تسریع فرایند یادگیری تقویتی با شکل‌دهی پاداش به‌کمک تحلیل گراف محیط

مریم مرعشی[1]، علیرضا خلیلیان[2] و محمد ابراهیم شیری[3]
[1]دانشگاه صنعتی امیرکبیر، marashi_maryam@aut.ac.ir
[2]دانشگاه علم و صنعت ایران، khalilian@comp.iust.ac.ir
[3]دانشگاه صنعتی امیرکبیر، shiri@aut.ac.ir

چکیده - یادگیری تقویتی، به مجموعه روش‌هایی گفته می‌شود که در آن عامل هوشمند با استفاده از تعامل پویا با محیط و دریافت سیگنال‌های تقویتی، رفتار خود را بهبود می‌بخشد. اما این فرایند اغلب بسیار زمان‌گیر، هزینه بر و گاهی پر خطر است. پاداش ساختگی، روشی موفق در افزایش سرعت یادگیری عامل در یادگیری تقویتی است. گرچه ایده اصلی این پاداش، دادن یک بازخورد عددی به غیر از پاداش محیط، به عامل یادگیرنده می‌باشد، اما چگونگی محاسبه این پاداش به نحو موثر در محیط‌های بزرگ و واقعی هنوز یک موضوع چالش برانگیز است. الگوریتم پیشنهادی در این مقاله، پاداش ساختگی جدیدی به عامل تزریق می‌کند تا بتواند سرعت یادگیری آن را افزایش دهد. این پاداش بر اساس ساخت گراف محیط، شناسایی اهداف میانی بر اساس معیار مرکزیت میانگی و شناسایی وضعیت‌های کم اهمیت با تحلیل خودکار گراف محیط تنظیم می‌شود. میزان موفقیت روش پیشنهادی روی محیط‌های محک مختلفی چون *maze* و برج هانوی آزمایش گردیده است. نتایج بدست آمده کارایی این روش را نشان می دهد.

کلید واژه- بازخورد مصنوعی ، پاداش ساختگی، یادگیری تقویتی، یادگیری کیو

1- مقدمه

یادگیری تقویتی، به مجموعه روش‌هایی گفته می‌شود که در آن عامل هوشمند با استفاده از تعامل با محیط پویا، رفتار خود را بهبود می‌بخشد. در یادگیری تقویتی، با عاملی روبرو هستیم که از طریق سعی و خطا با محیط تعامل کرده و یاد می‌گیرد تا عملی بهینه را برای رسیدن به هدف انتخاب نماید.

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 کنشی را در محیط انجام می‌دهد و وضعیت او در محیط تغییر می‌کند. عامل در ازای انجام این کنش، پاداشی دریافت می‌کند که این پاداش در بهبود کارایی رفتار عامل هوشمند به‌کار برده می‌شود. عامل در هر مرحله، کنشی را انتخاب می‌کند که در مجموع بیشترین پاداش را از محیط دریافت کند [1-2]. می‌توان ایده‌ی یادگیری تقویتی را تعامل عامل با محیط برای رسیدن به هدف دانست. اما در بسیاری از کاربردهای واقعی همانند مسابقه فوتبال، پاداش محیط با تاخیر بسیار زیادی به عامل داده می‌شود که در نتیجه عامل برای رسیدن به رفتار بهینه نیازمند صرف زمان بسیار است. راهکارهای مختلفی تا کنون برای غلبه بر این مشکل ارائه شده است [3-4-5]. یکی از راهکارها، ارائه پاداش ساختگی توسط طراح به عامل یادگیری تقویتی است [6]. ایده اصلی پاداش شکل داده شده، دادن یک بازخورد مصنوعی ایجاد شده توسط طراح به غیر از پاداش محیط، به عامل به‌منظور افزایش نرخ سرعت یادگیری است [7].

یکی از مشکلات این روش این است که تعیین مقادیر پاداش ساختگی عملاً برای محیط‌های واقعی و بزرگ غیر ممکن است. یکی از راه‌های غلبه بر این مشکل، ایجاد خودکار پاداش ساختگی می‌باشد [8-9-10]. در این مقاله رویکردی مبتنی بر گراف برای ایجاد خودکار پاداش ساختگی در یادگیری تقویتی ارائه شده است. در این رویکرد، عامل بر اساس تحلیل گراف حاصل از محیط، گلوگاه‌های محیط به‌عنوان وضعیت‌های موثر در رساندن عامل به هدف، ارزش بالایی داده شده و به وضعیت‌هایی که قرارگیری عامل در آنها، تنها منجر به افزایش زمان اکتشاف می‌گردد به‌عنوان وضعیت‌های کم اهمیت، ارزش کمتری داده می‌شود. در نهایت تمامی این ارزش‌های برآوردی به شکل پاداش

شکل داده به یادگیری تقویتی داده می‌شود.

ادامه مقاله بـه شـرح زیـر اسـت. در بخـش دوم بـه فرآینـد تصمیم‌گیری مارکوف و یادگیری تقویتی پرداخته شده اسـت، در بخش سوم شکل‌دهی پاداش در یادگیری تقویتی معرفی می‌شود، در ادامه در بخش چهارم روش پیشـنهادی ارائـه شـده اسـت، در بخش پنج آزمایشات تجربی و نتایج آن آورده شده است. تحلیـل و نتیجه گیری هم در بخش ششم ارائه شده است.

2- معرفی

یک نمـایش رسـمی بـرای مـدل نمـودن مسـائل یـادگیری تقویتی، فرآیند تصمیم‌گیری مارکوف است که به‌طور گسترده در محیط‌های گسسته به‌کار گرفته شده است.

2-1- فرایند تصمیم‌گیری مارکوف

فرآیندهای تصمیم‌گیری مارکوف (MDP)، که اولـین بـار در سال 1957 توسط بلمن معرفی گردید [11]، به‌صورت مجموعه پنج تایی (S, A, T, R, α) نشان داده مـی‌شـود کـه اجـزای ایـن مجموعه به‌صورت زیر تعریف می‌شود: S، مجموعه وضـعیت‌هـای ممکن در محیط، A، مجموعه کنش‌های ممکن برای یـک عامـل در هر مرحله از تصمیم‌گیری، T، نمایانگر احتمال گـذر عامـل از وضـعیت s بـه وضـعیت s' اسـت بـه شـرط انتخـاب عمـل a از مجموعه‌ی اعمال ممکن، $R(s,a,s')$، پاداش دریافتی عامل بعـد از انتخاب عمل a و انتقال از وضعیت s به وضعیت s' از محـیط، و α یک فاکتور کاهنده.

2-2- یادگیری تقویتی

در یادگیری تقویتی، هدف یافتن یک سیاسـت بهینـه بـرای مساله مدل شده به یکی از انواع فرآینـد مـارکوف اسـت. یکـی از راه‌های پیـدا کـردن سیاسـت بهینـه، اسـتفاده از الگـوریتم‌هـای برنامه‌نویسی پویـا اسـت. الگـوریتم‌هـای یـادگیری کیـو [12] و یادگیری سارسا [13] از معروف‌ترین روش‌های یادگیری تقویتی هستند. در الگوریتم یادگیری کیو در حالت گسسته بـه هـر زوج وضعیت-کنش یک مقدار $Q(s, a)$، معـرف مجمـوع پـاداش‌هـای دریافتی عامل به شرط شروع از حالت s، انجام کنش a در ادامـه پیروی از سیاست موجود، نسبت داده می‌شود. تـابع $Q(s, a)$، تـا زمانی که به مقدار بهینه همگرا شـود بـا اسـتفاده از فرمـول (1) بروزرسانی می‌شود:

$$Q_{t+1}(s,a) \leftarrow (1-\alpha)\, Q_t(s,a) + \alpha \left[r(s,a) + \gamma max_{a' \in A_s} Q_t(s',a') \right] \quad (1)$$

3- شکل‌دهی پاداش در یادگیری تقویتی

توجه الگوریتم‌های موجـود در بحـث یـادگیری تقـویتی بـه یافتن سیاستی با بیشترین مجموع پاداش اکتسابی است. بنابراین می‌توان گفت تابع پاداش به‌طور ضمنی رفتار بهینه را برای عامل توصیف می‌کند. لازم به ذکر است که گرچـه پـاداش شـکل داده شده ابزار قدرتمندی در افزایش سرعت یادگیری عامل است، امـا به همان نسبت به‌کارگیری غلط آن مـی‌توانـد در گمـراه نمـودن عامل تاثیر شایانی داشته باشد [14]. بررسی‌های مختلف صورت گرفته روی چگونگی تعریف و مقداردهی پاداش ساختگی از سال 1992 را می‌تـوان بـه دو دسـته تعیـین و مقـدار دهـی توسـط متخصص حوزه و برآورد آن به صورت خودکار تقسیم‌بندی نمود.

3-1- استفاده از دانش متخصص حوزه

در سال 1996، Bishop از ایده شبکه عصبی چند لایه برای مقـداردهی بـه پـاداش سـاختگی اسـتفاده کـرد [15]. در سـال 2003، بـر اسـاس ایـده Laud & Dejong [16] عامـل تنهـا قسمتی از محیط که از نظر فرد خبره ، احتمال بیشـتری جهـت رسیدن به هدف و یافتن سیاست بهینه دارد را مورد بررسی قـرار می‌دهد. سپس Wiewiora [17] در سال 2003 به‌صورت عملی ثابت کرد که اگر و تنها اگر پاداش شکل داده شده به‌صورت تابع پتانسیلی به صورت فرمول (2) در اختیـار عامـل قـرار گیـرد، در نهایت به سیاست بهینه همگرا خواهد شد.

$$F(s, s') = \gamma\, Q(s') - Q(s) \quad (2)$$

در سال 2005، Abbeel و Ng [18] روشی ارائه نمود کـه در آن عامـل بـا مشـاهده رفتـارهـای نشـان داده شـده توسـط متخصص محیط و اجتناب از تعامل با محیط تنها بـا اسـتفاده از دانش قبلی جهت تعیین چگـونگی کـار، نـه چگـونگی اکتشـاف، سعی در رسیدن به هدف و یافتن راه حل بهینه می‌کند.

3-2- محاسبه پاداش شکل داده شده به‌صورت خودکار

در سال 2007، Marthi برای اولین بـار بحـث شـکل‌دهـی اتوماتیـک پـاداش را مطـرح کـرد [19]. او الگـوریتمی معرفـی می‌کند که مجموعه وضعیت‌های ممکن محیط را به‌عنوان ورودی دریافت کرده و وضعیت‌هایی که عملکرد مشابه‌ای درخصوص یک وظیفه دارند را تجرید می‌نماید و بـر اسـاس آنهـا تـابع پتانسـیل جدیـدی تعریـف کـرده و از آن در سـاختار شـکل‌دهـی پـاداش استفاده می‌کند. در سـال 2010، Grzes و Kudenko [20] بـه ارائه راه‌کاری بـرای یـادگیری تـابع پتانسـیل بـه مـوازات فرآینـد

یادگیری تقویتی بر اساس الگوریتم R-max پرداخته‌اند. الگوریتم ارائه شده به ساختن و تعریف مدلی پویا از محیط برای یـادگیری تابع پتانسیل به‌صورت بر خط می پردازد. اما با توجه به وابستگی این تابع در تمامی الگوریتم‌های پیشنهادی، به دانـش متخصـص حوزه و یا ارزش وضعیت‌ها بر اساس دانش یادگرفته شده‌ی عامل از اکتشاف محیط، هیچ‌کدام نتوانستند تأثیر به‌سزایی در افـزایش سرعت یادگیری به‌خصوص در محیط‌های بزرگ و واقعـی داشـته باشند.

4- روش پیشنهادی برای تخمین پاداش ساختگی

در این بخش الگوریتم جدید پیشنهادی برای تخمـین تـابع پاداش ساختگی، مسـتقل از دانـش متخصـص حـوزه و اکتشـاف عامل در محیط ارائه می‌شود. در قسمت اول، نحوه نگاشت محیط به گـراف و تحلیـل آن‌را بیـان کـرده، در قسـمت دوم الگـوریتم پیشنهادی به همراه توضیح آن آورده شده اسـت. ایـن الگـوریتم مبتنی بر گراف است و در آن تـابع پـاداش سـاختگی جدیـد بـر اساس ارزش بدست آمده از بررسی و تحلیل گـراف، یـافتن زیـر گراف‌های مستقل، شناسایی گره‌های گلوگاه (اهداف میانی) و نیز گره های کم اهمیت، تعریف می‌شود.

4-1- نگاشت محیط به گراف و تحلیل آن

برای ساخت گراف غیر جهـت‌دار و بـدون وزن $G(V, E)$ بـر مبنای فرآیند تصمیم‌گیری مارکوف، هر وضعیت $s \in S$ به‌عنـوان یک گره v در گراف و هر انتقال $s \to s'$ در تابع انتقال فرآیند، بـه عنوان یک یال از $v \to v'$ در نظر گرفته می‌شود. سپس با تحلیـل خودکار گراف، ارزش تقریبی اکتشاف هر وضعیت برآورد شـده و تابع پاداش شکل داده شده بر اساس آن بدست می‌آید. ایـن امـر باعث می‌گردد تـا مقـداردهی پـاداش شـکل داده شـده نـه تنهـا وابسته به دانش متخصص حوزه که قطعاً ضریب خطـای بـالایی دارد نباشد، بلکه مستقل از ارزش به‌دست آمده از اکتشاف محیط توسط عامل بوده و عامل نیـازی بـه یـادگیری و بـرآورد آن نیـز نخواهد داشت.

قدم اول ایـن تحلیـل، شناسـایی اهـداف میـانی بـه عنـوان وضعیت‌هایی از محیط است که رسیدن به آنها برای عامل مفیـد خواهد بود. برای این منظور، روش مبتنی بر گراف بـا اسـتفاده از خاصیت مرکزیت میانگی گراف بکارگرفته شـده اسـت [21]. بـا شناسایی این نقاط و تعریف پاداش مجازی بیشتر بـرای آنهـا، در قدم دوم، به شناسایی نقاط کم اهمیت (کمتـرین همسـایگی بـه شرط هدف میانی نبودن) بر اسـاس مـاتریس مجـاورت گـراف و تخصیص ارزش کمتر می پـردازد. در نهایـت ارزش مجـازی هـر وضعیت مشخص شده و بر اساس آن ماتریس پـاداش سـاختگی، ساخته می‌شود. هرچند تزریق این پاداش، به‌تنهـایی بـه فرمـول یادگیری کیو تاثیر چشم‌گیـری در بـالا بـردن سـرعت یـادگیری عامل دارد، اما آزمایشات ثابت کرده‌اند که ترکیب این پـاداش بـا تابع پتانسیل، تاثیر به‌سزایی در سرعت یـادگیری عامـل موجـب می‌گردد که در قسمت نتایج نشان داده شده است.

4-2- معرفی الگوریتم

با تحلیل گراف حاصل از محیط، به‌صورت خودکار ارزش مجازی برای رسیدن به هر وضعیت محاسبه می‌گردد. این پاداش مجازی به‌دست آمده که به عنوان $Manual\ Reward(s, a)$ تعریف شده است، همانطور که در الگوریتم شکل 1 دیده می‌شود، با یک ضریب کاهنده در فرمول یادگیری کیو قرار می‌گیرد که باعث می‌گردد به مرور زمان و کسب دانش عامل از محیط، تأثیر خود را از دست بدهد؛ تا چنانچه خطایی در ارزش‌دهی خودکار وضعیت‌ها رخ دهد منجر به گمراه نمودن و به هدف نرسیدن عامل نشود و الگوریتم همچنان همگرا باقی بماند. فرمول جدید تعریف شده برای به‌روزرسانی مقدار $Q(s, a)$ عبارتست از:

$$Q(s, a) = Q(s, a) + \alpha * [r + \max_{a'} Q(s', a') - Q(s, a) + \beta * manual\ Reward(s, a) + \gamma * \max_{a'} Q(s', a') - Q(s, a)] \quad (3)$$

$$\beta = e^{-\frac{t^2}{\sigma^2}}$$

همانطور که در فرمول (3) دیده می‌شود، با توجه به اعمـال ضریب کاهنده بر پـاداش سـاختگی، عامـل در اپیزودهـای اولیـه کاملاً تصادفی حرکت کرده؛ اما به مرور زمان و با کسـب دانشـی قابل قبول از محیط، حرکات خود را به صورت حریصانه بر مبنای بیشترین پاداش اکتسابی انتخاب خواهد کرد.

عامل در اپیزودهای اولیه پاداشـی از طـرف محـیط دریافـت نمی‌کند. لذا الگوریتم پیشنهادی مقـدار پـاداش سـاختگی را بـه صورت خودکار از تحلیل گراف به‌دست می‌آورد و به عامل تزریق می‌نماید. عامل از این پاداش مصنوعی بـرای بهبـود رفتـار خـود استفاده می‌کند تا عملیات اکتشاف مورد نظـر بـرای شناسـایی و ارزش دهی به وضعیت‌ها کاهش یابد. بـا توجـه بـه اعمـال تـابع پتانسیل و عدم نیاز به منتظر ماندن عامل تا لحظـه رسـیدن بـه هدف نهایی و نیز دریافت بازخورد از محیط، ایـن امـر مـی‌توانـد سرعت یادگیری را افزایش دهد. با توجه به ضریب تعریـف شـده برای این پاداش، علاوه بر افزایش سرعت یادگیری عامـل، نقـش

دانش یادگرفته شده عامل از اکتشاف محیط و همگرایی الگوریتم یادگیری تقویتی حفظ می‌شود.

5- آزمایش‌ها و ارزیابی

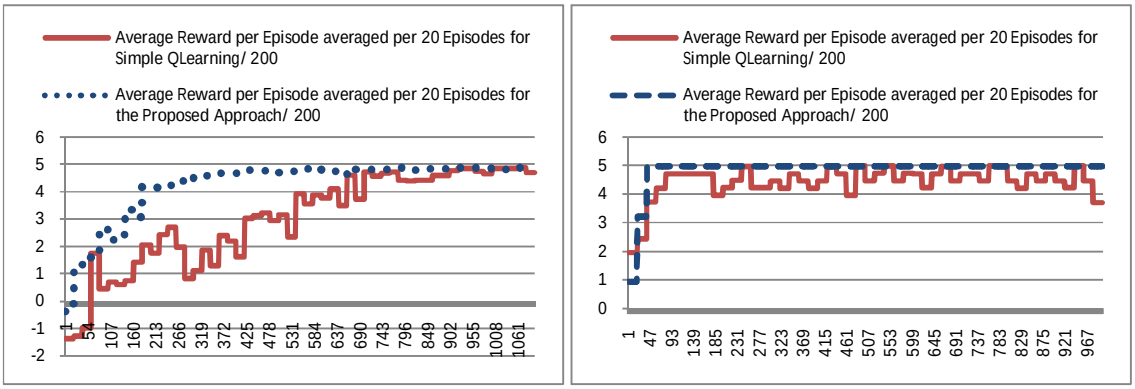
مؤثر بودن الگوریتم پیشنهادی در افزایش سرعت یادگیری و نرخ همگرایی، با انجام آزمایش روی محیط محک maze با 625 حالت و برج هانوی با 27 حالت نشان داده شده است. برج هانوی یک محیط دارای ویژگی غیر ناوبری است. در حالیکه محیط maze یک محیط ناوبری شمرده می‌شود. محیط ناوبری محیطی است که عامل یک شی را از نقطه الف به نقطه ب انتقال می‌دهد. در آزمایش‌های انجام شده مقادیر زیر مورد استفاده قرار گرفته است: α=0.5, γ=0.9, σ=0.9. همچنین مقدار پاداش و جریمه برای اهداف میانی و حالت‌های کم اهمیت 1 درصد پاداش هدف در نظر گرفته شده است. علامت مقدار پاداش اهداف میانی مثبت و علامت مقدار جریمه حالت کم اهمیت منفی است. آزمایش‌های هر دو محیط برای 2000 اپیزود انجام شده‌اند. همچنین، در 50 اپیزود اول شبیه‌سازی از روش اکتشاف تصادفی [1-21] و در سایر اپیزودها از اکتشاف با روش ε-greedy [1-21] استفاده شده است. از آنجائیکه نمودار روش پیشنهادی در هر دو آزمایش حدوداً از اپیزود 1000 به بعد همگرا شده است، برای صرفه‌جویی در فضا و نمایش بهتر، نمودار دو محیط فقط تا اپیزود 1000 رسم شده‌اند.

شکل 2 نتایج شبیه‌سازی‌ها را به‌صورت میانگین مجموع پاداش دریافتی و همگرایی به سیاست بهینه بر حسب تعداد اپیزودها به‌ترتیب روی محیط‌های آزمایشی برج هانوی (قسمت الف) و maze (قسمت ب) نشان داده‌اند. در این شکل‌ها نتایج روش پیشنهادی با نتایج یادگیری کیو مقایسه شده است. طبق آنچه در شکل‌ها مشاهده می‌شود، سرعت همگرایی توسط روش پیشنهادی به‌طور قابل ملاحظه‌ای بهبود یافته است. برای تعیین اینکه آیا بهبود به‌دست آمده توسط روش پیشنهادی از لحاظ آماری معنی‌دار است، از یک روش آماری به‌نام *آزمون فرض برای تفاضل دو میانگین* استفاده شده است [22]. با استفاده از فرمول (4)، مقدار z را محاسبه کرده و در جدولی از مقادیر بحرانی جستجو می‌کنیم. با محاسبه روی داده‌های آزمایش، مقدارهای 12/08 و 7/75 به‌ترتیب در محیط‌های maze و برج هانوی برای z محاسبه شد. با جستجو در جدول مقادیر بحرانی [22]، فرض صفر با احتمال بیشتر از 99/99 درصد برای هر دو محیط رد می‌شود. رد فرض صفر دلالت می‌کند بر اینکه بهبود به‌دست آمده توسط روش پیشنهادی از نظر آماری معنی‌دار است.

$$\frac{\overline{X_1}-\overline{X_2}-\delta}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} \qquad (4)$$

input: The matrices *Q*, *reward*, and *next* for a given environment
output: The final *r* value for each episode of the simulation
declare:
Beta*(episode)*: A function to reduce the effect of the domain expert reward values, it is defined as follows: exp[- (episode * episode) / (sigma * sigma)] in which sigma is a constant value between 0 and 1.
Modify(*manualReward*): A function that finds sub goals and bad states automatically, and then modifies the *manualReward* matrix according the these special states.
algorithm QLearningWithShapingAndManualReward **begin**
Initialize *manualReward(s, a)* with *reward*(*s*, *a*)
Modify(*manualReward*)
for each episode **do begin**
Initialize *s*, e.g. randomly
r := 0
repeat (for each step, action, of episode)
Choose *a* from *s* using policy derived from *Q* (e.g., ε-greedy)
s' := *next*(*s*, *a*)
r := *r* + *reward*(*s*, *a*)
$Q(s, a) := Q(s, a) + \alpha * [r + \max_{a'} Q(s', a') - Q(s, a) + \text{Beta}(episode) * manualReward(s, a) + \gamma * \max_{a'} Q(s', a') - Q(s, a)]$
s := *s'*
until *s* is terminal **or** the number of actions exceeds a threshold
Store the number of current episode and the corresponding final *r* value.
endfor
end QLearningWithShapingAndManualReward

شکل 1: شبه کد الگوریتم پیشنهادی



(ب) محیط maze با 625 حالت (الف) محیط برج هانوی سه حلقه‌ای با 27 حالت

شکل 2: مقایسه میانگین پاداش دریافتی در هر 20 اپیزود و تعداد 1000 اپیزود شبیه‌سازی برای دو روش یادگیری کیو (خط صاف) و روش پیشنهادی (خط نقطه چین)

6- نتیجه‌گیری

در این مقاله، روشی جدید با استفاده از تحلیل گراف محیط و به‌دست آوردن ماتریس پاداش مجازی به‌عنوان پاداش ساختگی پیشنهاد شد. آزمایش‌های تجربی انجام شده افزایش سرعت یادگیری و نرخ همگرایی را توسط روش پیشنهادی نشان می‌دهد. با توجه به آنکه مدل‌سازی هر محیط به گراف امکان‌پذیر بوده و زمانی چندجمله‌ای را می‌طلبد، به نظر می‌رسد این الگوریتم روی اغلب محیط‌های بزرگ و واقعی قابل پیاده‌سازی خواهد بود.

مراجع

[1] S.Sutton & A.G.Barto, Reinforcement Learning : An Introduction, 1998

[2] L.P. Kaelbling, et al. Reinforcement Learning :A Survey ,Journal Of Artificial Intelligence Research, vol., page 237-285, 1996

[3] M. J. Mataric. Reward functions for accelerated learning. In Proceedings of the 11th International Conference on Machine Learning(ICML), pages 181-189, 1994.

[4] A.Epshteyn and G.Dejong, Qualitative Reinforcement Learning, Appearing in Proceedings of the 23 rd International Conference on Machine Learning, Pittsburgh, PA, 2006.

[5] Andrew Y.Ng, Shaping and police search in reinforcement learning, PhD Thesis, University of California, Berkeley, 2003

[6] A.LAUD, Theory And Application Of Reward Shaping In Reinforcement Learning, PhD Thesis, University of Illinois at Urbana-Champaign, 2004

[7] P.Alstrom and J.Randlov, learning to drive a bicycle using reinforcement learning and shaping. In Proceedings of the 15th international conference on machine learning(ICML), Pages 463 – 471, 1998

[8] B.Marthi, Automatice shaping and Decomposition of Reward function, Proceedings of the 24th international conference on Machine learning (ICML), Pages 601 – 608 , 2007

[9] J.Asmuth And M.L.Littman And R.Zinkov, Potential-based shaping in model based reinforcement learning, Proceeding AAAI'08 Proceedings of the 23rd national conference on Artificial intelligence - Volume 2 , Pages 604-609, 2008

[10] Marek Grzes, Daniel Kudenko, Learning Shaping Rewards in Model-based Reinforcement Learning, Proc. of AAMAS 2009 ,Workshop on Adaptive Learning Agents, ALA 2009

[11] L.P. Kaelbling, et al., Reinforcement Learning : A Survey ,Journal Of Artificial Intelligence Research, vol.4, pp.237-285, 1996

[12] Watkin, Watkin Proof of Q-learning Convergence, 1992

[13] G. A. Rummery and M. Niranjan, "On-Line Q-Learning Using Connectionist Systems," Cambrige University Engineering Depatrment, 1994.

[14] Randløv and p.Alstrom, Learning to drive a bicycle using reinforcement learning and shaping, J. (Ed.), In Proceedings of the 15th international conference on machine learning, pages 463-471,Morgan Kaufmann,CA., 1998

[15] C. M. Bishop, Neural networks for pattern recognition. Oxford University Press., 1996

[16] A.Laud, and G.Dejong, The influence of reward on the speed of reinforcement learning: An analysis of shaping. Proceedings of the 20th International Conference on Machine Learning (ICML), pages 440-447, 2003

[17] Wiewiora, E., Potential-based shaping and Q-value initialization are equivalent. Journal of Artificial Intelligence Research., page 205-208, 2003

[18] Pieter Abbeel and Andrew Y.Ng, Exploration and apprenticeship learning in reinforcement learning, Appearing in Proceedings of the 22 th International Conference on Machine Learning (ICML), pages1-8, 2005

[19] B.Marthi, S.Russell, Automatice shaping and Decomposition of Reward function., In Proceedings of the 24th International Conference on Machine Learning(ICML), pages 601-608, 2007.

[20] Marek Grze, Daniel Kudenko., Online learning of shaping rewards in reinforcement learning., Department of Computer Science, University of York, York, YO10 5DD, UK., 2010

[21] P.Moradi, PhD Thesis, Amirkabir University Iran, 2011

[22] J. E. Freund, *Mathematical statistics*, 5th ed., Prentice-Hall, 1992