علوم رایانش و فناوری اطلاعات
نشریه علمی انجمن کامپیوتر ایران
مجلد ۱۶، شماره ۲، ۱۳۹۷
صفحات ۱۰- ۲۰
مقاله عادی

نشریه علوم رایانش و فناوری اطلاعات

بررسی عملکرد عامل باتجربه در تیم‌های اقتضایی پهپادی

رقیه حیدری[۱] محسن افشارچی[۲] رضا خان‌محمدی[۳]

[۱]شرکت توزیع نیروی برق استان زنجان، زنجان، ایران
[۲]دانشکده مهندسی، دانشگاه زنجان، زنجان، ایران
[۳]شرکت برق منطقه‌ای زنجان، زنجان، ایران

چکیده

در تیم‌های رباتیک خودمختار در کاربردهای دنیای واقعی لازم است عامل‌ها برای رسیدن به بیشترین سود با هم همکاری کنند. تصمیم‌گیری و چگونگی عملکرد و همکاری آنان با هم به دلیل پویا بودن محیط، پیوستگی برخی از پارامترها، غیرقطعی بودن محیط و ناشناخته بودن هم‌تیمی‌ها فرآیندی پیچیده محسوب می‌شود. در این مقاله مأموریت نظارت مداوم پهپادها به عنوان یک سیستم چندعامله در دنیای واقعی معرفی شده است. در ادامه، مسئله تصمیم‌گیری برخط عامل‌ها در شرایطی که اعضای تیم و محیط به طور کامل شناخته شده نیستند، مطرح شده است . این مسئله با کمک فرآیند تصمیم‌گیری مارکوف مدل شده و در نهایت یک الگوریتم حریصانه برای تصمیم‌گیری برخط عامل در تیم ارائه شده است. آزمایش‌های انجام شده نشان می‌دهند این روش تصمیم‌گیری که مبتنی بر دانش آموخته شده قبلی عامل است، عملکرد تیم را در محیط ناشناخته بهبود می‌دهد.

کلمات کلیدی: همکاری، تیم اقتضایی، فرآیند تصمیم‌گیری مارکوف، پهپاد.

۱- مقدمه

امروزه از تیم‌های رباتیک شامل عامل‌های بدون سرنشین خودمختار[۱] در کاربردهای بسیاری استفاده می‌شود. در این مسائل با عامل‌هایی مواجه هستیم که باید در هر لحظه بسته به موقعیت، بتوانند عمل بعدی خود را به درستی انتخاب نمایند و با سایر عامل‌های موجود در تیم همکاری[۲] نمایند تا وظیفه واگذار شده به مجموعه به بهترین نحو انجام گیرد. از جمله مسائل کاربردی در این زمینه می‌توان به عملیات جستجو و نجات، سیستم‌های مدیریت بحران و مانیتورینگ ترافیک شهری اشاره نمود. هر چه شناخت عامل‌ها از یکدیگر و محیط کمتر باشد، یافتن بهترین عمل در هر وضعیت کار پیچیده‌تری خواهد بود. فرض کنید در یک بحران اتفاق افتاده عامل‌های هوشمند ساخته شده توسط چندین شرکت مختلف، در منطقه برای امدادرسانی اعزام شده‌اند. اگرچه هدف کلی همه عامل‌ها امدادرسانی است، ولی از آنجا که عامل‌ها سازندگان متفاوتی دارند و قادر به برقراری ارتباط با یکدیگر نیستند، ممکن است متعارض با یکدیگر عمل نموده و در نهایت از هدف اصلی خود بازمانند. در چنین شرایطی عامل‌ها می‌بایست از روی رفتار یکدیگر یاد بگیرند که چگونه درکنار هم کار کنند تا بتوانند وظیفه خود را درست انجام دهند [۱].

اگر کل مجموعه عامل‌ها را یک تیم فرض کنیم، همکاری بیشینه هم‌تیمی‌ها در آن بهترین نتیجه را برای تیم در بر خواهد داشت. اگرچه برای به دست آوردن بهترین عملی که هر عامل در هر موقعیت می‌تواند انجام دهد روش‌های مختلفی وجود دارد، ولی می‌توان کل آن‌ها را از نظر برخط بودن تصمیم‌گیری عامل‌ها به دو گونه کلی دسته‌بندی کرد:

- یادگیری عامل‌ها در مرحله یادگیری اتفاق می‌افتد و از دانش به دست آمده در دنیای واقعی بدون توجه به تغییرات محیط استفاده می‌گردد.
- یادگیری عامل‌ها در مرحله یادگیری اتفاق می‌افتد و با کمک دانش به دست آمده، برحسب تغییرات اتفاق افتاده در محیط و تیم، تصمیم‌گیری و انتخاب عملی از عمل‌های آموخته شده، به شکل برخط صورت می‌گیرد.

در این مقاله با رویکرد دوم، الگوریتمی برای به دست آوردن بهترین عملی که هر عامل می‌تواند بسته به وضعیت کل محیط و تیم انجام دهد، ارائه شده است.

در سیستم‌های چندعامله هوشمند، مسئله‌های بسیاری مطرح شده که در آنها به انجام بهترین عمل ممکن برای هر عامل در یک تیم پرداخته شده است [۲، ۳]. برخی از آنها تحت عنوان همکاری انعطاف‌پذیر عامل‌ها[۳] [۴] یا همکاری انعطاف‌پذیر در کار تیمی[۴] [۵] به موضوع پویا بودن پارامترهای مسئله و چالش تصمیم‌گیری درخصوص بهترین عمل ممکن برای تیمی از عامل‌ها در این شرایط پرداخته‌اند. در [۶] موضوع تصمیم‌گیری متوالی اعضای تیم در شرایط غیرقطعی با کمک مدل MDP مطرح شده است. آنها مفهوم MDP را به سیستم چندعامله بسط داده‌اند و جواب به دست آمده از حل آن را به عنوان بهترین سیاستی که تیم همکارانه می‌تواند اتخاذ کند، معرفی کرده‌اند. مقاله [۷] در خصوص امکان استفاده از الگوریتم‌های یادگیری تقویتی مانند Q-Learning برای به دست آوردن بهترین عمل ترکیبی[۵] همه عامل‌ها در تیم صحبت کرده است. در این مقاله با کمک مفهوم بازی‌های همکارانه، نشان داده شده است که الگوریتم‌های یادگیری تقویتی مانند Q-Learning در سیستم تک‌عامله بسیار بهتر عمل می‌کنند.

در زمینه همکاری در سیستم‌های چندعامله، تحت عنوان کار تیمی، مقالات بسیار زیادی وجود دارد. مشخص کردن مرز بین مفهوم تیم معمولی و تیم اقتضایی در این بین، کار ساده‌ای نیست ولی با معرفی عنوان ad hoc team این بررسی‌ها از سال ۲۰۰۹ رنگ تازه‌ای به خود گرفته است. این مبحث به طور جدی در [۸] مطرح و به این مورد تأکید شده است که استفاده از چنین تیم‌هایی به سرعت افزایش خواهد یافت. در سال ۲۰۱۰ پیتر استون[۶] و همکارانش در [۹] به معرفی این نوع تیم‌ها پرداخته و با رویکرد نظریه بازی به همکاری بدون هماهنگی اشاره کرده‌اند. آنها با اشاره به این نکته که در برخی مسائل نیازی به هماهنگی قبلی بین عامل‌ها نیست، مسئله یافتن بهترین عمل در بازی‌های تکراری را حل کردند. سپس در همان سال در [۱۰] به معرفی دقیق‌تر تیم اقتضایی و نمونه مسائلی در این حیطه پرداختند. آنها در این مقاله به میزان پیچیدگی همکاری عامل‌ها وقتی که هیچ هماهنگی بین عامل‌ها قبل از تشکیل تیم صورت نمی‌گیرد، اشاره کرده‌اند.

تصمیم‌گیری برخط یک عامل در یک تیم اقتضایی در [۱۱] بررسی شده است. در روش استفاده شده در آن مقاله، در هر گام یک بازی همکارانه مدل شده و به طور تقریبی بهترین عمل ممکن محاسبه می‌گردد. ما در این مقاله با همین رویکرد درخصوص تصمیم‌گیری برخط، به جای حل یک مسئله با دیدگاه یک بازی همکارانه، از تجارب قبلی خود که با حل MDP در شرایط مختلف به دست آورده‌ایم، استفاده می‌کنیم. در [۱۲] به ایجاد مدلی از هم‌تیمی‌ها توسط یک عامل اقتضایی و انتقال مدل‌های یادگرفته شده قبلی پرداخته شده است[۷]. در [۱۳، ۱۴] چارچوبی برای ارزیابی قابلیت‌های تیم تشکیل شده ایجاد گردیده و روش معرفی شده به صورت تجربی در مسئله شکارچی برای ارزیابی تیم‌های تشکیل شده، به کار بسته شده است. در [۱۴] و به طور کامل‌تری در [۱] به چالش‌های مختلف یادگیری در یک تیم اقتضایی پرداخته شده است و سطوح دانش عامل از محیط و هم‌تیمی‌ها از یکدیگر مجزا شده است. آنها بحث روی تیم‌های اقتضایی را به سه شاخه تقسیم کرده‌اند:

- دانش یک عامل از عمل هم‌تیمی‌هایش در هر وضعیت
- دانش یک عامل از میزان احتمال گذار از وضعیتی به وضعیت دیگر
- دانش یک عامل از میزان تأثیر عمل وی بر کار تیم

در [۱] الگوریتم PLASTIC (Planning and Learning to Adapt Swiftly to Teammates to Improve Cooperation) ارائه شده است. در این روش از دانش اندوخته شده عامل‌ها در مرحله یادگیری طوری استفاده می‌شود که در مرحله تست، آنها بتوانند خود را سریعتر با عامل‌های جدید و ناشناخته هماهنگ کنند. در مقاله مذکور کمتر به استفاده از این روش‌ها در مسائل دنیای واقعی پرداخته شده است. همچنین در حالتی که شرایط محیط متغیر باشد، مسئله مورد بررسی قرار نگرفته است. ما با یک رویکرد مشابه PLASTIC-Policy، در شرایطی که پارامترهای محیط متغیر باشد، از سیاست‌های آموخته شده قبلی استفاده کرده و مسئله تصمیم‌گیری بهترین عمل در هر لحظه را با کمک آن بهبود داده‌ایم. از طرف دیگر مسئله مورد مطالعه در مقاله ما، یک مسئله واقعی و عملی می‌باشد که در ادامه به معرفی آن می‌پردازیم.

مسائل بسیاری در زمینه یادگیری تقویتی با کمک فرآیند تصمیم‌گیری مارکوف مدل شده‌اند. مسئله مورد مطالعه جهت انجام آزمایشات تجربی در مقاله حاضر "مأموریت نظارت مداوم"[۸] است؛ به این صورت که نظارت مداوم یک منطقه توسط پهپادها[۹] انجام می‌پذیرد به نحوی که هزینه مأموریت کمینه شود. وضعیت سلامت و سوخت پهپادها در این مسئله مدل شده است و به دلیل کمبود سوخت یا خرابی عملگرها یا سنسورهای هر پهپاد در مأموریت، ممکن است با وقفه مواجه گردیم.

این مسئله در [۱۵-۱۷] مدل شده است و در آن از روشی برای تخمین وضعیت سایر عامل‌ها استفاده شده است و الگوریتم جدیدی برای کاهش فضای مسئله ارائه گردیده است، ولی درخصوص پیروی عامل‌ها از یک سیاست خاص جهت تشکیل تیم، بررسی خاصی انجام نشده است. در این مقاله ضمن استفاده از مدل موجود در [۱۵-۱۷] به بررسی عملکرد تیم حاصل از عامل‌هایی که قبلاً سیاست خاصی را یادگرفته‌اند، مانند [۱، ۱۸]، پرداخته شده است. در بخش ۳-۱، یک نمونه تیم اقتضایی را در آزمایش اول تشکیل داده‌ایم. سپس نتایج حاصل از کار تیمی آنها را در مسئله مورد مطالعه خود، ارزیابی کرده و نشان داده‌ایم که درصورت داشتن آگاهی تک‌تک عامل‌ها از سیاست کلی[۱۰]، عملکرد تیم حاصل از آنها، همچنان قابل‌قبول می‌باشد. با این رویکرد در مسائلی که پارامترهای مسئله از جمله شرایط محیط، کاملاً شناخته شده نیستند، می‌توان عامل‌ها را از قبل تحت شرایط مشابه آموزش داد و سپس به تشکیل تیمی از عامل‌های آموزش داده شده پرداخت. در آزمایش دوم در بخش ۳-۲ به موضوع تصمیم‌گیری برخط یکی از هم‌تیمی‌ها پرداخته‌ایم و به صورت تجربی نشان داده‌ایم که تیم حاصل، در شرایطی که پارامترهای محیط متغیر هستند، عملکردی بهتر از زمانی که کل تیم با پارامترهای مشابه آموزش داده شده‌اند، دارد.

این مقاله در ادامه به این صورت بخش‌بندی شده است: مفاهیم و تعاریف اولیه مسئله در بخش ۲ آمده است. بخش‌های ۲-۱ تا ۲-۳ به معرفی سه مفهوم اساسی استفاده شده در این مقاله اختصاص دارند و در بخش ۲-۴ به طرح مسئله پرداخته شده است. در بخش ۲-۵ روش پیشنهادی ارائه گردیده است. در بخش ۳ به آزمایش تجربی و نتایج آن پرداخته شده است به این ترتیب که در بخش ۳-۱ و ۳-۲ به آزمایش‌های تجربی پرداخته و در بخش ۳-۳ به نحوه حل مسئله و برنامه‌نویسی آن اشاره کرده‌ایم. در بخش ۴ تحلیل نتایج و در بخش ۵ نتیجه‌گیری و کارهای آینده ذکر شده است.

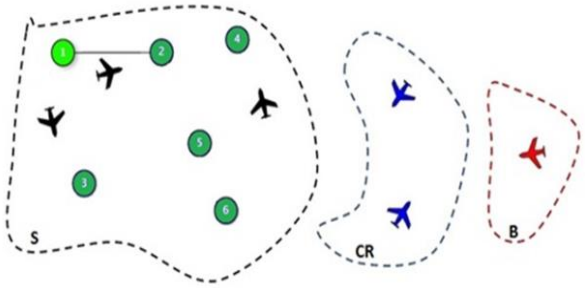
۲- مفاهیم و تعاریف اولیه

سه مفهوم اساسی در این مقاله به عنوان اساس کار استفاده شده‌اند که عبارتند از مسئله مأموریت نظارت مداوم، فرآیند تصمیم‌گیری مارکوف و کار تیمی و تیم اقتضایی. در ادامه به معرفی هر کدام می‌پردازیم.

۲-۱- مسئله مأموریت نظارت مداوم

در مأموریت نظارت مداوم، پهپادها باید منطقه اهداف را طوری نظارت کنند تا هیچ هدفی از دید آنها پنهان نماند. در مسئله‌ای که در این مقاله به آن پرداخته شده و در [۱۵] به طور کامل شرح داده شده است، کل منطقه به سه بخش مجزا تقسیم

شده است: منطقه نظارت، منطقه ارتباطی و پایگاه مرکزی. اهداف متحرکی در منطقه نظارت درحال حرکت هستند و مأموریت جستجو و دنبال کردن مداوم این اهداف بطور پیوسته توسط پهپادهای خودمختار[۱۱] انجام خواهد گرفت. در منطقه ارتباطی حتماً باید یک عامل وجود داشته باشد. این عامل وظیفه برقراری ارتباط بین دو منطقه پایگاه مرکزی و نظارت را برعهده دارد. از طرفی هر یک از عامل‌های موجود در منطقه نظارت، ممکن است به دلیل احتمال خرابی سنسور یا کمبود سوخت دچار مشکل شوند. در این حالت‌ها عاملی که در منطقه ارتباطی قرار دارد، می‌تواند به عنوان جایگزین عمل نماید. برای مثال عاملی که فقط سنسورش آسیب دیده باشد، همچنان می‌تواند وظیفه برقراری ارتباط را انجام دهد. فرض بر این است که اگر عامل خود را به پایگاه مرکزی برساند، هر دو عمل تعمیر و سوخت‌گیری مجدد برای وی بطور کامل انجام خواهد گرفت.



شکل ۱- شمای کلی مسئله: S منطقه نظارت، CR منطقه ارتباطی و B پایگاه مرکزی

اتمام سوخت در میانه مأموریت، عدم‌وجود پهپاد در منطقه ارتباطی و نظارت یا کمبود تعداد پهپاد مورد نیاز در منطقه نظارت، هزینه‌هایی را برای مأموریت در پی خواهد داشت. اتمام سوخت هر پهپاد در میانه مأموریت، منجر به نابود شدن[۱۲] آن پهپاد می‌گردد و در مدل ما، این اتفاق هزینه بالایی را به سیستم تحمیل می‌کند. فقدان پهپاد در رله ارتباطی یا منطقه نظارتی موجب شکست[۱۳] در مأموریت می‌گردد و کمبود تعداد پهپادها از تعداد مورد نیاز جهت پایش منطقه نظارتی، سیستم را متحمل هزینه شکاف[۱۴] می‌نماید. حال می‌خواهیم نظارت مداوم منطقه با کمترین هزینه صورت گیرد. در شکل ۱ شمای کلی مسئله نشان داده شده است. این مسئله در [۱۷] به صورت متمرکز برای ۳ دستگاه UAV (Unmanned Aerial Vehicle) با کمک مدل مارکوف، حل شده و نتایج مورد ارزیابی قرار گرفته‌اند.

۲-۲- فرآیند تصمیم‌گیری مارکوف

وقتی مسئله‌ای با کمک MDP مدل می‌شود، به این معنی است که وضعیت بعدی که سیستم با انجام عمل مشخصی به آن وارد خواهد شد، تنها به وضعیت فعلی وابسته است و از وضعیت‌های قبلی مستقل می‌باشد [۱۵]. یک MDP[۱۵] (Markov Decision Process) یا فرآیند تصمیم‌گیری مارکوف) شامل یک حالت اولیه s از مجموعه حالات S، یک تابع پاداش $r: S \rightarrow R$ و یک تابع گذار $T(s, a, s')$ است [۱۹]. بنابراین در فرآیند تصمیم‌گیری مارکوف باید تمامی پارامترهای زیر را مشخص کنیم:

- فضای حالت‌ها یا وضعیت‌ها: S
- فضای عمل‌ها: A
- تابع احتمال گذار: T
- تابع هزینه/پاداش: R/C

در مسئله PSM ما برای هر وضعیت یک هزینه در نظر می‌گیریم. بنابراین مولفه چهارم MDP، با عنوان تابع هزینه (C) معرفی خواهد شد. جزئیات مدل MDP برای مسئله PSM، در بخش ۲-۵ آمده است. با مشخص شدن همه پارامترهای MDP می‌توان به حل آن پرداخت. وقتی یک MDP حل شود، یک سیاست نزدیک به بهینه $\pi: S \rightarrow A$ به دست خواهیم آورد، که مشخص می‌کند در هر وضعیت بهترین عملی که باید انجام شود کدام است.

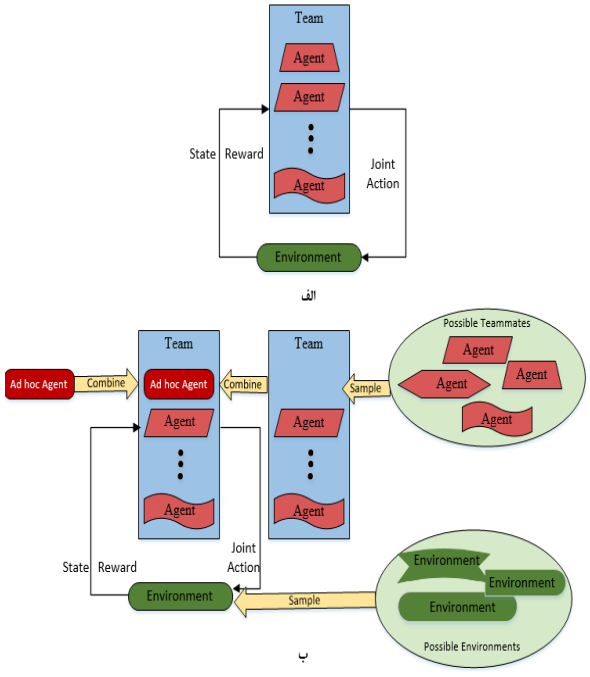
۲-۳- کار تیمی و تیم اقتضایی

در [۲۰] به ایده تشکیل تیم‌های اجتماعی که دنبال کردن و رسیدن به یک هدف خاص است، اشاره شده است. برای مثال هدف بازیکنان یک تیم فوتبال جدا از رعایت کردن قوانین حاکم بر بازی، "برنده شدن" است. رویکرد ما به کار تیمی به این صورت است که اعضای تیم:

- ✓ باید با هم و در کنار هم کار کنند تا به هدف برسند.
- ✓ بطور پیوسته وضعیت کل تیم را پایش می‌کنند.
- ✓ به همدیگر کمک می‌کنند.
- ✓ در کار هم دخالت نمی‌کنند.
- ✓ برای رسیدن به هدف بین آنها رقابتی وجود ندارد [۲۰].

در [۱] به این موضوع اشاره شده است که در یک تیم ممکن است علاوه بر اینکه هم‌تیمی‌ها برای هم شناخته شده نیستند، شناختی از محیط نیز نداشته باشند. هرچند در ادامه مقاله تنها در مورد ناشناخته بودن هم‌تیمی‌ها مسائلی مطرح و حل شده است. تفاوت بین یک تیم معمولی و یک تیم اقتضایی در شکل ۲ نمایش داده شده است.

در شکل ۲ در قسمت (الف) یک تیم معمولی و در قسمت (ب) یک تیم اقتضایی به نمایش درآمده است. همان‌طور که ملاحظه می‌شود، در این تیم هماهنگی قبلی بین هم‌تیمی‌ها و یا هماهنگی از قبل جهت شناسایی محیط صورت نمی‌گیرد.



شکل ۲- تیم معمولی در مقابل تیم اقتضایی. شکل الف مربوط به یک تیم معمولی است و شکل ب مربوط به تیم اقتضایی است [۱]

در شرایطی که موارد ناشناخته‌ای در مسئله وجود داشته باشد و یا هماهنگی بین عامل‌ها از قبل صورت نگیرد، پیدا کردن بهترین عملی که مجموعه عامل‌ها باید در هر وضعیت انجام دهند، یک مسئله بسیار پیچیده خواهد بود [۱۰، ۲۱].

یک تیم اقتضایی به طور کلی در [۱۸، ۲۲] به این صورت تعریف شده است:

"تیمی از عامل‌ها که در آن هم‌تیمی‌ها بدون هر گونه هماهنگی قبلی با هم کار می‌کنند تا به هدف مشترک دست یابند".

در هر تیمی عملکرد تک تک اعضای تیم بر روی برآیند فعالیت تیم تاثیر خواهد گذاشت. در یک تیم اقتضایی ممکن است اعضای متفاوتی وجود داشته باشند. برای بررسی تأثیر عملکرد هر عضو در تیم، یکی از ساده‌ترین روش‌ها این است که آن عضو را از تیم حذف کرده و نتیجه به دست آمده توسط تیم را با حالتی که آن عضو در تیم وجود داشته، مقایسه کنیم. ساموئل برت و همکارانش در [۱۴] از این روش جهت ارزیابی عملکرد هر عامل هوشمند استفاده کرده‌اند.

در این مقاله نیز در ارزیابی میزان تأثیر عامل آموزش دیده در عملکرد تیم در بخش ۳-۲، عملکرد تیم را یک بار با حضور این عامل و بار دیگر بدون حضور وی، اندازه‌گیری و مورد مقایسه قرار داده‌ایم.

در آزمایش اول که هیچ هماهنگی‌ای از قبل بین عامل‌ها انجام نمی‌گیرد ما یک تیم اقتضایی داریم. بعلاوه در آزمایش دوم که در این مقاله به آن پرداخته‌ایم، شرایط محیط متغیر است و از ابتدا برای عامل‌ها شناخته شده نیست. بنابراین با یک تیم اقتضایی مواجه هستیم که هیچ هماهنگی قبلی بین آنها جهت شناسایی محیط صورت نمی‌گیرد. در این آزمایش هم تأثیر تصمیم‌گیری برخط عامل آموزش دیده را مورد بررسی قرار داده‌ایم.

۲-۴- طرح مسئله

در حل مسئله نظارت مداوم [۱۵-۱۷]، کلیه پهپادها به یک شکل آموزش داده می‌شوند و در حالت متمرکز نهایت همکاری با یکدیگر را انجام می‌دهند تا هزینه مأموریت را به حداقل برسانند.

ایده ما بر این است که هزینه‌های در نظر گرفته شده برای یک مأموریت با مأموریت دیگر متفاوت است. مثلاً در یک مأموریت تمایلی به وجود شکاف در حین نظارت نداریم و اتفاق افتادن آن را به اندازه شکست در مأموریت مخرب می‌دانیم ولی در مأموریت دیگری شکاف در نظارت اهمیت کمتری دارد. این هزینه‌ها ممکن است به هزینه ساخت پهپادهای متفاوت نیز وابسته باشد، به طوری که هزینه نابود شدن در یک پهپاد نسبت به پهپاد نوع دیگر بسیار پایین‌تر باشد. از طرف دیگر، شرایط محیط متغیر است. ممکن است پهپادها به خاطر شرایط بد آب و هوا سوخت بیشتری مصرف کنند. احتمال خرابی سنسور و عملگر آنها نیز علاوه بر نوع ساخت پهپاد، به شرایط جوی از جمله میزان سرعت وزش باد، میزان رطوبت و دمای هوا و وضعیت بارش باران وابسته است. همه این موارد موجب ایجاد تیم‌های اقتضایی در PSM خواهند شد.

مسئله‌ای که در این مقاله به آن پرداخته شده است، تشکیل دادن تیم‌های مختلف جهت انجام مأموریت مداوم است. ابتدا با تغییر دادن مقادیر مختلف برای پارامترهای محیط یا هم‌تیمی‌ها، تیم‌هایی از پهپادها را آموزش می‌دهیم تا با سیاست کلی حاکم بر مسئله PSM آشنا شوند. این تیم‌ها را بین MMDP1 تا MMDP10 نامگذاری کرده‌ایم. در آموزش این تیم‌ها، نسبت هزینه شکاف به شکست یا نابودی و حدود تقریبی احتمالات خرابی پهپادها نزدیک به مقادیر محیط شبیه‌سازی در نظر گرفته شده است. با این کار، فرض کرده‌ایم سیاست کلی حاکم بر مسئله از ابتدای امر مشخص است و ما قادریم آن را به پهپادها آموزش دهیم. با انتساب یک مقدار بسیار بزرگ به هزینه نابودی و سپس انتساب مقدار نسبتاً بزرگ هزینه شکست به شکاف، پهپادها یاد می‌گیرند که باید در مأموریت از نابودی اجتناب کنند و تا می‌توانند تعداد شکست را به حداقل برسانند تا هزینه تجمعی کل مأموریت کمینه شود. ما این سیاست کلی را در PSM، سیاست محافظه‌کارانه نام نهاده‌ایم.

تعریف سیاست کلی محافظه‌کارانه:

- ✓ نسبت هزینه نابود شدن هر پهپاد به هزینه شکاف در نظارت، حداقل ۵۰۰ است.
- ✓ نسبت هزینه شکست مأموریت به شکاف در نظارت، حداقل ۱۰ است.
- ✓ نسبت هزینه نابود شدن هر پهپاد به شکست در مأموریت، حداقل ۲۵ است.

پس از تعریف سیاست کلی حاکم بر مسئله، به دنبال بررسی عملکرد تیم‌های اقتضایی مختلف در PSM هستیم. در آزمایش اول، اعضایی از تیم‌های مختلف MMDP، بدون هماهنگی قبلی با یکدیگر را برای تشکیل تیم درکنار هم قرار داده‌ایم. عملکرد این تیم‌های اقتضایی با تیم‌هایMMDP مقایسه شده است.

در آزمایش دوم، تیم‌های اقتضایی بدون شناسایی قبلی پارامترهای دقیق محیط تشکیل داده شده‌اند. در مرحله یادگیری، یک عامل در این تیم‌ها رفتار خود را با عامل‌های دیگر و در شرایط مختلف محیط هماهنگ می‌کند و به اصطلاح سیاست‌های مختلفی را آموزش می‌بینند. سپس در مرحله تست، با مشاهده رفتار سایر عامل‌ها تصمیم می‌گیرد که چه عملی را انجام دهد تا تیم متحمل کمترین هزینه گردد. در نهایت عملکرد این تیم با عملکرد تیم MMDP در محیط متغیر مقایسه شده است.

۲-۵- روش پیشنهادی

جهت تشکیل تیم‌های اقتضایی مختلف به پهپادهایی نیاز داریم که سیاست کلی را آموزش دیده باشند. به همین منظور، ابتدا چندین نوع از پهپادهای محافظه‌کار را با تغییر دادن پارامترهای هزینه‌های نابودی، شکست و شکاف و احتمال‌های مربوط به خرابی پهپادها آموزش می‌دهیم. در جدول ۱ مقادیر مختلفی که در آموزش پهپادهای متفاوت استفاده کرده‌ایم، آمده است. سپس با کمک الگوریتم Value-Iteration [۲۳]، به حل MMDP ایجاد شده می‌پردازیم. به این ترتیب پهپادهای محافظه‌کار، با تعریفی که در بخش ۲-۴ ارائه کردیم، را آموزش می‌دهیم.

جدول ۱- پارامترهای یادگیری تیم‌های مختلف

عنوان تیم	احتمال خرابی عملگر	احتمال خرابی سنسور	هزینه شکاف	هزینه شکست	هزینه نابودی
MMDP۱	۰,۰۲۵	۰,۰۵	۱	۲۰	۵۰۰
MMDP2	۰,۰۵	۰,۱۵	۱	۱۵	۷۵۰
MMDP۳	۰,۰۵	۰,۱	۱	۲۰	۷۵۰
MMDP4	۰,۰۵	۰,۱	۱	۱۲	۸۵۰
MMDP۵	۰,۱	۰,۱۵	۱	۱۵	۸۵۰
MMDP۶	۰,۰۵	۰,۱	۱	۲۰	۱۰۰۰
MMDP۷	۰,۰۵	۰,۱	۱	۲۰	۱۰۰۰
MMDP۸	۰,۰۱	۰,۰۲	۱	۲۰	۱۲۰۰
MMDP۹	۰,۰۰۰۱	۰,۰۰۵	۱	۲۰	۱۲۰۰
MMDP۱۰	۰,۲۵	۰,۴۵	۱	۲۰	۱۲۰۰

ابتدا چگونگی مدل نمودن مسئله PSM را به کمک MDP توضیح می‌دهیم و ۴ مولفه فضای وضعیت، فضای عمل، احتمال گذار و تابع هزینه را در مسئله PSM مشخص می‌نماییم.

فضای وضعیت (S): فضای وضعیت مسئله برای یک پهپاد شامل وضعیت مکانی l، سلامت h و سوخت f می‌باشد که به صورت $< l.h.f >$ نمایش داده می‌شود. هر یک از مجموعه‌های مذکور اعضای خاصی دارند که در روابط (۱) و (۲) و (۳) مشخص شده است.

$$l \in L = \{base.communication.surveillance\} \quad (۱)$$

$$h \in H = \{healthy.sensorfail.actuatorfail\} \quad (۲)$$

$$f \in F = \{0.1.\dots.8\} \quad (۳)$$

از آنجا که در $l = base$ میزان سوخت ۸ و عامل در وضعیت سلامت، یعنی healthy خواهد بود و مدل مصرف سوخت از هر منطقه به منطقه دیگر حداقل یک واحد است، لذا تعداد اعضای مجموعه S برابر ۴۶ است. ما جهت سهولت نمایش وضعیت پهپاد، اعضای مجموعه‌های L و H را به ترتیب از ۱ تا ۳ شماره‌گذاری کرده‌ایم. مثلاً $l = 1$ به معنی base است و $h = 2$ به معنی sensorfail است.

برای دو عامل ما باید حاصلضرب دکارتی مجموعه‌های اشاره شده را در نظر بگیریم و وضعیت ترکیبی دو عامل در کنار هم را به دست آوریم. در نهایت برای n عامل، فضای مسئله معادل 46^n خواهد بود. ملاحظه می‌شود که با ازدیاد عامل‌ها حتی برای ۶ عامل این عدد بسیار بزرگ است. برای مثال وضعیت یک سیستم چندعامله با سه پهپاد با یک ۹تایی، به صورت عبارت (۴) مشخص می‌شود:

$$< l_{a1}.h_{a1}.f_{a1}.l_{a2}.h_{a2}.f_{a2}.l_{a3}.h_{a3}.f_{a3} > \quad (۴)$$

که در آن منظور از a_i عامل iام است. بنابراین سه مولفه اول مربوط به وضعیت پهپاد اول، سه مولفه دوم مربوط به وضعیت پهپاد دوم و سه مولفه سوم، مربوط به وضعیت پهپاد سوم است. به عنوان یک مثال عملی وضعیت $< 1.1.8.2.3.5.3.1.6 >$ نشان دهنده این است که عامل اول در base است و بنابراین در وضعیت healthy بوده و سوخت حداکثر، یعنی ۸ را داراست. عامل دوم در منطقه ارتباطی است. وضعیت سلامت این عامل actuator fail بوده و ۵ واحد سوخت دارد. عامل سوم در منطقه نظارت است و وضعیت سلامتش healthy است و ۶ واحد سوخت دارد.

فضای عمل (A): فضای عمل‌های ممکن برای هر پهپاد یکی از حرکت‌های پیشروی، ماندن و عقبگرد است که از مجموعه $\{-1.0.1\}$ برای نمایش این اعمال استفاده می‌شود. در این مجموعه، عدد ۱ به معنی پیشروی، عدد ۰ به معنی بی‌حرکت ماندن و عدد ۱- به معنی عقبگرد است. باید توجه داشت که برای عاملی که در منطقه base واقع است عقبگرد و برای عاملی که در منطقه surveillance واقع است پیشروی معنا ندارد.

برای سه عامل در هر وضعیت ما به یک سه‌تایی برای مشخص کردن عملی که باید عامل‌ها انجام دهند، نیازمندیم. به عنوان مثال در وضعیت $< 1.1.8.2.3.5.3.1.6 >$ عمل (1,-1,0) به این معنی است که عامل اول باید به منطقه ارتباطی پیشروی کند، عامل دوم به منطقه پایگاه مرکزی برگردد و عامل سوم در منطقه نظارت باقی بماند.

تابع احتمال گذار: تابع احتمال گذار $T(s.a.s')$ مشخص می‌کند که اگر پهپاد در وضعیت s عمل a را انجام دهد، با چه احتمالی به حالت 's گذار خواهد کرد. با توجه به میزان احتمال خرابی سنسور یا عملگر پهپاد که مربوط به وضعیت سلامت هستند و غیر قطعی بودن میزان مصرف سوخت که ممکن است یک یا دو گالن در هر گام شبیه‌سازی باشد، این احتمال‌ها قابل محاسبه هستند. مکان بعدی پهپاد با انتخاب هر عمل بطور قطعی مشخص می‌شود.

در مدل مارکوف فرض بر این است که آنچه احتمال رسیدن به 's را تعیین می‌کند، وضعیت فعلی و عملی است که عامل در این وضعیت انجام می‌دهد. برای مثال در حالتی که سیستم به طور کامل شناخته شده باشد، احتمال خرابی سنسور و یا عملگر و میزان مصرف سوخت مشخص خواهد بود. فرض کنید احتمال خرابی سنسور را با PS و احتمال خرابی عملگر را با PA و احتمال اینکه پهپاد یک واحد یا دو واحد سوخت مصرف نماید را با PF نشان بدهیم. اگر روابط (۵) همیشه برقرار باشند، می‌توانیم احتمال گذار از یک وضعیت به وضعیت دیگر را با انجام یک عمل مشخص، محاسبه کنیم.

مثلاً در سیستمی با مشخصات روابط ۵ و برای یک عامل، احتمال گذار از وضعیت $< 2.1.7 >$ به وضعیت $< 3.3.6 >$ با انجام عمل پیشروی برابر ۰٫۰۲۵ است.

$$PA = 0.05 \;.\; PS = 0.1 \;.\; PF = 0.5 \quad (۵)$$

تابع هزینه (C): تابع هزینه تعیین می‌کند که در هر وضعیت، سیستم متحمل چه هزینه‌ای می‌شود. همان‌طور که در بخش ۲-۱ توضیح داده شد، ما برای پهپادی که در میانه مسیر سوخت تمام کند، هزینه نابود شدن، برای حالتی که هیچ پهپادی در منطقه ارتباطی یا نظارت نباشند، هزینه شکست و برای حالتی که تعداد پهپادهای مورد نیاز در منطقه نظارت کمتر از حد توقع است، هزینه شکاف، به میزان کمبود تعداد پهپادهای لازم، در نظر گرفته‌ایم.

در روابط (۶) این تابع به شکل دقیق تعریف شده است.

$$C(s) = \begin{cases} C_{crash}. & f = 0 \;.\; l \neq 1 \\ C_{fail}. & com = 0 \text{ or } n_s = 0 \\ \alpha C_{gap}. & com = 1 \;.\; n_s > 0 . n_d - n_s = \alpha \\ 0. & com = 1 \;.\; n_s = n_d \end{cases} \quad (۶)$$

در روابط (۶)، منظور از C_{crash} هزینه نابودی، C_{fail} هزینه شکست، C_{gap} هزینه شکاف، f سوخت و l مکان هر پهپاد است. عبارت $com = 0$ به این معنی است که در رله ارتباطی پهپادی که بتواند وظیفه برقراری ارتباط را انجام دهد وجود ندارد و در $com = 1$ وظیفه برقراری ارتباط به خوبی انجام می‌شود. ضمناً منظور از n_s تعداد پهپاد healthy موجود در منطقه نظارتی است و منظور از n_d تعداد پهپاد مورد نیاز جهت نظارت منطقه است.

برای مثال برای یک مأموریت با ۳ پهپاد، یک پهپاد برای انجام وظیفه برقراری ارتباط و ۲ پهپاد برای انجام مأموریت نظارت در نظر گرفته می‌شوند. وقتی هیچ پهپادی در منطقه ارتباطی یا نظارتی وجود نداشته باشد، هزینه شکست برای کل مجموعه لحاظ می‌شود. از طرفی وقتی یک پهپاد در منطقه ارتباطی و یک پهپاد در منطقه نظارتی وجود دارد، هزینه شکاف در نظر گرفته می‌شود. بنابراین هزینه هر یک از وضعیت‌ها قابل محاسبه است. به عنوان مثال هزینه وضعیت $< 1.1.8.2.3.5.3.1.6 >$ معادل C_{fail} است، چون عامل دوم در وضعیت actuatorfail قرار دارد و نمی‌تواند وظیفه برقراری ارتباط را به درستی انجام دهد.

- **نحوه تشکیل تیم‌های اقتضایی و انجام آزمایشات**

در آزمایش تجربی اول، تیم‌های متفاوتی را از پهپادهای محافظه‌کار، بدون هماهنگی با یکدیگر، تشکیل می‌دهیم. در این تیم‌ها هر عامل از سیاست خود پیروی می‌کند. سپس عملکرد این تیم‌ها را با حالتی که همه پهپادها به یک نحو

در کنار هم آموزش دیده‌اند، مقایسه می‌کنیم. خواهیم دید که هزینه تحمیلی به این دو نمونه تیم تفاوت قابل ملاحظه‌ای با هم ندارند.

با توجه به اینکه شرایط محیط در دنیای واقعی متغیر است، در آزمایش تجربی دوم، پارامترهای احتمال خرابی ثابت در نظر گرفته نشده و در گام‌های مختلف شبیه‌سازی به صورت تصادفی تغییر می‌کنند. در آزمایش تجربی دوم، عامل سوم از سیاست‌های یاد گرفته قبلی استفاده می‌کند و به صورت لحظه‌ای از روی رفتار هم‌تیمی‌ها تصمیم می‌گیرد که کدام عمل را باید انجام دهد. خروجی رفتار این تیم با تیم‌های قبلی مقایسه شده است. ما در این آزمایش روشی ارائه کرده‌ایم که تصمیم‌گیری رفتار بعدی عامل یادگیرنده در لحظه مأموریت به صورت برخط اتفاق بیفتد.

روش پیشنهادی جهت تصمیم‌گیری برخط عامل آموزش دیده در آزمایش دوم، یک الگوریتم حریصانه است که در الگوریتم ۱ و ۲ آمده است.

در الگوریتم ۱ بهترین سیاست در هر لحظه انتخاب می‌شود. روش کار به این صورت است که هزینه اجرای هر سیاستی که عامل قبلاً فرا گرفته است، در هر لحظه محاسبه می‌شود و سیاستی که منجر به کمترین هزینه گردد، انتخاب می‌شود.

الگوریتم ۲ برای محاسبه هزینه هر سیاست در هر لحظه مورد استفاده قرار می‌گیرد. ابتدا در خط شماره (۲)، محاسبه می‌شود که از وضعیت فعلی، احتمال گذار به کدام وضعیت‌ها وجود دارد. به عبارتی تمام وضعیت‌هایی که از وضعیت فعلی می‌توان با احتمال غیرصفر به آنها گذار کرد مشخص می‌شوند. این کار با کمک بررسی احتمال مصرف سوخت و خرابی پهپادها انجام می‌گیرد. برای مثال احتمال گذار از وضعیت < 2.3.5 > با انجام عمل پیشروی، به وضعیت < 3.1.4 > صفر است. چون پهپاد ناسالم تنها با برگشتن به پایگاه مرکزی تعمیر می‌گردد. به همین ترتیب، احتمال گذار از همان وضعیت به وضعیت < 3.3.5 > نیز صفر است. چون در هرگام از شبیه‌سازی، یک یا دو واحد از سوخت هر پهپاد مصرف می‌شود.

در خط شماره ۴، برای هر وضعیت بعدی ممکن (Next_States)، هزینه وضعیت بعدی در میزان احتمال گذار به آن وضعیت ضرب می‌شود و در نهایت به صورت تجمعی برای کلیه وضعیت‌های بعدی ممکن، به عنوان خروجی بازگردانده می‌شود.

الگوریتم ۱- انتخاب بهترین سیاست در هر لحظه

```
Function Policy-Selection (s is a state) returns policy
1.    for each policy which learned before
2.          Calculate Cost (State)
3.    end for
5.    return policy which makes cost minimum
```

الگوریتم ۲- محاسبه هزینه برای هر وضعیت

```
Function CalculateCost (s, policy) returns Total_Cost
1.    Total_Cost=0
2.    Create_Next (s, policy);
3.    for each state in Next_States
4.          Total_Cost = Transition (State, next_state, policy)
      * Cost(next_state) + Total_Cost
5.    end for
6.    return Total_Cost
```

به طور کلی الگوریتم حریصانه Policy-Selection، در هر وضعیت، سیاستی را انتخاب می‌کند که به صورت احتمالاتی، هزینه کمتری را در گام بعدی در بر داشته باشد.

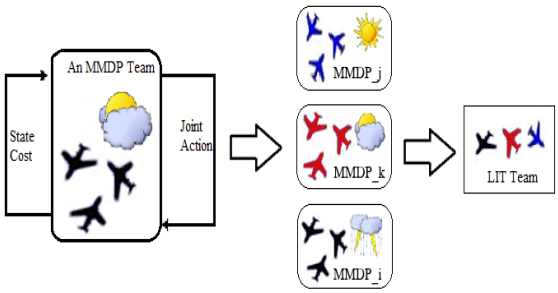
۳- آزمایش‌های تجربی

در این قسمت توضیح می‌دهیم که مسئله مدل شده در بخش ۲-۲، چطور حل شده و در نهایت نتایج به دست آمده را گزارش می‌نماییم.

۳-۱- هم‌تیمی‌های ناشناخته

در آزمایش اول تیم ما شامل عامل‌هایی است که گرچه همگی سیاست محافظه‌کارانه‌ای دارند ولی قبلاً در کنار هم آموزش ندیده‌اند. بنابراین یک تیم با اعضایی داریم که هر یک، سیستم را به صورت خاصی شناخته‌اند. این عامل‌ها یادنگرفته‌اند که در کنار عامل‌های دیگر تیم، بهترین عمل را انجام دهند. به عبارت دیگر، سیاست آموخته شده به هر عامل تیم،جهت پیروی داده می‌شود. بنابراین تیم حاصل از این عامل‌ها هیچ هماهنگی قبلی با یکدیگر ندارند و هر یک به صورت جداگانه از سیاست خود پیروی می‌کنند. در شکل ۳ شمای کلی این آزمایش آمده است.

در تشکیل تیم از عامل‌هایی استفاده شده است که قبلاً با پارامترهای موجود در جدول ۱ آموزش دیده‌اند. در این تیم‌ها احتمال خرابی سنسور و عملگر و هزینه شکست و نابودی و یا نرخ یادگیری با هم متفاوت هستند. به این ترتیب تیم‌هایی که همه اعضای آن درکنار هم آموزش دیده‌اند را با MMDP و تیم‌هایی که از تشکیل عامل‌های متفاوت و به صورت اقتضایی تشکیل داده‌ایم را LIT نامیده‌ایم. تیم‌های مختلفی از نوع LIT، تشکیل داده‌ایم. در جدول ۲ به این تیم‌ها و اعضای آنها اشاره شده است.



شکل ۳- شمای کلی آزمایش اول. از ترکیب اعضای تیم‌های مختلف MMDP، یک تیم جدید ایجاد می‌شود

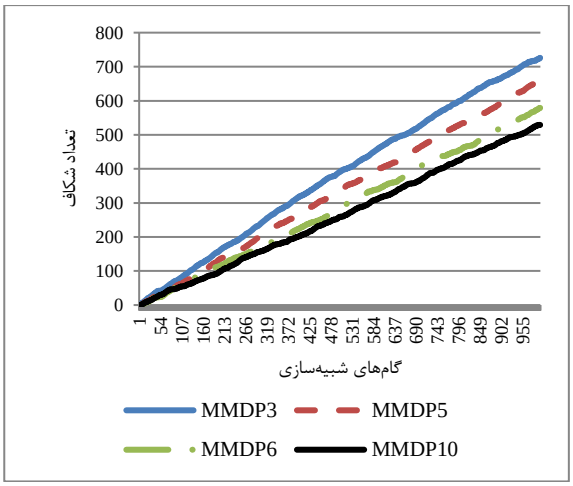
در سمت چپ شکل ۳ ملاحظه می‌شود که هر تیم MMDP به صورت جداگانه آموزش دیده است. این همان آموزش سیاست کلی محافظه‌کارانه است که قبلاً به آن اشاره شد. در ستون وسط نشان داده شده است که تیم‌های متفاوت MMDP در شرایط مختلف آموزش دیده‌اند. در این شکل، سه نمونه تیم MMDP به نام‌های MMDP_i، MMDP_j و MMDP_k نشان داده شده است. سپس در گام بعدی در سمت راست شکل، یک عضو از هر یک از این تیم‌ها را وارد تیم LIT نموده‌ایم. بنابراین این پهپادها نسبت به هم ناهمگون[۱۶] هستند و مانند هم رفتار نخواهند کرد. هیچ یک از پهپادها دقیقاً مطلع نیست که پهپاد هم‌تیمی او در یک وضعیت مانند s چطور عمل خواهد کرد. به عبارت دیگر پهپادها نسبت به هم ناشناس هستند و تیم تشکیل شده یک تیم اقتضایی می‌باشد.

ما خروجی عملکرد این تیم‌ها را از نظر میزان شکاف و شکست اتفاق افتاده در طول ۱۰۰۰ گام شبیه‌سازی مأموریت، با تیم‌های MMDP مقایسه کرده‌ایم.

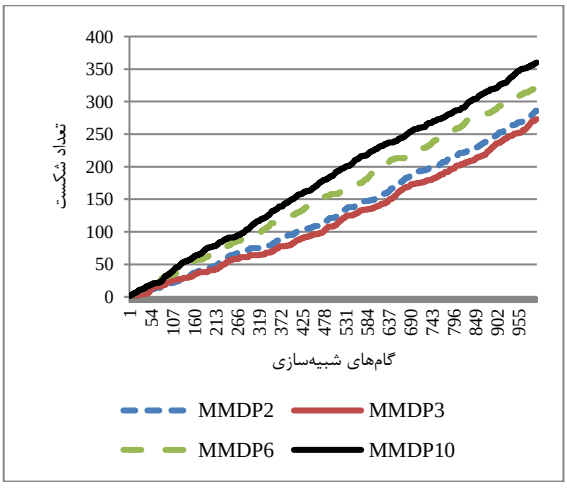
جدول ۲- اعضای تیم‌های مختلف LIT

عنوان تیم	عامل اول	عامل دوم	عامل سوم
LIT1	MMDP1	MMDP4	MMDP8
LIT2	MMDP1	MMDP5	MMDP3
LIT3	MMDP1	MMDP2	MMDP6
LIT4	MMDP2	MMDP5	MMDP6
LIT5	MMDP3	MMDP5	MMDP7
LIT6	MMDP8	MMDP9	MMDP10
LIT7	MMDP9	MMDP9	MMDP10
LIT8	MMDP9	MMDP10	MMDP10

در شکل ۴ نمودار تعداد شکاف و در شکل ۵ نمودار تعداد شکست در طول ۱۰۰۰ گام شبیه‌سازی شده، برای برخی از ۱۰ تیم MMDP قابل مشاهده است. در این شکل‌ها به دلیل گویا بودن نمودارها، از ترسیم نمودار باقی تیم‌ها که اعدادی بین MMDP3 و MMDP10 هستند، پرهیز شده و فقط به دو نمونه بسنده کرده‌ایم. عملکرد بقیه تیم‌ها بین این دو نمودار است. از طرفی کمترین شکست مربوط به MMDP3 و بیشترین شکست مربوط به MMDP10 است و نمودار میزان شکست بقیه تیم‌ها مابین این دو نمودار است.

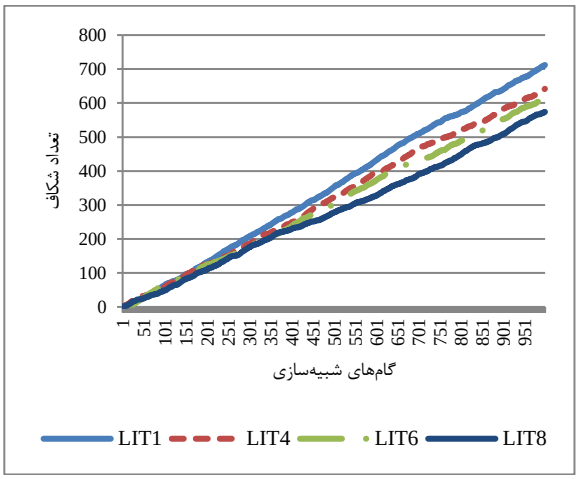


شکل ۴- نمودار تعداد شکاف بین تیم‌های MMDP

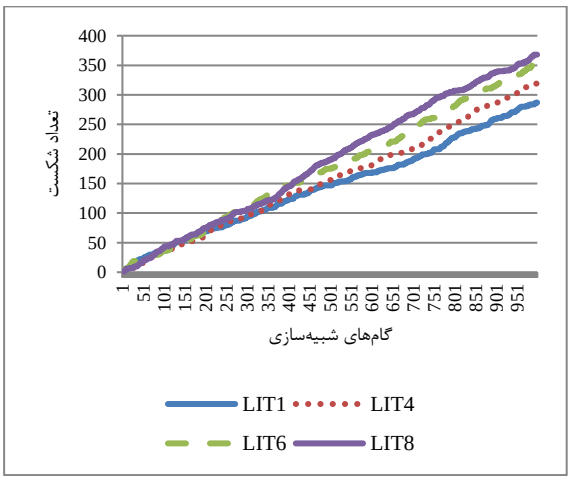


شکل ۵- نمودار تعداد شکست بین تیم‌های MMDP

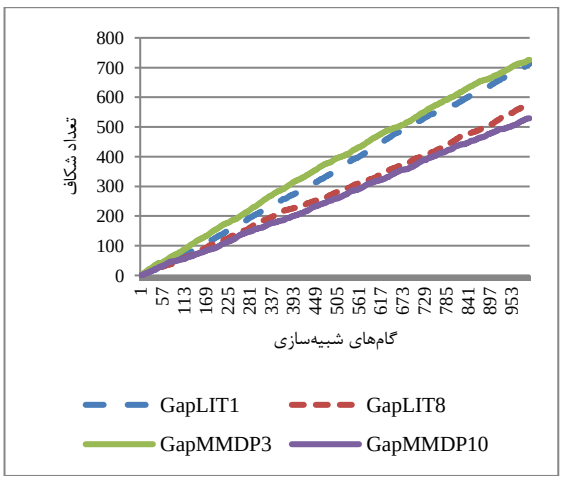
در شکل ۶ و ۷ نمودار میزان شکاف و شکست اتفاق افتاده در تیم‌های LIT نمایش داده شده است. تعداد شکاف بین ۵۷۴ تا ۷۱۲ و تعداد شکست بین ۲۸۷ تا ۳۶۸ متغیر است. در این تیم‌ها بیشترین تعداد شکست در تیم LIT8 و کمترین تعداد در LIT1 و بیشترین تعداد شکاف در LIT1 و کمترین تعداد شکاف در LIT8 دیده می‌شود. به دلیل گویا بودن نمودار، تنها دو نمونه نمودار میانی رسم شده است.



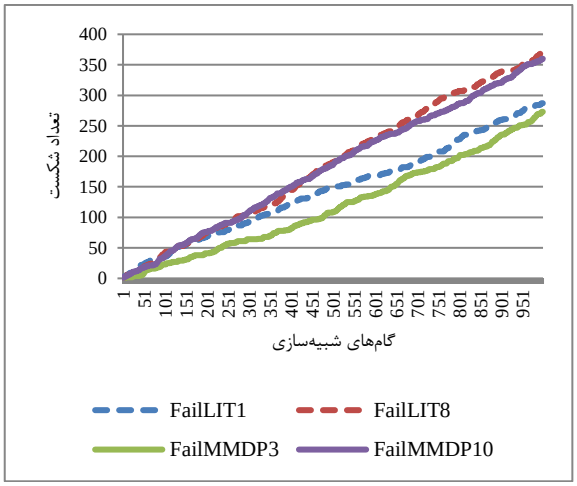
شکل ۶- نمودار تعداد شکاف بین تیم‌های LIT



شکل ۷- نمودار تعداد شکست تیم‌های LIT



شکل ۸- نمودار مقایسه تعداد شکاف تیم‌های LIT و MMDP



شکل ۹- نمودار مقایسه تعداد شکست تیم‌های LIT و MMDP

۳-۲- انتخاب برخط سیاست در محیط ناشناخته

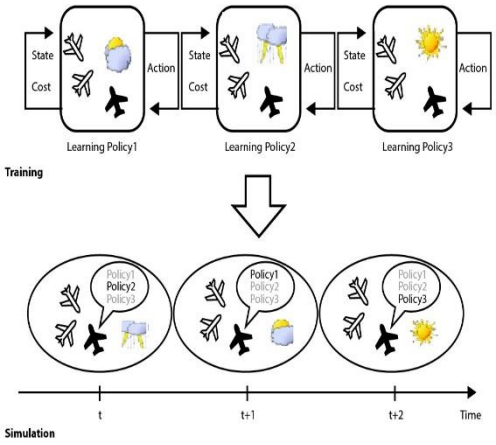
در آزمایش تجربی دوم، تمرکز بر روی تصمیم‌گیری برخط عامل‌ها می‌باشد. در این مرحله پارامترهای محیط به صورت متغیر در نظر گرفته شده‌اند. به این معنی که بسته به شرایط جوی ممکن است احتمال خرابی پهپادها نیز تغییر نماید. با این ایده، عاملی را در تیم‌های مختلف آموزش می‌دهیم به نحوی که بتواند چند سیاست در شرایط جوی مختلف را آموزش ببینند. ما در این آزمایش از تیم MMDP7 آزمایش قبل، دو عامل را در کنار یک عامل یادگیرنده، در یک تیم قرار داده‌ایم. در این شرایط دو عامل دیگر سیاست ثابتی را که در شرایط PS=0.1 و PA=0.05 یاد گرفته‌اند، اجرا می‌کنند. در اینجا منظور از PS احتمال خرابی سنسور و PA احتمال خرابی عملگر است.

عامل سوم، تحت شرایط متفاوتی، هماهنگی با محیط‌های مختلف را یاد می‌گیرد، در جدول ۳ این شرایط آمده است. شمای کلی این آزمایش در شکل ۱۰ نشان داده شده است. سیاست‌های آموخته شده توسط عامل سوم، در محیط شبیه‌سازی به عنوان ورودی وجود دارند و این عامل برحسب تغییر در شرایط محیط و از روی رفتار هم‌تیمی‌هایش، تصمیم می‌گیرد از کدام سیاست پیروی کند. الگوریتم انتخاب سیاست در هر وضعیت و الگوریتم محاسبه هزینه‌های هر سیاست در هر گام شبیه‌سازی، به ترتیب در الگوریتم شماره ۱ و ۲ در بخش ۲-۵ آمده است.

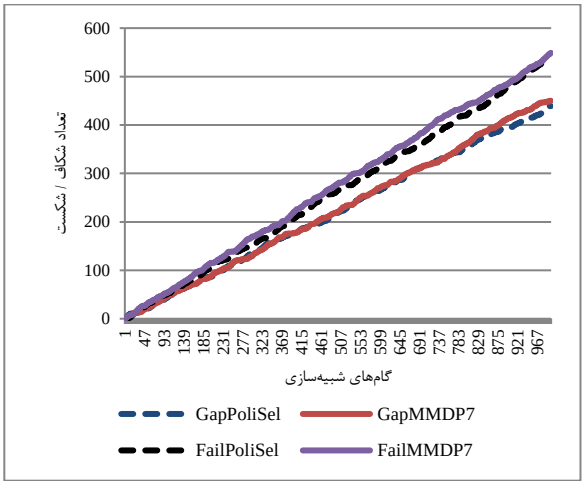
جدول ۳- شرایط مختلف یادگیری عامل سوم

سیاست آموخته شده	احتمال خرابی عملگر	احتمال خرابی سنسور	هزینه شکست	هزینه نابودی
Policy1	۰,۱	۰,۱	۲۰	۱۰۰۰
Policy2	۰,۲۵	۰,۴	۱۵	۱۰۰۰
Policy3	۰,۰۵	۰,۰۵	۱۵	۱۰۰۰

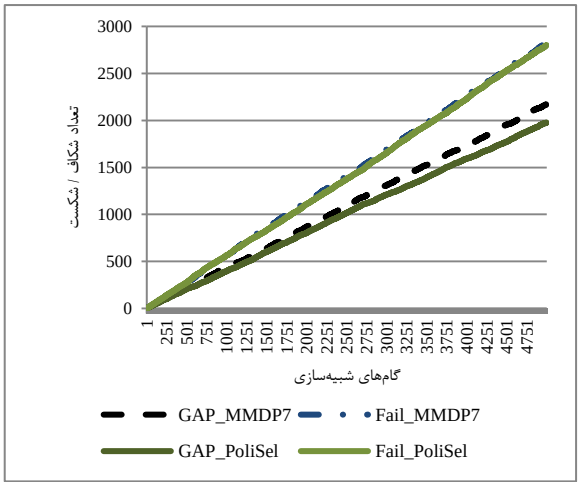
در قسمت بالایی شکل ۱۰، عامل با رنگ متفاوت در حال آشنا شدن با محیط‌های مختلف است. او یاد می‌گیرد درحالی‌که سایر هم‌تیمی‌هایش سیاست ثابتی را اجرا می‌کنند، بهترین عملی که در آن شرایط محیطی باید اجرا نماید، کدام است. در قسمت پایینی این شکل، محیط شبیه‌سازی به تصویر کشیده شده است. در این مرحله، پارامترهای محیط متغیر هستند و عامل آموزش دیده با کمک الگوریتم ۱ یکی از سیاست‌هایی را که قبلاً فرا گرفته است، اجرا می‌کند.



شکل ۱۰- شمای کلی آزمایش دوم. آموزش یک عامل در محیط‌های مختلف (قسمت بالایی شکل)، امکان انتخاب سیاست بهتر در زمان شبیه‌سازی (قسمت پایین شکل) را فراهم می‌آورد



شکل ۱۱- نمودار تعداد شکست و شکاف تیم PoliSel در مقایسه با MMDP7 در حالتی که پارامترهای احتمال خرابی در محیط مطابق جدول ۳ تغییر می‌کنند



شکل ۱۲- نمودار تعداد شکست و شکاف با الگوریتم PoliSel در مقایسه با MMDP7 در حالتی که پارامترهای احتمال خرابی در محیط به صورت تصادفی تغییر می‌کنند

در این آزمایش نیز، نتیجه حاصل با MMDP مقایسه شده است. تیم مقایسه شده MMDP7 است. سیاست‌های آموخته شده توسط عامل آموزش دیده، در محیط شبیه‌سازی به عنوان دانش قبلی وی وجود دارند و این عامل برحسب تغییر شرایط محیط و از روی رفتار هم‌تیمی‌هایش، تصمیم می‌گیرد از کدام سیاست پیروی کند. ابتدا شرایط محیط را به طور یکنواخت بین Policy1 و Policy2 و Policy3 تغییر داده‌ایم و تیم PoliSel را معرفی کرده‌ایم که در آن در هر لحظه عامل سوم، یکی از سیاست‌هایی که آموخته است را برمی‌گزیند. سپس خروجی PoliSel را با خروجی MMDP7 در ۱۰۰۰ گام شبیه‌سازی مقایسه کرده‌ایم. نتایج حاصل در شکل ۱۱ برای تعداد شکاف و شکست ترسیم شده است. این نمودارها نشان می‌دهند که میزان شکاف تقریباً ۲٫۵٪ و میزان شکست ۲٫۸٪ کاهش داشته است. در این آزمایش کاهش هر دو مقدار هزینه، نشان دهنده عملکرد بهتر سیستم است.

در ادامه آزمایش، محیط شبیه‌سازی به صورت تصادفی از توزیع یکنواخت، مقادیری را برای PA و PS اختیار می‌کند. در این حالت نیز خروجی الگوریتم PoliSel را با MMDP7 در ۵۰۰۰ گام شبیه‌سازی در شکل ۱۲ می‌بینید. ملاحظه می‌شود که تعداد شکاف در حالت PoliSel بهتر از MMDP7 و تعداد شکست نیز تقریباً با هم برابر است. این در حالی است که به دلیل متغیر بودن احتمال خرابی، تعداد خرابی اتفاق افتاده برای سنسور و عملگر در شبیه‌سازی PoliSel به ترتیب ۲۵۵۰ و ۹۴۹ بوده است و در شبیه‌سازی MMDP7 این مقادیر به ترتیب معادل ۲۴۲۹ و ۹۰۴ بوده‌اند. بنابراین با وجود این که در شبیه‌سازی MMDP7 خرابی کمتری اتفاق افتاده ولی خروجی PoliSel هم در تعداد شکست‌ها و هم تعداد شکاف‌های اتفاق افتاده در طول مأموریت بر MMDP7 برتری دارد.

۳-۳- پیاده‌سازی و شبیه‌سازی

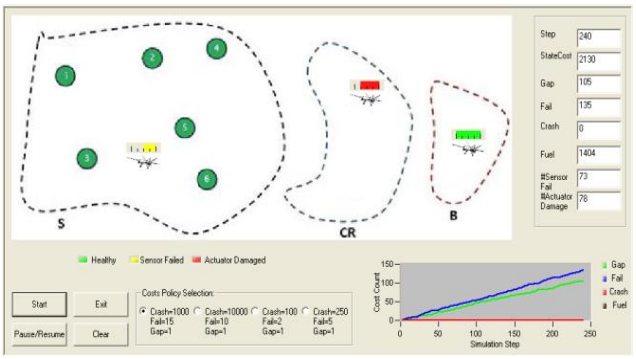
روش به کار رفته در یادگیری، برای حل فرآیند تصمیم‌گیری مارکوف، برنامه نویسی پویا و روش Value Iteration می‌باشد. پیچیدگی محاسباتی هر تکرار در این روش، نسبت به فضای حالت‌ها از درجه دوم و نسبت به فضای عمل‌ها خطی است [۲۳]. در مسئله حاضر با توجه به اینکه با انتخاب یک عمل در یک وضعیت، احتمال گذار تنها به تعداد محدودی وضعیت دیگر غیرصفر است، پیچیدگی محاسباتی هر تکرار در این الگوریتم، نسبت به فضای حالت نیز خطی است.

پهپادهای متفاوت از طریق Value Iteration آموزش داده شده‌اند. خروجی برنامه‌های یادگیری، بهترین عملی است که در هر حالت هر پهپاد می‌تواند انجام دهد. ما این خروجی‌ها را در فایل‌هایی به نام Policy نگهداری کرده‌ایم.

برنامه یادگیری در محیط Visual Studio 2015 و با زبان C++، روی یک سیستم با پردازنده Intel Core i5-3320 با سرعت 2.60GHz و با 8GBRAM و سیستم عامل Windows 7 64-bit کدنویسی شده است.

ما در این کار با کمک ابزارهای توسعه MSTD[۱۷] که شامل ابزارهای خاصی جهت تسهیل نمایش و کارهای گرافیکی مورد نظر در یک سیستم شبیه‌سازی است، محیط شبیه‌سازی را کدنویسی کرده‌ایم. علت این که از شبیه‌سازهای موجود برای سیستم‌های چندعامله مانند GAMA، MASON یا JADE استفاده نشده، استفاده ساده‌تر از سیستم شبیه‌سازی و انطباق کامل آن با نیازهای موجود در آزمایشات بوده است. برنامه شبیه‌سازی نیز در Visual Studio 6 و با زبان C++ بر روی ماشین مجازی با سیستم عامل XP روی همان ماشین کدنویسی و اجرا گردیده است. فایل‌های Policy به عنوان ورودی برنامه شبیه‌سازی هستند.

در شکل ۱۳ نمونه یک تصویر از شبیه‌سازی در حالت MMDP7، مشاهده می‌شود. در گام ۲۴۰ شبیه‌سازی، تعداد شکاف اتفاق افتاده ۱۰۵ مورد و تعداد شکست‌ها ۱۳۵ مورد است.



شکل ۱۳- تصویر برنامه شبیه‌سازی، MMDP7

۴- تحلیل نتایج

در آزمایش اول تک‌تک عامل‌ها تحت یک سیاست کلی ولی با پارامترهای مختلف آموزش دیده‌اند. در این آزمایش هم‌تیمی‌ها برای هم ناشناخته بوده و هیچ هماهنگی قبلی بین عامل‌ها صورت نگرفته است. نمودار شکل‌های ۴ تا ۹ نشان می‌دهند که مجموع تعداد شکاف و شکست اتفاق افتاده، تغییر چندانی با حالتی که همه اعضای تیم به یک شکل و از قبل در کنار هم آموزش دیده باشند، ندارد.

در آزمایش دوم نتایج حاصل نشان می‌دهند که آموزش پارامترهای مختلف محیط به یک عامل قبل از شروع مأموریت و بالا بردن دانش عامل از شرایط محیطی مختلف، موجب عملکرد بهتر تیم در زمان اجرا خواهد شد. در این تیم‌ها دو عامل اول یک دیدگاه نسبت به شرایط محیط دارند، در صورتی که عامل سوم تحت پارامترهای متفاوتی با محیط‌های مختلف آشنا شده است.

لذا می‌توان نتیجه گرفت که در شرایطی که پارامترهای محیط حاکم بر مسئله از ابتدا روشن نیست و یا مدام در حال تغییر است، یادگیری شرایط مختلف حتی فقط به یک عامل، می‌تواند منجر به عملکرد بهتر تیم گردد. بنابراین وقتی با تیم‌های اقتضایی مواجه هستیم و در آنها کنترل همه عامل‌ها در اختیار ما نیست، می‌توانیم با ارائه آموزش‌های وسیع‌تر به عامل تحت کنترل، به سمت بهبود عملکرد تیم پیش برویم.

در مأموریت نظارت مداوم، دامنه عمل‌های ممکن برای یک عامل بسیار محدود است. هرچه این دامنه وسیع‌تر باشد، تفاوت بین عملکرد تیم‌ها آشکارتر خواهد بود.

۵- نتیجه‌گیری و کارهای آینده

در این مقاله نشان دادیم که می‌توانیم بدون اینکه عامل‌ها را دقیقاً با پارامترهای دنیای واقعی و به صورت هماهنگ آموزش دهیم، به رفتار آنها در یک تیم امیدوار باشیم؛ به شرطی که در یادگیری هر عامل به طور جداگانه، از سیاست کلی حاکم بر مسئله پیروی کرده باشیم.

از طرف دیگر، در آزمایش دوم دیدیم که وقتی پارامترهای دنیای واقعی برای ما در حین یادگیری دقیقاً مشخص نباشند، عملکرد عامل‌هایی که تجربه قبلی بیشتری دارند، موجب بهبود عملکرد تیم خواهد شد. در این آزمایش، نشان دادیم که تصمیم‌گیری برخط یک عامل در تیم، عملکرد کل تیم را بهبود می‌دهد.

در این مقاله تنها برای یک تیم با سه عامل یادگیری انجام شد ولی ما می‌خواهیم همین رویکرد را برای تیم‌های بزرگتر نیز بسنجیم. در آن صورت بایستی به روش‌های تقریبی در یادگیری مانند [۱۵] متوسل شویم تا بتوانیم مسئله را مهار نماییم.

در این مقاله بحث تشکیل تیم و تصمیم‌گیری برخط عامل‌ها، تنها در مسئله مأموریت نظارت مداوم مورد بررسی قرار گرفت. ما می‌خواهیم در ادامه این بررسی را در مسائل مشابهی از سیستم‌های چندعامله که در آنها دامنه عمل‌های ممکن وسیع‌تر باشد، انجام دهیم. در این صورت تفاوت بین عملکرد تیم‌ها آشکارتر خواهد بود.

مراجع

[1] S. Barrett, A. Rosenfeld, S. Kraus, and P. Stone, "Making friends on the fly: Cooperating with new teammates," Artificial Intelligence, vol. 242, pp. 132-171, 2017.

[2] D. V. Pynadath, and M. Tambe, "The communicative multiagent team decision problem: Analyzing teamwork theories and models," Journal of artificial intelligence research, vol. 16, pp. 389-423, 2002.

[3] W. Ren, R. W. Beard, and E. M. Atkins, "A survey of consensus problems in multi-agent coordination," In Proceedings of the American Control Conference, 2005, pp. 1859-1864.

[4] O. Obst, and J. Boedecker, "Flexible coordination of multiagent team behavior using HTN planning," In Robot Soccer World Cup, pp. 521-528. Springer, 2005.

[5] M. Tambe, "Towards flexible teamwork," Journal of artificial intelligence research, vol. 7, pp. 83-124, 1997.

[6] C. Boutilier, "Sequential optimality and coordination in multiagent systems," In IJCAI, vol. 99, pp. 478-485, 1999.

[7] C. Claus, and C. Boutilier, "The dynamics of reinforcement learning in cooperative multiagent systems," AAAI/IAAI 1998, pp. 746-752, 1998.

[8] D. Virginia, "Handbook of Research on Multi-Agent Systems: Semantics and Dynamics of Organizational Models," IGI Global, 2009.

[9] P. Stone, G. A. Kaminka, and J. S. Rosenschein, "Leading a best-response teammate in an ad hoc team," In Agent-mediated electronic commerce, Designing trading strategies and mechanisms for electronic markets, pp. 132-146, Springer, 2010.

[10] P. Stone, G. A. Kaminka, S. Kraus, and J. S. Rosenschein, "Ad Hoc Autonomous Agent Teams: Collaboration without Pre-Coordination," In 24th AAAI Conference on Artificial Intelligence, 2010.

[11] F. Wu, S. Zilberstein, and X. Chen, "Online planning for ad hoc autonomous agent teams," In 22th International Joint Conference on Artificial Intelligence, 2011, pp. 439-445.

[12] S. Barrett, P. Stone, S. Kraus, and A. Rosenfeld, "Learning teammate models for ad hoc teamwork," In AAMAS Adaptive Learning Agents (ALA) Workshop, 2012, pp. 57-63.

[13] S. Barrett, and P. Stone, "An analysis framework for ad hoc teamwork tasks," In Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-vol. 1, pp. 357-364, 2012.

[14] S. Barrett, P. Stone, and S. Kraus, "Empirical evaluation of ad hoc teamwork in the pursuit domain," In the 10th International Conference on Autonomous Agents and Multiagent Systems- vol. 2, pp. 567-574, 2011.

[15] J. D. Redding, "Approximate multi-agent planning in dynamic and uncertain environments," PhD diss., Massachusetts Institute of Technology, 2011.

[16] B. Bethke, J. How, and J. Vian, "Multi-UAV Persistent Surveillance with Communication Constraints and Health Mangement," In AIAA Guidance, Navigation, and Control Conference, p. 5654, 2009.

[17] J. D. Redding, N. K. Ure, J. P. How, M. A. Vavrina, and J. Vian, "Scalable, MDP-based planning for multiple, cooperating agents," In American Control Conference (ACC), 2012, pp. 6011-6016.

[18] S. Barrett, and P. Stone, "Cooperating with Unknown Teammates in Complex Domains: A Robot Soccer Case Study of Ad Hoc Teamwork," In 29th AAAI conference on artificial intelligence, 2015, pp. 2010-2016.

[19] J. M. Vidal, "Fundamentals of Multi-Agent Systems with NetLogo, " Available: http://multiagent.com/, 2010.

[20] B. Dunin-Keplicz, and R. Verbrugge, "Teamwork in multi-agent systems: A formal approach," vol. 21, John Wiley & Sons, 2011.

[21] D. V. Pynadath, and M. Tambe, "Multiagent teamwork: Analyzing the optimality and complexity of key theories and models," In Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 2, 2002, pp. 873-880.

[22] K. Genter, N. Agmon, and P. Stone, "Role-Based Ad Hoc Teamwork," In Workshops at the 25th AAAI Conference on Artificial Intelligence, 2011.

[23] R. A. Howard, "Dynamic programming and Markov processes," NEW YORK: JOHN-WILEY, 1964.

رقیه حیدری در سال ۱۳۸۳ مدرک کارشناسی علوم کامپیوتر خود را از دانشگاه شهید بهشتی تهران و در سال ۱۳۹۶ مدرک کارشناسی ارشد علوم کامپیوتر گرایش سیستم‌های هوشمند خود را از دانشگاه تحصیلات تکمیلی علوم پایه زنجان دریافت نمود. وی در حال حاضر در شرکت توزیع نیروی برق زنجان مشغول به کار است. زمینه‌ی تحقیقاتی مورد علاقه‌ی وی سیستم‌های چندعامله هوشمند و یادگیری ماشین می‌باشد. خانم حیدری مدارک کارشناسی و کارشناسی ارشد خود را با درجه ممتازی کسب کرده است.

آدرس پست‌الکترونیکی ایشان عبارت است از:

r.heidari@zedc.ir

محسن افشارچی در سال ۱۳۷۵ مدرک کارشناسی ارشد مهندسی کامپیوتر خود را از دانشگاه علم و صنعت ایران و در سال ۱۳۸۵ مدرک دکترای هوش مصنوعی خود را از دانشگاه کالگری کانادا دریافت نمود. دکتر افشارچی از سال ۱۳۷۵ مدرس دانشگاه زنجان بوده است. وی هم‌اکنون دانشیار دانشکده مهندسی دانشگاه زنجان و محقق و استاد مدعو دانشگاه تحصیلات تکمیلی علوم پایه زنجان می‌باشد. ایشان از سال ۱۳۸۵ هدایت آزمایشگاه سیستم‌های چندعامله دانشگاه زنجان را بر عهده داشته‌اند. زمینه‌ی تحقیقاتی مورد علاقه‌ی وی Multi-agent Learning، Probabilistic Reasoning و Distributed Constraint Optimization می‌باشد.
آدرس پست‌الکترونیکی ایشان عبارت است از:
afsharchi@znu.ac.ir

رضا خان‌محمدی مدرک کارشناسی و کارشناسی ارشد خود را به ترتیب از دانشگاه زنجان در مهندسی کامپیوتر و دانشگاه آزاد قزوین در رشته مکاترونیک دریافت نموده است. وی هم‌اکنون عضو شرکت دانش بنیان عصر رایانه می‌باشد و هم‌زمان در شرکت برق منطقه‌ای زنجان مشغول به کار است. زمینه‌ی تحقیقاتی مورد علاقه ایشان یادگیری ماشین، پردازش تصویر و رباتیک می‌باشد.
آدرس پست‌الکترونیکی ایشان عبارت است از:
reza.km@asrerayaneh.com

اطلاعات بررسی مقاله:

تاریخ ارسال: ۱۳۹۶/۴/۱۷
تاریخ اصلاح: ۱۳۹۶/۱۰/۲۰
تاریخ قبول شدن: ۱۳۹۷/۰۴/۱۱
نویسنده مرتبط: دکتر محسن افشارچی، دانشکده مهندسی، دانشگاه زنجان، زنجان، ایران.

---

[1]Autonomous Unmanned Agents
[2]Cooperation
[3]Flexible Coordination of Multi-Agent
[4]Flexible Teamwork
[5]Joint-Action
[6]Peter Stone
[7]Transfer Learning
[8]Persistent Surveillance Mission
[9]Unmanned Aerial Vehicle
[10]General Policy
[11]Autonomous UAV
[12]Crash Cost
[13]Fail Cost
[14]Gap Cost
[15]Markov Decision Process
[16]Heterogeneous
[17]Measurement Studio