دانشگاه صنعتی امیرکبیر
(پلی‌تکنیک تهران)
دانشکده مهندسی کامپیوتر و فناوری اطلاعات

پایان‌نامه کارشناسی ارشد هوش مصنوعی

عنوان :
یادگیری تقویتی مبتنی بر نقشهٔ خودسازمان‌ده تطبیقی با زمان

توسط:
حسام منتظری

استاد راهنما:
دکتر رضا صفابخش

بهار ۸۵

بسمـه تعالـــی

دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)
معاونت پژوهشی

فرم اطلاعات پایان نامه
کارشناسی ارشد و دکترا

تاریخ: .....

پیوست: .....

نام و نام خانوادگی:	حسام منتظری	**دانشجوی آزاد** ■	**بورسیه** □	**معادل** □
شماره دانشجوئی:	۸۳۱۳۱۲۳۱	**دانشکده:** مهندسی کامپیوتر	**رشته تحصیلی:** هوش مصنوعی	
نام و نام خانوادگی استاد راهنما:	دکتر رضا صفابخش			
عنوان پایان نامه به فارسی:	یادگیری تقویتی مبتنی بر نقشهٔ خودسازمان‌ده تطبیقی با زمان			
عنوان پایان نامه به انگلیسی:	Reinforcement Learning using Time-Adaptive Self Organizing Map			
پنوع پروژه:	**کارشناسی ارشد** ■ / **دکترا** □	**کاربردی** ■	**بنیادی** □	**توسعه ای** □ **نظری** □
تاریخ شروع:	۸٤/۸/۱	**تاریخ خاتمه:** ۸۵/۱۰/۳۰	**تعداد واحد:** ۶	
سازمان تأمین کننده اعتبار:	-			
واژه های کلیدی به فارسی:	یادگیری تقویتی، نقشهٔ خودسازمان‌ده رشدیابنده، نقشهٔ خودسازمان‌ده تطبیقی با زمان، کنترل ترافیک			
واژه های کلیدی به انگلیسی:	Reinforcement Learning, Growing SOM, TASOM, Traffic Light Control			

نظرها و پیشنهادها به منظور بهبود فعالیت های پژوهشی دانشگاه:

استاد راهنما:

دانشجو:

امضاء استاد راهنما: **تاریخ:**

نسخه ۱: معاونت پژوهشی
نسخه ۲: کتابخانه و به انضمام دو جلد پایان نامه به منظور تسویه حساب با کتابخانه و مرکز اسناد و مدارک علمی

تقدیم به

پدر و مادر مهربان

و

همسر عزیزم

که در همه حال پشتیبان و یاور من بوده‌اند.

از استاد راهنمای خود، جناب آقای دکتر رضا صفابخش، به پاس راهنمایی‌های گرانقدر ایشان در طول مدت تحصیل در دورهٔ کارشناسی ارشد و بویژه طی انجام این پژوهش، کمال تشکر را دارم.

چکیده

یادگیری تقویتی، نگاشت وضعیت‌ها به عمل‌ها با هدف ماکزیمم کردن سیگنال پاداش دریافتی را بررسی می‌کند. در این نوع یادگیری، به عامل گفته نمی‌شود که چه عملی را انتخاب کند، بلکه عامل باید عملی را انتخاب کند که پاداش دریافتی از محیط را بیشینه کند. در چالش برانگیزترین حالات، پاداش عمل‌ها بلافاصله مشخص نمی‌شود. یادگیری تقویتی، از یک سو دارای پشتوانهٔ قوی از قضایا و اثبات‌های ریاضی است؛ و از سویی دیگر، این روش در مسائل مختلفی همچون مسیریابی ربات، اجتناب از مانع، تصمیم-گیری در بازی‌ها، مسائل مهارت‌ها در روبوکاپ، و کنترل ترافیک به طور موفق عمل کرده است. یکی از مسائل مهمی که در مورد این روش مطرح می‌شود، بسط و توسعه روش به مسائلی با فضای وضعیت پیوسته است. برای حل مسائل با فضای وضعیت پیوسته، روش-های مختلفی مانند شبکه‌های عصبی پرسپترونی چند لایه، کیمک، درخت‌های تصمیم، و نقشه‌های خود سازمان‌ده ارائه شده است. نشان داده شده است که یادگیری تقویتی با استفاده از نقشه‌های خود سازمان‌ده استاندارد در حل بسیاری از مسائل با فضای وضعیت پیوسته و حتی فضای عمل پیوسته موفق بوده‌اند. اما نقشهٔ خودسازمان‌ده استاندارد نمی‌تواند یک تابع هدف متغیر را به خوبی ارائه کند و برای توابع هدفی که با توپولوژی نقشه همخوانی ندارد مورد استفاده واقع نمی‌شود. در این پایان‌نامه، یادگیری تقویتی مبتنی بر نقشهٔ خودسازمان‌ده تطبیقی برای حل مشکل تابع هدف متغیر ارائه شده است. تابع هدف متغیر در یادگیری تقویتی منحصر به داده-های فضای ورودی نیست، بلکه داده‌هایی که به عنوان ورودی نقشهٔ خروجی داده می‌شود همیشه توزیع چگالی متغیر با زمان دارد. باید توجه داشت در یادگیری تقویتی عامل با گذشت زمان عملکرد خود را بهبود می‌دهد، در نتیجه داده‌های ورودی به نقشهٔ خروجی با گذشت زمان تغییر می‌کنند و توزیع چگالی آن ناایستا است. روش دیگری که در این پایان‌نامه ارائه شده است روش یادگیری تقویتی با استفاده از نقشهٔ خودسازمان‌ده رشدیابنده است. این روش برای حل هر دو مشکل ذکر شده ارائه شده است. ترکیب یادگیری تقویتی با نقشه‌های خودسازمان‌ده رشدیابنده به سادگی امکان‌پذیر نیست و ترکیب این نوع نقشه با الگوریتم‌هایی که جدول کیو آن در طول زمان ثابت است، میسر نمی‌باشد. در این پایان‌نامه الگوریتم جدیدی مبتنی بر نقشه‌های خودسازمان‌ده رشد یابنده ارائه شده است که جدول کیو آن در طول زمان بزرگ و کوچک می‌شود. نشان داده شده است این الگوریتم در حل مسائل مختلف از بقیهٔ روش‌ها موفق‌تر بوده است.

فهرست مطالب

فهرست اشکال

فهرست جدول‌ها

فصل اول

مقدمه

تکنیک‌های یادگیری ماشین معمولاً به سه دستهٔ اصلی تقسیم می‌شوند: یادگیری با نظارت، یادگیری بی‌نظارت، و یادگیری تقویتی. یادگیری با نظارت به نمونه‌هایی از ورودی و خروجی تابع هدف (داده‌های آموزشی) نیازمند است. در یادگیری بی‌نظارت دادهٔ هدف وجود ندارد و فقط بر روی داده‌های ورودی پردازش صورت می‌گیرد. یادگیری تقویتی از نظر ارزیابی از سوی محیط، در حد فاصل دو روش قبلی است. یادگیری تقویتی مانند روش بانظارت خروجی صحیح را دریافت نمی‌کند، اما سیگنالی را از محیط دریافت می‌کند که صرفاً خوب بودن انجام عمل قبلی را نشان می‌دهد. اثبات شده است که بسیاری از روش‌های یادگیری تقویتی همیشه به سیاست بهینه همگرا خواهند شد [۳، ۲۸]. از جنبهٔ کاربردی نیز روش یادگیری تقویتی در مسائل مختلفی همچون اجتناب از مانع در مسیریابی ربات‌ها [۲۴]، کنترل ترافیک [۱، ۴] تصمیم‌گیری در بازی‌ها [۳۲]، مهارت‌های عامل‌ها در لیگ‌های روبوکاپ [۲۷]، و غیره به طور گسترده مورد استفاده قرار گرفته است.

یکی از مسائل مهمی که در مورد یادگیری تقویتی مطرح می‌شود، حل مسائل با فضای وضعیت و عمل پیوسته است. بیشتر نظریه‌هایی که در مورد یادگیری تقویتی است، مبتنی بر فرض گسسته بودن فضای وضعیت است. برای حل مسائل پیوسته یادگیری تقویتی، روش‌های مختلفی همچون شبکه‌های عصبی پرسپترونی چند لایه [۱۴،۳]، کیمک [۲۶،۲۸،۲]، و نقشه‌های خود سازمان‌ده مطرح می‌شود [۲۴، ۳۱].

این پایان‌نامه استفاده از نقشه‌های خودسازمان‌ده در یادگیری تقویتی را مورد توجه قرار داده است. کارهای مختلفی در این زمینه صورت گرفته است. Q-Kohon روش تقویتی مبتنی بر این نقشه است که توسط توزئت ارائه شده است [۱۸]. در Q-Kohon جدول وضعیت-عمل با نقشه خودسازمان‌ده جایگزین شده است. هر نورون شامل سه‌گانه وضعیت، عمل، و ارزش کیو است. در این روش، انتخاب نورون برای اجرا مبتنی بر مقدار وضعیت و ارزش کیو انجام می‌شود، اما انتخاب نورون برای یادگیری بر اساس مقدار وضعیت و عمل است. روش دیگری که برای حل مسأله یادگیری تقویتی از نقشه خودسازمان‌ده استفاده کرده است، روش نقشه موتوری (Motoric Map) است [۳۱]. در این روش، یک نقشه خودسازمان‌ده در ورودی قرار می‌گیرد. هر نورون این نقشه شامل وزن عمل و مقدار کیو است. در نقشه موتوری تخمین مقدار کیو با استفاده از روش سارسا(sarsa) انجام می‌شود و عمل بهینه متناظر هر نورون با استفاده از روشی به نام یادگیری کوواریانسی محاسبه می‌گردد.

مهمترین کاری که در این زمینه صورت گرفته است، روش اسمیت است که از یادگیری کیو به عنوان تکنیک یادگیری تقویتی استفاده می‌کند و فضای ورودی و خروجی را به طور مجزا با استفاده از نقشه‌های خودسازمان‌ده استاندارد گسسته می‌کند [۲۴، ۲۵]. جدول کیو بر روی فضای ورودی و خروجی گسسته شده تشکیل می‌شود و مطابق با الگوریتم استاندارد یادگیری کیو(Q-Learning) بروز می‌شود. نقشهٔ ورودی با ورودی‌هایی که از محیط می‌گیرد، به روز می‌شود. اما برای بروزرسانی نقشهٔ خروجی دادهٔ مستقیمی وجود ندارد. اسمیت با مطرح کردن ایدهٔ نویزی کردن عمل اولیه (نورون برنده در فضای خروجی)، مشکل کاوش در فضای عمل را حل نموده است. نشان داده شده است که عملکرد این روش در کاربردهایی از رباتیک موفق بوده است [۲۵] .

کاری که در این پروژه صورت گرفته است، توسعه‌ای بر روش اسمیت است. همانگونه که گفته شد روش اسمیت از دو نقشهٔ خودسازمان‌ده استاندارد برای ارائه فضای ورودی و خروجی استفاده می‌کند. اما نقشهٔ خودسازمان‌ده استاندارد از سویی نمی‌تواند تابع هدف متغیر را به خوبی ارائه کند و از سویی قادر نیست توابع هدفی که با توپولوژی نقشه همخوانی ندارد را ارائه کند. در این پایان‌نامه، یادگیری تقویتی مبتنی بر نقشهٔ خودسازمان‌ده تطبیقی برای حل مشکل تابع هدف متغیر ارائه شده است. تابع هدف متغیر در یادگیری تقویتی منحصر به داده‌های فضای ورودی نیست، بلکه داده‌های ورودی به نقشهٔ خروجی همیشه توزیع چگالی متغیر با زمان دارد. باید توجه داشت در یادگیری تقویتی عامل با گذشت زمان عملکرد خود را بهبود می‌دهد، در نتیجه داده‌های ورودی به نقشهٔ خروجی با گذشت زمان تغییر می‌کنند و توزیع چگالی آن ناایستا است. روش یادگیری تقویتی با استفاده از نقشهٔ خودسازمان‌ده رشدیابنده برای حل هر دو مشکل ذکر شده ارائه شده است. ترکیب یادگیری تقویتی با نقشه‌های خودسازمان‌ده رشدیابنده به سادگی امکان‌پذیر نیست و ترکیب این نوع نقشه با الگوریتم اسمیت که جدول کیو آن در طول زمان ثابت است، امکانپذیر نیست. در این پایان‌نامه الگوریتم جدیدی مبتنی بر نقشه‌های خودسازمان‌ده رشد یابنده ارائه شده است که جدول کیو آن در طول زمان بزرگ و کوچک می‌شود. نشان داده شده است این الگوریتم در حل مسائل مختلف از بقیهٔ روش‌ها موفق‌تر بوده است.

نتایج حاصل از اعمال روش‌های پیشنهادی بر روی مسائل کلاسیکی از رباتیکز بررسی شده است. همهٔ روش‌ها در مسائل ایستا عملکرد تقریباً مناسبی دارند. اما روش رشدیابنده نسبت به دیگر روش‌ها عملکرد بهتری دارد. از سویی میانگین پاداش دریافتی این روش بیشتر از دیگر روش‌ها است و از سوی دیگر روش نسبت به دیگر روش‌ها پایدارتر است. نقشه‌های ورودی و خروجی در روش رشدیابنده بهتر از روش‌های دیگر ارائه کنندهٔ توزیع ورودی هستند. در مسائل ناایستا، یادگیری تقویتی با استفاده از نقشهٔ خودسازمان‌ده استاندارد نسبت به دیگر روش‌ها عملکرد بدتری دارد. این روش پایداری کمی در برابر تغییرات محیط دارد. اما دو روش یادگیری تقویتی با نقشهٔ خودسازمان‌ده تطبیقی و رشدیابنده نسبت به تغییرات محیط پایدارتر هستند و عملکرد قابل قبولی را ارائه می‌کنند.

مسألهٔ کنترل چراغ راهنما در قلمرو ترافیک شهری کاربرد دیگری است که برای ارزیابی مدل‌های پیشنهادی استفاده شده است. مشکل ترافیک از مسائل پیچیدهٔ شهری است که به عوامل مختلفی همچون تردد بیش از حد خودرو از جاده‌ها، عدم کنترل صحیح چراغ‌های راهنما، تصادفات، عدم پیروی رانندگان و عابران پیاده از قوانین راهنمایی و رانندگی، ورودی‌های خیابان‌های فرعی به اصلی، و دوربرگردان‌ها بستگی دارد. جهت انجام آزمایش‌ها یک شبیه‌ساز ترافیکی پیاده‌سازی شد که وضعیت‌های مختلف ترافیکی همانند ترافیک سبک و سنگین را ایجاد کند. استفاده از روش‌های پیشنهادی این پایان‌نامه در حل مسألهٔ کنترل چراغ راهنما نشان داد که از ترکیب چندین عامل یادگیری تقویتی می‌توان برای حل مسائل پیچیده استفاده نمود. در بین روش‌های بررسی شده، عملکرد مدل رشدیابنده در سه وضعیت ترافیکی سبک، متوسط، و سنگین از روشهای دیگر بهتر است. لازم به ذکر است عملکرد مدل تطبیقی نیز در این مسأله قابل قبول است و کارایی و قابلیت اطمینان این روش در وضعیت‌های مختلف ترافیکی از روش اسمیت بهتر است.

این پایان‌نامه از هشت فصل تشکیل شده است. فصل دوم به معرفی روش یادگیری تقویتی می‌پردازد. در این فصل مسألهٔ یادگیری تقویتی و روش‌های کلاسیک حل آن بررسی می‌شود. نقشه‌های خودسازمان‌ده و انواع آن در فصل سوم شرح داده می‌شوند. در این فصل نقشه‌های خودسازمان‌ده استاندارد، تطبیقی با زمان، و رشدیابنده ارائه می‌شوند. فصل چهارم پیشینهٔ پژوهش در ترکیب یادگیری تقویتی با نقشه‌های خودسازمان‌ده را بررسی می‌کند. مدل‌های پیشنهادی ترکیب یادگیری تقویتی با نقشهٔ خودسازمان‌ده تطبیقی و نقشهٔ خودسازمان‌ده رشدیابنده در فصل پنجم مورد بررسی واقع می‌شوند. فصل‌های ششم و هفتم نتایج تجربی مدل‌ها را بررسی می‌کند. فصل ششم دو کاربرد از رباتیکز که در کارهای قبلی مورد توجه قرار گرفته است را بررسی می‌کند. در فصل هفتم مسألهٔ ترافیکی

معرفی و نتایج حاصل از مدل‌های پیشنهادی ارائه و مقایسه می‌شود. این پایان‌نامه در فصل هشتم با بیان نتیجه‌گیری و پیشنهادات به پایان می‌رسد.

فصل دوم

یادگیری تقویتی

در این فصل به معرفی و بیان روش یادگیری تقویتی می‌پردازیم. در ابتدای این فصل، مسألهٔ یادگیری تقویتی[1] را تعریف می‌کنیم. در ادامه، روش تفاضل زمانی[2] معرفی می‌شود. روش تفاضل زمانی یکی از مهمترین روش‌های حل مسأله یادگیری تقویتی است که در قسمت دوم این فصل مورد بررسی قرار می‌گیرد.

۲-۱ مسأله یادگیری تقویتی

در این قسمت به بیان مسألهٔ یادگیری تقویتی می‌پردازیم. از آنجایی که همهٔ روش‌های حل این مسأله از جمله روش‌های یادگیری تقویتی به شمار می‌روند، این مسأله تعریف کنندهٔ شاخهٔ یادگیری تقویتی است. در این قسمت همچنین حالت ایده‌آل ریاضی مسألهٔ یادگیری تقویتی را بیان می‌کنیم. در انتهای این قسمت اشاره‌ای به روابط عناصر اصلی آن مانند توابع ارزش و معادلات بلمن می‌شود.

۲-۱-۱ رابط[3] عامل و محیط

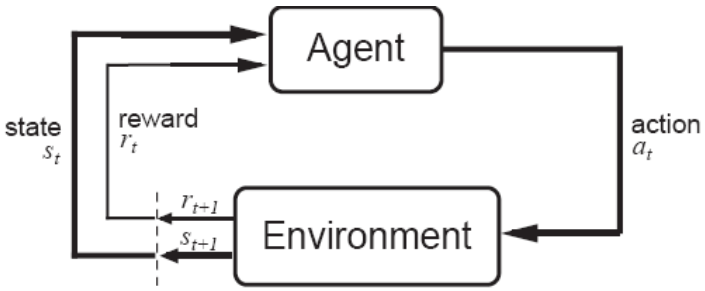
مسألهٔ یادگیری تقویتی، چارچوب مناسبی را برای یادگیری در حین تعامل (با محیط) و رسیدن به هدف ایجاد می‌کند. یادگیرنده یا تصمیم گیرنده را عامل گویند و آن چیزی که با آن تعامل می‌شود، شامل هر چیزی خارج از عامل، محیط نامیده می شود. این تعامل به طور پیوسته صورت می‌پذیرد و عامل عملی را انتخاب می‌کند و محیط پاسخ آن را می‌دهد و عامل را در وضعیت جدیدی قرار می-دهد. محیط همچنین پاداش[4] عمل‌های عامل را می‌دهد. پاداش مقدار عددی است که عامل باید آن را در طول زمان ماکزیمم کند [۲۸].

عامل و محیط در بازه‌های زمانی گسسته $t = 0, 1, 2, \ldots$ با یکدیگر تعامل می‌کنند و در هر بازه زمانی t، عامل وضعیتی[5] (s_t) را از محیط دریافت می‌کند که $s_t \in S$ و S همهٔ وضعیت‌های ممکن محیط است و بر مبنای آن عملی[6] (a_t) را انتخاب می‌کند که $a_t \in A(s_t)$ و $A(s_t)$ مجموعه عمل‌هایی است که در وضعیت s_t موجود می‌باشند. در بازه زمانی بعد، و در نتیجهٔ عملش، عامل مقدار عددی $r_{t+1} \in R$ را دریافت می‌کند و در وضعیت جدید s_{t+1} قرار می‌گیرد. در شکل ۲-۱ چگونگی تعامل عامل و محیط ترسیم شده است.

در هر بازه زمانی، عامل نگاشتی از وضعیت‌ها به احتمال انتخاب هر عمل ممکن را انجام می‌دهد. این نگاشت سیاست[7] عامل نامیده و با π_t داده می‌شود که $\pi(s,a)$ احتمال انتخاب عمل a در وضعیت sاست و یا احتمال $a_t = a$ در وضعیت $s_t = s$ است. روش‌های یادگیری تقویتی چگونگی تغییر سیاست عامل در اثر کسب تجربه را بررسی می‌کنند.

---

[1] Reinforcement Learning Problem
[2] Temporal Difference
[3] Interface
[4] Reward
[5] state
[6] action
[7] policy



شکل ۲-۱ تعامل عامل و محیط در یادگیری تقویتی [۲۸]

این چارچوب کاملاً مجرد و انعطاف‌پذیر است و می‌تواند به مسائل مختلف و با روش‌های گوناگون اعمال شود. برای مثال، نیازی به ثابت بودن طول بازه‌های زمانی نیست و بازه‌ها می‌توانند مراحل متوالی و دلخواه تصمیم‌گیری و عمل‌کردن باشند. اعمال می‌توانند کنترل سطح پایین، مانند ولتاژ اعمالی به بازوی ربات، یا تصمیم‌گیری سطح بالا مانند صرف نهار یا رفتن به دانشگاه باشند. به طور مشابه، وضعیت‌ها می‌توانند اشکال گوناگونی داشته باشند. وضعیت‌ها می‌توانند با ورودی‌های سطح پایین، همچون دریافت‌های مستقیم حس‌گر[۱]، و یا سطح بالای مجرد مانند توصیف سمبلیک یک شیء در اتاق توصیف شوند. وضعیت‌ها می‌توانند بر مبنای حافظه‌ای از دریافت‌های گذشته و یا حتی کاملاً ذهنی باشند. برای مثال، عامل می‌تواند در وضعیتی باشد که مکان شیء خاصی را نداند. به همین ترتیب بعضی از عمل‌ها کاملاً ذهنی یا محاسباتی هستند. مثلاً بعضی از عمل‌ها کنترل می‌کنند که عامل درباره چه چیزی فکر کند. به طور کلی، عمل‌ها می‌توانند هر تصمیمی باشند که می‌خواهیم چگونگی اتخاذ آن را یاد بگیریم، و وضعیت‌ها هر چیزی که در اتخاذ چنین تصمیمی مفید هستند می‌باشند.

وضعیت‌ها و عمل‌ها از کاربردی تا کاربرد دیگر تغییر می‌کنند و چگونگی ارائه[۲] آن‌ها تاثیر زیادی در کارایی دارد. در یادگیری تقویتی همانند دیگر روش‌های یادگیر، انتخاب نوع ارائه بیش از آن که علم باشد، هنر است.

۲-۱-۲ هدف‌ها و پاداش‌ها

در یادگیری تقویتی، هدف[۳] عامل براساس سیگنال پاداش که از محیط دریافت می‌کند فرموله شده است. در هر بازهٔ زمانی، پاداش، عددی حقیقی ($r_t \in R$) می‌باشد و هدف عامل ماکزیمم کردن مجموع پاداش‌ها در طول زمان است که به این معنی است که پاداش بلاواسطه را ماکزیمم نمی‌کند، بلکه مجموع پاداش دریافتی در طول زمان[۴] را ماکزیمم می‌کند[۱۱].

استفاده از سیگنال تقویتی، ویژگی متمایز یادگیری تقویتی است. در ابتدا ممکن است این روش فرموله کردن، محدود کننده به نظر برسد، ولی در عمل اثبات شده است که کاملاً انعطاف‌پذیر و عملی است. برای اینکه چگونگی استفاده از سیگنال تقویتی کاملاً مشخص شود به مثال‌های زیر توجه شود. برای مثال، رباتی می‌خواهد فرار از ماز[۵] را یاد بگیرد. به این منظور، پاداش همیشه مقدار صفر دارد مگر اینکه عامل از ماز خارج شود که پاداش ۱+ می‌گیرد. روش دیگر پاداش دهی مسألهٔ ماز عبارت است از پاداش ۱- در هربازهٔ زمانی و پاداش صفر در صورت فرار از ماز. که این نحوهٔ پاداش دهی باعث فرار ربات از ماز در کمترین زمان ممکن می‌شود. برای عاملی که

[1] sensor
[2] representation
[3] goal
[4] long run
[5] maze

می‌خواهد شطرنج یاد بگیرد. پاداش‌دهی معمول عبارت است از: ۱+ برای برد، ۱- برای باخت، و پاداش صفر در صورت تساوی و همهٔ حالت‌های غیر پایانی دیگر.

در مثال‌های بالا، عامل ماکزیمم کردن مقادیر پاداش را یاد می‌گیرد. اگر بخواهیم که عامل مسأله‌ای را حل کند باید پاداش‌دهی را به گونه‌ای تعریف کنیم که ماکزیمم شدن آن در طول زمان، باعث حل مسأله شود بنابراین تعیین نوع پاداش‌دهی مناسب برای حل مسأله کاملاً ضروری است. یکی از نکات دیگر پاداش‌دهی این است که سیگنال‌های تقویتی نباید دانش ما را به عامل منتقل کنند و همچنین نباید چگونگی رسیدن به هدف را مشخص نمایند ، بلکه تنها باید هدف و آن چیزی را که عامل می‌خواهد به آن برسد را نشان دهند.

نکتهٔ دیگری که وجود دارد این است که محاسبهٔ پاداش در خارج از عامل و در محیط صورت می‌گیرد. به عبارت دیگر پاداش‌دهی خارج از کنترل عامل است و عامل نباید قادر به تغییرات این سیگنال باشد. البته عامل می‌تواند پاداش‌های داخلی برای خود تعریف کند(بسیاری از روش‌های یادگیری تقویتی از چنین پاداش‌هایی استفاده می‌کنند.)

۲-۱-۳ بازگشت[1]

گفته شد هدف عامل، ماکزیمم کردن مجموع پاداش دریافتی در طول زمان است، اما چگونه آن را به صورت فرمول نشان می‌دهیم؟ اگر دنبالهٔ پاداش‌های دریافتی بعد از بازه زمان t را r_{t+1}، r_{t+2}، r_{t+3} و ... تعریف کنیم، بطور کلی می‌خواهیم مقدار بازگشت مورد انتظار[2] را ماکزیمم کنیم. بازگشت دارای تعاریف مختلفی است که در ساده‌ترین حالت عبارت است از:

$$R_t = r_{t+1} + r_{t+2} + r_{t+3} + \ldots + r_T$$

که T بازهٔ زمانی پایانی اپیزود است. این تعریف در کاربردهای عملی، بسیار مورد استفاده قرار می‌گیرد. زیرا معمولاً ارتباط عامل و محیط به زیردنباله‌هایی از بازه‌ها تقسیم می‌شوند که اپیزود[3] نام دارند. مثالی از آن انجام بازی فرار از ماز، و یا هر تعامل تکراری دیگر است. هر اپیزود در وضعیت خاصی که وضعیت پایانی نامیده می‌شود، پایان می‌یابد و ابتدای اپیزود بعدی، در وضعیت شروع استانداردی قرار می‌گیرد. مسائل با ویژگی‌های فوق را مسائل اپیزودیک گویند. بعضی اوقات در مسائل اپیزودیک مجموعهٔ وضعیت‌های غیر پایانی را با S و مجموعهٔ وضعیتهای غیر پایانی و پایانی را با S^+ نشان می‌دهند. غیر از تعامل اپیزودیک، تعامل غیر اپیزودیک عامل و محیط نیز وجود دارد که در [۱۱، ۲۸] شرح داده شده است و ارتباط این دو حالت تعامل با محیط مورد بررسی قرار گرفته است.

نرخ کاهندگی[4] مفهوم دیگری است که بایستی تعریف شود. اگر از نرخ کاهندگی در تعریف بازگشت استفاده شود، عامل عمل‌هایی را انتخاب می‌کند که مجموع پاداش‌های کاهش یافته[5] که در آینده دریافت می‌کند را ماکزیمم کند. در این حالت رابطهٔ بازگشت عامل به صورت زیر تعریف می‌شود:

$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \ldots = \sum_{k=0}^{T} \gamma^k r_{t+k+1}$$

---

[1] return
[2] expected return
[3] episode
[4] discounting
[5] discounted

که پارامتر γ، $(0\leq\gamma\leq1)$ نرخ کاهندگی است. نرخ کاهندگی، مقدار فعلی پاداش‌های آینده را مشخص می‌کند. پاداشی که در k بازهٔ زمانی بعد دریافت می‌شود γ^{k-1} برابر پاداشی است که مستقیماً دریافت می‌شود.

۲-۱-۴ خصوصیت مارکف

مناسب است که سیگنال وضعیتی، همهٔ دریافت‌های گذشته را خلاصه کند و همهٔ اطلاعات مربوط به آنها را نگه‌داری کند. به صورت عادی برای تحقق این مورد، نیاز به چیزی بیش از یک دریافت مستقیم است. البته این نیاز فراتر از کل دریافت‌های گذشته نمی‌شود. سیگنال وضعیتی که موفق در نگه‌داری همهٔ اطلاعات گذشته می‌باشد، دارای خصوصیت مارکف است. برای مثال، مکان مهره‌های شطرنج دارای خصوصیت مارکف است. زیرا دنبالهٔ کل حرکاتی که به این وضعیت رسیده است را خلاصه می‌کند. اگرچه اکثر اطلاعاتی که منجر به این وضعیت شده است از دست رفته‌اند، ولی همهٔ نکات مهم برای ادامهٔ بازی نگه‌داری شده است. به این خصوصیت گاهی خصوصیت مستقل از مسیر[1] هم گویند که به این معنا است که وضعیت فعلی، مستقل از مسیر و تاریخچهٔ سیگنال‌های گذشته است.

با فرض تعداد محدود وضعیت و مقادیر پاداشها، تعریف ریاضی خصوصیت مارکف را بیان می‌کنیم. توجه کنید که یک محیط بطور کلی در زمان t+1 به عمل زمان t پاسخ می‌دهد و در کلی‌ترین حالت، این پاسخ ممکن است به همهٔ پاسخ‌های گذشته ارتباط داشته باشد. در این حالت دینامیک محیط با توزیع احتمال کامل زیر مشخص می‌کنیم:

$$\Pr\{s_{t+1}=s', r_{t+1}=r \mid s_t, a_t, r_t, s_{t-1}, a_{t-1}, \ldots, r_1, s_0, a_0\} \qquad (۱-۲)$$

برای همهٔ r, s ها و همهٔ مقادیر ممکن رویدادهای گذشته: $s_0, a_0, r_1, s_1, a_1, \ldots, r_t, a_t, s_t$. ولی در حالتی که سیگنال وضعیت، دارای خصوصیت مارکف باشد، پاسخ محیط در زمان t+1 فقط بستگی به وضعیت و عمل در زمان t دارد که در این حالت دینامیک محیط فقط با رابطهٔ زیر بیان می‌شود:

$$\Pr\{s_{t+1}=s', r_{t+1}=r \mid s_t, a_t\} \qquad (۲-۲)$$

برای همهٔ مقادیر s', r, s_t, a_t. به عبارت دیگر می‌گوییم سیگنال وضعیت دارای خصوصیت مارکف است، در صورتی که دو رابطهٔ ۲-۱ و ۲-۲ برای همهٔ مقادیر r, s و گذشته‌های $s_0, a_0, r_1, s_1, a_1, \ldots, r_t, a_t, s_t$ معادل یکدیگر باشند. در این حالت اصطلاحاً گفته می‌شود که محیط دارای خصوصیت مارکف است. محیط‌هایی که دارای خصوصیت مارکف هستند باعث سادگی حل مسأله می‌شوند. حتی برای سیگنال وضعیتی که خصوصیت مارکف ندارد می‌توانیم فرض کنیم که سیگنال تقریبی از خصوصیت مارکف را دارد.

۲-۱-۵ فرآیند تصمیم‌گیری مارکف

نمونه‌ای از مسأله یادگیری تقویتی که دارای خصوصیت مارکف است را فرآیند تصمیم‌گیری مارکف[2] گویند. اگر فضای وضعیت‌ها و اعمال محدود باشند، آنگاه فرآیند تصمیم‌گیری مارکف متناهی[3] است. فرآیند تصمیم‌گیری مارکف متناهی برای درک روش یادگیری تقویتی بسیار مهم می‌باشد. فرآیند تصمیم‌گیری مارکف متناهی توسط مجموعه وضعیت‌ها و اعمالش در دینامیک یک مرحله بعد محیط تعریف می‌شود. برای هر وضعیت و عمل داده شدهٔ r, s، احتمال انتقال به وضعیت بعدی s' برابر است با:

$$P_{ss'}^{a} = \Pr\{s_{t+1}=s' \mid s_t=s, a_t=a\}$$

---

[1] independence of path
[2] Markov Decision Process (MDP)
[3] finite MDP

این مقادیر را احتمال‌های انتقال گویند. بطور مشابه برای هر وضعیت و عمل فعلی r,s و با وضعیت بعدی s'، امید ریاضی مقدار حالت بعدی عبارت است:

$$R^a_{ss'} = E\{r_{t+1} | s_t = s, a_t = a, s_{t+1} = s'\}$$

این مقادیر $P^a_{ss'}$ و $R^a_{ss'}$، کاملاً ویژگی‌های مهم دینامیک فرآیند تصمیم‌گیری مارکف متناهی را بیان می‌کنند.

۲-۱-۶ توابع ارزش

همهٔ الگوریتم‌های یادگیری تقویتی، بر مبنای تخمین توابع ارزش[1] هستند. توابع ارزش، توابعی از وضعیت‌ها (یا زوج‌های وضعیت عمل) هستند که خوب بودن عامل در وضعیتی خاص را تخمین می‌زنند (یا خوب بودن انجام عمل داده شده در وضعیتی خاص را تخمین می‌زنند). خوب بودن را با پاداش‌های دریافتی در آینده تعریف می‌کنیم. واضح است پاداش‌هایی که عامل، انتظار دریافت آن را در آینده دارد وابسته به عمل فعلی است. بنابراین، توابع ارزش با توجه به سیاست انتخابی تعریف می‌شوند.

همانگونه که ذکر شد π نگاشتی از وضعیت $s \in S$ و عمل $a \in A(s)$ به احتمال انتخاب عمل a در وضعیت s یا $\pi(s,a)$ است. ارزش وضعیت s در سیاست π، $V^\pi(s)$، امید ریاضی بازگشت است در صورتی که حالت اولیهٔ s باشد و سیاست π در ادامهٔ اپیزود انتخاب شود. تابع ارزش وضعیت برای فرآیند تصمیم‌گیری مارکف به صورت زیر تعریف شده است:

$$V^\pi(s) = E_\pi\{R_t | s_t = s\} = E_\pi\{\sum_{k=0}^{T} \gamma^k r_{t+k+1} | s_t = s\}$$

که $E_\pi\{\}$ نشان‌دهندهٔ امید ریاضی بازگشت در صورت اتخاذ سیاست π است. توجه کنید که ارزش وضعیت‌های پایانی در صورت وجود، صفر می‌باشد. V^π را تابع ارزش وضعیت برای سیاست π می‌نامیم.

بطور مشابه، $Q^\pi(s,a)$ ارزش انتخاب عمل a در وضعیت s در صورت اتخاذ سیاست π است. در واقع $Q^\pi(s,a)$ امید ریاضی بازگشت است در صورتی که از وضعیت s شروع و عمل a را انتخاب کنیم و پس از آن سیاست π را پیش بگیریم.

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

Q^π را تابع ارزش عمل برای سیاست π می‌نامیم.

می‌توان توابع ارزش V^π و Q^π را با استفاده از تجربه، تخمین زد. برای مثال، فرض کنید عامل سیاست π را اتخاذ کرده باشد و برای هر وضعیتی که مواجه می‌شود میانگینی از بازگشت واقعی که پس از آن وضعیت دریافت می‌کند را نگه دارد. در صورت زیاد شدن تعداد دفعات مشاهدهٔ وضعیت، این میانگین به ارزش وضعیت $V^\pi(s)$ همگرا خواهد شد. اگر میانگین‌های مجزایی برای هر زوج وضعیت-عمل نگه‌داری شود، این میانگین‌ها به مقادیر عمل $Q^\pi(s,a)$ همگرا خواهند شد. به این گونه روش‌های تخمین، روش‌های مونت کارلو [۱۱، ۲۸] گویند. البته، اگر تعداد وضعیت‌ها زیاد شود، آنگاه عملی نیست که برای هر وضعیت مستقلاً، میانگینی نگه‌داری شود. در مقابل V^π و Q^π را به صورت توابعی پارامتری بیان می‌کنیم و پارامترها را به گونه‌ای که بهتر با بازگشت‌های مشاهده شده منطبق باشند، تعیین می‌کنیم.

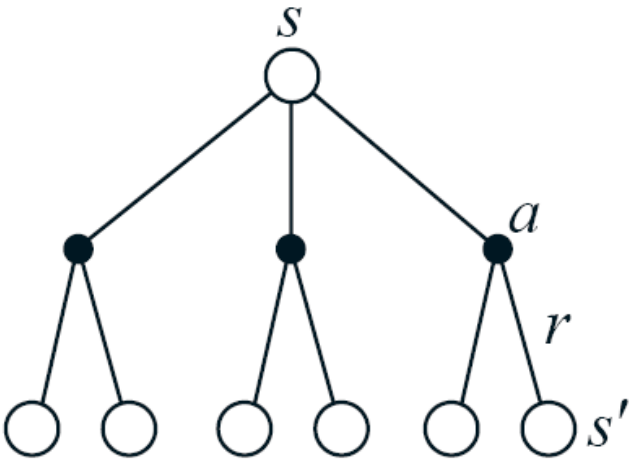
دارا بودن رابطهٔ بازگشتی، یکی از خصوصیت اساسی توابع ارزش می‌باشد. برای هر سیاست π و هر وضعیت s، رابطهٔ زیر بین ارزش وضعیت s و ارزش وضعیت‌های ممکن بعدی آن وجود دارد:

---

[1] value function

$$
\begin{aligned}
V^{\pi}(s) &= E\{R_t \mid s_t = s\} = \\
&= E_{\pi}\{\sum_{k=0}^{T} \gamma^k r_{t+k+1} \mid s_t = s\} \\
&= E_{\pi}\{r_{t+1} + \gamma \sum_{k=0}^{T} \gamma^k r_{t+k+2} \mid s_t = s\} \\
&= \sum_{a} \pi(s,a) \sum_{s'} P^{a}_{ss'} [R^{a}_{ss'} + \gamma E_{\pi}\{\sum_{k=0}^{T} \gamma^k r_{t+k+2} \mid s_{t+1} = s'\}] \\
&= \sum_{a} \pi(s,a) \sum_{s'} P^{a}_{ss'} [R^{a}_{ss'} + \gamma V^{\pi}(s')]
\end{aligned}
\qquad (۳-۲)
$$

که $a \in A(s)$ و $s' \in S^{+}$ است. معادلهٔ (۲-۳) را معادلهٔ بلمن[1] برای V^{π} گوییم که رابطهٔ بین ارزش یک وضعیت و ارزش وضعیتهای بعدی آن را بیان می‌کند. تغییر از یک وضعیت به وضعیت‌های ممکن بعدی در شکل ۲-۲ نشان داده شده است. در این شکل هر دایرهٔ سفید نمایانگر یک وضعیت و هر دایرهٔ سیاه نشان دهندهٔ زوج وضعیت-عمل است. با شروع از وضعیت s، گره ریشه در شکل ۲-۲، عامل می‌تواند مجموعه‌ای از عمل‌ها را انتخاب کند (سه عمل در شکل نشان داده شده است). با انتخاب هر کدام از عمل‌ها، محیط به یکی از وضعیت‌های بعدی، s'، می‌رود و پاداش r را به عامل می‌دهد. معادلهٔ بلمن (۲-۳) میانگین وزن‌دار همهٔ حالات ممکن و احتمال وقوع را به صورت وزن دار بیان می‌کند. ارزش وضعیت اولیه برابر امید ریاضی (کاهش‌یافته) وضعیت بعدی، به اضافهٔ امید ریاضی پاداش دریافتی تا زمان بعدی است. این نمودار را نمودار پشتیبان[2] می‌گویند و برای شرح تصویری الگوریتم‌ها استفاده می‌شود.



شکل ۲-۲ نمودار پشتیبان برای V^{π}

این نمودار را از آن جهت پشتیبان گویند که شامل عملیات پشتیبان یا به روز رسانی می‌باشد که هستهٔ اصلی روش‌های یادگیری تقویتی است. عمل پشتیبان ارزش وضعیت‌های بعدی را به وضعیت قبلی منتقل می‌کند.

[1] Bellman equation
[2] backup