس از آن عامل با دریافت تقویت r و درک حالت جدید محیط (s_{t+1})، $Q(s_t,a_i)$ را با نرخ یادگیری $0 \le a \le 1$ و ضریب تضعیف $0 \le g \le 1$ با رابطه (۲) به‌روزرسانی می‌کند:

$$Q(s_t,a_i) \leftarrow (1-a)Q(s_t,a_i) + a(r + gV(s_{t+1}))$$
$$V(s) = \max_{b \in A_t} Q(s,b) \qquad (۲)$$

به این صورت تقریب‌های جدول Q پس از گذشت زمان و حصول تجربه کافی به اندازه واقعی خود نزدیک شده، عامل توان تصمیم‌گیری مناسب‌تر را پیدا خواهد کرد.[۱۳و۱۲و۷]



شکل۱- یادگیری تقویتی

با توجه به توصیف روال یادگیری، تنظیم تشویق و تنبیه (با توجه به نوع وظیفه در محیط) از نکات مهم مطرح در تنظیم پارامترها می‌باشد. ما در کلیه آزمایش‌ها خود در مدل فصلی بیشتر مبتنی بر تنبیه و در مدل عطفی بیشتر مبتنی بر تشویق عمل می‌کنیم. حصول هدف در یادگیری تقویتی چندعامله در محیط فصلی تنها حاکی از این است که حداقل یکی از عامل‌ها کار درست انجام است بنابراین هنگام دادن پاداش تیمی اطمینانی در این که کدام عامل یا عامل‌ها کار درست را انجام داده وجود ندارد. به همین ترتیب هنگام شکست تیم این اطمینان وجو دارد که همه کار نادرست انجام داده‌اند و با اطمینان می‌توان همه را تنبیه کرد. بنابراین در محیط فصلی فقط با داشتن امتیاز منفی می‌توان با یقین یادگیری سیستم را پیش برد. اما در محیط عطفی وقتی سیستم به هدف می‌رسد که همه عامل‌ها کار درست انجام داده باشند. بنابراین هنگام تنبیه تیمی اطمینانی در این که کدام عامل یا عامل‌ها کار درست را انجام نداده وجود ندارد. اما هنگام حصول هدف می‌توان اطمینان داشت که همه کار درست انجام داده‌اند و همه را تشویق کرد. بدین ترتیب در محیط عطفی با داشتن فقط امتیاز مثبت می‌توان با یقین یادگیری سیستم را پیش برد.

با ورود عامل‌های افزونه در محیط عطفی(محیط ترکیبی عطفی و فصلی) وقتی سیستم به هدف می‌رسد که حداقل به تعداد گروه عطفی عامل‌های درست کار وجود داشته باشد و حداکثر به تعداد عامل‌های افزونه عامل خطاکار موجود باشد. بنابراین هنگام تشویق تیمی اطمینانی در این که کدام عامل‌ها کار درست و کدام کار نادرست انجام داده‌اند وجود ندارد. هنگام شکست نیز فقط مشخص است که تعداد عامل‌های خطا کار بیشتر از تعداد عامل‌های افزونه است. بدین ترتیب در دادن امتیازات منفی و مثبت هیچ یقینی وجود ندارد، اما با توجه به این که تعداد عامل‌های افزونه معمولاً کمتر از تعداد عامل‌های اصلی است می‌توان گفت که محیط به مدل عطفی نزدیک‌تر است و در نتیجه در دادن امتیاز مثبت یقین بیشتری نسبت به دادن امتیاز منفی وجود دارد.

بدین ترتیب برای محیط فصلی یادگیری مبتنی بر تنبیه و برای محیط عطفی یادگیری مبتنی بر پاداش مناسب است. علاوه بر این مطلب بر اساس نتایج آزمایش‌هایی که در اینجا گزارش نشده‌اند یادگیری تیمی در محیط فصلی نسبت به امتیاز مثبت، مقاوم‌تر از یادگیری تیمی در محیط عطفی نسبت به امتیاز منفی است. توجیه این امر برای محیط فصلی در این است که امتیاز مثبت روی یادگیری فردی عامل‌ها تأثیر نامناسب دارد اما تأثیر زیادی روی واگرا شدن یادگیری تیمی ندارد. بدین ترتیب در مدل ترکیبی عطفی و فصلی نیز بهتر است بیشتر شبیه مدل عطفی عمل کنیم و از یادگیری مبتنی بر پاداش همراه با تنبیه جزئی استفاده کنیم.

حال با این تفاسیر و پی بردن به وجود نایقینی در امتیازدهی در محیط‌های مختلف بر آنیم تا از پارامترهای دیگر یادگیری نیز در این راستا استفاده کرده و با تنظیم مناسب آنها در کاهش این نایقینی بکوشیم. یکی از این پارامترها تنظیم دمای یادگیری در این شرایط است که هدف اصلی این مقاله را تشکیل می‌دهد.

ذکر این نکته در این جا لازم است که اگر قانون تقسیم امتیاز مناسبی داشته باشیم و امتیاز همراه با اطمینان تخصیص پیدا کند، مباحث مطرح در این مقاله برای تنظیم دما که ریشه در جستجوگرانه و انتفاع‌گرانه بودن روش دارد بسیار کمرنگ‌تر می‌شود.

۱-۴ ساختار مقاله

در ادامه مبحث تنظیم دما با توجه به نوع وظیفه محیط مطرح می‌شود. این موضوع برای محیط‌های فصلی، عطفی و محیط ترکیبی عطفی و فصلی به صورت مجزا بررسی می‌شود. سپس در بخش شبیه‌سازی محیط آزمایش‌ها، الگوریتم یادگیری و تنظیم برخی پارامترهای الگوریتم یادگیری مورد تحلیل قرار می‌گیرد. در ادامه آزمایش‌ها و نتایج آنها به تفصیل بررسی می‌شود. در انتها نیز به نتیجه‌گیری و پیشنهاد برای ادامه روند این پژوهش خواهیم پرداخت.

۲- تنظیم دما با توجه به نوع وظیفه

روش تصمیم‌گیری در انتخاب عمل با توجه به مقادیر جاری جدول ارزش اعمال و حالات[۵]، از قسمت‌های اساسی مطرح در الگوریتم یادگیری تقویتی می‌باشد. وقتی از الگوریتم نوع Q همراه با روش انتخاب عمل ماشین بولتزمن[۶] استفاده می‌کنیم، مقدار اولیه دما و سرعت کاهش آن، جستجوگری و انتفاع‌گری الگوریتم را در خلال مراحل مختلف یادگیری تنظیم می‌کند. در این روش احتمال انتخاب هر عمل به صورت نمایی متناسب با میزان ارزش آن است و از پارامتر T به عنوان دمای سیستم برای کنترل میزان جستجو برای شناسایی و تجربه حالات جدید استفاده می‌شود. بین مقدار دما و رفتار جستجوگرانه رابطه مستقیم وجود دارد و هر چه مقدار اولیه دما بیشتر و سرعت کاهش آن کمتر باشد رفتار جستجوگرانه‌تر است. رابطه (۳) احتمال انتخاب را در این روش نشان می‌دهد.

$$\Pr(a_i \mid s = s_t) = \frac{e^{Q(s_t,a_i)/T}}{\sum_{a_j \in A_t} e^{Q(s_t,a_j)/T}} \qquad (۳)$$

به طور معمول دما معمولاً از یک مقدار بیشینه شروع شده و بر اساس رابطه (۴) به مرور زمان کم می‌شود. گذشت زمان در قالب تعداد تلاش‌های انجام شده و تعداد مراحل به‌روز شدن جدول دانش عامل نشان داده می‌شود. برای جلوگیری از صفر شدن دما می‌توان مقدار کمینه‌ای ($T_{\min}$) برای آن در نظر گرفت.

$$T = \max\left(\frac{T_{\max}}{\log(trial\ number + 1)}, T_{\min}\right) \qquad (۴)$$

بدین ترتیب با توجه به تأثیر دما در روند یادگیری، در ادامه به مطالعه آن بر اساس نوع وظیفه محیط می‌پردازیم.

۲-۱ محیط فصلی

در حالت فصلی اغلب سیستم سریع به هدف می‌رسد و نرسیدن به هدف معدود است. در این محیط عامل‌ها دیر به دیر تنبیه م ی‌شوند. فرکانس به هدف رسیدن تیم بیشتر از فرکانس به هدف رسیدن یک عامل است. الگوریتم یادگیری به گونه‌ای است که موفقیت پی‌درپی تیم را به سمت رفتار انتفاعی پیش می‌برد. با گرفتن پاداش برای عملی خاص مقدار ارزش آن رشد می‌کند و این کار خود باعث انتخاب مجدد آن در ادامه کار می‌شود و به همین ترتیب به صورت مداوم، رفتار انتفاعی‌تر می‌شود. با کمی دقت درمی‌یابیم که در محیط فصلی ممکن است به عنوان نمونه تنها یک عامل اعمال درست را یاد بگیرد و بقیه همه اعمال اشتباه انجام دهند و غلط یاد بگیرند، زیرا حتی با این کار نیز تیم تشویق می‌شود. اگر الگوریتم درستی انتخاب نشود، یادگیری‌های فردی تک تک عامل‌ها دچار اخلال می‌شود. بدین ترتیب با استفاده از دمای اولیه زیاد و کم کردن سرعت کاهش دما می‌توان ماهیت انتفاعی بودن الگوریتم را تعدیل کرده، آن را به سمت جستجوگرا بودن سوق دهیم.

۲-۲ محیط عطفی

در محیط عطفی همه باید هم‌زمان کار درست حالت جاری خود را انتخاب کرده و انجام دهند تا تیم به هدف برسد، بدین ترتیب رسیدن به هدف معدود است. بارها عامل عمل درست انجام می‌دهد ولی تشویق نمی‌شود. بنابراین چون سیستم دیر به هدف می‌رسد امتیازها دیر به دیر می‌رسند و تفاوت بین ارزش اعمال تشدید نمی‌شود. وقتی تفاوت ارزش حالات تشدید نمی‌شود، رفتار سیستم بیشتر به روش جستجوگرانه نزدیک است. در واقع کاهش فرکانس تشویق سیستم را جستجوگرا می‌کند. حال که رفتار طبیعی سیستم جستجوگراست، برای برقراری توازن باید سعی شود تا در الگوریتم آن را انتفاع‌گرا کنیم. بدین ترتیب با دمای اولیه کمتر و افزایش سرعت کاهش دما می‌توان به نتایج بهتری دست یافت.

۲-۳ محیط ترکیبی عطفی و فصلی

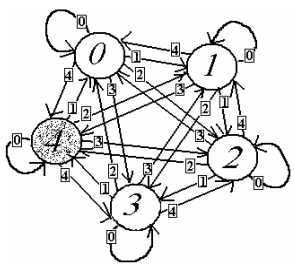
در این حالت توضیحات مذکور برای مدل عطفی و مدل فصلی جنبه بینابینی پیدا می‌کنند. بسته به این که تعداد عامل‌های گروه عطفی با تعداد عامل‌های افزونه چه نسبتی داشته باشند به یکی از مدل‌های مذکور نزدیک می‌شود. اگر عامل‌های افزونه کمتر باشند، بیشتر شبیه مدل عطفی و در غیر این‌صورت شبیه مدل فصلی خواهد بود.

۳- محیط شبیه‌سازی

در این آزمایش‌ها محیطی بسیار ساده و انتزاعی در نظر گرفته شده است تا به طور واضح و روشن یادگیری در محیط عطفی و فصلی ترکیبی بررسی شود. تعدادی عامل داریم هر یک دارای متغیر حالت پنج گزینه‌ای برای تشخیص محیط هستند. مجموعه حالات با صفر تا چهار نشان داده می‌شود. در هر حالت نیز پنج عمل مختلف وجود دارد که با اعداد صفر تا چهار مشخص می‌شود. عدد متناظر با عمل در هر حالت مقدار شیفت عامل را در حرکت بین حالت‌ها بیان می‌کند، بدین ترتیب که :

$$S_{t+1}(S_t, a) = (S_t + a) \bmod 5 \qquad (۵)$$

بنابراین شکل ۲ انتقال بین حالات مختلف با توجه به عمل جاری را برای هر عامل نشان می‌دهد.



شکل ۲- انتقال حالات محیط

یادگیری انفرادی هر عامل این است که با انتخاب عمل مناسب در هر حالت بتواند به حالت شماره چهار برسد.

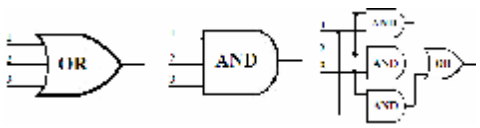
$$S_g = 4 \qquad (۶)$$

وقتی به تعداد m عامل یا بیشتر به حالت شماره چهار خود برسند، سیستم به هدف رسیده است. وقتی تنها m عامل داریم وظیفه محیط عطفی است و در صورتیکه بیش از m عامل داشته باشیم، وظیفه محیط حالت ترکیبی عطفی و فصلی خواهد بود. برای داشتن مدل فصلی نیز، بنا را بر این می‌گذاریم که با داشتن n عامل اگر هر کدام از عامل‌ها به هدف رسید سیستم به هدف رسیده است. بدین ترتیب با محیط ساده مذکور هر سه مدل وظیفه عطفی، فصلی و ترکیبی عطفی و فصلی را خواهیم داشت.

با این توصیف اگر n عامل در محیط داشته باشیم و شرط به هدف رسیدن سیستم، رسیدن به هدف به صورت همزمان برای m عامل باشد، آنگاه می‌توان نوع وظیفه محیط را بر مبنای عبارات زیر بیان کرد.

$$\begin{aligned} &if\ (n>1,\ m=1) \Rightarrow OR\ type\ task \\ &if\ (n>1,\ m=n) \Rightarrow AND\ type\ task \\ &if\ (n>1,\ n>m>1) \Rightarrow AND-OR\ type\ task \end{aligned} \qquad (۷)$$

شکل ۳ در راستای توصیف بهتر نمونه‌ای از هرکدام از انواع وظایف محیط را در قالب مدار عطف و فصل نمایش می‌دهد. در اشکال زیر n=3 انتخاب شده است. در حالت ترکیبی عطفی و فصلی m=2 انتخاب شده است. بدین ترتیب در این محیط سه عامله تعداد گروه عطفی برابر دو بوده، یک عامل افزونه نیز داریم. در تیمی سه عامله که یک عامل آن افزونه باشد، هرگاه دوتا از عامل‌ها به صورت هم‌زمان به هدف برسند، تیم به هدف می‌رسد.



شکل۳- توصیف مداری انواع محیط

در این محیط ساده رسیدن از هر حالت به هدف با یک گام امکان پذیر است. ضریب انتقال ارزش حالت (۷)، در آزمایش‌های مربوط به این محیط برابر صفر در نظر گرفته خواهد شد.

با توجه به توصیف عامل‌ها جدول Q مربوط به ارزش حالات و اعمال را می‌توان به صورت ۵×۵ نمایش داد. پس از یادگیری کامل عامل، ارزش خانه‌های روی قطر فرعی این جدول باید بیشینه و ارزش بقیه خانه‌ها به‌طور مساوی کمینه باشد. شکل ۴ این جدول را نشان می‌دهد. رنگ تیره نمایانگر ارزش عمل است.

شکل۴- جدول دانش فردی

در ادامه کار با توجه به شبیه بودن جدول دانش بدست آمده برای عامل به شکل بالا، در مورد میزان یادگیری او نتیجه‌گیری خواهیم کرد. بدین ترتیب هرچه خانه‌های روی قطر فرعی تیره‌تر و بقیه خانه روشن‌تر باشند، یادگیری بهتر صورت پذیرفته است.

برای بررسی میزان یادگیری انفرادی عامل‌ها هنگام یادگیری تیمی، بعد از تعداد معینی رویداد در تیم در واقع باید عامل را از تیم جدا کرده و یادگیری آن را نیز متوقف کنیم. سپس عملکرد آن را مورد آزمایش قرار دهیم و میزان به هدف رسیدن آن در محیط تک‌عامله را بررسی کنیم. اما ما به عنوان جایگزینی برای این کار جدول دانش عامل را بازبینی کرده و در مورد میزان دانش او نتیجه‌گیری می‌کنیم. این روش بعضاً می‌تواند دقیق‌تر از روش قبلی باشد. در واقع با متوقف کردن یادگیری و حریصانه کردن مدل عملکرد عامل، تنها با بیشینه بودن مقدار خانه روی قطر فرعی در هر سطر (هر حالت) نسبت به بقیه خانه‌های آن سطر، عامل همواره درست عمل خواهد کرد. در صورتیکه با بررسی و بازبینی جدول دانش می‌توان به نتیجه دقیق‌تری در مورد سطح دانش عامل‌هایی که با انتخاب عمل حریصانه، عملکرد یکسان دارند دست یافت.

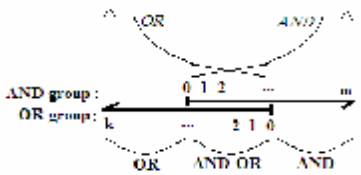
در خلال آزمایش‌های این تحقیق بعد از تعدادی رویداد در یادگیری تیمی با مراجعه به جدول دانش عامل‌ها، میزان یادگیری فردی آنها را مورد ارزیابی قرار می‌دهیم. در خلال مراحل یادگیری تیمی نیز با توجه به نمودار طول رویداد می‌توان وضعیت دانش جاری عامل‌ها در بازه زمانی خاص را تخمین زد.

۴- یادگیری فردی و تیمی

در محیط عطفی معمولاً عملکرد تیمی مناسب، نشان‌گر دانش مناسب فردی همه عامل‌ها می‌باشد، ولی درست عمل نکردن تیم لزوماً به معنی دانش نامناسب در همه نیست. وقتی تیم پیاپی به هدف نمی‌رسد تنها می‌توان مطمئن بود که حداقل یک عامل در حداقل یک حالت، عمل درست را بلد نیست. در محیط فصلی دقیقاً عکس این موضوع صادق است. درست عمل کردن تیم حاکی از دانش مناسب همه نیست و تنها این را بیان می‌دارد که در هر حالت حداقل یک عامل عمل درست را بلد می‌باشد. اما هنگام عملکرد نامناسب تیم در محیط فصلی می‌توان به نامناسب بودن دانش فردی تک‌تک عامل‌ها پی برد.

اشاره به مطلب دیگری در مورد شدت عطفی یا فصلی بودن محیط نیز ضروری می‌نماید. هر چه تعداد عامل‌ها در وظیفه عطفی یا فصلی بیشتر باشند شدت رفتار خاص آن نوع محیط بیشتر می‌شود. به عنوان نمونه مباحث مطرح در محیط عطفی با شش عامل، بسیار مشهودتر و بغرنج‌تر از محیط عطفی با سه عامل است. در واقع با در نظر گرفتن احتمالات موفقیت و شکست تیمی در این محیط‌ها به نسبت حالت تک‌عامله می‌توان گفت شدت رفتار عطفی یا فصلی محیط به صورت نمایی با تعداد عامل‌های محیط تناسب دارد. وقتی محیط حالت ترکیبی پیدا می‌کند دیگر شدت بروز رفتار فصلی یا عطفی به تعدد عامل‌ها بستگی ندارد و تنها با توجه به نسبت تعداد عامل‌های گروه عطف به تعداد عامل‌های افزونه قابل پیش‌بینی می‌باشد. البته ما در صدد اثبات ادعاهای اخیر برنخواهیم آمد.

شکل ۵ به گونه‌ای شدت عطفی یا فصلی بودن محیط را با توجه به توضیحات مذکور نمایش می‌دهد. اعداد زیر محور افقی تعداد عامل‌های گروه فصلی و اعداد بالای محور تعداد عامل‌های گروه عطفی را نشان می‌دهد. بنابراین در قسمت همپوشانی این دو گروه (ناحیه وسط محورافقی) مدل ترکیبی عطفی و فصلی داریم. هر چه به قسمت راست محور پیش برویم مدل عطفی‌تر و هرچه به سمت چپ پیش برویم مدل فصلی‌تر می‌شود. شدت عطفی و فصلی شدن محیط در قالب منحنی نمایش داده است. بر اساس مباحث مذکور تغییرات شدت عطفی یا فصلی بودن محیط در ناحیه ترکیبی (ناحیه وسط محورافقی) به صورت خطی رسم شده است.



شکل ۵- شدت عطفی یا فصلی بودن محیط

بنابراین مباحثی که برای تنظیم پارامترها مطرح می‌شوند با توجه به میزان شدت عطفی یا فصلی بودن محیط تشدید می‌شوند.

۵- الگوریتم یادگیری

الگوریتمی که در این تحقیق بکار خواهیم گرفت به گونه‌ای مستقل از نوع وظیفه عمل خواهد کرد. این الگوریتم بسیار ساده و مشابه با یادگیری تقویتی نوع Q در محیط تک عامله در نظر گرفته شده است. در این الگوریتم تنها قسمت مربوط به بررسی رسیدن به هدف تیمی با توجه به نوع عطفی، فصلی و یا ترکیبی بودن وظیفه محیط متفاوت خواهد بود. روال الگوریتم را می‌توان به صورت زیر توصیف کرد.

۱. ابتدا عامل‌ها به صورت تصادفی در حالت‌های مختلف قرار گرفته و رویداد جدید آغاز می‌شود.
۲. همه عامل‌ها با توجه به وضعیت دانش و قانون تصمیم‌گیری، عمل خود را انتخاب کرده هر کدام بعد از انجام عمل، حالت جدید را درک می‌کنند.
۳. سپس محیط رسیدن به هدف تیمی را بررسی کرده با توجه به آن امتیاز تیمی را صادر می‌کند.
۴. این امتیاز به صورت مساوی به همه عامل‌ها ارائه شده و همه بر مبنای آن یادگیری می‌کنند. (اگر از قانون تقسیم امتیاز خاصی استفاده کنیم، بر مبنای آن امتیازدهی صورت می‌پذیرد.)
۵. با این کار به تعداد تلاش‌های رویداد جاری یکی اضافه می‌شود.
۶. اگر تیم به هدف نرسیده است در صورتیکه تعداد تلاش از مقدار بیشینه در نظر گرفته شده (۳۰۰) تجاوز نکرده باشد، پروسه به شماره ۲ برمی‌گردد و در غیر این صورت رویداد همراه با شکست خاتمه می‌یابد.
۷. اگر تیم به هدف رسیده باشد تعداد تلاش انجام شده به عنوان طول رویداد جاری ثبت می‌شود.
۸. به تعداد رویدادهای صورت پذیرفته از ابتدای شبیه‌سازی یکی اضافه می‌شود.
۹. اگر نیاز به ادامه شبیه‌سازی باشد پروسه از شماره ۱ از سر گرفته خواهد شد.

به منظور شبیه‌سازی محیط واقعی و حل مدل مسأله برنامه‌ای جامع نوشته شد. در این برنامه که با استفاده از مدل شیءگرا و زبان java طراحی و پیاده‌سازی شد، هسته‌ای با قابلیت استفاده مجدد برای کارهایی که نیاز به یادگیری تقویتی دارند طراحی و تعبیه شد. کلیه آزمایش‌های مختلف در طول پروژه به همراه توسعه مجموعه اعمال و حالات مختلف با استفاده از همین برنامه انجام شده است.

معماری نرم‌افزار مذکور به گونه‌ای تدوین شد که با پیاده‌سازی کلاس‌هایی که تنها از واسط خاصی تبعیت می‌کنند، می‌توان آن را در محیط‌های مختلف برای شبیه‌سازی یادگیری بکار برد. هر کدام از این محیط‌ها مشخصات خاص خود را به همراه نمایش محیط فعالیت عامل‌ها دارا خواهند بود. از جمله قابلیت‌های این نرم‌افزار می‌توان به مواردی مثل ارائه الگوریتم موازی یا سری برای یادگیری، انواع روش‌های تصمیم‌گیری برای انتخاب عمل، کار براساس انواع خبرگی، بکارگیری انواع قوانین تقسیم امتیاز، ذخیره و بازیابی وضعیت و دانش‌عامل‌ها در فایل، قابلیت پی‌گیری یادگیری بر اساس وضعیت ذخیره شده در اجراهای قبلی، نمایش لحظه به لحظه جدول دانش عامل‌ها، نمایش همزمان نمودار طول رویداد در زمان یادگیری و تنظیم سرعت عمل کردن عامل‌ها در محیط به صورت پویا به منظور مشاهده جزئیات و مراحل انجام کار توسط آنها اشاره کرد.

۶- نتایج آزمایش‌ها

۶-۱ آزمایش در محیط فصلی

محیط ساده توصیف شده را به صورت فصلی در نظر می‌گیریم و با رسیدن یک عامل به هدف سیستم به هدف می‌رسد. چهار عامل در محیط حضور دارند. الگوریتم را به صورت موازی بکار می‌گیریم. هنگام موفقیت تیمی ۵ امتیاز مثبت و هنگام شکست ۱۰ امتیاز منفی به تیم داده می‌شود. امتیاز تیمی به همه عامل‌ها داده می‌شود. نمودارهای زیر طول متوسط رویداد برای رسیدن به هدف را طی مراحل یادگیری نشان می‌دهد. الگوریتم یادگیری Q-Learning و ضریب γ برابر صفر در نظر گرفته شده است. در کلیه آزمایش‌های ما نرخ یادگیری بر مبنای رابطه زیر در طول یادگیری تغییر خواهد کرد. مقادیر $a_{initial}$ ،$a_{\min}$ و $T_{\min}$ به ترتیب برابر ۰،۷، ۰،۰۵ و ۰،۵ انتخاب می‌شوند.

$$a = \max(a_{initial} - \frac{\log(trialnumber+1)}{10}, a_{\min}) \quad (۸)$$

نتایج آزمایش با استفاده از دمای اولیه ۱، ۵، ۱۰ و ۵۰ طبق رابطه توصیف شده برای تغییرات دما (رابطه ۳) بررسی شدند. در اشکال ۶ و ۷ دو نمونه از نمودارهای طول یادگیری برای دو حالت دمای اولیه ۱ و ۵۰ نشان داده شده است.

به منظور تلخیص، نمودارهای همه آزمایش‌ها ترسیم نشده‌اند و در اینجا تنها به نمایش دو نمونه بسنده کرده‌ایم. بر اساس نتایج آزمایش‌ها افزایش دما از ۱ به ۵ و از ۵ به ۱۰ در متوسط طول رویداد و به هدف رسیدن تیم تفاوت چندانی محسوس نیست (نمودار طول یادگیری برای این موارد شبیه شکل ۶ است). اما همانطور که در شکل ۷ مشهود است با زیاد کردن خیلی زیاد دما، یادگیری نامناسب شده و تیم کمی انتفاع‌گری خود را از دست می‌دهد. حال باید یادگیری فردی عامل‌ها را برای دماهای اولیه ۱، ۵ و ۱۰ مقایسه کنیم، تا بتوان در مورد دما در محیط فصلی نتیجه گرفت. شکل۸ جدول Q عامل‌های شرکت کننده را بعد از ۵۰۰ تلاش نشان می‌دهد. توجه کنید که هر چه خانه‌های روی قطر فرعی سیاه‌تر و بقیه خانه‌ها روشن باشند، یادگیری بهتر بوده است.

بدین ترتیب نتایج آزمایش حاکی از بهتر شدن یادگیری فردی عامل‌ها با افزایش دما می‌باشد. البته اگر دمای اولیه را خیلی زیاد کنیم(۵۰)، یادگیری تیمی انتفاع‌گری خود را از دست می‌دهد و نتیجه نامطلوب می‌شود.

طی آزمایش‌های مذکور نتیجه انتخاب مقادیر کم یا زیاد برای دمای اولیه را بررسی کردیم. در ادامه در آزمایش ساده دیگری تاثیر نرخ (سرعت) کاهش دما برای محیط فصلی را بررسی می‌کنیم. در روال آزمایش‌های فوق دما بر مبنای رابطه (۴) در طول تلاش‌های یادگیری کم می‌شود. بر این اساس سرعت کاهش دما در ابتدا بسیار زیاد بوده و از یک مقدار اولیه بالا به سرعت کم شده و در ادامه با سرعت بسیار کمتری به یک مقدار حدی کاهش می‌یابد. بر خلاف رابطه (۴) برای کم کردن سرعت کاهش دما در مراحل اولیه دما می‌توانیم برای نمونه از مدل خطی استفاده کنیم که به مرور گذشت زمان یادگیری دما از مقدار اولیه خود به صورت خطی و یکنواخت تا مقدار حدی کمینه خود کاهش یابد.

به عنوان نمونه آزمایش بالا را با دمای اولیه ۵ ولی با سرعت کاهش دمای متفاوت دوباره انجام می‌دهیم. این بار در طول یادگیری دما را به صورت خطی کم می‌کنیم. به این ترتیب که دما با مقدار اولیه ۵ آغاز شده و در طول ۳۰۰ رویداد به صورت خطی کاهش پیدا کرده به مقدار کمینه خود ($T_{\min}$) رسیده و از آن پس ثابت می‌ماند. نکته این که تنها منظور استفاده از مدل خطی، کم‌کردن سرعت کاهش دما می‌باشد و می‌توان از هر مدل دیگری که این منظور را تأمین کند استفاده کرد. با استفاده از این مدل تنظیم دما، نمودار طول رویداد و یادگیری فردی در اشکال ۹ و ۱۰ ترسیم شده است. در شکل ۹ جدول بالا برای کاهش سرعت خطی و جدول پایین برای کاهش سرعت لگاریتمی رسم شده است.

همانطور که مشاهده می‌شود سرعت یادگیری تیمی به خاطر کم کردن انتفاع‌گری تیم، تنها کمی کمتر شده اما یادگیری فردی عامل‌ها بهتر شده است. بنابراین در مدل فصلی بهتر است سرعت کاهش دما نیز کم باشد.

۶-۲ آزمایش در محیط عطفی

محیط توصیف شده را به صورت عطفی در نظر می‌گیریم و با رسیدن همه عامل‌ها به صورت هم‌زمان به هدف، سیستم به هدف می‌رسد. چهار عامل در محیط حضور

دارند. الگوریتم را به صورت موازی بکار می‌گیریم. برای رسیدن به هدف ۱۰ امتیاز مثبت و برای نرسیدن به هدف ۲ امتیاز منفی به تیم داده می‌شود. امتیاز تیمی به همه عامل‌ها داده می‌شود. نمودارهای زیر طول متوسط رویداد برای رسیدن به هدف را طی مراحل یادگیری نشان می‌دهد. الگوریتم یادگیری Q-Learning و ضریب γ برابر صفر در نظر گرفته شده است. نتایج آزمایش با استفاده از دمای اولیه ۱، ۵، ۱۰ و ۵۰ بررسی شدند. اشکال ۱۱ و ۱۲ نمودارهای طول یادگیری برای مقادیر ۱و ۵۰ را نشان می‌دهد.

نتیجه آزمایش‌ها حاکی از این است که با دمای کم نتیجه بهتر می‌توان گرفت و با زیاد شدن دما یادگیری با این شرایط واگرا می‌شود. البته ذکر این نکته الزامی است که اگر تنبیه وجود نداشت با دمای بالا نیز یادگیری واگرا نمی‌شد. همانطور که انتظار می‌رود کیفیت یادگیری فردی مطابق شکل ۱۳ است. با این شرایط یادگیری در محیط عطفی حساسیت زیادی به اکثر پارامترها از جمله دما دارد.

۳-۶ آزمایش در محیط ترکیبی عطفی و فصلی

محیط توصیف شده را به صورت ترکیبی عطفی و فصلی در نظر می‌گیریم و با رسیدن تعدادی از عامل‌ها به هدف به صورت همزمان، سیستم به هدف می‌رسد. الگوریتم را به صورت موازی بکار می‌گیریم. برای رسیدن به هدف ۱۰ امتیاز مثبت و برای نرسیدن به هدف ۵ امتیاز منفی به تیم داده می‌شود. امتیاز تیمی به همه عامل‌ها داده می‌شود. نمودارهای زیر طول متوسط رویداد برای رسیدن به هدف را طی مراحل یادگیری نشان می‌دهد. الگوریتم یادگیری Q-Learning و ضریب γ برابر صفر در نظر گرفته شده است. نکته اینکه اگر امتیاز مثبت و منفی از لحاظ مقداری با هم برابر باشند آزمایش‌هایی که در اینجا گزارش نشده‌اند نشان از واگرا شدن یادگیری در محیطی که متمایل به مدل عطفی باشد دارند.

این آزمایش را در دو مرحله انجام می‌دهیم. در مرحله اول چهار عامل در محیط حضور دارند و با انجام عمل درست دو عامل، تیم به هدف می‌رسد. بدین ترتیب تعداد عامل‌های گروه عطفی برابر دو و تعداد عامل‌های گروه افزونه نیز برابر دو است. در آزمایش مرحله بعد شش عامل در محیط حضور دارند و با انجام عمل درست چهار عامل، تیم به هدف می‌رسد. بدین ترتیب تعداد عامل‌های گروه عطفی برابر چهار و تعداد عامل‌های گروه افزونه برابر دو است. در مرحله دوم محیط بیشتر شبیه مدل عطفی خواهد بود.

نتایج آزمایش را برای مدل چهار عامله با استفاده از دمای اولیه ۱، ۵، ۱۰ و ۵۰ بررسی کردیم. در شکل‌های ۱۴ و ۱۵ نمودارهای یادگیری دو حالت دمای اولیه ۱ و ۵۰ رسم شده‌اند.

این محیط حالت بینابینی عطف و فصل را دارد، زیرا تعداد عامل‌های افزونه برابر گروه عطفی است. بدین ترتیب حالتی بینابینی برای دمای اولیه مناسب است. در نمودارهای حاصل(که همه لزوما در اینجا رسم نشده‌اند)، دمای کمتر نتیجه کمی مطلوب‌تر نشان می‌دهد اما یادگیری فردی عامل‌ها که در شکل ۱۶ مشهود است حاکی از این است که با دمای اولیه ۵ نتیجه بهتری می‌گیریم.

در آزمایش مرحله بعد که تعداد عامل‌های افزونه نسبت به گروه عطفی کمتر است، به نمودارهای نشان داده شده در اشکال ۱۷ و ۱۸ می‌رسیم.

در این آزمایش‌ها باید مقدار تنبیه را کم کنیم، در غیر این صورت یادگیری همگرا نمی‌شود. مقدار تنبیه را دو امتیاز منفی در نظر می‌گیریم. از آنجایی‌که محیط مذکور شبیه محیط عطفی شده است، لذا افزایش دما نیز نتیجه مطلوبی ندارد. بنابراین بهترین نتیجه در دمای کم‌تر است. نتیجه یادگیری فردی عامل‌ها نیز در دماهای اولیه مختلف در شکل ۱۹ نشان داده شده است.

۷- نتیجه‌گیری

با توجه نتایج تحقیقات گذشته عمدتاً یادگیری در محیط عطفی مبتنی بر تشویق و در محیط عطفی مبتنی بر تنبیه صورت می‌پذیرد. اما در محیط عطفی با حضور عاملهای افزونه حالت بینابینی پدید آمده و تنبیه و تشویق باید به نسبت مناسب انتخاب شود.

در شرایط یکسان وقتی وظیفه محیط به صورت فصلی باشد اغلب سیستم سریع به هدف می‌رسد، فرکانس به هدف رسیدن تیم بیشتر از فرکانس به هدف رسیدن یک عامل است، عامل‌ها دیر به دیر تنبیه می‌شوند، با گرفتن پاداش برای عملی خاص مقدار ارزش آن رشد می‌کند و این کار خود باعث انتخاب مجدد آن در ادامه کار می‌شود و به همین ترتیب به صورت مداوم رفتار انتفاعی‌تر می‌شود و در نتیجه الگوریتم یادگیری به گونه‌ای است که موفقیت پی‌درپی تیم را به سمت رفتار انتفاعی پیش می‌برد.

اما اگر وظیفه عطفی باشد همه عاملها باید عمل مناسب حالت جاری خود را انجام دهند تا تیم به هدف برسد، بدین ترتیب رسیدن به هدف معدود است. بارها عامل عمل درست انجام می‌دهد ولی به خاطر اشتباه دیگران تشویق نمی‌شود. بنابراین چون سیستم دیر به هدف می‌رسد امتیازها دیر به دیر می‌رسند و تفاوت بین ارزش اعمال تشدید نمی‌شود و در نتیجه رفتار سیستم بیشتر به روش جستجوگرانه نزدیک است. در واقع کاهش فرکانس تشویق سیستم را جستجوگرا می‌کند.

بدین ترتیب همانطور که نتایج آزمایشها هم نشان می‌دهند در محیط عطفی برای انتفاع‌گر کردن رفتار، دمای اولیه کم و در محیط فصلی برای جستجوگر کردن رفتار، دمای اولیه بالا نیاز است. در حالت ترکیبی عطفی و فصلی با توجه به نزدیک بودن به مدل عطفی یا فصلی این مورد حالت بینابینی پیدا می‌کند. نکته مورد توجه این که در محیط فصلی با این که با افزایش دما نتیجه یادگیری فردی عامل‌ها بهتر می‌شود اما اگر دما را خیلی زیاد کنیم یادگیری تیمی انتفاع‌گرایی خود را از دست می‌دهد و نتیجه عکس می‌شود.

علاوه بر این بر اساس نتایج آزمایش‌ها، پارامتر دما و تغییرات آن در یادگیری تقویتی در محیط عطفی معمولاً روی یادگیری تیمی و در محیط فصلی روی یادگیری فردی تأثیر بیشتر دارند. در واقع در مدل عطفی لازمه یادگیری تیمی یادگیری فردی خوب عامل‌هاست، اما در مدل فصلی این چنین نیست. بنابراین وقتی یادگیری تیمی مناسب در مدل عطفی داشته باشیم می‌توان از یادگیری فردی مناسب نیز، مطمئن بود ولی در مدل فصلی با یادگیری خوب تیمی ممکن است یادگیری فردی نامناسبی داشته باشیم.

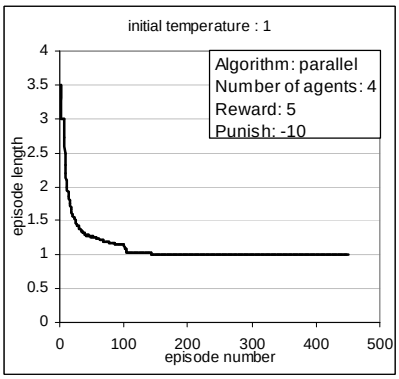
برای بررسی نتایج محیطی بسیار ساده را انتخاب کردیم تا به سادگی مستقیماً آثار تغییرات پارامترها مشهود باشد. اما در قسمتی پژوهشهای دیگر این نتایج را در کنار مباحث دیگر برای محیط‌های پیچیده‌تر مانند مهار اجسام[۷] بکار برده‌ایم. [۲و۱] تشریح این موارد در منابع مذکور قابل بررسی و از مجال مقاله جاری خارج است.

ما در ادامه این پژوهش به بررسی روش تقسیم امتیازی که باعث کاهش نایقینی و در نتیجه کم‌رنگ‌تر شدن مشکلات موجود در رفتار الگوریتم یادگیری خواهیم پرداخت. از دیگر گام‌های ادامه این پژوهش، محاسبه فاکتوری عددی می‌باشد که با توجه به نوع وظیفه محیط مقداردهی شده و در تنظیم پارامترهای یادگیری به عنوان متغیری جدید دخیل شود. در مرحله بعد باید بررسی شود آیا می‌توان این فاکتور را بدون دانستن شرایط محیط به صورت خودکار با توجه به مشاهده روند یادگیری به صورت پویا محاسبه و تنظیم کرد.

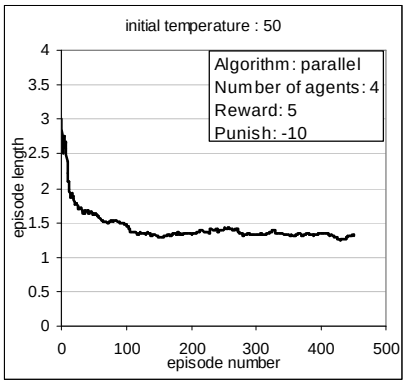
منابع

[1] A. H. Elahibakhsh, M. Nili Ahmadabadi, F. Janabi Sharifi and B. N. Araabi, “Passive Form Closure around a Rectangular Object Using Four Distributed Q_Learning Agents”, *Proc. 9th Annual Computer Society of Iran Conference*, Tehran, Iran, 2004.

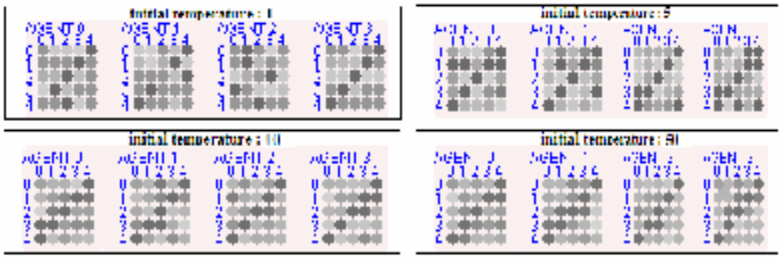
[2] A. H. Elahibakhsh, M. Nili Ahmadabadi, F. Janabi Sharifi and B. N. Araabi, "Distributed From Closure for Convex Planar Objects through Reinforcement



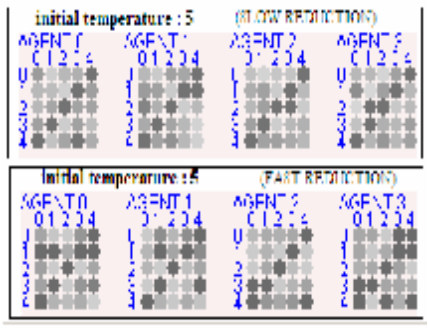
شکل۶- محیط فصلی و دمای اولیه ۱



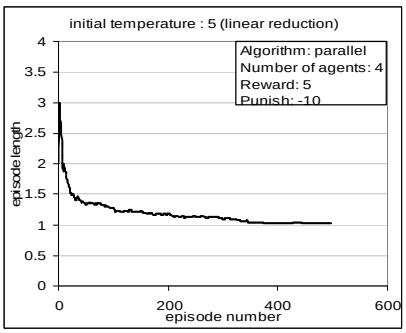
شکل۷- محیط فصلی و دمای اولیه ۵۰



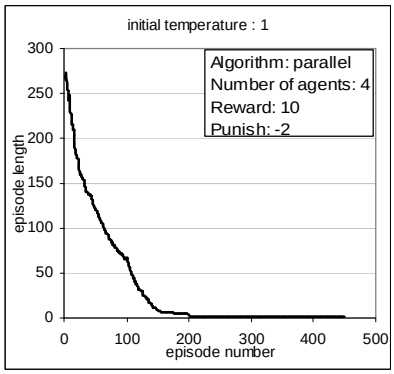
شکل ۸- مقایسه دانش فردی در محیط فصلی



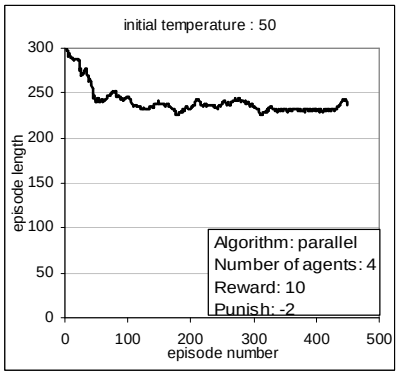
شکل۹- مقایسه دانش فردی



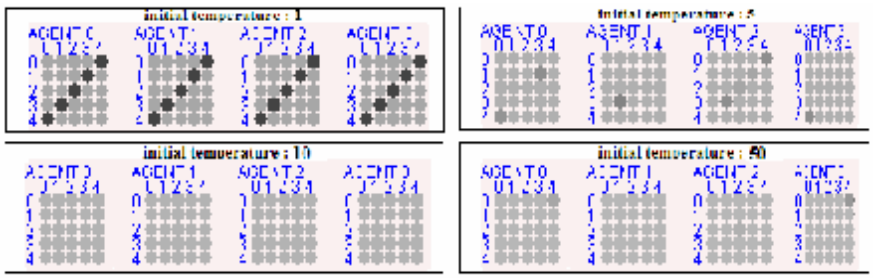
شکل ۱۰- نمودار طول رویداد با سرعت کاهش خطی دما



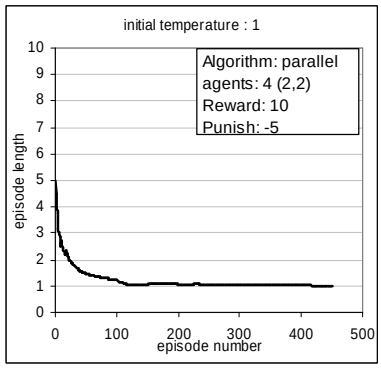
شکل۱۱- محیط عطفی و دمای اولیه ۱



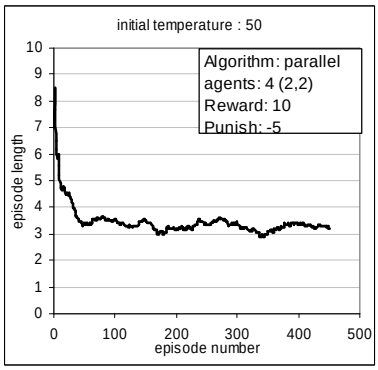
شکل ۱۲- محیط عطفی و دمای اولیه ۵۰



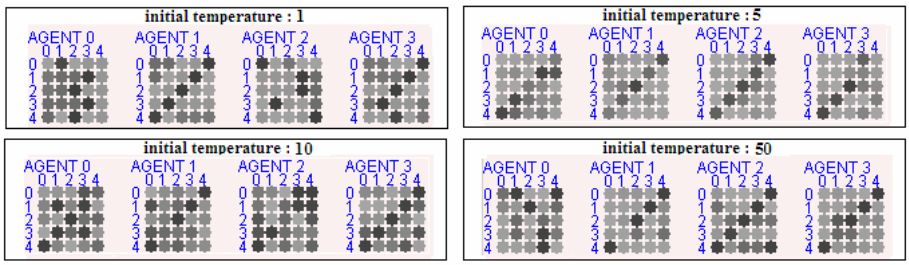
شکل ۱۳- مقایسه دانش فردی در محیط عطفی



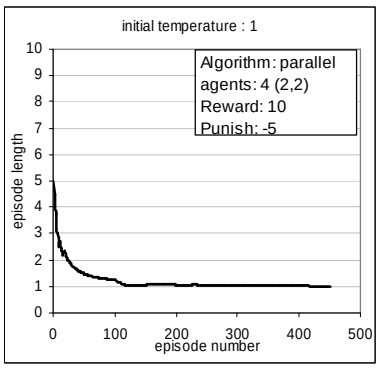
شکل۱۴- محیط ترکیبی (۲،۲) و دمای ۱



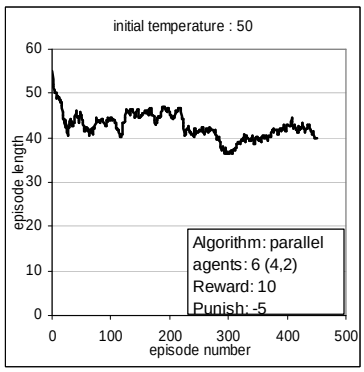
شکل ۱۵- محیط ترکیبی (۲،۲) و دمای ۵۰



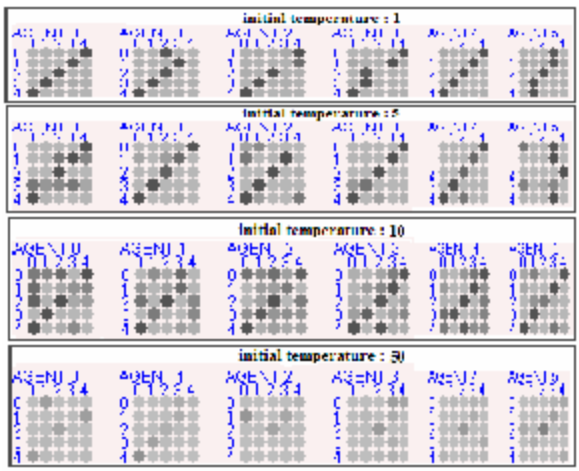
شکل ۱۶- مقایسه دانش فردی در محیط ترکیبی (۲،۲)



شکل۱۷- محیط ترکیبی (۲،۴) و دمای ۱



شکل ۱۸- محیط ترکیبی (۲،۴) و دمای ۵۰



شکل ۱۹- مقایسه دانش فردی در محیط ترکیبی (۲،۴)

Learning with Local Information," *Proc. 9th Annual Computer Society of Iran Conference*, Tehran, Iran, 2004." *Proc. 9th Annual Computer Society of Iran Conference*, Tehran, Iran, 2004.

[3] A. H. Elahibakhsh, M. Nili Ahmadabadi, F. Janabi Sharifi and B. N. Araabi, "Learning Distributed Grasp in Presence of Redundant Agents", *Proc. IEEE/ASME Intl. Conf. Intelligent Mechatronics: AIM 2005*, Monterey, California, USA, 2005.

[4] A. Harati, and M. Nili Ahmadabadi, "A New Approach to Credit Assignment in a Team of Cooperative Q-Learning Agents," *Proc. IEEE Conf. Systems, Man & Cybernetics (SMC'2002)*, Hammamet, Tunisia, 2002.

[5] A. Harati and M. Nili Ahmadabadi, "Certainty and Expertness-Based Credit Assignment for Cooperative Q-Learning Agents with an AND-Type Task," *Proc. 9th Int. Conf. Neural Information Processing (ICONIP'2002)*, Nov. 2002, pp. 306-310.

[6] A. Harati and M. Nili Ahmadabadi, "Experimental Analysis of Knowledge Based Multiagent Credit Assignment," *in Neural Information Processing: Research and Development*, J. C. Rajapakse, L. Wang (eds.), Springer - Verlag, Studies in Fuzziness and Soft Computing Series, vol. 152, pp. 437-459, 2003.

[7] L. P. Kaelbling, M. L. Littman and A. W. Moore, "Reinforcement Learning: A Survey," *Journal of Artificial Intelligence Research*, pp. 237-285, 1996.

[8] K. Miyazaki and S. Kobayashi, "Rationality of Reward Sharing in Multi-Agent Reinforcement Learning," *Second Pacific Rim Int. Workshop on Multi-Agents*, pp. 111-125, 1999.

[9] M. Nili Ahmadabadi and M. Asadpour, "Expertness Based Cooperative Q-Learning," *in IEEE Trans. On SMC*, Part B, vol. 32, no. 1, pp. 66-76, 2002.

[10] M. Nili Ahmadabadi and M. Asadpour, "Cooperative Q-Learning: The Knowledge Sharing Issue," *Advanced Robotics*, vol. 15, no. 8, pp. 779-878, 2001.

[11] P. Stone, M. Veloso, "Team-Prtitioned, Opaque-Transition Reinforcement Learning," *3rd Annual Conference on Autonomous Agents*, Washington, United States, pp. 206 – 212, 1999.

[12] R. S. Sutton (Editor), "Machine Learning: Special Issue on Reinforcement Learning," *Machine Learning*, vol. 8, 1992.

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

[14] C. Watkins, J. Christopher, and P. Dayan, "Q- Learning," *Technical note in* [Sutton, 1992], pp. 55-68, 1992.

[1] AND-type
[2] OR-type
[3] Redundant
[4] Q-Learning
[5] Q-table
[6] Boltzmaan Machine
[7] Grasp

امیرحسین الهی‌بخش لیسانس خود را در رشته کامپیوتر گرایش نرم‌افزار از دانشگاه صنعتی شریف در سال ۱۳۷۹ اخذ کرد. سپس در سال ۱۳۸۰ وارد دانشگاه تهران شده، در سال ۱۳۸۳ کارشناسی ارشد خود را در گرایش هوش ماشین و رباتیک به اتمام رسانید. وی در طول دوره کارشناسی ارشد با پژوهشکده علوم شناختی، پژوهشگاه دانش‌های بنیادین همکاری داشته است. ایشان از سال ۱۳۷۹ تا ۱۳۸۳ در شرکت نبراس انفورماتیک به عنوان مدیر گروه جاوا و از سال ۱۳۸۳ تا ۱۳۸۵ در شرکت داده پردازان دوران به عنوان مدیر تولید ERP مشغول به کار بوده است. همچنین وی از اردیبهشت ۱۳۸۲ تاکنون عضو کمیته تخصصی کامپیوتر انجمن فارغ التحصیلان دانشگاه صنعتی شریف و از اواخر ۱۳۸۴ تاکنون عضو هیئت امنای انجمن فارغ التحصیلان دانشگاه صنعتی شریف می‌باشد.

مجید نیلی احمدآبادی لیسانس خود را در رشته مهندسی مکانیک از دانشگاه صنعتی شریف در سال ۱۳۶۹ و کارشناسی ارشد و دکتری خود را در علوم اطلاعات به ترتیب در سالهای ۱۳۷۳ و ۱۳۷۶ از دانشگاه توهوکو ژاپن اخذ نمود. او در همان سال به عنوان استادیار به استخدام دانشگاه توهوکو در آمد و در سال بعد به دانشکده مهندسی برق و کامپیوتر دانشگاه تهران پیوست. دکتر نیلی هم اکنون دانشیار آن دانشکده و بنیان گذار و سرپرست آزمایشگاه رباتهای متحرک آن می باشد. ایشان همچنین به‌عنوان پژوهشگر وابسته با پژوهشکده علوم شناختی، پژوهشگاه دانش‌های بنیادین همکاری دارد. زمینه های اصلی تحقیقاتی دکتر نیلی شامل یادگیری چند عامله، مدلسازی شناخت در موجودات، روشهای هوشمند الهام گرفته شده از طبیعت، سیستمهای گسترده رباتیکی و رباتهای متحرک می باشد.

بابک نجار اعرابی در سال ۱۳۴۸ در تهران متولد و در سال‌های ۱۳۷۱، ۱۳۷۴ و ۱۳۸۰ دوره‌های کارشناسی، کارشناسی ارشد و دکتری مهندسی برق را به ترتیب در دانشگاه‌های صنعتی شریف، تهران و Texas A&M به اتمام رسانید. دکتر اعرابی از زمستان سال ۱۳۸۰ همکاری خود را با دانشکده مهندسی برق و کامپیوتر دانشگاه تهران آغاز کرد و در حال حاضر به‌عنوان دانش‌یار و مدیر گرایش کنترل در این دانشکده مشغول به پژوهش و تدریس است. ایشان از فروردین ۱۳۸۱ تا کنون به‌عنوان پژوهشگر وابسته با پژوهشکده علوم شناختی، پژوهشگاه دانش‌های بنیادین همکاری داشته است. زمینه‌های اصلی تحقیقاتی دکتر اعرابی عمدتا در حوزه علوم و مهندسی اطلاعات قرار داشته، در چند سال اخیر بیشتر به پژوهش در زمینه یادگیری ماشینی در سیستم‌های چند عامله، بازشناخت الگو در سیستم‌های مبتنی بر بینایی و مدلسازی و پیش بینی نروفازی پرداخته است. حاصل پژوهش‌های دکتر اعرابی تا کنون در بیش از ۱۰۰ مقاله ژرنال و کنفرانس معتبر بین المللی ارائه شده است. برای کسب اطلاعات بیشتر می‌توانید به eng.ut.ac.ir/ece/araabi/ مراجعه کنید.