دانشگاه خوارزمی تهران

دانشکده فنی و مهندسی

پایان نامه کارشناسی ارشد مهندسی کامپیوتر،گرایش علوم تصمیم و مهندسی دانش

عنوان :

یادگیری تقویتی در فضای پیوسته و تقریب تابع به وسیله شبکه عصبی خودسازمان ده DIGNET

دانشجو :

نجمه علی بابایی

استاد راهنما :

دکتر میرمحسن پدرام

پاییز ۱۳۹۱

یادگیری تقویتی درفضای پیوسته

و تقریب تابع به وسیله شبکه عصبی خودسازمان‌ده

DIGNET

ای دو جهان از قلمت یک رقم

بی قلمت لوح دو عالم عدم

در کف من مشعل توفیق نه

ره به نهانخانه تحقیق ده

شمع زبانم سخن افروز ساز

شام من ازصبح سخن روز ساز

تقدیم به پدر ومادر

بزرگوارانی که همواره

مدیون زحماتشان

شرمنده محبتشان

ومنتظر دعای خیرشان

هستم.

با سپاس ازسه وجود مقدس:

آنان که ناتوان شدند تا ما به توانایی برسیم...

موهایشان سپید شد تا ما روسفید شویم...

و عاشقانه سوختند تا گرمابخش وجود ما و روشنگر راهمان باشند...

پدر

مادر

استاد

چکیده

یادگیری تقویتی عبارت است از قالب بندی یک مسئله به فرم یادگیری از طریق تعامل برای رسیدن به هدف. زمانی که فضای حالات و یا کنش‌ها پیوسته و یا خیلی بزرگ شود استفاده از عناصر حافظه برای نگه‌داری ارزش حالات بسیار زیاد خواهد شد. این مسئله در رابطه با زمانی که در طول یادگیری ارزش کنش‌ها به دست آورده می‌شود، بحرانی تر خواهد شد. علاوه بر حافظه مصرفی مسئله ، داده و زمان لازم برای پر کردن آنها نیز مهم است. بنابراین مسئله تعمیم پیش خواهد آمد.روش پیشنهاد شده برای مسائل یادگیری تقویتی با فضای پیوسته‌ای از حالت‌ها و کنش‌ها مناسب است.

در این روش از شبکه عصبی خودسازمان ده *DIGNET* برای نمایش فضای حالت و کنش استفاده شده‌است.استفاده از این شبکه سبب می‌شود در یک فضای پیوسته، عامل بتواند با استفاده از یک حافظه مصرفی مناسب ،میزان داده و زمان قابل قبول به هدف دست یابد.در پیشنهاد این پایان نامه مفاهیم اساسی یک مسئله یادگیری تقویتی و همچنین یک شبکه عصبی خودسازمان‌ده DIGNET بررسی می‌شود، و ساختار این شبکه به‌عنوان یک روش خوشه‌بندی کارا برای حل چالش پیوستگی فضای حالات و کنش‌ها ،در مسئله یادگیری تقویتی پیشنهاد می‌شود و کارایی آن بررسی و مقایسه می‌شود.

کلمات کلیدی : یادگیری ماشین ،یادگیری تقویتی، مسیرهای شایستگی ،تعمیم، یادگیری Q ، شبکه‌عصبی‌خودسازمان ده DIGNET

فهرست مطالب

فهرست جدول‌ها، شکل‌ها و نشانه‌های اختصاصی

فصل ۱ : مقدمه

چالش اساسی یک تیم روبوکاپ بر سر توسعه فوتبالیستهای مستقل و خودمختار است . گاهی هدف اصلی یک مساله مشخص است اما راه رسیدن به آن دشوار و یا معلوم نیست در این مواقع یادگیری تقویتی به عنوان یک ابزار مناسب می‌تواند با اینگونه مسائل روبرو شود. ایده اصلی این تئوری برپایه جعبه پرپیچ وخم[۱] است.

هدف اساسی یک تیم روبوکاپ توسعه رباتهایی است که بتواند در مسابقه فوتبال روباتها برنده جام جهانی شوند. در این زمینه روش‌های زیادی استفاده شده است و یکی از آنها یادگیری تقویتی است. ربات به وسیله این روش می‌تواند خود مختاری[۲] و استقلال را از طریق تجربیات خود بیاموزد. این نوع یادگیری بر پایه پاداش است همانند روشی که برای تعلیم سگها و آموزش ترفندها به آنان به کار می‌برند.

از جمله روشهای حل یک مسئله یادگیری تقویتی می‌توان به برنامه‌ریزی‌پویا[۳] ، روش‌های مونت‌کارلو[۴] و یادگیری اختلاف زمانی[۵] اشاره کرد. هر یک از این روش‌ها مزایا و معایب خاص خود رادارند. مثلاً نقطه قوت روش‌های برنامه‌ریزی‌پویا خود را در توسعه آن به‌صورت ریاضی نشان می‌دهد درحالی‌که نقطه ضعف آن ریشه در نیاز آن به یک مدل کامل و دقیق از محیط دارد.

۱–۱ : هدف و دستاوردها

زمانی که فضای حالات و یا عمل‌ها پیوسته و یا خیلی بزرگ شود استفاده از عناصر حافظه برای نگه داری ارزش حالات بسیار زیاد خواهد شد و این مسئله در رابطه با زمانی که ارزش کنش‌ها به دست می‌آید، بحرانی‌تر خواهد شد. علاوه بر حافظه مصرفی مسئله ، داده و زمان لازم برای پر کردن آنها نیز مهم است.

[1] Maze
[2] Autonomous
[3] Dynamical programming
[4] Monte Carlo
[5] Temporal Difference Learning

بنابراین مسئله تعمیم پیش خواهد آمد. آیا می‌توان با تعداد محدودی حالت یک مسئله را تجربه کرد و نتایج را به مسئله بزرگتر تعمیم داد و تقریبی برای مقادیر به دست آورد.تعمیم در این بحث همان ترکیب روش‌های تعمیم و روش‌های یادگیری تقویتی است.

در مبحث یادگیری تقویتی ، تعمیم به عنوان مسئله ای سخت و شاق مطرح می‌شود. به این ترتیب زمانی‌که یادگیری تقویتی در بسیاری از زمینه‌ها به کار می‌رود، حالاتی پیش می‌آید که پیش از این با آنها برخوردی صورت نگرفته است.این حالت زمانی رخ می‌دهد که فضاهای کنش یا حالت دارای متغیرهای پیوسته و ظواهر پیچیده باشند و تنها راه برای یادگیری این است که عامل بتواند اطلاعاتی که پیش از این آموخته است را برای حالاتی که تا به حال ندیده است ، تعمیم دهد.

در مساله پیش رو پویایی محیطی موجود معلوم نیست.بنابراین از الگوریتم‌های برنامه ریزی پویا نمی‌توان استفاده کرد.دوم اینکه تاحد امکان هم متغیرهای حالت پیوسته و هم متغیرهای کنش پیوسته درنظرگرفته شوند.سوم اینکه با توجه به اینکه در الگوریتم‌هایی مثل مونت کارلو و... که پویایی محیط در آنها مشخص نیست ممکن است عامل با حالاتی مواجه شود که پیش از این هرگز با آنها مواجه نشده است. پس باید به‌گونه ای عمل کرد که در هر لحظه با مواجه شدن با موقعیت جدید ، عامل بتواند نتایج حاصل از تجربه این موقعیت را به دانش خود بیفزاید.روش پیشنهاد شده برای مسائل یادگیری تقویتی با فضای پیوسته‌ای از حالت‌ها و کنش‌ها مناسب است.در این روش از شبکه عصبی خودسازمان ده *DIGNET* برای نمایش فضای حالت و کنش استفاده شده است.استفاده از این شبکه سبب می‌شود در یک فضای پیوسته، عامل بتواند با استفاده از یک حافظه مصرفی مناسب ،میزان داده و زمان قابل قبول به هدف دست یابد.

۱–۲: مروری بر فصل‌های رساله

محورهای اصلی مطرح شده در این پایان نامه بر سه قسمت کلی استوار است. درقسمت اول بررسی روش‌های موجود در یادگیری تقویتی و جزئیات مربوط به آن است، تئوری یادگیری تقویتی مطرح شده در این مکتوب براساس کتاب" *Reinforcement Learning: An introduction*"(۱) است. قسمت دوم شبکه عصبی

خودسازمان‌ده *DIGNET* را معرفی و نکات موجود در آن را مطرح می‌کند و در نهایت قسمت سوم ترکیب دو قسمت پیش را برای به‌دست آوردن روش پیشنهای ما و پیاده سازی و نتایج کار را نشان می‌دهد.

درفصل دوم پس از معرفی مفاهیم اولیه سعی می‌شود به جزئیات روش‌های برنامه‌ریزی‌پویا برای حل یک مسئله یادگیری تقویتی پرداخته شود. پیش از آن خاصیت مهم مارکوف[6] به عنوان یک پیش نیاز برای تئوری‌های بعدی معرفی می‌شود .به علاوه المان‌های کلیدی در ساختار ریاضی مسئله نظیر توابع مقدار[7] و معادله بلمن[8] نیز توضیح داده خواهد شد. به علاوه روش تکرار سیاست[9] برای به‌دست آوردن توابع ارزش و اساس روش یادگیری از اختلاف زمانی عنوان خواهد شد.

در فصل سوم به معرفی شبکه عصبی خودسازمان‌ده *DIGNET* و چند اصل ریاضی به کار رفته در آن ، تشریح روال اصلی این روش در ساختار خوشه‌بندی الگوها ، پایداری ،همگرایی ، ظرفیت و معیارهای شباهت مختلفی که می‌تواند در آن به کار رود پرداخته می‌شود.در فصل چهارم نیز ساختار شبکه عصبی *DIGNET* را درون چند مسئله یادگیری تقویتی به کار برده و ضمن ارائه روش پیشنهادی، نتایج به‌دست آمده ارزیابی می‌شود.درنهایت در فصل پنجم به جمع بندی و پیشنهادها پرداخته می شود.

[6] Markov
[7] Value Functions
[8] Bellman Equation
[9] Policy Iteration

فصل ۲: مروری بر یادگیری تقویتی و شبکه عصبی خود سازمان ده DIGNET

۲–۱ : مقدمه

یادگیری ماشین[10] علمی‌است که در آن سعی بر طراحی وگسترش الگوریتم‌هایی است که از طریق آنها یک ماشین بتواند یاد بگیرد.(۲) انواع روشهای یادگیری در یادگیری ماشین عبارتند از :

۱. یادگیری نظارت شده[11]
۲. یادگیری بدون نظارت[12]
۳. یادگیری تقویتی[13]

۱. یادگیری نظارت شده :

در این نوع یادگیری ، تعدادی ورودی و خروجی معلوم وجود دارد و در طی این یادگیری نگاشتی از ورودی‌ها به خروجی انجام می‌شود سپس رفتار مناسب تثبیت می‌شود و عناصر یادگیری سعی می‌کنند که این رفتار را برای موارد مشاهده نشده کپی کنند(۲). یک مثال برای این روش الگوریتم پس انتشار خطا[14] می‌باشد که در شبکه‌های پرسپترون[15] برای آموزش ، مورد استفاده قرار می‌گیرد .

۲. یادگیری بدون نظارت:

در این نوع یادگیری ، سعی بر یافتن ارتباطات موجود بین ورودی‌ها است و تنها از ورودی‌ها برای ارائه رفتار خروجی استفاده می‌کنند. نمونه‌هایی از این نوع یادگیری در مساله خوشه‌بندی[16] و طبقه‌بندی و تخمین ابعادو مواردی همچون تشخیص چهره[17] مشاهده می‌شود. این روش برای تفسیر و پردازش

[10] Machine learning
[11] Supervised learning
[12] Unsupervised learning
[13] Reinforcement Learning
[14] Error Backpropagation
[15] Perceptron Network
[16] Clustering
[17] Face recognition

مقادیر زیادی داده به کار رفته و عنوان می‌شود که هنگامی‌که بتوان حواس پنجگانه دریافتی را تفسیر و پردازش کرد به نوعی فرمی از هوشمندی وجود دارد(۲).یک مثال برای این روش الگوریتم کوهنن[۱۸] می‌باشد.

۳. یادگیری تقویتی

قالب بندی یک مسئله به فرم یادگیری از طریق تعامل برای رسیدن به هدف(۳) . آنچه که واقعاً در یک مسئله یادگیری تقویتی روی می‌دهد به شرح ذیل است : یک یادگیرنده و تصمیم گیرنده با نام عامل[۱۹] شناخته می‌شود. این عامل مرتب در مقاطع زمانی با محیط تعامل دارد . به این ترتیب که عامل در ابتدا در یک حالت[۲۰] از محیط قرار دارد. سپس باتوجه به حالتی که در آن قرار دارد یک کنش[۲۱] از میان کنش‌های ممکن در آن حالت را انجام می‌دهد. سپس محیط، یک حالت جدید را به عامل نشان می‌دهد. به علاوه یک مقدار پاداش[۲۲] هم به ازای کنش مورد نظر به عامل اختصاص می‌دهد.هدف یک عامل بیشینه ساختن این پاداشها است که در طی یک دوره بلند مدت به‌دست می‌آید.

در هر گام ، عامل یک نگاشت از حالات به احتمال انتخاب کنش‌ها انجام می‌دهدکه این نگاشت با نام سیاست[۲۳] شناخته می‌شود و با π_t نشان داده می‌شود. به عبارت دیگر $\pi_t(s,a)$ یعنی اینکه اگر حالت جاری s باشد ($s_t = s$) احتمال اینکه کنش a باشد برابر است با $\pi_t(s,a)$.

۲-۲: ایده کلی

در یک مسئله یادگیری تقویتی عامل سعی می‌کند بر اساس تجربیات خود یاد بگیرد و از طرف دیگر ناظری نیز وجود ندارد.یعنی عاملی وجود داردکه به طور همزمان هم برای به‌دست آوردن دانش جدید تلاش می‌کند و هم تصمیمات خود را براساس دانش موجود بهینه می‌کند .

[18] Kohonen
[19] Agent
[20] State
[21] Action
[22] Reward
[23] Policy

در همه این مسائل عملی یک موضوع اساسی وجود دارد و آن هم به وجود آوردن نوعی تعادل بین به‌دست‌آوردن حداکثر پاداش بر اساس دانشی که قبل به‌دست آمده است و همینطور تلاش برای انجام یک کنش جدید است به طوری که بتواند دانش بعدی را افزایش دهد که این مسئله با نام اکتشاف در مقابل بهره‌برداری[۲٤] در یادگیری تقویتی شناخته می‌شود. به عبارت دیگر در این مسئله نیاز به جست و جو در فضای کنش‌ها می‌باشد. لذا گفته می‌شود که بازخورد در این مسئله صرفاً ارزیابی[۲۵] است و این برخلاف حالتی است که در مسئله یادگیری نظارت شده وجود دارد.

۲–۳: معادله بلمن[۲۶]

تقریباً تمامی الگوریتم‌های یادگیری تقویتی بر پایه تخمین توابع ارزش هستند و آنها توابعی از حالت‌ها و یا حالت – کنش‌ها هستند . منظور این است که برآورد شود که چقدر خوب است که یک عامل در یک حالت معلوم باشد. یا اینکه برآورد شود که چقدر خوب است که عامل در یک حالت معلوم یک کنش مشخص را انتخاب کند(٤).

ارزش اتخاذ کنش ، a ، در یک حالت s و تحت سیاست ، π ، با $Q^{\pi}(s,a)$ نشان داده می‌شود.به‌صورت زیر برای هر π و هر حالت s :

$$\sum_a \pi(s,a) \sum_{s'} P^a_{ss'} [R^a_{ss'} + \gamma V^{\pi}(s')] \qquad ۱–۲$$

معادله بالامعادله بلمن نامیده می‌شود. معادله بلمن ارتباط بین ارزش یک حالت و ارزش حالت‌های جایگزینش را نشان می‌دهد.معادله بلمن میانگین وزن دار همه حالات ممکن را براساس احتمال رخداد هریک محاسبه می‌کند.

[24] Exploration vs exploitation
[25] Evaluation
[26] Bellman equation

همیشه حداقل یک سیاست وجود دارد که بهتر یا مساوی همه سیاست‌ها باشد اما ممکن است چندین سیاست برتر وجود داشته باشد به هر حال این سیاست‌های بهینه با π^* نمایش داده می‌شودو برای آنها یک تابع ارزش حالت بهینه مشترک به نام V^* استفاده می‌شود:

$$V^*(s) = max_\pi V^\pi(s) \quad \forall s \in S \qquad ۲-۲$$

و همچنین یک تابع ارزش کنش بهینه مشترک به نام Q^* به‌صورت زیر

$$Q^*(s,a) = max_\pi Q^\pi(s,a) \quad \forall s \in S \ , a \in A(s) \qquad ۳-۲$$

معادله بهینگی بلمن برای Q^* به‌صورت زیر است:

$$Q^*(s,a) = \ E\left\{r_{t+1} + \gamma \max_{a'} \ Q^*(s_{t+1}, a') \middle| s_t = s \, , a_t = a\right\}$$
$$= \sum_{s'} P^a_{ss'} [R^a_{ss'} + \gamma \, max_{a'} Q^*(s', a')] \qquad ٤-۲$$

اگر V^* موجود باشد می‌توان سیاست بهینه را مشخص کرد، به این صورت که برای هر حالت ، s ، یک یا چند کنش وجود دارد که در معادله بهینگی بلمن مقدار حداکثر را به‌دست آورده است.

$$V^*(s) = \ max_{a \in A(s)} \sum_{s'} P^a_{ss'} [R^a_{ss'} + \gamma V^*(s')] \qquad ۵-۲$$

هر سیاست که به این کنش‌های ماکسیمم احتمال غیر صفری را نسبت دهد یک سیاست بهینه است.

۲-۴: بهره برداری و اکتشاف

اغلب برای انجام یک کنش یا انتخاب یک گزینه نمی‌توان هم اکتشاف[۲۷] و هم بهره برداری[۲۸] انجام داد و این موضوع با عنوان جنگ میان شناسایی و اکتشاف شناخته می‌شود. اینکه اکتشاف در یک حالت خاص بهتر است

[27] Exploration
[28] Exploitation

یا بهره برداری به روش پیچیده ای براساس مقادیر غیر مبهم از مقادیر برآورد شده ، عدم قطعیت‌ها و تعداد مراحل باقیمانده بستگی دارد(۵) . در زمینه تعادل و توازن بهره برداری و اکتشاف‌ها روشهای توانمندی وجود دارد که بیشتر آنها از این فرض استفاده می‌کنند که اطلاعات قبلی و ثابت معلومی در این کاربردها و مسائل یادگیری تقویتی وجود ندارد.

۲–۵: مفاهیم مورد نیاز

۲–۵–۱: مدل مارکوف

در چهار چوب یک مسئله یادگیری تقویتی ، یک عامل تصمیمات خود را به‌صورت تابعی از یک سیگنال محیط یعنی حالت می‌گیرد.در یک پرتاب گلوله مکان آن و سرعت و شتاب آن مهم است ولی این موضوع اهمیتی ندارد که این سرعت و شتاب چگونه و از کجا آمده است. لذا به این موضوع با نام "استقلال از مسیر"[۲۹] اشاره می‌شود، چرا که تمامی چیزهایی که اهمیت دارند در سیگنال حالت فعلی هست وبا این حال از مسیر یا تاریخچه مستقل است. اما توضیح ریاضی خاصیت مارکوف :

یک کنش در زمان t اتخاذ می‌شود و محیط در زمان $t+1$ به آن پاسخ می‌دهد.اما شاید پاسخی که در گام زمانی فعلی صادر شده است وابسته به مواردی باشد که قبلاً انجام شده باشد به عبارتی این مطالب حاکی از درنظرگرفتن پویایی مسئله به وسیله توزیع احتمال کل است.

$$Pr\{s_{t+1} = s', r_{t+1} = r \mid s_t, a_t, r_t, s_{t-1}, a_{t-1}, \dots, r_1, s_0, a_0\} \qquad ۶-۲$$

برای همه s' و rها و همه مقادیر ممکن برای رویدادهای گذشتهٔ $s_t, a_t, r_t, s_{t-1}, a_{t-1}, \dots, r_1, s_0, a_0$.

اما اگر سیگنال حالت ما خاصیت مارکوف داشته باشد، پاسخی که محیط در زمان $t+1$ می‌دهد فقط وابسته به‌حالت و کنش زمان t است ، لذا پویایی محیط به وسیله احتمال زیر نشان داده می‌شود.

$$Pr\{s_{t+1} = s', r_{t+1} = r \mid s_t, a_t\} \qquad ۷-۲$$

[29] Independence of Path

برای همه s' و r و s_t و a_tها.

یک سیگنال حالت خاصیت مارکوف دارد یا به عبارتی یک حالت مارکوف است هرگاه، به ازای تمام s' و r به علاوه $s_t, a_t, r_t, ..., r_1, s_0, a_0$ دومعادله اخیر برابرباشد. در چنین صورتی وظیفه[30] و محیط به همراه هم خاصیت مارکوف دارند. نتیجه سودمندخاصیت مارکوف برای یک محیط می‌تواندر موارد زیر دیده شود:

۱. به سادگی می‌توان حالت بعدی را باتوجه به اطلاعات موجود حالت فعلی و در نتیجه خاصیت مارکوف پیش بینی کرد
۲. خاصیت مارکوف برای یک حالت می‌تواند تشکیل دهنده پایه مناسبی برای انتخاب کنش باشد ، به‌این‌معنا که بهترین سیاست برای انتخاب یک کنش به‌صورت تابعی از یک حالت مارکوف می‌تواند به‌خوبی یک سیاست باشد که تابعی از تاریخچه کامل است(٦).

فرایند تصمیم مارکوف : هر وظیفه یادگیری تقویتی که در آن خاصیت مارکوف وجود داشته باشد با نام فرایند تصمیم مارکوف شناخته می‌شود. اگر فضای کنش‌ها و حالت‌ها محدود و متناهی باشد آن گاه فرایند مذکور با نام فرایند تصمیم مارکوف محدود شناخته می‌شود(۷).

۲-۵-۲: تکرار سیاست

هنگامی که با استفاده از V^{π} ، سیاست ، π به عنوان سیاست بهتر به‌دست می‌آید می‌توان $V^{\pi'}$ را محاسبه کرد و مجدداً از آن استفاده کردو باز سیاست بهتر π'' را به‌دست آورد و به همین ترتیب یک توالی از سیاست‌هایی را به‌دست آورد که مرتب بهبود یافته اند یعنی :

نماد نشان دهنده ارزیابی سیاست[31] $\xrightarrow{E}$ نماد نشان دهنده بهبود سیاست[32] $\xrightarrow{I}$

[30] Task
[31] Policy Evaluation
[32] Policy Improvement

به این روش یافتن سیاست بهینه، تکرار سیاست گفته می‌شود که از تابع مقدار مربوط به سیاست قبلی شروع می‌شود و نهایتاً همگرایی سریعتری از ارزیابی سیاست به دنبال خواهد داشت.

۲-۵-۳: تکرار مقدار

براساس آنچه که درتکرار سیاست گفته شد پس از اینکه π تعیین می‌شود روند ارزیابی سیاست اجرا می‌شود و این روند پس از چند دنباله اجرا به V^{π} همگرا می‌شود. پس از آن در فاز $\xrightarrow{I}$ آن بهبود می‌یابد و مجدداً $\xrightarrow{E}$ برروی π' اجرا می‌شود. همین محاسبات و دوره اجرای مربوط به $\xrightarrow{E}$ می‌تواند کار را طولانی کند و این موضوع اشکالی است که بر تکرار سیاست وارد می‌شود.

در حقیقت کوتاه کردن گام $\xrightarrow{E}$ در تکرار سیاست به روشهای مختلفی امکان پذیر است ، بدون اینکه در این میان خللی به تضمین همگرایی آن وارد شود . در یکی از مهمترین آنها $\xrightarrow{E}$ تنها پس از یک جریان متوقف می‌شود که به این الگوریتم تکرار مقدار گفته می‌شود.

۲-۵-۴: روش‌های مقدار کنش

در واقع بحث بر سر نحوه برآورد مقداربرای کنش‌ها است به عبارت دیگر منظور برآورد میانگین برای توزیع مربوط به هر گزینه است. یک روش ساده برای این منظور استفاده از میانگین پاداش‌هایی است که پیش از این به‌دست آمده است. به این روش میانگین نمونه‌ها[۳۳] گفته می‌شود. این روش یکی از روش‌های ساده است اما لزوماً بهترین نیست . روشهای جایگزین دیگری وجود دارد که در آنها در اکثر مواقع اقدام حریصانه انجام داده می‌شود، اما هر از گاهی با احتمال $\in$ به‌صورت یکنواخت و تصادفی از بین کنش‌های موجود بدون در نظر

[33] Sample-Average Method