انتقال دانش میان عاملهای غیرهمسان به کمک اشتراک رویداد و اشتراک دانش

سید محمد رضا میرفتاح مجید نیلی احمدآبادی بابک نجار اعرابی

قطب علمی کنترل و پردازش هوشمند و آزمایشگاه هوش مصنوعی و رباتیک، گروه مهندسی برق و کامپیوتر، دانشکدهٔ فنی، دانشگاه تهران

پژوهشکدهٔ علوم شناختی، مرکز تحقیقات فیزیک نظری و ریاضیات

araabi@ ut.ac.ir mnili@ ut.ac.ir mirfataa@ yahoo.com

چکیده: موضوع این نوشتار، ارائهٔ روشهایی است که همکاری در یادگیری را میان عاملهای هوشمند غیرهمسان میسر می‌سازند. پیش از این، مفید بودن همکاری در یادگیری میان عاملهای همسان به اثبات رسیده است. لیکن چگونگی همکاری در یادگیری میان عاملهای غیرهمسان که از لحاظ توانایی های عملکردی متفاوت هستند، به تازگی مورد توجه محققین علم هوش مصنوعی قرار گرفته است. در این مقاله، روشهایی برای همکاری مفید میان عاملهای غیرهمسان به هنگام یادگیری، ارائه شده است. به این ترتیب، عاملهایی که دارای اعمال مختلف هستند، می توانند با یکدیگر ارتباط معنی دار برقرار کرده و به نحو مطلوبی از دانش و تجارب دیگران برای ارتقاء دانش و کارایی خود استفاده کنند.

کلمات کلیدی: عاملهای غیرهمسان، اَبَرجدول نگاشت اعمال، همکاری در یادگیری تقویتی، اشتراک رویداد، اشتراک دانش

۱. مقدمه

اگر به چگونگی فرایند آموختن بیندیشیم، احتمالاً اولین ایده‌ای که به نظرمان می‌رسد، آموختن از طریق تعامل با محیط اطراف به همراه دریافت پاداش یا تنبیه از آن است. در این نوع یادگیری که یادگیری تقویتی[۱] نام دارد، عامل در جستجوی حالتهایی از محیط است که پاداشهای کسب شده را حداکثر و تنبیه‌ها را حداقل کند. یکی از روشهای پیاده سازی این نوع یادگیری، Q-Learning می‌باشد. در این روش، دانش عامل توسط یک جدول نشان داده می‌شود. بازای هر حالت محیط، یک سطر و بازای هر عملی که عامل توانایی انجام آن را داشته باشد، یک ستون در این جدول قرار دارد. مقادیر موجود در خانه‌های این جدول نشان می‌دهند که در هر حالت، میزان مطلوبیت هر عمل چقدر است. هر چه این مقدار بیشتر باشد، عمل مربوطه مطلوبتر بوده و مسیر بهتری را برای رسیدن به هدف نشان می‌دهد. عامل با سعی و خطا و انجام دادن اعمال مختلف و گرفتن پاداش یا تنبیه از محیط، مقادیر موجود در جدول را بروز می کند.

تجربه کردن همه اعمال برای همه وضعیت ها، ممکن است از توان یک عامل خارج بوده یا لااقل مستلزم صرف زمان طولانی برای او باشد. برای بالا بردن کارایی عاملها می‌توان از دانش کسب شده توسط دیگران نیز استفاده کرد. همکاری در یادگیری در جوامع انسانی و حیوانی دارای نمونه های فراوان است. در محیط عاملهای هوشمند نیز کارایی این روش نشان داده شده است (مثلا در [۱] نشان داده شده که استفاده بجا از داده های بدست آمده از حواس دیگر عاملها و یا به اشتراک گذاشتن تجارب کسب شده توسط آنها موجب تسریع فرایند یادگیری می‌شود.)

در بیشتر تحقیقات انجام شده روی سیستمهای چند عامله فرض بر همسان[۲] بودن عاملها قرار دارد. اما آنچه واقعاً در طبیعت مشاهده می‌شود، غیر از این است؛ حتی رباتهای تجاری به ظاهر یکسان نیز از لحاظ ویژگیهای حسی و حرکتی دقیقاً مانند هم نیستند. بعد از کار Ming Tan در سال ۱۹۹۳ [۱] که یادگیری تقویتی را بین عاملهای همسان بررسی کرد، در سال ۱۹۹۶ مقاله‌ای مشابه در مورد عاملهای غیر همسان منتشر شد [۲]. در این کار نویسندگان

---

[۱] Reinforcement learning

[۲] Homogeneous

با متفاوت انتخاب کردن نرخ یادگیری عاملها، آنها را غیرهمسان کرده‌اند. [۳] در سال ۲۰۰۰ به بررسی تقلید از یک عامل متفاوت در توانایی های فیزیکی (حرکتی) پرداخته است. در این کار عامل مقلد (آموزنده) با انجام تست امکان پذیری و تست جایگزینی، از مربی که عاملی است خبره، برای سریعتر پرکردن جدول Q خود کمک می‌گیرد. نتایج تحقیق دیگری که بدنبال پاسخ به این سؤال بوده که چگونه گروهی از عاملهای غیرهمسان، که در حل مسائل مشابهی دخیل هستند، می‌توانند به کمک مبادله مصلحت، همکاری کرده و کاراییشان را افزایش دهند، در[۴] ارائه شده است. غیرهمسانی مورد نظر در این کار ناشی تفاوت در روشهای یادگیری عاملها بوده است. جدیدترین مقالات در زمینه همکاری در یادگیری میان عاملهای غیرهمسان از لحاظ توانایی‌های فیزیکی، به چگونگی استفاده از چارچوب ریاضی ارائه شده برای تطابق[۳] [۵]، در یادگیری اعمال به کمک تقلید پرداخته است. در این تحقیقات مکانیزم ALICE[۴] به عنوان روشی جامع برای مشخص کردن تطابقات با استفاده از هر روش ایجاد رفتار تقلیدی، معرفی شده است[۶].خلاصه جامعی از تحقیقات مرتبط با موضوع این مقاله در [۷] ارائه شده است.

سؤال اصلی این تحقیق در مورد چگونگی انتقال و اشتراک دانش در گروهی از عاملهای یادگیر غیرهمسان[۵] است. غیرهمسانی این عاملها در اعمالی است که قابلیت اجرای آن را دارند. روش یادگیری عاملها، One-step Q-learning بوده و مسألهٔ مورد نظر، توسط فرایند تصمیم گیری مارکوف[۶] مدل شده است. پیش از این [۱][۲]، انتقال دانش میان عاملهای همسان مورد بررسی قرار گرفته و روشهای خوبی نظیر اشتراک وزندار استراتژی نیز برای این منظور طراحی شده است[۸]. لیکن صحت عملکرد تمام این روشها منوط به همسان بودن عاملها و یکسان بودن دانش بهینهٔ آنها است. اگر عاملها از لحاظ مجموعهٔ اعمالی که در اختیار دارند متفاوت باشند، دانش نهایی آنها متفاوت بوده و در نتیجه موضوع انتقال دانش با مشکل روبرو می‌شود. در واقع باید پاسخی برای چنین پرسشهایی یافت: آیا دانش یک عامل برای عامل غیرهمسان دیگر قابل استفاده است یا نه؟ آیا می‌توان با بکارگیری روشهایی، دانشهای غیرهمسان را به اشتراک گذاشت؟

در فصل بعد به تعریف مساله مورد نظر خود پرداخته و ضمن فصل بعد از آن، دو راه حل پیشنهادی خود را ، ارائه می نمائیم. در فصل ۴ به ذکر نتایج آزمایش‌هایی که نشان دهنده صحت عملکرد روش پیشنهادی است، پرداخته شده است. جمع بندی و بیان نتایج در فصل ۵ آمده است.

۲. تعریف مساله

مساله مورد بررسی در این مقاله قابلیت مدل شدن توسط فرایندهای تصمیم‌گیری مارکوف را دارد. در این مساله فرض شده که عاملها از هم مستقل بوده و عملکرد آنها با هم تداخل ندارد. در MDPهای مورد بررسی این تحقیق[۷]، فرضیات زیر در نظر گرفته شده‌اند:

الف) فضای حالت (S) برای تمام عاملها یکسان است. ب) هر عامل از مجموعهٔ اعمال خود (A) مطلع است. این اطلاع تنها به معنای دانستن تعداد اعمال و تمایز قائل شدن بین آنهاست. عامل از نتیجهٔ عمل خود از قبل اطلاعی ندارد. پ) تابع انتقال حالت (T) برای عاملها مشخص نبوده و عاملها بدنبال یادگیری آن نیستند (یادگیری غیرمبتنی بر مدل). همچنین احتمال وجود اغتشاش در محیط وجود دارد. ت) تابع پاداش (R) برای عاملها معلوم نیست اما برای تمام عاملها یکسان است. این تابع، بصورت تابعی از حالت عامل (و نه حالت-عمل) تعریف شده است. مسأله مورد بررسی دارای یک هدف بوده و تنها بازای رسیدن به آن، پاداشی (مثبت) به عاملها تعلق می‌گیرد. در سایر حالات، عامل، تنبیه (با مقادیر مختلف منفی) می‌شود (مثلاً بازای اتلاف انرژی، یا رفتن به حالات مخاطره آمیز). بستر آزمایشی که برای یادگیری در نظر گرفته شده، حل یک ماز[۸] است. در این مساله ، هدف ، پیدا کردن مسیر مناسبی از میان موانع ، برای رساندن عامل به یک هدف از پیش تعیین شده می‌باشد. دو عامل غیرهمسان بطور مستقل برای یادگیری این ماز تلاش می‌کنند. یکی از این دو عامل در لحظات مختلف از دانش عامل دیگر استفاده می‌کند.

مساله غیرهمسانی در دو نوع مورد بررسی قرار گرفته است. در نوع اول، مجموعه اعمال هر دو عامل یکسان بوده و تنها ترتیب آنها متفاوت است. به عنوان مثال، مجموعه{U, D, R, L} را برای حرکت در چهار جهت اصلی در نظر گرفته، برای یک عامل ترتیب (U, D, L, R) و برای دیگری ترتیب (D, R, U, L) را در نظر می‌گیریم. در نوع دوم مجموعه اعمال یک عامل زیرمجموعه اعمال دیگری است. در ضمن تأثیر عمل (یا اعمال) اضافی، توسط جایگشتی از بعضی اعمال مشترک، قابل بازسازی است.به عنوان مثال، یکی دارای مجموعه اعمال پنج حرکتی{U, D, R, L, UR} بوده و دیگری دارای مجموعه اعمال{U, R, L, D} است. پیش از این در [۹][۱۰] چگونگی آموختن نگاشت بین اعمال در هر دو حالت غیرهمسانی

---

[۳] Correspondence

[۴] Action Learning for Imitation via Correspondence between Embodiments

[۵] Heterogeneous

[۶] ویژگی این مدل این است که هر انتقال حالت و سیگنال تقویتی که عامل با آن مواجه می‌شود، تنها به حالت فعلی و عملی که در آن انتخاب کند، بستگی داشته و به حالات قبلی ارتباطی ندارد.

[۷] $M = \langle S, A, T, R \rangle$

[۸] Maze Problem

فوق، مورد بررسی قرار گرفته و برای این منظور استفاده از جدول نگاشت اعمال[۹] پیشنهاد شده است. اطلاعات کامل در مورد این روش در [۷] ارائه شده است.

انگیزهٔ اصلی این مقاله، گسترش روشهای همکاری در یادگیری (من جمله WSS [۸]) به دامنهٔ عاملهای غیرهمسان است. در اینجا می‌خواهیم با استفاده از مقدماتی که فراهم شده، انتقال دانش بین عاملهای غیرهمسان را بصورتی مفید و کارا مورد بررسی قرار دهیم. در این مقاله فرض شده که عاملهای همکار، نگاشت بین اعمال یکدیگر را در اختیار داشته و می خواهند برای بالا بردن سطح دانش خود، در حد امکان از دانش دیگران بهره‌برداری کنند. واضح است در صورتی که غیرهمسانی نوع اول به کمک داشتن AMT مناسب، رفع شده باشد، انواع روشهای همکاری در یادگیری که برای عاملهای همسان پیشنهاد شده، برای چنین عاملهای غیرهمسان نیز قابل استفاده است. در واقع با بکارگیری AMT، می‌توان عاملها را همسان در نظر گرفت. مسألهٔ اصلی این مقاله در مورد تأثیر غیرهمسانی نوع دوم و در اصل حضور اعمال اضافی در مجموعه اعمال بعضی از عامهای همکار است.

در این مقاله فرض می کنیم عاملی که دارای عمل (یا اعمال) اضافی است، از عامل دیگر خبره‌تر است. در نتیجه، آنها را خبره و غیرخبره می‌نامیم. هدف این است که با طراحی روشهایی، با کمک دانش خبره، سرعت یادگیری غیرخبره را افزایش دهیم. در مورد فرض معکوس، یعنی خبره‌تر بودن عامل دیگر به [۷] مراجعه کنید.

۳. انتقال دانش

عاملهایی که در این مقاله مورد توجه قرار دارند، از روش Q-Learning استفاده می کنند. در این روشِ غیرمبتنی بر مدلِ حل MDPها، بازای هر حالت و هر عمل، یک ارزش (Q(s, a)) محاسبه می‌شود. این ارزش، مقدار مطلوبیت انتخاب عمل مربوطه را در آن حالت نشان می‌دهد. در تحقیق حاضر از نمایش جدولی مقادیر Q(s, a) استفاده شده است.

نکته‌ای که در اینجا اشاره به آن مورد نظر بوده، این است که در روش Q-Learning، داشتن اعمال اضافی باعث بزرگتر شدن (احتمالی) مقادیر ارزش Q(s, a)ها می‌شود. به تعبیر حسی‌تر، اگر گزینه‌های جدیدی که در رسیدن به یک هدف، در اختیار انسان قرار می گیرند، هیچ مزیتی نسبت به گزینه‌های قبلی نداشته باشند، مسلماً در انجام عمل، از آنها استفاده نخواهد شد؛ یعنی اینکه استفاده از آنها وقتی صورت می گیرد که در رسیدن به هدف کمک مؤثری داشته باشند. برای اثبات مطلب فوق، به رابطهٔ بروز رسانی مقادیر Q توجه می کنیم:

$$Q(s,a)=(1-\beta)Q(s,a)+\beta(r+\gamma V(t)) \qquad \text{رابطه ۳-۱}$$

این رابطه بازای انجام عمل a در حالت s و رفتن به حالت t و دریافت سیگنال تقویتی r، روی Q(s, a) اعمال می‌شود. در این رابطه، β نرخ یادگیری و γ پارامتر کاهش اثر مطلوبیت حالات آتی بوده و V(t) از رابطهٔ زیر بدست می‌آید:

$$V(t)=\max_b Q(t,b) \qquad \text{رابطه ۳-۲}$$

عملگر max در رابطهٔ اخیر، سبب می‌شود تا ارزش بهترین عمل در حالت t به حالت s تسرّی پیدا کند.در نتیجه اگر در حالتی، عمل اضافی مورد بحث، به عنوان بهترین عمل انتخاب شود، نسبت به عاملی که این عمل اضافی را ندارد، مقدار بزرگتری را برای انتشار در جدول Q ایجاد می کند. این مقدار بزرگتر می‌تواند در جاهای مختلفی از جدول Q تأثیر گذاشته و ارزش آنها را بالاتر ببرد. این، دقیقاً همان مشکلی است که عمل اضافی در روند همکاری در یادگیری طراحی شده برای عاملهای همسان ایجاد می کند: دستِ بالا گرفتن ارزش بعضی عمل–حالت ها.

در روشهای معمول همکاری در یادگیری مانند WSS، فرض می‌شود که جدول بهینهٔ دانش عاملها، یکسان است. به این ترتیب، ترکیب دانش عاملها باعث دست بالا گرفتن ارزش جایی نخواهد شد. اما با حضور عمل اضافی، احتمال غیریکسان شدن دانشهای بهینه و در نتیجه، مخدوش شدن دانش عاملها در اثر همکاری در یادگیری وجود دارد. بعنوان مثال، دانش بهینهٔ دو عامل غیرهمسان از نوع دوم که در یک صفحهٔ مشبکِ با مانع، برای رسیدن به هدفی واحد آموزش دیده‌اند، در جدول ۳-۱ دیده می‌شود. مقدار موجود در هر خانه، ارزش بهترین عمل را در آن خانه نشان می‌دهد.

فرض شده که γ =۰/۵ و پاداش رسیدن به هدف برابر ۱۰۲۴ و در غیر این صورت، صفر باشد. اگر دانش عامل دوم در اختیار عامل فاقد عمل اضافی قرار گیرد، عامل از خانهٔ مشخص شده با دایره، با اجرای سیاست حریصانه، نمی‌تواند خارج شود؛ زیرا، ارزش این خانه بدلیل یک انتقال غیرممکن برای این عامل، بالا رفته است[۱۰]. می‌توان ایده‌هایی را که در این مقاله ارائه می‌شود، برگرفته از عبارات زیر دانست:

«اگر عمل اضافی در مسیر بهینه تا هدف، مورد استفاده قرار نگرفته باشد، دانش عامل در طی این مسیر، متأثر از وجود عمل اضافی نبوده و می‌توان آن را براحتی در اختیار عامل فاقد آن، قرار داد. در غیر این صورت، دانش عامل در طی مسیر تحت تأثیر عمل اضافی بوده و ممکن است موجب

---

[۹] Action Map Table (AMT)

[۱۰] نمونه‌ای از وابستگی دانش به ابزار کسب آن

سردرگمی عامل دیگر شود. بنابراین در چنین مواردی باید اثر این عمل را حذف یا جبران کرد.»

جدول ۳-۱ : نمونه‌ای از مخدوش شدن عملکرد عاملهای غیرهمسان در اثر همکاری در یادگیری

$A_1 = \{\leftarrow, \rightarrow, \uparrow, \downarrow\}$

۱۶		۱۰۲۴	هدف
۳۲	۶۴		۱۰۲۴
۶۴	۱۲۸	۲۵۶	۵۱۲

$A_2 = A_1 \cup \{\nearrow\}$

۱۲۸		۱۰۲۴	هدف
۲۵۶	(۵۱۲)		۱۰۲۴
۲۵۶	۲۵۶	۵۱۲	۵۱۲

۳-۱. اشتراک رویداد میان عاملهای غیرهمسان

هر رویداد تشکیل شده از رشته‌ای از سه‌تایی‌های (سیگنال تقویتی، عمل، حالت). این رشته از سه‌تایی‌ها با رسیدن به هدف و یا طی شدن تعداد مشخصی از گامهای جستجو در فضای مسأله خاتمه می‌یابد. عامل خبره (که دارای عمل اضافی است) در هر دور[11] از عملکرد خود، رویدادی را تولید می‌کند. این رویداد به عامل غیرخبره داده می‌شود. عامل غیرخبره، این رویداد را از انتها به ابتدا (بصورت معکوس زمانی)، در ذهن خود تکرار می‌کند[12]؛ گویی که خود او این رویداد را تجربه کرده است. فرض کنیم که رویداد مورد نظر بصورت زیر باشد:

$$(s_1, a_1, r_1), \ldots, (s_i, a_i, r_i), \ldots, (s_n, a_n, r_n) \quad \text{رابطه ۳-۳}$$

در نتیجه، برای بروز رسانی ارزش در حالت i خواهیم داشت:

$$Q'(s_i, a_i') = (1-\beta)Q'(s_i, a_i') + \beta(r_i + \gamma V'(s_{i+1})) \quad \text{رابطه ۳-۴}$$

در نوشتن این روابط، از متغیرهای بدون اندیس برای عامل خبره استفاده شده است. اگر a_i یکی از اعمال مشترک باشد، a_i' عمل معادل در عامل غیرخبره برای آن است (که با توجه به AMT بدست می آید). اگر a_i عمل اضافی عامل خبره باشد، بررسی دو حالت لازم است: یا اینکه در حالت s_i ، هیچ یک از معادلهای a_i که در HAMT موجودند، قابل اعمال نیستند و یا اینکه لااقل یکی از آنها قابل اجراست. برای مثال در شکل ۳-۱، قسمتی از یک صفحهٔ مشبکِ با مانع دیده می‌شود. در شکل سمت راست، عمل$\nearrow$ در حالت ۳ هیچ معادل قابل اجرایی ندارد. اما در شکل سمت چپ، می توان با اجرای $\longrightarrow$ و سپس $\uparrow$ به نتیجهٔ مشابه رسید.

در حالتی که هیچ یک از معادلها ممکن نباشند، از سه‌تایی مربوطه صرفنظر کرده و به سه‌تایی بعدی پرداخته می‌شود. در حالتی که لااقل یکی از معادلها قابل اجرا باشند، رابطهٔ زیر برای بروز رسانی ارزش مربوطه بکار برده می‌شود:

$$Q'(s, a_i'') = (1-\beta)Q'(s_i, a_i'') + \beta(p + \gamma[r_i + \gamma V'(s_{i+1})]) \quad \text{رابطه ۳-۵}$$

در این رابطه a_i'' ، اولین عمل از عمل ترکیبی معادل ممکن و p کوچکترین سیگنال تقویتی است.

$A_1 = \{\leftarrow, \rightarrow, \uparrow, \downarrow\}$

	۱
۳	

$A_2 = A_1 \cup \{\nearrow\}$

	۱
(۳)	۲

شکل ۳-۱ : راست- معتبر نبودن معادلها، چپ- معتبر بودن یک معادل

صحت این رابطه را ضمن یک مثال بررسی می‌کنیم. در شکل ۳-۲ عامل خبره برای رفتن از حالت ۳ به حالت ۱، عمل را بکار برده است. از آنجایی که اجرای این عمل برای غیرخبره ممکن نیست، عامل غیرخبره می‌تواند با در نظر گرفتن معادل ترکیبی، $\not\nearrow$ قبل خبره را دنبال کند. با توجه به اینکه در مسألهٔ مورد نظر تنها یک پاداش و آنهم به ازای رسیدن به هدف به عامل تعلق می‌گیرد، می توان مطمئن بود که رفتن به حالت ۲، با پاداش روبرو نخواهد بود[13]. اما عامل غیرخبره از حالت ۲ بی‌اطلاع است (زیرا عمل $\longrightarrow$ را واقعاً اجرا نمی کند). در نتیجه نمی‌تواند از سیگنال تقویتی مربوط به آن هم مطلع باشد. راه حلی که در اینجا پیشنهاد می‌شود، بکار گرفتن کوچکترین سیگنال تقویتی ممکن برای حالت ۲ می‌باشد. ارزش حالت ۲ را نیز می‌توان با $r_3 + \gamma V'(1)$ جایگزین کرد (توجه شود که در مسأله مورد نظر ما، r تابعی است از حالت جدید عامل).

---

[11] Trial

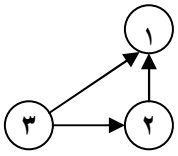
[12] Mental replay

[13] اگر رویداد تولید شده، بهینه نباشد، ممکن است که حالت ۲، همان هدف مسأله باشد. در این صورت باز هم تنبیه شدن عامل باعث دست بالا گرفتن ارزش حالت ۳ نخواهد شد.

در نتیجه رابطهٔ زیر برای بروز رسانی ارزش عمل $\longrightarrow$ در حالت ۳ مورد استفاده قرار می‌گیرد:

$$Q'(3,\rightarrow) = (1-\beta)Q'(3,\rightarrow) + \beta\left(p + \gamma[r_3 + \gamma V'(1)]\right) \quad \text{رابطه ۳- ۶}$$

تعمیم رابطهٔ ۳-۵ به حالتی که معادل عمل اضافی، از ترکیب بیش از دو عمل حاصل شود، براحتی قابل انجام است .



شکل ۳-۲ : انتقال ارزش به کمک معادل قابل اجرا

۳-۲. اشتراک دانش

روش اشتراک رویداد در جاهایی که رویدادهای غیر بهینه تولید شده باشند، قابل استفاده بوده و کارایی خود را نشان داده است. اما گاهی اوقات ترجیح می‌دهیم تا به جای آنکه مجدداً چرخ را اختراع کنیم، از آنچه وجود دارد، استفاده کنیم. دانش ایجاد شده در طی مدت جستجوی یک عامل در فضای مسأله، بسیار ارزشمندتر از تجارب مقطعی اوست. همانطور که قبلاً بیان شد، این دانش متأثر از ابزار کسب آن بوده و در نتیجه احتمالاً تحت تأثیر حضور اعمال اضافی، تغییراتی داشته است. اگر بدون انجام ملاحظاتی، روشی مثل WSS برای عاملهای غیرهمسان بکار برده شود، ممکن است که دانش عاملها از اعتبار خارج شود[۱۱].

اطلاع مورد استفاده در کنار جدول Q، در این تحقیق، مسیر حرکت عامل خبره در فضای مسأله است. این مسیر را می‌توان مستقلاً از روی مشاهدهٔ خبره و یا از روی رویداد مربوطه استخراج کرد. به عبارت دیگر جدول Q روی مسیر حرکت عامل خبره، برای انتقال دانش به عامل غیرخبره مورد نیاز است. فرض کنیم که رویداد مورد نظر بصورت رابطه ۳-۳ باشد. اگر این رویداد بصورت بهینه تولید شده باشد[۱۴]، بهترین جواب از این روش حاصل خواهد شد. در صورتی که رویداد، ناشی از حرکت غیربهینه در محیط باشد، از انتقال دانش معتبر (فاقد دست بالا گرفتن بعضی ارزشها) نمی‌توان مطمئن بود. این مطلب در آزمایشهای انجام شده نیز مشاهده شده است.

روند انتقال دانش از عامل خبرهٔ تولید کنندهٔ رویداد به عامل غیرخبره و فاقد عمل اضافی در جدول ۳-۲ ارائه شده است. این روند، حالت کلی انتقال دانش را مورد توجه قرار داده و حتی نحوهٔ انتقال دانش در مورد رویدادهای غیربهینه نیز در آن لحاظ شده است. البته همانطور که بیان شد، نمی‌توان اطمینان داشت که دانشی که از رویداد غیربهینه منتقل می‌شود، بدون عیب باشد. روال کاهش ارزش در جدول ۳-۲ مشاهده می شود.

بیان این نکته ضروری است که، انتقال دانش همواره باید روی مسیر بهینه (با توجه به دانش عامل[۱۵]) صورت گیرد. انتقال دانش روی مسیر غیربهینه می‌تواند در بعضی شرایط موجب ایجاد دانش معیوب گردد. در واقع باید گفت که تضمینی برای ایجاد دانش غیرمعیوب در این حالت نیست؛ زیرا اعتبار روش فوق الذکر به درست شمردن اعمال اضافی مؤثر در دانش انتقالی و حذف تأثیرات آن است. حال اگر مسیر طی شده، غیربهینه باشد، دیگر نمی‌توان به مقدار شمارنده استناد کرد (ممکن است مقدار حقیقی شمارنده بیش از مقدار بدست آمده بوده و در نتیجه دانش دست بالا گرفته شده‌ای منتقل شود). نکتهٔ دیگری که در روال انتقال دانش از آن استفاده شده، مقایسهٔ دانش جدید با دانش قبلی است. در شبیه‌سازیهای انجام شده، ارزش تمام حالات با کوچکترین مقدار ممکن آنها آغازین دهی شده است. در نتیجه، روند تغییرات ارزش هر عمل-حالت همواره صعودی خواهد بود. اگر دانش انتقالی مقداری کمتر از دانش موجود در ذهن عامل داشته باشد (به علت کمی دانش عامل پیشنهاد دهنده)، با مقایسه آنها از انتقال دانش کم ارزش جلوگیری شده و روند صعودی آن دچار خلل نمی‌شود.

٤. آزمایشها

آزمایشات بسیاری برای اطمینان از صحت و چگونگی عملکرد روشهای بیان شده در شرایط گوناگون، انجام شد. آنچه که در ادامه خواهد آمد، تنها به عنوان نمونه‌ای برای نشان دادن کارایی این روشها است.

---

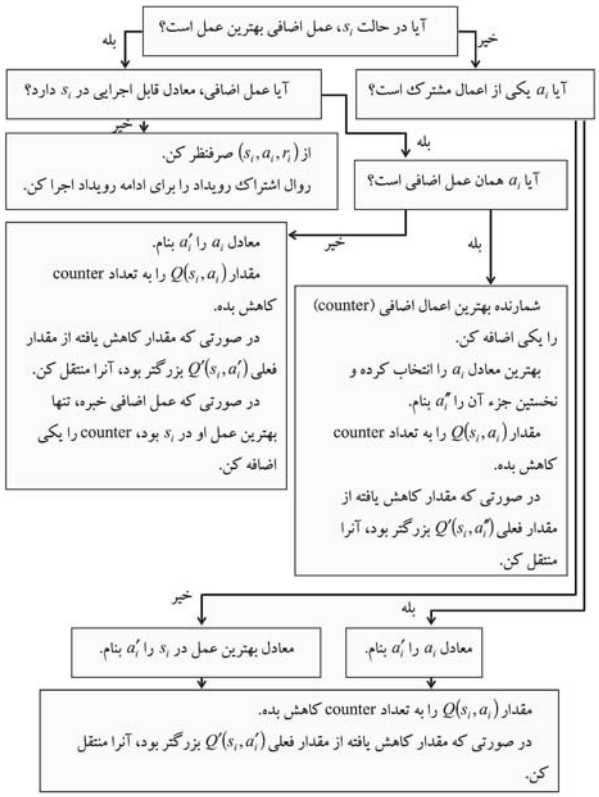
[۱۴] این رویداد می‌بایست به هدف ختم شده باشد، در غیر این صورت تنها می‌توان از روش قبلی (اشتراک رویداد) استفاده کرد.

[۱۵] حتی اگر خبره نباشد.

جدول ٤-١ : روال کاهش ارزش v به اندازه counter مرتبه

```
double Discount(double v, int counter)
{   for (int k=۱; k < counter+۱; k++)
      v = γ*v + p
    return v }
```

جدول ٤-٢ : روند انتقال دانش از عامل خبرهٔ دارای عمل اضافی به عامل فاقد آن



در این آزمایشات از دو عامل غیرهمسان با مجموعهٔ اعمال موجود در جدول ۴-۳ استفاده شده است. این عاملها در صفحهٔ مشبک دارای مانع[16] قرار گرفته و با روش Q-learning به یادگیری پرداخته‌اند. هر دو عامل از روش انتخاب عمل ε-greedy استفاده کرده‌اند. برای γ مقدار ۰/۹ در نظر گرفته شده است. در صورت دستیابی هر عامل به هدف (که با * در تصاویر مربوطه مشخص شده)، محیط ۱۰۰ امتیاز به عنوان پاداش به عامل داده و عامل با قرارگیری تصادفی در محلی دیگر، دور تازه‌ای را آغاز می‌کند[17]. بازای برخورد به موانع، عامل با امتیاز ۱۰- تنبیه شده و در جای قبلی خود می‌ماند. رفتن به هر خانهٔ دیگر، با ۱- امتیاز تنبیهی از طرف محیط پاسخ داده خواهد شد.

جدول ٤-٣ : مجموعه اعمال عاملهای مورد آزمایش

شماره اندیس عمل	۱	۲	۳	۴	۵
مقصود عمل در عامل اول	←	→	↓	↑	↗
مقصود عمل در عامل دوم	→	←	↑	↓	

٤-١. کارایی روشهای اشتراک رویداد و انتقال دانش

برای نشان دادن کارایی روشهای مطرح شده، عاملها را در صفحهٔ مشبک شکل ۴-۱ و **شکل ٤-٢**، قرارداده و نتیجه یادگیری فردی عامل دوم (بدون عمل اضافی) را با نتایج روشهای فوق که در آنها از عامل اول (که دارای عمل اضافی است) کمک گرفته شده، مقایسه می‌کنیم (هر نمودار میانگین ۱۰

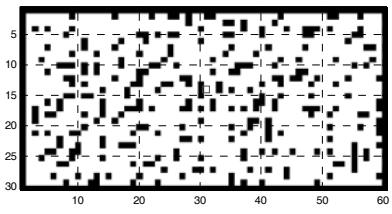
---

[16] Maze

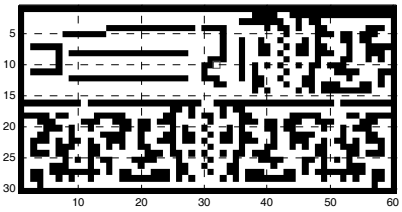
[17] در صورتی که عامل طی ۱۰۰۰ گام جستجو در محیط، به هدف دست نیابد، آن دور خاتمه یافته و عامل از خانه‌ای دیگر، دور تازه‌ای را آغاز می‌کند.

بار اجراست). شکل ۴-۱ با توجه به مجموعه اعمالی که عاملها در اختیار دارند، بعنوان صفحه ای که در آن معادلهای عمل اضافی، همواره قابل اجرا هستند، انتخاب شده است. **شکل ٤-۲**، بعنوان مسأله‌ای که در آن موانع محدود کننده (برای معادلهای عمل اضافی) وجود داشته و از پیچیدگی بالایی برخوردار است، انتخاب شده است. (آزمایشهای مشابهی بر روی صفحات مشبک دیگر انجام شده که شرح کامل آنها در [۷] آمده است.)

در این آزمایشات، عامل اول از دانش بهینه برخوردار بوده است. برای به اشتراک گذاشتن رویداد، عامل اول با ε=۰/۲ در صفحه حرکت کرده و رویدادهای مشاهده شده را به عامل دوم منتقل می‌کند. نرخ یادگیری عامل دوم ۰/۷ می‌باشد. برای انتقال دانش، عامل اول با حرکت حریصانه (ε=۰) روی صفحه، مسیرهایی را طی کرده و به اطلاع عامل دوم می‌رساند. برای یادگیری فردی، عامل دوم از نرخ یادگیری ۰/۷ و ε=۰/۲ استفاده کرده است (جدول ۴-۴). عامل دوم از HAMT صحیح مطلع است. از آنجا که در روشهای مورد بررسی، بروز رسانی به ترتیب عکس زمانی (از انتها به ابتدای مسیر) انجام می‌شود، نوع وارونهٔ یادگیری فردی[۱۸] که در آن عامل دوم پس از اتمام هر دور یادگیری، بروز رسانی ارزشهای مسیر مربوطه را انجام می‌دهد، نیز مورد مقایسه قرار می‌گیرد.



شکل ٤-۱ : صفحهٔ مشبک ٦٠ در ۳۰ ، بدون مانع بد، هدف * (Maze1)



شکل ٤-۲ : صفحهٔ مشبک ٦٠ در ۳۰ ، دارای مانع بد و پیچیدگی زیاد، هدف * (Maze 2)

نقاطی که در آن ارزیابی‌های مربوط به همکاریها انجام شده، در دورهای ۱۰، ۵۰، ۱۰۰، ۲۰۰، ۳۰۰، ۴۰۰، ۵۰۰، ۶۰۰، ۱۰۰۰، ۲۰۰۰، ۷۰۰۰ و ۱۲۰۰۰ بوده است. ارزیابی دانش ناشی از یادگیریهای فردی در دورهای ۱۰، ۵۰، ۱۰۰، ۲۰۰، ۳۰۰، ۴۰۰، ۵۰۰، ۱۰۰۰، ۲۰۰۰، ۴۰۰۰، ۸۰۰۰ و ۱۲۰۰۰ انجام شده است. معیار اولی که برای مقایسهٔ دانشها اطلاعات مفیدی را دربردارد، فاصلهٔ دانش عامل از مقدار بهینهٔ آن بر حسب تعداد گامهای طی شده می‌باشد (رابطهٔ ۴-۸).

از شکل ۴-۳ (که برای تمایز بهتر درمنحنی‌ها، از مقیاسهای لگاریتمی برای هر دو محور آن استفاده شده) مطلب زیر استخراج می‌شود: روش اشتراک دانش به علت تأثیر گذاری روی مقدار ارزش بهترین عمل، و در نظر نگرفتن سایر اعمال، نمی‌تواند به دانش بهینه برسد. از آنجا که روش اشتراک رویداد محدود به حرکت بهینه در صفحه نیست، می‌تواند دانش مربوط به سایر اعمال (غیر از بهترین) را ارتقاء دهد.

$$\sum_{s}\left|Q'_{opt}(s,a)-Q'(s,a)\right| \qquad \text{رابطه ٤- ۷}$$

جدول ٤-٤ : پارامترهای مربوط به سنجش کارایی روشهای انتقال دانش

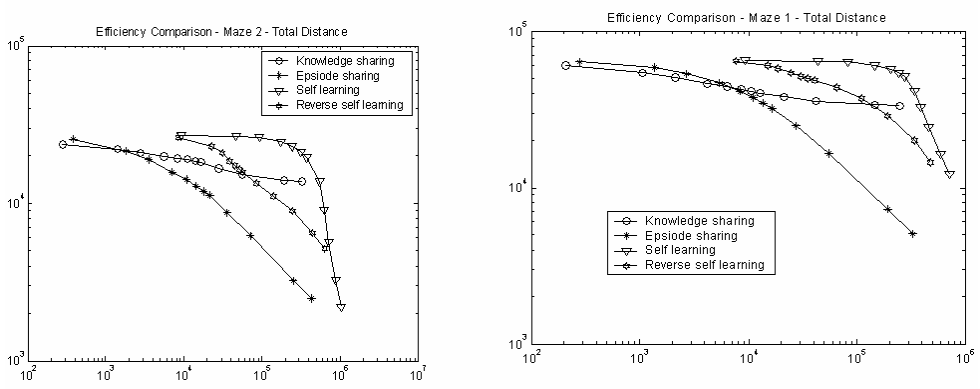
Reverse Self Learning	Self Learning	Episode Sharing	Knowledge Sharing
$\alpha_2 = 0.7$	$\alpha_2 = 0.7$	$\alpha_2 = 0.7$	$\alpha_2 = 0.7$ (فقط در قسمت ES)
$\varepsilon_2 = 0.2$	$\varepsilon_2 = 0.2$	$\varepsilon_1 = 0.2$	$\varepsilon_1 = 0.0$

شکل ۴-۴ نشان دهندهٔ نتیجهٔ ارزیابی دانشها با استفاده از معیار رفتاری، بر حسب گامهای سپری شده در یادگیری است. برای محاسبهٔ این معیار، حداقل فاصلهٔ هر خانه تا هدف بدست آمده و میانگین آنها مورد توجه قرار می‌گیرد. آنچه در شکل نشان داده شده، مقدار نرمال این میانگین با توجه به مقدار بهینهٔ آن می‌باشد (محور افقی بصورت لگاریتمی نشان داده شده است). اگر نتوان از خانه‌ای به هدف رسید، فاصلهٔ آن، ۱۰۰۰ گام در نظر گرفته می‌شود.

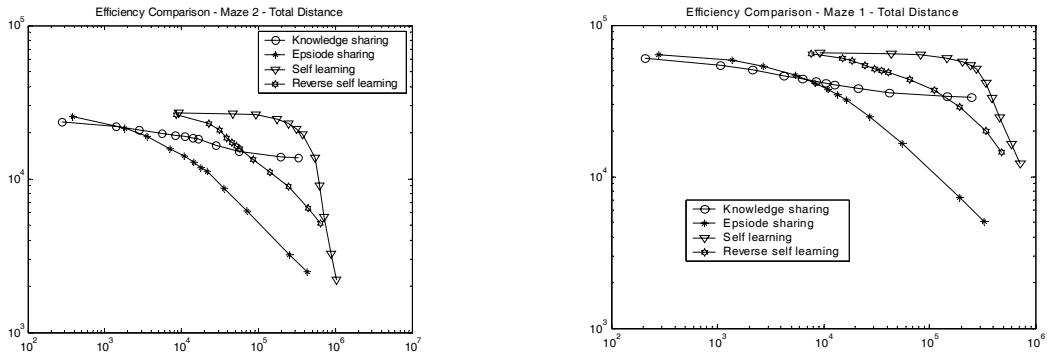
از این شکل چند مطلب استخراج می‌شود: روشهای یادگیری فردی به مقدار بهینهٔ خود خواهند رسید؛ هر چند که این کار نیاز به گذشت زمان طولانی‌تری دارد. بروز رسانی وارونهٔ ارزش خانه‌ها در مسیر، باعث تسریع در رسیدن به رفتار بهینه می‌شود. روشهای اشتراک دانش و اشتراک رویداد توانسته‌اند در ابتدای کار از یادگیری فردی پیشی بگیرند. به علت یک بودن تلویحی نرخ یادگیری در روش اشتراک دانش، این روش از اشتراک رویداد، سریعتر همگرا می‌شود. تفاوت مقدار نهایی روش اشتراک دانش و اشتراک رویداد، عمدتاً بدلیل تفاوت تعداد خانه‌هایی است که از آنها می‌توان به هدف رسید (توجه کنید که هر خانه‌ای که از آن نتوان به هدف رسید، معادل ۱۰۰۰ گام خواهد بود). حداکثر دانشی که از روشهای اشتراک دانش و اشتراک رویداد حاصل می‌شود، لزوماً به اندازه بهینهٔ خود نیست. علت این امر حضور موانع و تاثیر عمل اضافی در مسیرهایی است که به عنوان مسیر بهینه مشخص می‌شوند. در نتیجه، از بعضی از خانه‌ها، حتی با حذف کردن اثر عمل اضافی، نمی‌توان بهترین مسیری را که در صورت نبود این عمل

[۱۸] Reverse single learning

بدست می‌آید را بدست آورد.



شکل ٤-٣ : مقایسه فاصلهٔ ارزش بهترین عمل در هر حالت از مقدار بهینهٔ آن بر حسب گامهای یادگیری



شکل ٤-٤ : مقایسه فاصلهٔ دانش از مقدار بهینهٔ آن بر حسب گامهای یادگیری

٤-٢. نتایج دیگر

یکی از نکاتی که در روش اشتراک دانش بسیار حائز اهمیت است، دریافت مسیر بهینهٔ ناشی از دانش فعلی عامل کمک کننده است. همانطور که بیان شد با دور شدن این عامل از حرکت بهینه، احتمال ورود دانش معیوب به ذهن عامل دریافت کنندهٔ دانش بوجود می‌آید. صحت این مطلب در آزمایشهای انجام شده مشاهده شد. در این آزمایشها با مقایسهٔ نتیجهٔ حاصل از اشتراک دانش بازای ε های مختلف، مشاهده شد که دانشی که برای بهترین عمل منتقل می‌شود، در اثر افزایشε ، دچار عیب می‌شود.همانطور که انتظار می‌رود با افزایش میزان جستجو در صفحه، دانش کلی به نحو مطلوبتری کامل می‌شود، اما همین دانش نیز در اثر ادامهٔ این روند، می‌تواند دچار عیب شود.

انتظار این است که با بیشتر شدن مقدار جستجو در محیط، روش اشتراک رویداد بتواند دانش بیشتری را منتقل کند. سرعت انتقال دانش در اثر افزایش ε، بیشتر می‌شود. البته واضح است که انتقال دانش روی بهترین عمل ضعیفتر می‌شود. در ضمن می‌توان دید که با افزایش ε، هر دور از گامهای بیشتری تشکیل می‌شود. مشاهده شد که معیار رفتاری دانش به تغییرات ε، حساسیت چندانی ندارد.

٥. جمع بندی و نتیجه گیری

نشان داده شد که بکارگیری روشهای موجود که برای انتقال دانش میان عاملهای همسان طراحی شده، در اینجا خیلی کارامد نبوده و چه بسا مشکل آفرین باشد. بنابرین با توجه به دانستن نگاشت بین اعمال عاملهای غیرهمسان، سعی کردیم تا روشهایی را برای انتقال دانش بین عاملهای غیرهمسان (بویژه مسأله اصلی در نوع دوم غیرهمسانی خود نمایی می‌کند) بدست آوریم. نتیجهٔ کار، دو روش اشتراک رویداد و اشتراک دانش بود.

دو روش ارائه شده در این مقاله، زمینه های کاربرد مخصوص به خود را دارند. در هر دو روش، انتقال دانش از عاملی خبره که دارای عمل اضافی است به عامل غیرخبره و فاقد آن مورد توجه بوده است. هنگامی که عامل فرستنده اطلاعات، بصورت حریصانه، فضای مسأله را طی می‌کند، استفاده از روش اشتراک دانش که وابسته به انتقال بهینه دانش عامل است، بسیار کاراست. در این روش با استفاده از دانش عامل خبره (دارای عمل اضافی) بعلاوه مسیری که در هر دور طی می‌کند، دانش عامل غیرخبره بروز می‌شود. در این روش، مسیری که عامل خبره بصورت حریصانه طی کرده از انتها به ابتدا مورد توجه قرار می‌گیرد. با شمارش تعداد اثرات عمل اضافی در هر نقطه از مسیر، دانش مربوطه برای استفاده عامل دیگر تطبیق داده می‌شود. از آنجا که

تعداد صحیح این اثرات در عمل تطبیق بسیار مؤثر است، لذا می‌بایست این مسیر، بهترین مسیری باشد که عامل دارای عمل اضافی با توجه به دانش خود آنرا انتخاب کرده باشد. در غیر این صورت، دانش منتقل شده ممکن است معیوب باشد. با استفاده از این روش، حداکثر دانش ممکن بین عاملهای غیرهمسان انتقال داده می‌شود. در صورتی که عامل خبره، عامل فاقد دارای عمل اضافی باشد، نمی‌تواند در مورد این عمل، دانشی را منتقل کند؛ مگر آنکه خود عامل دارای این عمل با انجام یک بازنگری در دانش خود، برای این عمل دانشی را ایجاد کند (این کار را می‌توان یکی از گامهای آتی در این تحقیق دانست).

اگر عامل خبره (دارای عمل اضافی)، فضای مسأله را بصورتی غیر بهینه جستجو کند، بعلت احتمال انتقال دانش معیوب، از روش فوق نمی‌توان استفاده کرد. در این شرایط، روش اشتراک رویداد که برای عاملهای غیرهمسان توسعه داده شده، نتایج خوبی را نشان داده است. در این روش، عامل غیرخبره با تکرار ذهنی رویداد تولید شده توسط عامل خبره بصورت از انتها به ابتدا، دانش خود را بروز می‌کند. این روش با توجه به نرخ یادگیری لحاظ شده در بروزرسانی دانش، از روش قبل کندتر عمل می‌کند. همچنین هر چه میزان جستجو در مسأله بیشتر باشد (دمای بالا در انتخاب عمل بولتزمن و ε نزدیک به یک در روش ε-greedy)، دانشی که منتقل می‌شود، گستردگی بیشتری روی فضای عمل- حالت دارد. این ویژگی در روش قبل که تنها منحصر به انتقال دانش مربوط به بهترین اعمال بود، وجود ندارد.

هر دو روش به نوع مسأله‌ای که در آن بکار گرفته می‌شوند، بسیار وابسته‌اند. اگر مسأله بگونه‌ای باشد که در جاهایی از آن عمل اضافی دارای معادل نباشد، انتقال دانش بخصوص از روش اشتراک دانش کم بهره خواهد بود. البته انتظار دیگری هم نمی‌توان داشت. اگر شخصی در مورد توانایی‌های منحصر به فرد خود، دانشی کسب کرده باشد، نمی‌توان انتظار داشت این دانش در شخصی که فاقد آن توانایی‌هاست، بکار آید. هر چه تأثیرات عمل اضافی در دانش تولید شده بیشتر باشد، دانشی که می‌تواند به اشتراک گذاشته شود، کمتر خواهد بود.

بهر حال توانستیم با استفاده از دانش عامل خبره و مسیر بهینه و یا با استفاده از رویداد تجربه شده توسط عامل خبره، دانشی را به عامل غیرخبره منتقل کنیم. این انتقال دانش از جهاتی بر روشهای یادگیری فردی (عادی و وارونه) برتری خود را نشان داد. به نظر می‌رسد که انتقال و تطبیق دانشهای غیرهمسان بدون استفاده از اطلاعی دیگر (مثل مسیر حرکت) کار ممکنی نباشد. اما این مسأله بعنوان افقی از این کار می‌تواند موضوع تحقیقات دیگری باشد (هر چند که جستجو برای یافتن روشی که تنها از روی خود دانش بتواند، دانش منطبق با عامل غیرهمسان را مشخص کند، در این تحقیق به ثمر نرسید).

٦. مراجع

[١] Ming Tan, Multi-agent reinforcement learning: Independent vs. cooperative agents, in *Proceedings of the Tenth International Conference on Machine Learning*, ١٩٩٣.

[٢] Kawaishi, I., et al, Experimental Comparison of a Heterogeneous Learning Multi-Agent System with a Homogeneous one, *IEEE International Conference on Systems, Man and Cybernetics*, ١٩٩٦, pp. ٦١٣-٦١٨.

[٣] Bob Price & Craig Boutilier, Imitation and Reinforcement Learning in Agents with Heterogeneous Actions, *Proceedings of the Seventeenth International Conference on Machine Learning*, ٢٠٠٠.

[٤] Luís Nunes & Eugénio Oliveira, Advice Exchange in Heterogeneous Groups of Learning Agents, *Tech. report ١-٠٢ FEUP/ISCTE/LIACC-NIAD&R*, ٢٠٠٢.

[٥] Nehaniv, C. L. & Dautenhahn, K., The Correspondence Problem, in Nehaniv, C. L. & Dautenhahn, K., editors, *'Imitation in Animals and Artifacts'*, MIT Press, ٢٠٠٢.

[٦] Alissandrakis, A., Nehaniv, C. L. & Dautenhahn, K. , Solving the Correspondence Problem Between Dissimilarly Embodied Robotic Arms Using the ALICE Imitation Mechanism, *Proceedings of the Second International Symposium on Imitation in Animals & Artifacts*, The Society for the Study of Artificial Intelligence and Simulation of Behavior, ٢٠٠٣, pp. ٧٩-٩٢.

[٧] سید محمد رضا میرفتاح، همکاری در یادگیری میان عاملهای غیرهمسان، پایان نامه کارشناسی ارشد، دانشگاه تهران، ١٣٨٢.

[٨] Majid Nili Ahmadabadi & Masoud Asadpour, Expertness Based Cooperative Q-learning, *IEEE Transactions on Systems, Man and Cybernetics*, Part B: Cybernetics, Vol. ٣٢, No. ١, ٢٠٠٢, pp. ٦٦-٧٦.

[٩] S.M.Reza Mirfattah and Majid Nili Ahmadabadi, Cooperative Q-learning with heterogeneity in actions, *Proceedings of ٢٠٠٢ IEEE International Conference on Systems, Man and Cybernetics*, Hammamet, Tunisia, ٢٠٠٢.

[١٠] سید محمد رضا میرفتاح، مجید نیلی احمدآبادی و بابک نجار اعرابی، معادل یابی اعمال در عاملهای غیر همسان: گامی‌به سوی همکاری در یادگیری، مجموعه مقالات هشتمین کنفرانس بین المللی سالانه انجمن کامپیوتر ایران، ١٣٨١، صفحه ١٣٤-١٤١.

[١١] Mastour Eshgh, S., Araabi, B., Nili Ahmadabadi, M., Cooperative Q-learning through State Transitions: A Method for Cooperation Based on Area of Expertise, *Proceedings of the ٤th Asia-Pacific Conference on Simulated Evolution and Learning (SEAL'٠٢)*, Singapore, ٢٠٠٢.