بهبود سرعت آموزش در مسائل یادگیری تقویتی مبتنی بر انتقال دانش عصبی‌فازی

فاطمه سعادت‌جو[۱]، استادیار؛ عرفان قندهاری[۲]، کارشناس ارشد

۱- دانشکده مهندسی کامپیوتر- دانشگاه علم و هنر- یزد- ایران- saadatjou@sau.ac.ir
۲- دانشکده مهندسی کامپیوتر- دانشگاه علم و هنر- یزد- ایران- erfan.ghandehari@sau.ac.ir

چکیده: این مقاله به موضوع انتقال یادگیری در محیط‌هایی که بعضی از ویژگی‌های آن مشترک است می‌پردازد. چالش اصلی در این مبحث، نحوه انتقال دانش به‌دست‌آمده از محیط مبدأ به محیط مقصد است. در ایده ارائه‌شده با در نظرگرفتن ویژگی‌های مشترک در فضای عامل بین دو محیط، ابتدا مقدار ارزش - عمل در محیط مبدأ به‌دست می‌آید، سپس از یک شبکه عصبی- فازی برای تقریب مقدار تابع ارزش - عمل بهره برده می‌شود. در محیط مقصد، مقدار ارزش - عمل از ترکیب مقدار پیش‌بینی شبکه عصبی - فازی و مقدار به‌دست‌آمده در خود آن محیط استفاده می‌شود. به‌عبارت دیگر با توجه به آموزش انجام‌شده در محیط مبدأ، مقادیر ارزش - عمل در محیط مقصد از ترکیب مقادیر ارزش - عمل تقریب‌زده‌شده توسط شبکه عصبی - فازی و مقدار به‌دست آمده از الگوریتم یادگیری در آن محیط به‌دست می‌آید. شایان ذکر است که از الگوریتم یادگیری Q در محیط استفاده‌شده است. نتایج حاصل از ایده ارائه‌شده، حاکی از افزایش چشمگیر سرعت یادگیری می‌باشد.

واژه‌های کلیدی: یادگیری تقویتی، انتقال دانش، ویژگی مشترک، شبکه عصبی- فازی.

Improving the learning speed in reinforcement learning issues based on the transfer learning of neuro-fuzzy knowledge

F. Saadatjoo[1], Assistant Professor; E. Ghandehari[2], MSc

1- Computer Engineering Department, Science and Arts University, Yazd, Iran, Email: saadatjou@sau.ac.ir
2- Computer Engineering Department, Science and Arts University, Yazd, Iran, Email: erfan.ghandehari@sau.ac.ir

Abstract: This paper to the topic of transfer learning in environments that share some of its features. The main challenge in this topic is how to transfer knowledge from the source environment to the target environment. In the presented idea, taking into account the common features in the operating space between the two environments, the value of the operation in the source environment first is obtained and then it uses a neuro -fuzzy network to approximate the value of the value function of the operation. In the target environment, the value of the mode of operation is used to combine the predictive value of the neuro - fuzzy network and the amount received in the environment itself. In other words, according to the training carried out in the source environment, value-action values in the target environment are derived from the combination of value-action values approximated by the neuro - fuzzy network and the amount obtained from the learning algorithm in that environment. It is worth noting that the learning algorithm Q is used in the environment. The results of the proposed idea indicate a significant increase in learning speed.

Keywords: Reinforcement learning, transfer knowledge, common features, neuro - fuzzy network.

تاریخ ارسال مقاله: ۱۳۹۶/۰۷/۲۵
تاریخ اصلاح مقاله: ۱۳۹۷/۰۶/۰۵
تاریخ پذیرش مقاله: ۱۳۹۷/۰۸/۱۶
نام نویسنده مسئول: فاطمه سعادت‌جو
نشانی نویسنده مسئول: ایران- یزد- دانشگاه علم و هنر، دانشکده فنی و مهندسی، گروه مهندسی کامپیوتر.

۱ -مقدمه

یادگیری تقویتی[۱]، شاخه‌ای از یادگیری ماشین[۲] است، که هدف آن بیشینه‌کردن میزان پاداش عامل[۳] در محیط می‌باشد [۱]. به‌عبارت دیگر یادگیری تقویتی راهی است، برای آموزش عامل جهت انجام یک عمل از طریق دادن پاداش و جریمه، بدون آن‌که لازم باشد نحوه انجام عمل را برای عامل مشخص نمود. در هر مرحله از یادگیری، عامل عملی را در محیط انجام می‌دهد و وضعیت او در محیط تغییر می‌کند. عامل در ازای این عمل، پاداشی دریافت مینماید که این پاداش در بهبود کارایی رفتار عامل هوشمند به‌کار برده می‌شود. عامل در هر مرحله، عملی را انتخاب می‌کند که در مجموع بیشترین پاداش را از محیط دریافت نماید [۲]. در واقع یادگیری تقویتی مجموعه روش‌هایی می‌باشد که در آن عامل با استفاده از تعامل با محیط پویا، رفتار خود را بهبود می‌بخشد. در یادگیری تقویتی عاملی وجود دارد که از طریق آزمون و خطا با محیط تعامل کرده و یاد می‌گیرد تا عملی بهینه را برای رسیدن به هدف انتخاب نماید [۳، ۴]. مزایای فوق به‌اضافه قدرت مکاشفه بالای یادگیری تقویتی، آن را به یک الگوریتم قوی آموزشی برای سیستم‌های هوشمند تبدیل نموده است [۵].

با وجود مزایای فراوان آموزش تقویتی، یکی از چالش‌های مهم و اساسی در آن این است که عامل برای رسیدن به رفتار بهینه نیازمند صرف زمان بسیار زیادی است [۶]. یکی از دلایل مهم این مسئله، استفاده‌نکردن از دانش به‌دست‌آمده، در طول یادگیری در یک سیستم هوشمند[۴] می‌باشد. به این منظور اخیرا محققان توجه خود را به انتقال دانش[۵] در سیستم‌های هوشمند معطوف کرده‌اند.

مفهوم انتقال دانش این است که عمل یادگیری، در یک محیط انجام و دانش حاصل از آن در محیط‌های مشابه به‌کار گرفته شود. به‌عبارت دیگر انتقال دانش به عامل کمک می‌کند تا عملی که در محیط مبدأ[۶] انجام داده است را در محیط مقصد[۷]، سریع‌تر انجام دهد [۷]. در صورتی انتقال دانش موفق خواهد بود که با استفاده از دانش یاد گرفته‌شده از محیط مبدأ، یادگیری در محیط مقصد سریع‌تر و با عملکرد بهتری صورت گیرد [۸].

با توجه به این‌که راه‌کارهای ارائه‌شده به‌منظور حل چالش یادگیری تقویتی، کارایی آن را به میزان قابل توجهی تغییر نداده‌اند [۱۱- ۹]، بنابراین نیاز به توسعه رویکرد انتقال دانش به‌منظور رسیدن به سرعت و کارایی بالا در حوزه یادگیری تقویتی کاملا ضروری به‌نظر می‌رسد. در ادامه کارهای صورت‌گرفته در بحث انتقال دانش که بیشترین اهمیت را در بین سایر کارهای صورت‌گرفته دارند را مورد بررسی قرار خواهیم داد.

بر اساس پژوهش‌هایی که تاکنون در زمینه انتقال دانش صورت‌گرفته است، می‌توان دو حالت را به‌منظور انتقال دانش در نظر گرفت: در حالت اول گروهی از عامل‌ها که همزمان در حال حل مسئله هستند از دانش یکدیگر استفاده می‌کنند و در حالت دوم از دانش مسائلی که قبلا حل‌شده‌اند برای حل مسائل دیگر استفاده می‌شود. در این مقاله از حالت دوم استفاده‌شده است.

بر این اساس در [۱۲، ۱۳]، با استفاده از یک ماتریس تسهیم وزن‌دار[۸]، میزان مهارت عامل‌ها با استفاده از دانش به‌دست‌آمده از آن‌ها افزایش یافته و باعث بهبود فرآیند یادگیری در عامل‌هایی‌شده که تجربه‌های مختلف داشتند.

بیانچی و همکاران در سال ۲۰۱۵ یک الگوریتم اکتشافی، بر اساس الگوریتم یادگیری کیو[۹] پیشنهاد دادند [۱۴]. الگوریتم پیشنهادی از خبرگی عامل‌ها و روش شکل‌دهی پاداش، برای بهبود هر چه بیشتر یادگیری استفاده نموده است. این الگوریتم از معیار خبرگی، جهت سنجش میزان خبرگی عامل‌های هم‌تیمی خود، بهره برده است. تکنیک دیگر انتقال دانش بر اساس معیارهای مشارکت عامل‌ها است. یادگیری عامل‌ها از یکدیگر با مشارکت در یاگیری، یکی از موضوع‌های مهم در یادگیری تقویتی است. در [۱۵]، یادگیری همزمان دو عامل با اشتراک‌گذاشتن دانش به‌منظور بهبود سرعت یادگیری، در قالب دو ایده در بخش‌های گسسته و پیوسته ارائه‌شده است.

به‌منظور انتقال دانش از مسائل حل‌شده به مسئله پیچیده‌تر، روش‌های متفاوتی بر اساس مقدار ارزش - عمل بررسی‌شده که خیلی قابل دسته‌بندی نیست. برخی پژوهش‌ها مفهومی به نام انتقال یادگیری را معرفی نموده‌اند. در انتقال یادگیری، دانش از یک یا چند وظیفه که آن‌ها را به‌عنوان وظیفه مبدأ می‌نامند به یک یا چند وظیفه دیگر که آن‌ها را به‌عنوان وظیفه مقصد در نظر می‌گیرند انتقال می‌یابد [۱۶]. انتقال یادگیری در یادگیری تقویتی دارای سه قدم اصلی است که می‌توان آن سه قدم را به‌صورت زیر بیان نمود [۱۷، ۱۸]:

الف) ابتدا باید یک وظیفه مبدأ متناسب با وظیفه مقصد داده‌شده را انتخاب نمود. ب) ارتباط بین وظیفه مبدأ و وظیفه مقصد مشخص باشد. پ) انتقال دانش را از یک وظیفه مبدأ به وظیفه مقصد به‌صورت مناسب و مؤثر انجام گیرد.

بر اساس موارد بیان‌شده در [۱۹]، دو ایده مطرح‌شده است. ایده اول با این فرض است که ارتباط بین عمل‌های موجود و عمل‌های حذف‌شده وجود دارد. در این صورت مقدار اولیه ارزش - عمل را با ضریبی از مقادیر ارزش - عمل‌های حذف‌شده که حوزه تأثیرشان با حوزه تأثیر عمل‌های موجود مرتبط است، اصلاح می‌کند. نتایج ارائه‌شده، بهبود زمان و تعداد برخورد به موانع را نشان می‌دهد.

ژو و همکاران با مجردکردن دانش به‌دست‌آمده از حل مسئله، مشکلات پیداکردن شباهت بین مسائل و نگاشت بین آن‌ها را مرتفع کردند. به‌طوری که نیاز به نگاشت یک به یک بین حالت و عمل عامل در مسائل وجود نداشت. همچنین در آن روش ابتدای کار باید یادگیری تا حدی انجام شود تا عامل توانایی استفاده از مفاهیم را پیدا و بعد از آن، روش را استفاده کند. به‌عبارت دیگر در ابتدای کار که سرعت یادگیری کم است، سرعت یادگیری افزایش نمی‌یابد ولی در ادامه باعث بهبود پاداش به‌دست‌آمده و افزایش سرعت یادگیری می‌شود [۲۰].

در [۲۱] با هدف انتخاب سیاست مناسب و سرعت‌بخشیدن به فرآیند یادگیری توسط انتقال یادگیری، عامل سعی در انتخاب سیاست مناسب و سرعت‌بخشیدن به فرآیند یادگیری توسط انتقال یادگیری دارد. اگر عامل‌ها نتوانند سیاست‌های مناسبی را پیدا کنند، آن‌ها سیاست‌های منبع دیگری را انتخاب و کاوش را تکرار می‌کنند. در این روش نیز در ابتدای کار که سرعت یادگیری کم است باید تا حدی یادگیری، در محیط انجام شود.

بر اساس پژوهش کویندا‌ریس و همکاران [۲۲] ابتدا می‌توان نتیجه گرفت برای حل مسائل بزرگ می‌توان با تقسیم مسئله به مسائل کوچک‌تر، ابتدا مسائل کوچک‌تر را حل نموده بعد با استفاده از دانش آن‌ها مسئله بزرگ‌تر را حل نمود.

در [۲۳] جهت افزایش سرعت بازی Tic-tac-toe از روش یادگیری تقویتی استفاده‌شده است. در آنجا ساخت ویژگی‌ها بر اساس جستجوی پیشرو یک درخت بازی انجام می‌شود. ویژگی‌های بازی مستقل جهت انتقال اطلاعات مقدار حالت از یک بازی به بازی دیگر بکار برده‌شده است. این انتقال اطلاعالات باعث‌شده سرعت بازی از روش minmax پیشرو بیشتر شود. دلیل این کار نداشتن اطلاعات حریف بازی در روش پیشرو بر خلاف بازیگر در محیط انتقال دانش بیان‌شده است. اما روش بیان‌شده نتوانسته قدرت انتقال دانش را به خوبی نشان دهد. از طرفی این مقاله ئذ دو محیط یکسان عمل می‌کند و نتایج حاصل از احرای یک الگوریتم را در روش یادگیری تقویتی بکار می‌بندد نه در محیط مشابه و بزرگتر. از طرفی زمان اجرای الگوریتم مشخص نشده و از یک روش جستجوی پیشرو در انتقال دانش استفاده‌شده است.

در [۲۴] یک روش سیاست محور با تغییر در تابع ارزش پروتو[۱۰] ارائه و از ان در انتقال یادگیری استفاده‌شده است. عملکرد این تابع روی گراف حاصل از قدم‌زدن تصادفی مناسب بوده است. تعریف این گراف در انتقال یادگیری به‌ویژه زمانی که انتقال دانش در همه وظایف[۱۱] با تغییر هم در تابع پاداش و هم دامنه قدم‌زدن سر و کار دارد؛ جنبه کلیدی ایفا می‌نماید. دو تابع پروتو متناسب با هر تغییر طراحی‌شده و نتایج در محیط شبکه جهانی[۱۲] مورد آزمایش قرار گرفته است.

یکی از محدودیت‌های یادگیری ماشین این است که نیاز به تنظیم پارامترهای زیادی دارد و با تغییر هر وظیفه، یادگیری نیز باید از ابتدا شروع شود. برای غلبه بر این محدودیت روی روش‌های انتقال تجربه تمرکزشده است. اکثر این روش‌ها روی یادگیری با ناظر انجام‌شده است. به‌همین دلیل در [۲۵] انواع سناریوهایی که در یادگیری تقویتی موثر است مورد بررسی قرار گرفته است. ولی در هیچکدام از کارهای فوق از یک روش هوشمند برای انتقال یادگیری استفاده نشده است. تنها کار انجام‌شده در این زمینه، استفاده از شبکه عصبی در انتقال دانش به کمک ویژگی‌های مشترک درمساله یادگیری تقویتی بوده است [۲۶].

در آن پژوهش با استفاده از ویژگی‌های مشترک بین مسائل مرتبط، مقادیر ارزش - عمل آنها توسط عامل برای هر مسئله، به‌دست می‌آید. با استفاده از یک روش خطی، تابع ارزش اولیه‌ای برای مسائل جدید تقریب‌زده می‌شود. از چالش‌های مطرح در به‌کارگیری این روش، می‌توان به چگونگی ارائه ساختار مناسب برای تقریب تابع ارزش در محیط پیوسته و نحوه ترکیب ارزش تقریب‌زده‌شده با ارزش به‌دست‌آمده در حین یادگیری در محیط مقصد اشاره داشت.

آن‌چه در پژوهش‌های صورت‌گرفته به‌عنوان چالش محسوب می‌شود این است که نحوه انتقال دانش استخراج‌شده از محیط مبدأ به محیط مقصد نتوانسته کارایی لازم را در افزایش سرعت یادگیری داشته باشد و به‌منظور توسعه نسل بعدی روش‌های انتقال دانش لازم است که کارهای تحقیقاتی بیشتری صورت گیرد تا در خلال آن بتوان با ایجاد همگرایی بیشتر در بین قسمت‌های مختلف، امکان تطبیق آن را با مفهوم یادگیری تقویتی افزایش داد. در این مقاله با استفاده از شبکه عصبی - فازی راهکاری جهت انتقال دانش از محیط مبدأ به محیط مقصد داده می‌شود. این کار در سه مرحله صورت می‌گیرد. در مرحله اول عامل در محیط مبدأ شروع به یادگیری می‌کند و دانش حاصل از این یادگیری که همان مقادیر ارزش - عمل هستند در ماتریس تقویتی ذخیره و در مرحله دوم با استفاده از یک شبکه عصبی - فازی به آموزش عامل در محیط مبدأ پرداخته می‌شود. در مرحله سوم دانش حاصل از شبکه عصبی - فازی آموزش‌دیده در محیط مبدأ، جهت افزایش سرعت یادگیری عامل در محیط مقصد که شباهت زیادی با محیط مبدأ دارد استفاده می‌شود. نتایج حاصل از انجام مراحل بیان‌شده حاکی از افزایش سرعت یادگیری عامل در محیط مقصد نسبت به حالتی که از دانش محیط مبدأ استفاده نکند؛ می‌باشد.

در ادامه در بخش دوم به تشریح الگوریتم یادگیری تقویتی پرداخته‌شده است. سپس در بخش سوم مفهوم انتقال دانش معرفی می‌شود. در بخش چهارم به تشریح روش پیشنهادی پرداخته می‌شود. در بخش پنجم ارزیابی روش پیشنهادی ارائه می‌گردد و در نهایت در بخش ششم به نتیجه‌گیری و راه‌کارهای آینده اشاره خواهد شد.

۲ -یادگیری تقویتی

در یک مسئله یادگیری تقویتی، عامل از طریق سعی و خطا با محیط تعامل کرده و یاد می‌گیرد تا عملی بهینه را برای رسیدن به هدف انتخاب کند [۲۷]. در این نوع یادگیری هیچ ناظر خارجی وجود ندارد و عامل به تنهایی با محیط تعامل کرده، یاد می‌گیرد، تجربه کسب و پاداشی دریافت می‌کند [۲۸]. در یادگیری تقویتی، یادگیرنده و تصمیم‌گیرنده را عامل و آن چیزی که عامل با آن تعامل دارد شامل هر چیز خارج از عامل، محیط نامیده می‌شود. در هر قدم زمانی، عامل حالت جاری s_t را مشاهده نموده و عمل a_t را از مجموعه عمل‌های ممکن با توجه به سیاست اتخاذشده انتخاب کرده و به محیط اعمال می‌نماید. محیط با احتمال $P(s_t, a_t, s_{t+1})$ به حالت s_{t+1} می‌رود،

همچنین سیگنال تقویت یا جایزه آنی $r(s_t,a_t)$ را که به‌صورت r_{t+1} هم نشان داده می‌شود، توسط عامل دریافت می‌گردد [۲۹].

چارچوب ریاضی محیط استفاده‌شده در این مقاله یک مدل مسئله مارکوف (MDP[۱۳]) می‌باشد، یک MDP یک چندتایی $<S,A,R,P>$ است که در این رابطه، S یک مجموعه محدود حالت‌های گسسته محیط، A مجموعه محدود عمل‌های گسسته عامل، R امید آنی سیگنال‌های تقویت $R:S\times A\to r$ و P احتمالات گذرا از یک حالت با انجام عمل برای رفتن به حالت بعدی است $P:S\times A\times S\to[0,1]$ [۳۰].

قانونی که عامل با توجه به آن در هر حالت عمل را برای اجرا انتخاب می‌کند، سیاست می‌نامند و معمولا به‌صورت $\pi(s,a)$ نشان داده می‌شود که در آن s بیانگر حالت و a بیانگر عمل است. هدف عامل انجام عمل‌هایی است که به حالت‌هایی با بالاترین ارزش برسد نه به بالاترین پاداش چرا که جایزه‌ها مطلوبیت لحظه‌ای و ارزش مطلوبیت در بلند مدت را نشان می‌دهد. ارزش یک حالت کل مقدار پاداشی است که عامل می‌تواند انتظار داشته باشد بعد از شروع از آن حالت دریافت کند. تابع ارزش حالت تحت سیاست π به‌صورت $V^{\pi}(s)$ نشان داده می‌شود.

$$V^{\pi}(s)=E_{\pi}\{\sum_{k=0}^{\infty}\gamma^{k}r_{t+k+1}\,|\,s_t=s\} \qquad (۱)$$

$$0\le\gamma\le1$$

در رابطه (۱) γ فاکتور تخفیف و $E_{\pi}\{\}$ امید ریاضی می‌باشند. به‌طور مشابه ارزش عمل a در حالت s تحت سیاست π را تابع ارزش - عمل نامیده و با $Q^{\pi}(s,a)$ نشان می‌دهند.

$$Q^{\pi}(s,a)=E_{\pi}\{\sum_{k=0}^{\infty}\gamma^{k}r_{t+k+1}\,|\,s_t=s,a_t=a\} \qquad (۲)$$

۲-۱- یادگیری Q

یادگیری Q یک توسعه از یادگیری تقویتی است که بر پایه مدل می‌باشد. با این تفاوت که در هر حالت سعی در تخمین مقدار ارزش - عمل بهینه دارد. الگوریتم مبتنی‌بر مدل در حین یادگیری به کمک تابع ارزش - عمل، سیاست عامل را مشخص می‌کند. مدل تولیدشده از این فرایند یادگیری می‌تواند برای انجام شبیه‌سازی‌های لازم استفاده شود[۷].

این الگوریتم به تخمین مقادیر تابع ارزش - عمل می‌پردازد. با این تفاوت که هدف، تخمین ماکزیمم مقدار تابع ارزش - عمل روی همه سیاست‌های ممکن است. در واقع در تازه‌سازی مقادیر ارزش - عمل، عملی که عامل در حالت بعدی انتخاب می‌کند؛ نقشی ندارد. به‌همین دلیل روش را off-policy می‌نامند. مقادیر ارزش - عمل در روش مذکور به‌صورت زیر تازه‌سازی می‌گردند:

$$Q(s_t,a_t)\leftarrow Q(s_t,a_t)+\alpha_t[r_{t+1}+y\max_{b\in A}Q(s_{t+1},b)-Q(s_t,a_t)] \qquad (۳)$$

در رابطه (۳)، α نرخ یادگیری، s_t حالت γ فاکتور تخفیف و r مقدار جایزه آنی است. عبارت سوم در فرمول فوق $\max_{b\in A}Q(s_{t+1},b)$ برابر با مقدار ارزش حالت بعدی $V(s_{t+1})$ می‌باشد.

مقادیر ارزش- عمل روش یادگیری Q در صورتی که فضای مسئله به حد کافی کشف گردد و دو شرط زیر برقرار باشد به مقدار بهینه همگرا می‌شوند [۳۱]. شرط (۱): محیط MDP غیرنوسانی[۱۴]، کاهش‌ناپذیر[۱۵] و با سیگنال‌های تقویت محدود باشد.

شرط (۲): نرخ یادگیری مثبت، غیرافزایشی و در روابط (۴) صدق کند. در ادامه به تشریح روش‌های انتخاب عمل در یادگیری Q پرداخته می‌شود.

$$\sum_{t=0}^{\infty}\alpha_t^2<\infty,\sum_{t=0}^{\infty}\alpha_t=\infty \qquad (۴)$$

۲-۲- روش‌های انتخاب عمل در یادگیری Q

یکی از مواردی که در سرعت یادگیری تأثیر به‌سزایی دارد نحوه انتخاب عمل و برقراری تعادل بین کاوش و بهره‌برداری از تجربیات است. در این قسمت سه روش مهم انتخاب عمل در یادگیری Q معرفی می‌شود.

- روش *greedy*: یکی از ساده‌ترین روش‌های انتخاب عمل می‌باشد. در هر لحظه عمل دارای بالاترین مقدار ارزش - عمل تخمین‌زده انتخاب می‌شود:

$$a=\arg\max_{b\in A}(Q(s,b)) \qquad (۵)$$

- این روش امکان مکاشفه را کاهش داده و معمولا منجربه جواب بهینه نمی‌شود و پاسخ‌هایی در اکسترمم‌های محلی پیدا می‌کند.
- روش $\varepsilon-greedy$: در این روش با احتمال ε-۱ عمل دارای بالاترین مقدار تخمینی ارزش - عمل انتخاب و با احتمال ε با شانس مساوی یکی از عمل‌ها انتخاب می‌گردد.

$$P(a)=\begin{cases}1-\varepsilon+\dfrac{\varepsilon}{N} & a=\arg\max_{b\in A}(Q(s,b))\\ \dfrac{\varepsilon}{N} & otherwise\end{cases} \qquad (۶)$$

که در آن N، تعداد کل عمل‌ها است. نتایج به‌کارگیری این روش حاکی از راندمان بالاتر آن در اکثر موارد نسبت به روشgreedy است ولی تعادل مناسبی بین مکاشفه و استفاده از تجربه برقرار نمی‌کند، چرا که امکان انتخاب همه عمل‌ها به‌جز عمل دارای ارزش - عمل بالاتر یکسان می‌باشد حتی عمل‌هایی که دارای ارزش - عمل منفی هستند. لذا احتمال تعداد جریمه‌های دریافتی در این روش نسبتا بیشتر می‌شود.

- روش softmax: یکی از کاربردی‌ترین روش‌های انتخاب عمل می‌باشد که فرمول احتمال انتخاب هر عمل مانند a به‌صورت زیر تعریف می‌گردد:

$$P(a)=\frac{\exp(Q(s,a)/T)}{\sum_{b\in A}\exp(Q(s,b))/T} \quad (۷)$$

در رابطه (۷)، $T>0$ ضریب دما نامیده می‌شود. اگر $T\to 0$ به روش greedy همگرا می‌گردد و اگر $T\to\infty$ انتخاب‌ها کاملا تصادفی می‌گردند. معمولا در حین آموزش هر چه به‌سمت جلو می‌رود مقدار ضریب دما کاهش می‌یابد تا از تجربیات قبلی بیشتر استفاده گردد. مزیت عمده این روش ارتباط بین احتمال انتخاب عمل‌ها با مقادیر ارزش - عمل می‌باشد [۲۷].

در این مقاله، روش انتخاب عمل softmax، مورد استفاده قرار گرفته است. در ادامه به بررسی مفهوم انتقال دانش خواهیم پرداخت.

مشکل اصلی یادگیری تقویتی، پیچیدگی محاسباتی نمایی در انتخاب عمل می‌باشد و این موضوع باعث‌شده تا این الگوریتم مقیاس-پذیر نباشد. یادگیری تقویتی یک راه‌حل برای تصمیم‌گیری متوالی است که به‌صورت فرایند تصمیم‌گیری مارکوف مدل‌شده است در مسایل یادگیری تقویتی رسیدن به بهترین سیاست زمان‌بر می باشد و برای همین از انتقال دانش استفاده می‌گردد. از روش‌های انتقال دانش برای تسریع همگرایی استفاده می‌شود[۳۲].

انتقال دانش بین چندین عامل در یک محیط و یا چندین محیط مشابه با یک عامل صورت می‌گیرد. انتقال دانش در مجیط‌های چند عامله بین چندین عامل صورت گیرد. در این مقاله، تمرکز روی انتقال دانش برای حالتی است که محیط‌ها با هم تشابه دارند.

۳ -شبکه عصبی - فازی

به‌منظور تسهیل فرآیند یادگیری و انطباق، منطق فازی با شبکه عصبی مصنوعی ترکیب می‌شود. با استفاده از این شبکه عصبی تطبیقی مشکل اصلی استفاده از سیستم استنتاج فازی که همان به‌دست‌آوردن قواعد اگر- آنگاه فازی و بهینه‌سازی پارامترهای مدل می‌باشد؛ حل می‌گردد. این شبکه دارای پنج لایه است. لایه اول، لایه داده‌های ورودی است که توسط کاربر مشخص می‌شود. مجموعه عملیات مدل‌سازی در لایه دوم تا چهارم انجام می‌گیرد. در این سه لایه عملیات فازی‌سازی، ساخت قواعد فازی و نرمال‌سازی قواعد صورت می‌پذیرد. لایه آخر، لایه خروجی است و هدف آن حداقل‌کردن اختلاف خروجی به‌دست‌آمده از شبکه با خروجی واقعی است [۳۷-۳۳].

۴ -روش پیشنهادی

در این بخش روش پیشنهادی جهت افزایش سرعت یادگیری بیان می‌گردد. همان‌طور که قبلا بیان شد در این مقاله روشی جهت انتقال دانش از محیط مبدأ به محیط مقصد معرفی می‌گردد. پایه و اساس الگوریتمی که مورد استفاده قرار می‌گیرد الگوریتم یادگیری Q می-باشد. مراحل زیر چگونگی انجام روش پیشنهادی را نشان می‌دهد.

- **انجام یادگیری در محیط مبدأ:** در این مرحله، محیط مبدأ و عمل‌های قابل‌قبول به‌همراه تک تک حالت‌های موجود مشخص می‌شوند. عامل بر اساس سیاست مشخص‌شده شروع به انتخاب عمل می‌کند به‌گونه‌ای که مقدار سیگنال تقویتی دریافتی یا جایزه آنی را بیشتر نماید. در هر مرحله از یادگیری مقدار ارزش - عمل متناسب یا عمل انتخابی طبق رابطه (۳) محاسبه و در ماتریس Q ذخیره می‌گردد. در حقیقت دانش حاصل از یادگیری عامل در این ماتریس قرار دارد. جهت استفاده از این دانش و به‌کارگیری آن در محیط مبدأ ایده اصلی و نوآوری این مقاله در مرحله بعد آمده است.
- **آموزش عامل توسط شبکه عصبی - فازی از روی مقادیر ارزش - عمل:** در این مقاله از شبکه عصبی - فازی سوگنو استفاده‌شده است. پارامترهای هر دو بخش مقدم و تالی قواعد فازی با روش گرادیان نزولی آموزش داده می‌شوند. از روش خوشه‌بندی جزئی[۱۶] موجود در نرم‌افزار MATLAB 7.0 جهت خوشه‌بندی داده‌های ورودی و به‌دنبال آن تعیین قواعد فازی استفاده‌شده است. در این روش نوع تابع عضویت با توجه به اطلاعات ورودی و دسته‌بندی‌های موجود توسط خود مدل تعیین می‌شود. در روش خوشه-بندی محدوده نفوذ توسط کاربر تعیین می‌شود و معمولا بین صفر و یک می‌باشد. در این مقاله از مقدار نفوذ ۰/۵ استفاده‌شده است.

با تعیین شبکه عصبی - فازی برای این‌که بتوان دانش حاصل از ارزش – عمل به‌دست آمده توسط عامل در محیط مبدأ را به محیط مقصد انتقال داد k هدایت‌گر[۱۷] (هم‌جنس با حالت تعریف‌شده در محیط) $p_1, p_2, ..., p_k$ در محیط مبدأ در نظر گرفته می‌شود. این هدایت‌گرها نقش هدایت کننده عامل در محیط را دارند. فاصله عامل با هر یک از این هدایت‌گرها در هر حالت و قبل از انتخاب عمل محاسبه و در ماتریس D قرار می‌گیرد. لازم به توضیح است اگر محیط دو بعدی باشد این هدایت‌گرها، نقاطی در صفحه مختصات دو بعدی هستند. به‌همین ترتیب با توجه به بعد تعریف‌شده برای محیط می‌توان این هدایت‌گرها را تعریف نمود و متناسب با آن فاصله بین هر هدایت‌گر و عامل را محاسبه کرد. برای این‌که این اندازه‌گیری به نتیجه رسیدن عامل به هدف تعیین‌شده در محیط منجر شود؛ یکی از هدایت‌گرها روی حالت هدف قرار می‌گیرد و سایر هدایت‌گرها به‌صورت تصادفی در محیط پخش می‌شوند. از مقادیر ماتریس ارزش - عمل Q و ماتریس D جهت آموزش عامل در محیط مبدأ استفاده می‌شود. در هر بار یادگیری، عامل episode بار فرصت دارد تا از حالت شروع خود را به حالت هدف[۱۸] برساند. برای این‌که دانش حاصل از یادگیری به اندازه کافی در ماتریس Q ذخیره شود عمل یادگیری iteration بار تکرار می‌گردد. کل این عملیات در مجموع یک‌بار یادگیری عامل در نظر گرفته می‌شود.

فرض کنید عامل n بار به‌صورت مستقل در محیط مبدأ شروع به یادگیری نموده است. در این صورت n ماتریس Q و n ماتریس D به‌ازای هر بار یادگیری عامل در محیط وجود دارد. به‌راحتی می‌توان یک شبکه عصبی - فازی[19] جهت آموزش عامل در محیط تشکیل داد. کافی است ورودی و خروجی شبکه مشخص شوند. از آن‌جا که دانش نهایی حاصل از یادگیری عامل، ماتریس ارزش - عمل Q می‌باشد، سعی می‌شود تا عامل به گونه‌ای آموزش داده شود تا در هر حالت از محیط قرار گرفت، شبکه عصبی - فازی مقدار ارزش - عمل متناظر با آن حالت را تقریب بزند. به‌همین دلیل مقادیر موجود در ماتریس D به‌عنوان ورودی شبکه عصبی - فازی در نظر گرفته می‌شود. بدیهی است که تعداد پارامتر ورودی شبکه برای هر نمونه، n یا همان تعداد هدایت‌گرها خواهد بود. خروجی شبکه مقادیر ماتریس ارزش - عمل می‌باشد. تعداد پارامتر خروجی برابر با تعداد عمل انتخابی عامل در محیط است. بعد از آموزش شبکه عصبی - فازی، با توجه به فاصله عامل به هر یک از هدایت‌گرها و بدون انجام عمل یادگیری، مقدار ارزش - عمل متناظر با آن حالت پیش‌بینی می‌شود.

همان‌طور که قبلا بیان شد در دو محیط مبدأ و مقصد عمل‌های انتخابی عامل یکسان هستند از طرفی برای این‌که بتوان از این شبکه آموزش‌دیده در محیط مقصد استفاده نمود لازم است تا تعداد هدایت‌گرها نیز در هر دو محیط یکسان تعریف شوند. ضمنا مقادیر ماتریس D باید نرمال شوند تا مشکلی در ورودی پدید نیاید. از آن‌جا که مقادیر ارزش - عمل در هر حالت بین ۱- و ۱ هستند نیازی به نرمال‌سازی این مقادیر نیست.

- **انتقال دانش از محیط مبدأ به محیط مقصد:** همان‌طور که قبلا بیان شد هر چه محیط مقصد به محیط مبدأ شبیه‌تر باشد انتقال دانش بهتر صورت می‌گیرد. به‌همین دلیل بعضی از ویژگی‌های بین این دو محیط با هم مشترک انتخاب می‌شوند مثل عمل‌های انتخابی عامل، سیاست انتخاب حالت و مقدار سیگنال تقویتی. در محیط مقصد مقدار ارزش - عمل متناظر با هر حالت $Q(s_t, a_t)$ ابتدا توسط رابطه (۳) محاسبه می‌شود سپس فاصله عامل تا هدایت‌گرها به‌عنوان ورودی به شبکه عصبی - فازی آموزش‌دیده از مرحله قبل داده می‌شود و خروجی شبکه که همان مقدار ارزش - عمل تقریبی $N(s_t, a_t)$ یا دانش موجود در محیط مبدأ است، به‌دست می‌آید. جهت استفاده صحیح از این دانش از رابطه (۸) و (۹) استفاده می‌شود.

اگر $N(s_t, a_t) < \eta$ آنگاه:

$$Q(s_t, a_t) = Q(s_t, a_t) + \alpha_t [r_{t+1} + y \max_{b \in A} Q(s_{t+1}, b) - Q(s_t, a_t)] \quad (۸)$$

در غیر این صورت:

$$Q(s_t, a_t) = w\ Q(s_t, a_t) + (1-w)\ N(s_t, a_t) \quad (۹)$$

که در آن η عددی بین ۱- تا ۱ و w عددی بین صفر و یک می‌شود. به عبارت دیگر یک ترکیب وزن‌دار از دو مقدار ارزش - عمل متناظر با یک حالت جهت انتقال دانش استفاده می‌شود.

شکل (۱) الگوریتم روش پیشنهادی جهت انتقال دانش با استفاده از ویژگی‌های مشترک را نشان می‌دهد. در الگوریتم فوق، متغیرها بر اساس متغیرهای بیان‌شده در بخش چهارم تعریف‌شده‌اند. گام اول این الگوریتم به یادگیری عامل در محیط انتخابی می‌پردازد. در این محیط با توجه به تعداد تکرار مشخص‌شده عمل یادگیری را انجام می‌دهد. ماتریس تقویتی حاصل از این یادگیری به‌همراه ماتریس فاصله به‌دست‌آمده عامل تا هدایت‌گرها به‌عنوان ورودی شبکه عصبی- فازی در گام دوم انتخاب می‌شوند. بعد از آموزش شبکه با مقادیر ورودی، در گام سوم الگوریتم مجددا از یادگیری تقویتی جهت یادگیری عامل در محیطی مشابه و بزرگتر از محیط موجود در گام اول استفاده می‌شود. با این تفاوت که ماتریس تقویتی با استفاده از مقادیر آموزش‌دیده در گام دوم بهینه می‌شود

با توجه به آنچه بیان گردید ایده این پژوهش انتقال دانش از محیط مبدأ به محیط مقصد با استفاده از شبکه عصبی- فازی است. مقادیر تقریب‌زده‌شده ارزش - عمل‌ها توسط شبکه عصبی- فازی با دانش اکتسابی در محیط مقصد ترکیب‌شده و انتخاب عمل با استفاده از دانش ترکیب‌شده انجام می‌شود.

Input
- Source environment
- Target environment
-: A vector of Actions selection
-Start state
-Goal state
-Obstacles

Output
- Reinforcement Matrix
-Average number of steps to reach the goal
-Average number of steps traversed
-Average number of hits to obstacles

Initialization

Step 1: Q-Learning in Source Environment
get source environment // A Grid World
2: guides
3: //Reinforcement Matrix
4://Distance Matrix
5: start state
6: for each do
7: for each do
8: select action
9: determine a reward based on policy
10: update
11: select next state
12: update as follows:
13: until is not goal
14: until the learning is finished

Step 2: Start Training with Fuzzy Neural Network
1: network input $\leftarrow D$
2: network output $\leftarrow Q$
3: network training

Step 3: Knowledge Transfer

1: $E2 \leftarrow \rightarrow$ get target environment
2: $P \leftarrow$ guides
3: $Q \leftarrow 0$
4: $D \leftarrow 0$
5: $s \leftarrow$ start state
6: for each $episode$ do
7: for each $iteration$ do
8: $a \leftarrow$ select action
9: $r \leftarrow$ determine a reward based on policy
10: update D
11: $s' \leftarrow$ select next state
12: Compute $N(s,a)$ by trained fuzzy neural network in step 2
13: update Q as follows:

a) if $N(s,a) < \eta$ then

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma \max_{b \in A} Q(s',b) - Q(s,a)]$$

b) else

$$Q(s,a) = w\ Q(s,a) + (1-w)\ N(s,a)$$

14: until s' is not goal
15: until the learning is finished

شکل ۱: الگوریتم روش پیشنهادی جهت انتقال دانش

۵ -ارزیابی روش پیشنهادی

جهت بررسی کارایی روش پیشنهادی مسئله‌ای با توجه به الگوریتم بیان‌شده در بخش قبل شبیه‌سازی‌شده است. این مسئله، آموزش ربات یا عاملی است که می‌خواهد فقط با دوربین در یک محیط پر از مانع حرکت کند. فرض کنید این محیط دو بعدی در نظر گرفته شود. در ادامه جزئیات روش پیشنهادی تشریح‌شده است.

۵-۱- ویژگی‌های مشترک دو محیط

نمونه مسئله‌ای که ایده پژوهش را در آن به‌کار گرفته‌شده یک محیط مارپیچ ۲۰ است. محیط مبدأ یک مارپیچ با ابعاد $n_1 \times m_1$ است. بدیهی است که تعداد حالات یا نقاط انتخابی توسط عامل در این محیط $n_1 \times m_1$ خواهد بود. تعداد p مانع و n منبع یا نقطه انتشار یا همان دوربین در این محیط قرار می‌گیرد. ویژگی‌های مشترک هر دو محیط عبارتند از:

۱ - حالت: مختصات عامل

۲ - تعداد و نوع عمل انتخابی: حرکت عامل در چهار جهت اصلی

۳ - سیگنال تقویتی: اگر عامل با مانع برخورد کند (۱-) واحد جریمه می‌شود، اگر به هدف برسد (۱+) و در غیر این صورت مقدار جریمه (۰/۰۱-) را دریافت می‌کند. رابطه (۱۰) چگونگی محاسبه سیگنال تقویتی و به‌دنبال آن سیاست اتخاذشده را نشان می‌دهد.

$$r = \begin{cases} -1 & obstacle \\ -0.01 & otherwise \\ 1 & gole \end{cases} \qquad (۱۰)$$

۴ - تعداد هدایت‌گر: به‌جز موارد گفته‌شده دو محیط در سایر خصوصیات می‌توانند متمایز باشند. خصوصیات غیرمشترک دو محیط عبارتند از:

-ابعاد محیط: در اینجا محیط دوم از محیط اول بزرگتر انتخاب می‌شود.

-نقطه شروع و هدف

-تعداد و مکان موانع

-مکان نقاط انتشار به‌جز یک هدایت‌گر که روی هدف قرار می‌گیرد و فاصله عامل تا هدف را می‌دهد.

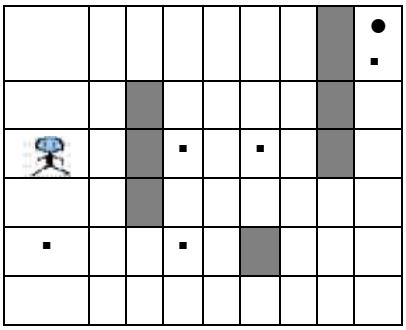
نمونه‌ای از مشخصات دو محیط در جدول (۱) آورده‌شده است. حالت در هر دو محیط، مختصات عامل و عمل‌ها چهار جهت جغرافیایی بالا، پایین، چپ و راست در نظر گرفته‌شده‌اند. نمونه‌ای از محیط مبدأ و مقصد به‌ترتیب در شکل‌های (۲) و (۳) نشان داده‌شده است.

طبق شکل‌های (۱) و (۲) عامل در نقطه شروع قرار گرفته و فرض بر این است که روی ربات یا عامل حسگری وجود دارد تا امواج فرستاده‌شده توسط هدایت‌گرها (که به‌صورت ▪ در شکل‌های (۱) و (۲) قابل‌مشاهده هستند) را دریافت نماید. فاصله اقلیدسی بین عامل و هدایت‌گر به عددی در بازه (۰,۱) تبدیل می‌شوند تا در محیط‌های بزرگتر نیز قابل استفاده باشند. به‌ازای هر حالت که عامل در آن قرار دارد این مقدار به‌ازای هر یک از هدایت‌گرها محاسبه می‌شود. در این شکل‌ها حالت هدف با ● مشخص‌شده است.

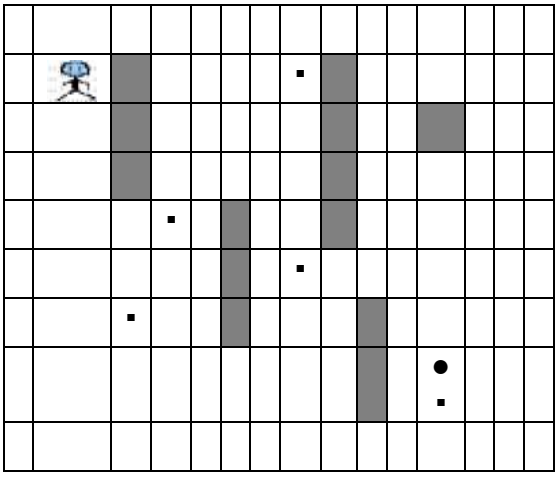
طبق الگوریتم پیشنهادی در بخش چهارم ابتدا عامل توسط الگوریتم یادگیری Q، در محیط مبدأ شروع به یادگیری می‌نماید. دانش حاصل از این یادگیری یا همان مقادیر ارزش - عمل در ماتریس تقویتی Q ذخیره می‌شود.

جدول ۱: مشخصات محیط مبدا و محیط مقصد

	بعد محیط	تعداد موانع	تعداد هدایت‌گر	نقطه شروع	نقطه هدف
محیط مبدأ	۹×۶	۷	۵	(۳و۱)	(۹ و ۱)
محیط مقصد	۱۵×۹	۱۳	۵	(۲و۲)	(۱۲و ۸)



شکل ۲: نمونه‌ای از محیط مبدأ



شکل ۳: نمونه‌ای از محیط مقصد

برای ایجاد مجموعه داده لازم آموزش و تست مورد نیاز جهت شبکه عصبی - فازی، عامل در محیط مبدأ برای ۱۰۰ بار به‌صورت مستقل اقدام به یادگیری نموده است. در هر بار یادگیری، ماکزیمم تعداد قدم‌های عامل از نقطه شروع تا نقطه هدف ۲۰۰ در نظر گرفته‌شده است. ضریب یادگیری $\alpha = ۰/۹$، فاکتور تخفیف $\gamma = ۰/۹۵$ و ضریب دما $T = ۰/۱$ در نظر گرفته‌شده است. برای انتخاب عمل از روش بیشینه نرم یا softmax استفاده‌شده است. ماتریس تقویتی Q و ماتریس فاصله D حاصل از هر بار یادگیری جهت ورودی و خروجی شبکه عصبی - فازی استفاده‌شده است.

۵-۲- استفاده از شبکه عصبی- فازی

جهت پیش‌بینی مقدار ارزش - عمل از شبکه عصبی - فازی معرفی‌شده در بخش قبل استفاده می‌شود. با انجام عمل یادگیری در محیط مبدأ توسط الگوریتم یادگیری Q به میزان ۱۰۰ بار و ذخیره داده‌های مورد نیاز نوبت به تقریب مقدار ارزش - عمل می‌رسد. ۸۰٪ درصد داده‌ها به‌عنوان داده تست و ۲۰٪ به‌عنوان داده آموزش به‌صورت تصادفی انتخاب می‌شوند. ورودی شبکه عصبی - فازی همان فاصله عامل تا هر یک از هدایت‌گرها می‌باشد. تعداد پارامتر ورودی برابر تعداد هدایت‌گر یعنی ۵ می‌باشد. مقادیر ارزش - عمل‌های حالت‌ها، به‌عنوان خروجی شبکه در نظر گرفته می‌شود. در اینجا تعداد خروجی برابر ۴ در نظر گرفته می‌شود. جداول (۲) و (۳) به‌ترتیب چهار نمونه از مقادیر ورودی و خروجی را نشان می‌دهند. در اینجا مقدار حسگر همان فاصله بین عامل با هر هدایت‌گر می‌باشد.

جدول ۲: چند نمونه از داده‌های جمع‌آوری‌شده فضای عامل

شماره داده ورودی	مقدار حسگر ۱	مقدار حسگر ۲	مقدار حسگر ۳	مقدار حسگر ۴	مقدار حسگر۵
۱	۰/۸۴	۰/۸۴	۰/۸۲	۰/۷	۰/۶۵
۲	۰/۶۹	۰/۷۲	۰/۷۶	۰/۷۹	۰/۷۲
۳	۰/۸۹	۰/۹۳	۰/۹۵	۰/۹۳	۰/۹
۴	۰/۷۵	۰/۹	۰/۹۵	۰/۹۵	۰/۹۱

جدول ۳: مقادیر ارزش چهار عمل متناظر با جدول (۲)

شماره داده خروجی	حرکت به بالا	حرکت به پایین	حرکت به چپ	حرکت به راست
۱	-۰/۰۶	-۰/۰۶۵	-۰/۰۶۷	-۰/۰۶۲
۲	-۰/۰۷	-۰/۰۷۵	-۰/۰۷۴	-۰/۰۷۲
۳	-۰/۰۴	-۰/۰۴	-۰/۰۴	۰/۳۶
۴	-۰/۰۴۶	-۰/۹۱	-۰/۰۴	۰/۶۳

پس از آموزش شبکه با این داده‌ها در محیط مبدأ، می‌توان برای تسریع یادگیری در محیط مقصد از آن به‌منظور تخمین اولیه ارزش - عمل‌ها استفاده نمود. مربعات خطا پس از آموزش این شبکه با تعداد ۴۳۲۰ داده، ۰/۱۵۱ شده است

۵-۳- انتقال دانش در محیط

در محیط مقصد از الگوریتم Q جهت یادگیری پایه عامل استفاده‌شده است. ترکیب دانش اکتسابی حاصل از این یادگیری با دانشی که از محیط مبدأ به‌دست می‌آید توسط رابطه (۸) صورت می‌گیرد. در اینجا $\eta = ۰/۲$ و $w = ۰/۳$ در نظر گرفته‌شده است. اگر در محیط مقصد، فاصله عامل تا هر هدایت‌گر به‌عنوان ورودی (یک بردار ۵ تایی) به شبکه عصبی- فازی آموزش‌دیده در زیربخش قبل داده شود؛ مقادیر اولیه‌ای برای ارزش - عمل‌ها در آن حالت از خروجی شبکه (یک بردار چهارتایی) به‌دست می‌آید. ترکیب این مقادیر اولیه تابع ارزش - عمل حاصل از شبکه با دانش اکتسابی مسئله مقصد طبق رابطه (۸)، تأثیر به‌سزایی در افزایش سرعت یادگیری مسئله مقصد خواهد داشت.

۵-۴- ارزیابی

همان‌طور که قبلا بیان گردید؛ در محیط مقصد، عامل هر حالتی را که ملاقات می‌کند، مقدار ارزش اولیه را با استفاده از شبکه عصبی آموزش داده‌شده، به‌دست می‌آورد. سپس ترکیب دانش انتقالی یعنی خروجی شبکه عصبی- فازی و دانش اکتسابی که از یادگیری Q در این حالت کسب کرده را به‌صورت جمع دو دانش با فرمول خطی طبق رابطه (۸) انجام می‌دهد. هدف اصلی از انتقال دانش در حل مسائلی که با یادگیری تقویتی سر و کار دارند؛ کاهش زمان اجرا و رفع مشکل مقیاس‌پذیری آن‌ها است. به‌همین دلیل جهت ارزیابی روش پیشنهادی

سه معیار که در اکثر مراجع به‌کاربرده شده است مورد بررسی قرار می‌گیرد [۱۹]:

- زمان یادگیری عامل در محیط: بیانگر سرعت یادگیری است.
- تعداد گام رسیدن عامل از حالت شروع به حالت هدف: هر چقدر تعداد گام پیمایش‌شده توسط عامل بعد از عمل یادگیری کمتر باشد نشان‌دهنده انتخاب مسیر مناسب‌تر توسط عامل می‌باشد.
- متوسط تعداد گام پیمایش‌شده در هر مرحله از یادگیری: این نیز مانند معیار اول نشان‌دهنده سرعت یادگیری است.

برای ارزیابی و مقایسه روش پیشنهادی چندین بار خصوصیات غیرمشترک محیط مقصد نسبت به محیط مبدأ مثل اندازه محیط و تعداد مانع تغییر داده شد. نقاط شروع و هدف همه این محیط‌ها برابر آن‌چه در جدول (۱) آمده ثابت در نظر گرفته‌شده است. در آن محیط‌ها، عامل یک‌بار بدون انتقال دانش از محیط مبدأ و یک‌بار با استفاده از روش انتقال دانش بیان‌شده در این مقاله به یادگیری در محیط پرداخته است. میانگین نتایج حاصل از ۴۰ اجرای مستقل در جداول (۴-۶) آورده‌شده است. در آزمایش اول کارایی روش پیشنهادی در دو محیط با خصوصیات معرفی‌شده د جدول (۱) بررسی‌شده است. نتایج حاصل از این ارزیابی در جدول (۴) قابل‌مشاهده می‌باشد. همان‌طور که در جدول فوق مشخص است الگوریتم پیشنهادی توانسته است سرعت یادگیری را افزایش دهد. در آزمایش دوم تأثیر تعداد موانع در محیط مقصد بررسی‌شده است. نتایج حاصل از این کار در جدول (۵) مشخص‌شده است. همان‌گونه که از جدول فوق پیداست هر چقدر تعداد موانع بیشتر باشد زمان آموزش افزایش می‌یابد. در آزمایش سوم تأثیر اندازه محیط روی روش پیشنهادی بررسی‌شده است. در این آزمایش تعداد موانع، نقطه شروع و پایان مانند شکل (۱) می‌باشد. با این آزمایش سعی در بررسی مقیاس‌پذیری روش پیشنهادی‌شده است. نتایج این آزمایش در جدول (۶) آمده است. نتایج حاکی از کاهش مشکل عدم مقیاس‌پذیری الگوریتم یادگیری تقویتی توسط روش پیشنهادی است. همان طور که از جداول فوق پیداست با روش انتقال دانش بیان‌شده، عامل سریع‌تر یاد می‌گیرد و تعداد گام رسیدن به هدف نیز کاهش یافته است این امر مدت زمان یادگیری را نیز کاهش داده است.

جدول ۴: ارزیابی روش پیشنهادی برای محیط ۱۵× ۹ با ۱۵ مانع

	بدون انتقال دانش	انتقال دانش با شبکه عصبی- فازی
زمان یادگیری (بر حسب میلی ثانیه)	۲۵۸/۲۳	۱۸۳/۶
متوسط تعداد گام دستیابی به هدف	۱۹/۶	۱۵/۷
متوسط تعداد گام پیمایش‌شده	۵۸/۲	۴۳/۸
متوسط تعداد برخورد به مانع	۷۰/۸	۳۵/۳

جدول ۵: تاثیر تعداد مانع در روش پیشنهادی برای محیط ۱۵× ۹

تعداد / مانع — پارامتر ارزیابی	۷	۱۵	۵۰
زمان یادگیری (بر حسب میلی ثانیه)	۱۰۲/۴	۱۸۳/۶	۲۴۸۶/۵
متوسط تعداد گام پیمایش‌شده	۴۰/۴	۴۳/۸	۶۱/۵
متوسط تعداد برخورد به مانع	۲۸/۲	۳۵/۳	۴۶/۲

جدول ۶: تاثیر اندازه محیط روی روش پیشنهادی

اندازه / محیط — پارامتر ارزیابی	۹ ×۱۵	۹ ×۲۵	۲۰ ×۲۵
زمان یادگیری (بر حسب میلی ثانیه)	۱۷۶/۲	۲۴۲/۶	۱۴۵۲/۴
متوسط تعداد گام دستیابی به هدف	۱۵/۷	۱۵/۸	۱۵/۶
متوسط تعداد گام پیمایش‌شده	۴۳/۸	۴۹/۶	۴۸/۵
متوسط تعداد برخورد به مانع	۵۳/۳	۳۶/۲	۳۷/۱

همان‌طور که در بخش مقدمه بیان شد تنها کار نزدیک به روش پیشنهادی این مقاله، کار بیان‌شده در [۲۶] می‌باشد. یکی از تفاوت‌های اصلی کار انجام‌شده در این مقاله با مرجع فوق تغییر در تابع بهنگام‌سازی ماتریس تقویتی است. در اینجا همیشه از مقدار دانش به‌دست‌آمده از شبکه عصبی - فازی در بهنگام‌سازی ماتریس تقویتی استفاده نمی‌شود و این بهنگام‌سازی بر طبق روابط (۸ و ۹) انجام می‌پذیرد.

جهت مقایسه در شرایط یکسان، محیط مقصد ۱۸× ۵ معرفی‌شده در [۲۶] و محیط مبدا شکل (۲) انتخاب‌شده است. نتایج حاصل از ۴۰ اجرای مستقل در جدول (۷) آمده است. همان‌طور که از نتایج مشخص است روش شبکه عصبی- فازی توانسته است سرعت یادگیری را با استفاده از انتقال دانش نسبت به روش شبکه عصبی کاهش دهد. این موضوع با کاهش متوسط تعداد گام پیموده‌شده قابل‌تشخیص می‌باشد.

جدول ۷: نتایج ۴۰ تست در محیط ۱۸×۱۵

	انتقال دانش با شبکه عصبی- فازی	انتقال دانش با شبکه عصبی
متوسط تعداد گام دستیابی به هدف	۳۰/۸۲	۳۸/۴۵
متوسط تعداد گام پیمایش‌شده	۴۸/۵۳	۵۴/۷۱
متوسط تعداد برخورد به مانع	۲۳/۳۷	۲۷/۸۲

همان‌طور که ملاحظه شد، در این بخش یک روش انتقال دانش از محیط مبدأ به محیط مقصد پیشنهادشده است. با در نظر گرفتن فضای عامل دارای ویژگی‌های مشترک بین محیط‌های مبدأ و مقصد، ابتدا مقدار ارزش - عمل در محیط مبدأ به‌دست‌آمده و سپس از یک شبکه عصبی- فازی برای تقریب مقدار تابع ارزش - عمل در محیط مقصد استفاده‌شده است. در محیط مقصد، ایده پیشنهادی ما برای ترکیب مقدار ارزش حالت- عمل پیش‌بینی شبکه و مقدار به‌دست‌آمده از تابع ارزش - عمل خود آن محیط که در حین یادگیری به‌روز

می‌شود، حاصل‌جمع وزنی، این دو مقدار در هر حالت بود. نتایج حاصل، بهبود استفاده از شبکه عصبی- فازی نسبت به روش بدون انتقال دانش، زمانی که ابعاد مسئله افزایش می‌یابد را نشان می‌دهد. با توجه به نتایج، متوسط تعداد قدم دستیابی به هدف، کاهش یافته که این امر حاکی از افزایش سرعت یادگیری است.

۶- نتیجه

یکی از مشکلات اصلی یادگیری تقویتی این است که با افزایش ابعاد یادگیری و پیچیده‌ترشدن محیط، تعداد پارامترهای تصمیم‌گیری نیز افزایش می‌یابد و در نتیجه فرآیند یادگیری کند می‌گردد و از سرعت یادگیری می‌کاهد. در ایده ارائه‌شده در این پژوهش سعی‌شده تا با ارائه روشی مؤثر زمانی که ابعاد مسئله افزایش می‌یابد و یا موانع جابه‌جا می‌شوند از کندشدن سرعت یادگیری جلوگیری شود. در شبیه‌سازی با محیط مارپیچ نشان داده شد که با استفاده از شبکه عصبی- فازی می‌توان بهبود چشمگیری نسبت به روش خطی جهت تخمین مقادیر ارزش اولیه‌ها را شاهد بود. با توجه به نتایج این شبیه‌سازی، روش ارائه‌شده، تعداد قدم رسیدن به هدف، را کاهش داده و سرعت یادگیری را افزایش می‌دهد. می‌توان گفت این پژوهش دارای ایده‌ها و مراحل زیر است:

الف) تعریف ویژگی‌های فضای عامل، ب) تنظیم اولیه پارامترهای ساختار شبکه عصبی- فازی بعد از تکمیل یادگیری تقویتی محیط‌های ساده با داده‌های جمع‌آوری‌شده جهت تقریب مقادیر ارزش - عمل‌ها، پ) تنظیم نهایی پارامترهای ساختار شبکه عصبی فازی در محیط‌های پیچیده‌تر. در این مرحله تصحیح وزن‌های شبکه عصبی- فازی تنها با خطای به‌دست‌آمده از خروجی شبکه و بدون یادگیری تقویتی محیط انجام می‌شود. ت) تست با استفاده از مقدار ارزش - عمل تقریب‌زده‌شده توسط شبکه عصبی- فازی به‌منظور رسیدن ربات به هدف در محیط جدید ناشناخته، بدون نیاز به یادگیری. در این مرحله ربات تا حدی مستقل از فضای حالت‌شده و تنها با استفاده از اطلاعات حسگری، یادگیری انجام می‌شود.

روش پیشنهادی در این پژوهش بهبود قابل‌توجهی در یادگیری محیط‌های پیچیده داده است، اما محیط می‌تواند از این هم پیچیده‌تر شده و یادگیری را سخت کند. برای مثال ممکن است پیداکردن شباهت بین برخی مسائل سخت باشد، یا این‌که عمل‌ها در مسائل متفاوت باشند. در این راستا تلاش در این زمینه بایستی ادامه یابد تا در زمان کمتر به آموزش مؤثرتر پارامترها دست‌یابیم. در ایده این پژوهش ترکیب به‌صورت رابطه خطی انجام شد. امید است در آینده، بتوان راه‌کاری جهت ترکیب مؤثر دانش‌ها ارائه داد به‌طوری که باعث بهبود بیشتر همه معیارها با هم گردد.

مراجع

[1] W.Böhmer, J. T.Springenberg, J.Boedecker, M.Riedmiller and K.Obermayer, Autonomous learning of state representations for control: An emerging field aims to autonomously learn state representations for reinforcement learning agents from their real-world sensor observations, KI-Künstliche Intelligenz, vol. 29, no. 4, pp. 353-362, 2015.

[2] J. Kober, J. A. Bagnell and J. Peters, *Reinforcement learning in robotics: A survey*, International Journal of Robotics Research, vol. 32, no. 11, pp. 1238-1274, 2013.

[3] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y.Tassa and D.Wierstra, *Continuous control with deep reinforcement learning*, arXiv preprint arXiv:1509.02971, 2015.

[۴] عادل اکبری مجد، حسین شایقی، حمید محمد نژاد، عبدالله یونسی، *کنترل‌کننده مقاوم تطبیقی بار فرکانس مبتنی بر یادگیری تقویتی برای یک سیستم قدرت به هم پیوسته شامل SMES* ، مجله مهندسی برق دانشگاه تبریز، جلد ۴۷، شماره ۲، ۱۳۹۶.

[5] Y. J. Liu, L. Tang, S. Tong, C. P. Chen and D. J. Li, Reinforcement learning design-based adaptive tracking control with less learning parameters for nonlinear discrete-time MIMO systems, IEEE Transactions on Neural Networks and Learning Systems, vol. 26, no.1, pp. 165-176, 2015.

[6] H. B.Ammar, E., Eaton, J. M., Luna and P.Ruvolo, *Autonomous cross-domain knowledge transfer in lifelong policy gradient reinforcement learning*, International Joint Conference on Artificial Intelligence, pp. 3345-3351, 2015.

[7] A. Fachantidis, I. Partalas, G. Tsoumakas and I. Vlahavas, *Transferring task models in reinforcement learning agents*, Neurocomputing, vol. 107, pp. 23-32, 2013.

[8] M.Ghavamzadeh S.Mannor, J.Pineau and A. Tamar, *Bayesian reinforcement learning: A survey*, Foundations and Trends® in Machine Learning, vol. 8, *no.* 5-6, pp. 359-483, 2015.

[9] A. Gupta, C. Devin, Y. Liu, P. Abbeel and S. Levine, *Learning Invariant Feature Spaces to Transfer Skills with Reinforcement Learning*, arXiv preprint arXiv: 1703.02949, 2017.

[10] O. Mohammed, G.Bailly and D.Pellier, *Acquiring Human-Robot Interaction skills with Transfer Learning Techniques*, Proceedings of the Companion on Human-Robot Interaction, pp. 359-360, 2017.

[11] F. L.da Silva and A. H. R. Costa, *Accelerating Multiagent Reinforcement Learning through Transfer Learning*, Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, pp. 5034-5035, 2017.

[12] M. N. Ahmadabadi and M. Asadpour, *Expertness based cooperative Q-learning*, IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 32, no. 1, pp. 66-76, 2002.

[13] K. Ito, A. Gofuku, Y. Imoto, and M. Takeshita, *A study of reinforcement learning with knowledge sharing for distributed autonomous system*, Proceedings of the International Symposium on Computational Intelligence in Robotics and Automation, pp. 16-20, 2003.

[14] R. A. Bianchi, L. A. Celiberto, P. E. Santos, J. P.Matsuura and R. L. deMantaras, *Transferring knowledge as heuristics in reinforcement learning: A case-based approach*, Artificial Intelligence, vol. 226, pp.102-121, 2015.

[15] Y. Hou, Y. S.Ong, L. Feng and J. M. Zurada, *An Evolutionary Transfer Reinforcement Learning Framework for Multi-Agent System*, IEEE Transactions on Evolutionary Computation, vol. 21, no. 4, pp. 601-615, 2017.

[16] Y. Duan, J. Schulman, X. Chen, P. L. Bartlett, I. Sutskever and P. Abbeel, *RL^2: Fast Reinforcement Learning via Slow Reinforcement Learning*. *arXiv preprint arXiv:1611.02779, 2016*.

[17] P. Tommasino, D. Caligiore, M. Mirolli and G. Baldassarre, *A Reinforcement Learning Architecture that Transfers Knowledge between Skills when Solving Multiple Tasks*, IEEE Transactions on Cognitive and Developmental Systems, 2016.

[18] G. F. Wang, Z. Fang, P. Li and B. Li, *Transferring knowledge from human-demonstration trajectories to reinforcement*

[19] R. Glatt, F. L. da Silva and A. H. R. Costa, *Towards Knowledge Transfer in Deep Reinforcement Learning*, 5th Brazilian Conference on Intelligent Systems, pp. 91-96, 2016.

[20] L. Zhou, P. Yang, C. Chen, Y. Gao, *Multiagent reinforcement learning with sparse interactions by negotiation and knowledge transfer*, IEEE transactions on cybernetics, vol. 47, no. 5, pp. 1238-1250, 2017.

[21] T. Takano, H. Takase, H. Kawanaka and S. Tsuruoka, *Preferential exploration method of transfer learning for reinforcement learning in same transition model*, 6th International Conference on Soft Computing and Intelligent Systems and 13th International Symposium on Advanced Intelligent Systems, pp. 2099-2103, 2012.

[22] G. Konidaris, I. Scheidwasser and A. Barto, *Transfer in reinforcement learning via shared features*, Journal of Machine Learning Research, pp. 1331-1371, 2012.

[23] B. Banerjee and P. Stone, *General Game Learning Using Knowledge Transfer*. IJCAI, pp. 672-677. 2007.

[24] E. Ferrante, A. Lazaric, and M. Restelli, *Transfer of task representation in reinforcement learning using policy-based proto-value functions*, Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems, Vol. 3, pp. 1329-1332, 2008.

[25] A. Lazaric, *Knowledge transfer in reinforcement learning*, PhD thesis, Politecnico di Milano, 2008.

[۲۶] سیده ملیحه اخلاقی هاشمی‌پور، *انتقال دانش در مسائل یادگیری تقویتی با ویژگی‌های مشترک*، پایان نامه کارشناسی ارشد، دانشکده برق و کامپیوتر، دانشگاه یزد، ۱۳۹۴.

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

[28] G. Yen and T. Hickey, *Reinforcement learning algorithms for robotic navigation in dynamic environment*, ISI Transaction, vol. 43,no. 2, pp. 217-230, 2004.

[29] A. Epshteyn and G. Dejong, *Qualitative Reinforcement Learning*, Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, 2006.

[30] J. Asmuth, M. L. Littman and R. Zinkov, *Potential-based shaping in model based reinforcement learning*, Proceedings of the 23rd AAAI conference on Artificial intelligence, pp. 604-609, 2008.

[31] H. Van Hasselt, A. Guez and D. Silver, *Deep Reinforcement Learning with Double Q-Learning*, Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, pp. 1-7, 2016.

[32] Q. Wang, L. Ruan and L. Si, *Adaptive Knowledge Transfer for Multiple Instance Learning in Image Classification*, Proceedings of the 28rd AAAI Conference on Artificial Intelligence, pp. 1334-1340, 2014.

[۳۳] حسین مرادی فراهانی، جواد عسگری، *طراحی کنترل‌کننده عصبی-فازی نوع-۲*، مجله مهندسی برق دانشگاه تبریز، جلد ۴۳، شماره ۱، ۱۳۹۲.

[34] A. Belaout, F. Krim, A. Mellit, B. Talbi and A. Arabi, Multiclass adaptive neuro-fuzzy classifier and feature selection techniques for photovoltaic array fault detection and classification, Renewable Energy, vol. 127, pp. 548-558, 2018.

[35] A. Z. Kamil, S. Rustamov, M. A. Clements and E. Mustafayev, Adaptive Neuro-Fuzzy Inference System for Classification of Texts, Recent Developments and the New Direction in Soft-Computing Foundations and Applications, pp. 63-70, 2018.

[36] S. V. R. Termeh, A. Kornejady, H. R. Pourghasemi and S. Keesstra, Flood susceptibility mapping using novel ensembles of adaptive neuro fuzzy inference system and metaheuristic algorithms, Science of the Total Environment, vol. 615, pp. 438-451, 2018.

[37] D. Wang, T. He, Z. Li, L. Cao, N. Dey, A. S. Ashour, ... and F. Shi, Image feature-based affective retrieval employing improved parameter and structure identification of adaptive neuro-fuzzy inference system, Neural Computing and Applications, vol. 29, no. 4, pp. 1087-1102, 2018.

زیرنویس‌ها

[1] Reinforcement Learning
[2] Machine Learning۲
[3] Agent
[4] Smart System
[5] Transfer Knowledge
[6] Source Environment
[7] Target Environment
[8] Weighted strategy sharing
[9] Q-Learning
[10] Proto Value Function
[11] Tasks
[12] Grid World
[13] Markov Decision Problems
[14] A periodic
[15] Irreducible
[16] Sub-Clustering
[17] Guide
[18] Goal
[19] ANFIS
[20] Maze