سیستم‌های فازی و کاربردها
سال اول، شماره دوم (زمستان ۱۳۹۷) صص. ۱۱ تا ۳۴
DOR: 20.1001.1.27174409.1397.1.2.2.5

مروری بر روش‌های یادگیری تقویتی فازی با معماری نقاد-تنها

ولی درهمی و فریناز اعلمی‌یان هرندی

دانشگاه یزد، پردیس فنی و مهندسی، گروه مهندسی کامپیوتر

چکیده

این مقاله به مرور روش‌های یادگیری تقویتی فازی با معماری نقاد-تنها می‌پردازد. یادگیری تقویتی فازی از ترکیب سیستم‌های فازی به عنوان تقریب‌زننده‌ی جامع و روش یادگیری تقویتی حاصل شده است. یادگیری تقویتی یک روش یادگیری قوی است که تنها با استفاده از سیگنال عددی پاداش یا جریمه پارامترهای سیستم را به صورت برخط تنظیم می‌نماید. در معماری نقاد-تنها یک سیستم فازی مدل سوگنو مرتبه‌ی صفر برای تقریب تابع ارزش- عمل استفاده می‌شود و عمل نهایی بر اساس مقدار ارزش عمل‌های نامزد در تالی هر قاعده‌ی فازی به‌دست می‌آید. در این مقاله دو روش پایه به نام‌های یادگیری کیو فازی (FQL) و یادگیری سارسای فازی (FSL) برای تنظیم ارزش عمل‌های نامزد قواعد بیان می‌شود. در این دو روش به ترتیب از تعمیم روش‌های یادگیری کیو استاندارد و یادگیری سارسای استاندارد بهره برده شده است. مهمترین برتری FSL بر FQL وجود تحلیل‌های مثبت ریاضی درخصوص همگرایی است در حالی‌که مثال‌هایی از واگرایی در FQL وجود دارد. روش‌های FQL و FSL و گسترش‌هایی از آنها در مسائل کنترلی زیادی همچون حرکت ربات، حرکت بازوی ربات، حرکت قایق، مسیریابی در شبکه‌های کامپیوتری، و کنترل نیروگاه بادی استفاده شده و کارآیی خود را نشان داده‌اند.

Mathematics Subject Classification (2010): - - - - ; - - - - , **Email:** f.alamiyan@yazd.ac.ir.

عبارات و کلمات کلیدی: سیستم فازی، مدل سوگنوِ، قاعده‌ی فازی، یادگیری تقویتی، معماری نقاد-تنها

۱ مقدمه

سیستم‌های استنتاج فازی ابزاری برای رسمی‌سازی عملکرد یک فرآیند به کمک قواعد اگر- آنگاه فازی هستند. این سیستم‌ها با بهره‌برداری از منطق فازی که اولین بار توسط پروفسور لطف‌علی عسگرزاده در سال ۱۹۶۵ در دانشگاه برکلی مطرح شد، ساختاری شفاف (جعبه‌سفید) برای استنتاج ارائه می‌دهند و شبیه‌سازی منطق و طرز تفکر انسانی را در یک پیاده‌سازی انعطاف‌پذیر فراهم می‌کنند. به بیان دیگر کد کردن دانش نایقین و غیر قطعی انسانی را ممکن می‌سازند. این سیستم‌ها به دلیل توانایی برجسته‌ی خود در تقریب توابع غیرخطی با درجه پیچیدگی متنوع به عنوان تقریب‌زننده‌های جامع مطرح هستند [۲۳، ۱۴]. هم اینک از سیستم‌های مبتنی بر قاعده‌ی فازی در بسیاری از مسائل همچون طبقه‌بندی و شناسایی الگو، تعیین ریسک معاملات، پیش‌بینی خرابی، مدل‌سازی سیستم‌های پردازش سیگنال، و کنترلگرهای رباتیکی استفاده می‌شود. چرا که این سیستم‌ها قادرند به خوبی مابین دقت نتایج و تفسیرپذیری قواعد مصالحه کنند [۶، ۲].

نقطه‌ی شروع سیستم‌های فازی از مجموعه‌های فازی است که در آن برای هر عضو یک درجه‌ی عضویت بین صفر تا یک لحاظ می‌شود در حالی‌که در مجموعه‌های قطعی[1] و غیر فازی درجه‌ی عضویت هر عضو می‌تواند یکی از دو مقدار یک یا صفر باشد. قواعد فازی نیز بر اساس همین تعریف ایجاد شدند. اولین ساختار کاربردی سیستم‌های استنتاج فازی توسط ممدانی در سال ۱۹۷۴ برای کنترل موتور بخار معرفی شد [۲۳]. این نوع از سیستم‌های فازی در بخشِ تالیِ قواعد همانند بخش مقدم از برچسب‌های فازی بهره‌برداری می‌کنند. این ساختار فازی به نام مدل ممدانی شناخته می‌شود.

یک حالت خاصِ پرکاربرد در سیستم‌های فازی ممدانی از ضرب و جمع به ترتیب به عنوان عملگرهای برداشت[2] و اجتماع[3]، و از روش مرکز ثقل[4] برای فازی‌زدایی استفاده می‌کند. رابطه‌ی به‌دست آمده برای محاسبه‌ی خروجی نهایی در این مدل نشان می‌دهد که تنها مساحت و مرکز ثقل توابع عضویت تالی در نتیجه‌ی این رابطه تاثیر گذارند و شکل و دیگر ویژگی‌های توابع عضویت تالی

[1] Crisp
[2] Implication
[3] Aggregation
[4] Centroid

تاثیری در مقدار خروجی ندارند. حال اگر مساحت تالی‌ها مساوی باشند آنگاه تنها مرکز ثقل تالی قواعد در رابطه‌ی خروجی ظاهر می‌شود. با الهام از این موضوع، تاکاگی، سوگنو و کانگ در سال ۱۹۸۸ نوع دیگری از سیستم‌های استنتاج فازی را تحت عنوان سیستم فازی تاکاگی-سوگنو-کانگ (TSK) یا به اختصار سوگنو معرفی نمودند [۲۳].

در سیستم‌های فازی مدل سوگنو، بخش مقدم قواعد فازی از برچسب‌های فازی تشکیل شده است در حالی‌که تالی قواعد فازی مقداری قطعی (به صورت غیر فازی) و به شکل یک چند جمله‌ای از متغیرهای ورودی است. در صورتی‌که متغیرهای ورودی موجود در چندجمله‌ای‌های تالیِ قواعد از درجه‌ی صفر باشند، یعنی بخش تالی از مقادیر ثابت تشکیل شده باشد، سیستم فازی حاصل یک سیستم سوگنو از مرتبه‌ی صفر نامیده می‌شود. اگر تالی قواعد فازی ترکیبی خطی از متغیرهای ورودی باشد، سیستم فازی سوگنو از درجه‌ی یک است. با توجه به کاربرد و انعطاف‌پذیری بالای مدل‌های فازی سوگنو، تمرکز این مقاله بر روش‌های تنظیم مبتنی بر یادگیری تقویتی[5] برای کنترلگر فازی مدل سوگنو مرتبه صفر است.

پایگاه قواعد فازی که دانش سیستم در آن قرار دارد به عنوان مهمترین بخش یک سیستم فازی مطرح است. در تولید قواعد فازی لازم است ساختار سیستم فازی تعیین شود. به عبارت دیگر باید فضای ورودی‌های سیستم به زیر فضاهای فازی تقسیم‌بندی شده و توابع عضویت ورودی و تالی در تعریف یک قاعده‌ی فازی برای هر یک از این زیر فضاها، معرفی شوند [۲].

راهکارهای مختلفی برای تخمین شکل و تعداد توابع عضویت در بخش مقدم و پارامترهای بخش تالی ارائه شده‌اند [۲۰]. در ایجاد زیر فضاهای فازی می‌توان از روش تقسیم‌بندی شبکه روی تک‌تک ابعاد فضای ویژگی استفاده نمود [۱۳، ۱۸، ۱۹، ۲۵]. همچنین می‌توان با استفاده از روش‌های خوشه‌بندی، توابع عضویت را همزمان روی چند بُعد فضای ویژگی تعریف کرد [۵]. در تعیین نوع و شکل توابع عضویت برچسب‌های فازی، گاه از توابع عضویت همگن روی تمام ابعاد فضای ورودی استفاده می‌شود؛ اما گاهی برای هر یک از ابعاد، از تعدادی مجموعه‌ی فازی از پیش تعریف شده بهره‌برداری می‌گردد که با توجه به دانش خبره بدست آمده‌اند.

تعیین مقدار تالی قواعد در کنترلگر های فازی سوگنو از پیچیدگی و اهمیت بسیار بیشتری

[5] Reinforcement Learning

نسبت به قسمت مقدم برخوردار است. مقدار دهی اولیه می‌تواند بر اساس دانش خبره انجام پذیرد، اما برای افزایش بازدهی لازم است از روش‌های یادگیری استفاده شود. این روش‌ها را می‌توان به دو دسته‌ی کلی روش‌های مبتنی بر یادگیری با ناظر و یادگیری بدون ناظر دسته‌بندی نمود.

یادگیری با ناظر، یکی از اولین الگوریتم‌های یادگیری است. این یادگیری از داده‌های آموزش استفاده کرده و پارامترها را با بهره‌برداری از روش‌های مبتنی بر گرادیانت با کاهش جمع مربعات خطا در خروجی‌ها تنظیم می‌کند [۱۴]. در این نوع یادگیری، چالش‌های قابل‌توجهی وجود دارد؛ برای مثال جمع‌آوری داده‌های آموزشی در بعضی مسائل به خصوص در کاربردهای واقعی با دشواری‌هایی روبه‌رو است و در برخی موارد تقریبا غیر ممکن است. ناسازگاری در داده‌ها، وجود داده‌های نویزی و خطاهای زیاد در آنها بسیار چالش‌برانگیز هستند و گاه کیفیت سیستم فازی را به شدت تحت تأثیر قرار می‌دهند [۷، ۱۲].

با توجه به ضعف‌های اشاره شده در یادگیری باناظر، استفاده از روش‌های هوشمند جهت یادگیری گسترش یافته است. یادگیری تقویتی یک الگوریتم مدرن هوشمند است که به جهت دارا بودن قابلیت‌هایی همچون عدم نیاز به خروجی مطلوب، یادگیری تنها با استفاده از یک معیار عددی بازدهی، امکان یادگیری برخط، و درجه کاوش بالا گزینه‌ی مناسبی جهت تنظیم پارامترهای یادگیرنده (کنترلگر فازی) است. در واقع در یادگیری تقویتی به یادگیرنده گفته نمی‌شود که عمل صحیح در هر وضعیت چیست، و فقط با استفاده از یک معیار عددی که سیگنال تقویتی نامیده می‌شود خوب یا بد بودن عمل به کنترلگرِ در حال یادگیری نشان داده می‌شود. به این ترتیب، یادگیرنده تلاش می‌کند سیاستی مناسب را برای دستیابی به اهداف تعیین‌شده در کاربرد تحت بررسی یاد بگیرد [۸]. در استفاده از این روش‌های یادگیری، رفتار بهینه به صورت مستقیم از تعامل با محیط و بدون نیاز به آگاهی از مدل دینامیکی محیط عملیات یاد گرفته می‌شود [۲۲].

به عبارت دیگر، یادگیری تقویتی به معنای یادگرفتن عملِ مناسب از میان مجموعه اَعمال مجاز برای یک موقعیت خاص بر اساس پاداش و جریمه‌های دریافتی است. در الگوریتم‌های یادگیری تقویتی، عامل (کنترلگر) یادگیرنده دنباله‌ای از حالت‌های محیط را پیگیری کرده و با مشاهده‌ی هر حالت عملی از اعمال مجاز خود را به محیط اعمال می‌کند و به کمک سیگنال تقویتی دریافت‌شده از محیط، سیاست رفتاری مناسب را یاد می‌گیرد [۴].

با ترکیب روش‌های یادگیری تقویتی و سیستم‌های فازی به عنوان تقریب‌زننده‌های تابع، سیستم‌های یادگیری تقویتی فازی ارائه شدند. معماری‌های یادگیری تقویتی فازی در نگاهی گسترده، به سه دسته‌ی معماری نقاد-تنها[۶]، عملگر-تنها[۷] و عملگر-نقاد[۸] تقسیم می‌شوند. هر یک از این معماری‌ها مزایا و معایب خاص خود را دارند. اما معماری نقاد-تنها به دلیل درجه کاوش بالاتر و شفافیت بیشتر در ارائه‌ی دانش خبره کاربرد وسیع‌تری دارد. به همین دلیل این مقاله به بررسی روش‌های مطرح در روش یادگیری تقویتی فازی با معماری نقاد-تنها می‌پردازد. دو روش پایه‌ای اصلی در این معماری یادگیری کیو فازی[۹] [۱۵] و یادگیری سارسای فازی[۱۰] [۱۰] نامیده شده‌اند. این روش‌ها بر اساس مدل فازی سوگنو مرتبه صفر هستند و راهکاری برای تنظیم برخط تالی قواعد ارائه می‌دهند. مهمترین ویژگی متمایز کننده‌ی این دو روش پایه وجود تحلیل همگرایی روش سارسا در مقابل امکان واگرایی روش یادگیری کیو است. در طول زمان بهبودهای متنوعی در هر یک از دو روش پایه ارائه شده اند که در این مقاله مورد بحث قرار خواهند گرفت.

ساختار این مقاله به شرح ذیل است. ابتدا در بخش دوم به معرفی مفاهیم پایه از جمله ساختار سیستم فازی سوگنو و یادگیری تقویتی پرداخته می‌شود. در بخش سوم، معماری نقاد-تنها در قالب یک معماری یادگیری تقویتی فازی تشریح می‌شود. بخش چهارم توسعه‌هایی از روش‌های یادگیری تقویتی فازی و کاربردهای آنها را مرور می‌کند. نهایتاً در بخش پنجم نتیجه‌گیری آمده است.

۲ مفاهیم پایه

دراین بخش، مفاهیم پایه در سیستم‌های فازی و یادگیری تقویتی به صورت خلاصه مرور می‌شوند. به طور خاص، مدل فازی سوگنو و الگوریتم‌های یادگیری تفاضل موقتی[۱۱] که در حل مسائل کنترل‌محور مرسوم هستند شرح داده می‌شوند.

---

[6] Critic Only
[7] Actor Only
[8] Actor-Critic
[9] Fuzzy Q-Learning
[10] Fuzzy Sarsa Learning
[11] Temporal Difference Learning

۱.۲ ساختار کلی سیستم‌های فازی

سیستم‌های استنتاج فازی در مسائل مختلف به عنوان پایگاه قواعد فازی، مدل‌های فازی، حافظه‌های تجمعی فازی و کنترلگرهای فازی به‌کار می‌روند. به طور کلی سیستم‌های فازی از پنج بخش اصلی به صورت زیر تشکیل می‌شوند [۲۳، ۱۴]:

۱. پایگاه قواعد، شامل تعدادی قواعد اگر-آنگاه فازی

۲. پایگاه داده فازی، شامل مشخصات توابع عضویت مجموعه‌های فازی در قواعد فازی

۳. موتور استنتاج فازی[۱۲]، واحد تصمیم‌گیری روی قواعد فازی

۴. فازی‌ساز[۱۳]، واحد تبدیل ورودی قطعی و غیر فازی به مجموعه فازی

۵. فازی‌زدا[۱۴]، واحد تبدیل خروجی فازی به مقدار قطعی به عنوان خروجی نهایی

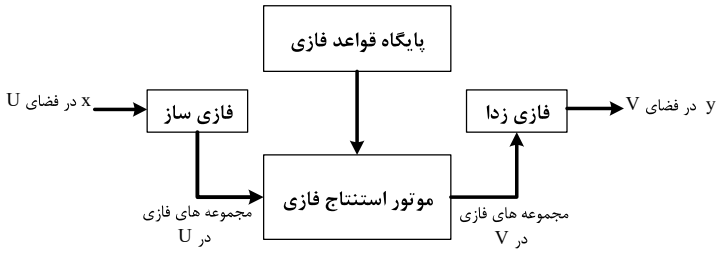
مراحل استدلال فازی (عملیات استنتاج فازی) به ترتیب عبارتند از: ۱) فازی‌سازی داده‌های ورودی، ۲) محاسبه‌ی درجه تطابق در هر بُعد از ابعاد مقدم قواعد فازی با توجه به عملگر اشتراک، ۳) محاسبه‌ی شدت آتش هر قاعده با استفاده از عملگر اشتراک، ۴) محاسبه‌ی مجموعه فازی منتجه از هر قاعده با بهره‌برداری از عملگر برداشت، ۵) محاسبه‌ی مجموعه فازی نهایی توسط عملگر اجتماع، ۶) فازی‌زدایی برای تولید خروجی نهایی سیستم فازی. شکل ۱ ساختار بلوکی سیستم فازی را نشان می‌دهد.

۱.۱.۲ سیستم فازی مدل سوگنو

ساختار قواعد فازی در یک سیستم استنتاج فازی سوگنو به صورت زیر است:

$$R_i : \text{if } x_1 \text{ is } L_{i1} \text{ and} \ldots \text{and } x_n \text{ is } L_{in},$$

[12] Fuzzy Inference Engine
[13] Fuzzifier
[14] Defuzzifier



شکل ۱: *نمودار بلوکی سیستم فازی [۲۳].*

$$\text{then } y = g_i(x_۱, x_۲, \ldots, x_n), \quad i = ۱, ۲, \ldots, R \tag{۱}$$

که در این رابطه x_k ورودی kاُم، n تعداد متغیرهای ورودی، R تعداد قواعد فازی، L_{ij} مجموعه فازی متناظر با jاُمین متغیر ورودی برای iاُمین قاعده‌ی فازی و g_i یک تابع چند جمله‌ای مرتبه‌ی اول یا مرتبه‌ی صفر از x_k می‌باشد. در سیستم فازی سوگنو مرتبه‌ی صفر g_i یک مقدار ثابت به فرم $g_i = o_i$ است. در نتیجه خروجی نهایی a بر اساس متوسط وزن‌دار به صورت زیر قابل بیان است و در آن μ_i شدت آتش قاعده‌ی iاُم است:

$$a = \frac{\sum_{i=۱}^{R} \mu_i \times o_i}{\sum_{i=۱}^{R} \mu_i} \tag{۲}$$

۲.۲ یادگیری تقویتی

یادگیری از طریق تعامل با محیط، ایده‌ای است که تقریباً زیربنای تمام نظریه‌های یادگیری و هوشمندسازی به حساب می‌آید. در یادگیری تقویتی، موضوع مورد چالش عبارت است از: چگونگی نگاشت موقعیت‌ها به عمل‌ها، به‌طوری‌که یک سیگنال پاداش عددی (یا سیگنال تقویتی) بیشینه شود. نکته مهم این است که در این روش برخلاف اکثر روش‌های یادگیری ماشین، عمل مطلوب به یادگیرنده گفته نمی‌شود بلکه یادگیرنده باید ”خودش“ با امتحان کردن عمل‌های مختلف در آن موقعیت، عمل مناسب را در جهت بیشینه کردن امید دریافت پاداش پیدا کند. شش عنصر اصلی

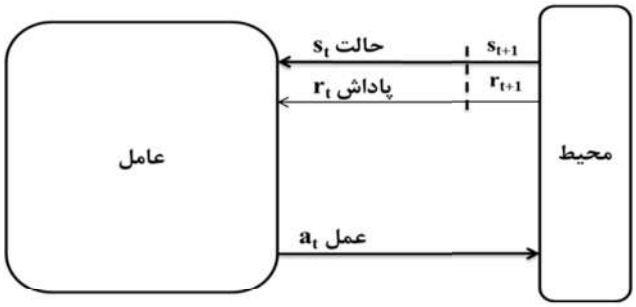
یک سیستم یادگیری تقویتی عبارتند از: عامل، حالت محیط، سیاست، تابع پاداش، تابع ارزش و در برخی از سیستم‌ها به صورت اختیاری مدلی از محیط [۴، ۲۲].

چارچوب ریاضی مرسوم در بیشتر مسائل یادگیری تقویتی، محیط را یک فرآیند تصمیم‌گیری مارکوف در نظر می‌گیرد [۲۲]. در یک سیستم مارکوف حالت بعدیِ محیط و پاداش دریافتی تنها به عمل و حالت قبلی عامل در محیط بستگی دارد. عامل و محیط در دنباله‌ای از گام‌های زمانی گسسته $t = ۱, ۲, ۳, \ldots$ با یکدیگر در تعامل هستند. در هر گام زمانی t ، عامل حالت محیط را دریافت می‌کند، این حالت با $s_t \in S$ بیان می‌شود که در آن S مجموعه‌ی تمام حالت‌های ممکن است. بر اساس این حالت درک‌شده، عامل یک عمل $a_t \in A(s_t)$ را از مجموعه عمل‌هایش انتخاب می‌کند، به‌طوری‌که $A(s_t)$ مجموعه‌ی تمام عمل‌های ممکنی است که عامل در حالت s_t می‌تواند انتخاب کند. در گام زمانی بعدی (یعنی گام $t + ۱$)، در پاسخ به عمل انتخاب‌شده‌ی a_t، محیط به حالت جدید $s_{t+۱}$ منتقل می‌شود و عامل پاداش عددی $r_{t+۱} \in R$ را از محیط دریافت می‌کند. شکل ۲ نمودار بلوکی تعامل عامل یادگیری تقویتی و محیط را نشان می‌دهد.

در هر گام زمانی t، عامل حالت‌ها را به احتمال انتخاب هر کدام از عمل‌های ممکن در آن حالت نگاشت می‌کند. این نگاشت سیاست عامل نامیده شده و با π_t نشان داده می‌شود؛ به‌طوری‌که $\pi_t(s, a)$ احتمال انتخاب عمل a در حالت s و گام زمانی t است. ارزش انجام عمل a در حالت s تحت سیاست π با $Q^\pi(s, a)$ نشان داده می‌شود و این تابع، ارزش عمل تحت سیاست π نامیده می‌شود. این ارزش عبارت است از بازگشت (امید ریاضی پاداش‌های تخفیف‌یافته) مورد انتظاری که با شروع از حالت s و انجام عمل a و ادامه دادن کار تحت سیاست π به‌دست می‌آید:

$$\begin{aligned} Q^\pi(s, a) &= E_\pi[R_t | s_t = s, a_t = a] \\ &= E_\pi[\sum_{k=۰}^{\infty} \gamma^k r_{t+(k+۱)} | s_t = s, a_t = a] \end{aligned} \qquad (۳)$$

به طوری که R معرف مفهوم بازگشت و $\gamma \in [۰, ۱]$، فاکتور تخفیف برای پاداش‌های آتی r است و $E_\pi[\,]$ به امید ریاضی پاداش یا همان ارزش مورد انتظاری اشاره دارد که با پیروی از سیاست π به‌دست می‌آید.



شکل ۲: *نمودار بلوکی تعامل عامل یادگیری تقویتی و محیط [۴].*

روش‌های حل مساله‌ی کنترل به کمک یادگیری تقویتی را می‌توان به دو دسته‌ی "بر سیاست[15]" و "برون‌سیاست[16]" تقسیم کرد. روش سارسا (Sarsa) یک روش کنترل بر سیاست است. این روش مقادیر ارزش عمل را تحت سیاست جاری تخمین می‌زند و در محاسبات خود، انتقال از یک جفت حالت–عمل به جفت حالت–عمل بعدی را لحاظ می‌کند. قاعده‌ی به‌روز رسانی در محاسبه‌ی تابع ارزش حالت–عمل به صورت رابطه‌ی (۴) است. این قاعده پس از هر انتقال از یک حالت غیرپایانی s_t اِعمال می‌شود و در صورتی‌که حالت s_{t+1} یک حالت پایانی باشد، مقدار $Q(s_{t+1}, a_{t+1})$ برابر با صفر خواهد بود.

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\Big(r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)\Big) \qquad (۴)$$

روش یادگیری کیو[17] مشهورترین الگوریتم برون‌سیاست برای حل مساله‌ی کنترل است. در این الگوریتم، تابع ارزش حالت–عمل یادگرفته‌شده‌ی Q به صورت مستقیم، تابع ارزش حالت–عمل بهینه را مستقل از سیاستی که بر اساس آن رفتار می‌کند، تقریب می‌زند و سیاست به کار برده شده در انتخاب عمل بعدی، در قاعده‌ی به‌روز رسانی مقادیر ارزش عمل (رابطه‌ی (۵)) نقشی ندارد.

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\Big(r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)\Big) \qquad (۵)$$

[15] On-Policy
[16] Off-Policy
[17] Q-Learning

۳ یادگیری تقویتی فازی با معماری نقاد-تنها

یکی از مهم‌ترین معماری‌های مطرح در یادگیری‌تقویتی پیوسته، معماری نقاد-تنها است. این معماری بر مبنای تخمین یا تقریب تابع ارزش عمل عمل می‌کند و انتخاب سیاست بر اساس این تابع ارزش انجام می‌شود. در حقیقت، استفاده از تقریب تابع ارزش (معمولاً تابع ارزش حالت – عمل) و عدم وجود یک تابع سیاست صریح، ویژگی خاصِ این معماری است. شکل ۳ معماری نقاد- تنها را نشان می‌دهد.

در یادگیری تقویتی پیوسته با معماری نقاد-تنها با استفاده از تقریب‌زننده‌های تابع، نگاشتِ میان فضای حالت و فضای عمل تقریب زده می‌شود. با توجه به ویژگی‌های سیستم‌های استنتاج فازی، روش‌های مطرح در معماری نقاد- تنها، غالباً از این سیستم‌ها به عنوان تقریب‌زننده‌ی تابع استفاده می‌کنند. سیستم‌های استنتاج فازی به عنوان تقریب‌زننده‌های جامع، گزینه‌ی مناسبی برای ذخیره‌ی مقادیر ارزش محسوب می‌شوند و با در برگرفتن دانش قبلی و تعبیه‌ی آن‌ها در قواعد فازی، سرعت یادگیری را به طور قابل ملاحظه‌ای افزایش می‌دهند.

۱.۳ یادگیری سارسای فازی (FSL)

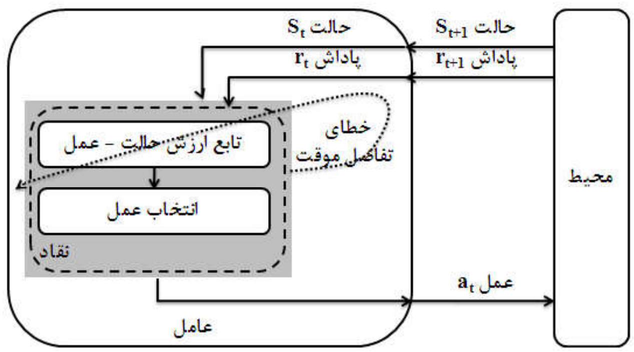
روش FSL یک الگوریتم یادگیری تقویتی فازی با معماری نقاد-تنها است که از ترکیب روش سارسا با سیستم‌های فازی به عنوان یک تقریب‌زننده‌ی تابعیِ خطی حاصل شده است. روش FSL یک روش بر سیاست را پیشنهاد می‌دهد که تالی قواعد سیستم فازی را به صورت برخط تنظیم می‌کند [۹].

ساختار قواعد در FSL برای یک سیستم فازی سوگنوی مرتبه صفر با n ورودی و یک خروجی که دارای R قاعده است به صورت زیر می‌باشد:

$$R_i : \text{if } x_1 \text{ is } L_{i1} \text{ and} \dots \text{and } x_n \text{ is } L_{in},$$
$$\text{then } (o_{i1} \text{ with value } w^{i1}) \text{ or} \dots \text{or } (o_{im} \text{ with value } w^{im}) \qquad (۶)$$

که در آن $s = [x_1, \dots, x_n]$، بردار n بُعدی متغیرهای حالت ورودی، $L_i = L_{i1} \times \cdots \times L_{in}$

شامل n مجموعه‌ی فازی محدبِ اکیداً نرمال با مرکزهای یکتا برای iاُمین قاعده، m تعداد عمل‌های گسسته‌ی ممکن برای هر قاعده، o_{ij}، jاُمین عمل نامزد و w^{ij}، ارزش تقریب‌زده‌شده برای عمل jاُم در قاعده‌ی iاُم است. مقدار تالی هر قاعده در هر گامِ زمانی با توجه به مقادیر w^{ij} انتخاب می‌شود. توجه شود که تعداد عمل‌های نامزد و مقدار آن‌ها جزء پارامترهای طراحی است که توسط طراح با توجه به مساله تعیین می‌گردد. هدف یادگیری، به‌روز رسانی بر خط وزن w^{ij} است به گونه‌ای که بهترین سیاست انتخاب عمل بر مبنای آن‌ها حاصل گردد.



شکل ۳: *نمودار بلوکی معماری نقاد-تنها [۴].*

شدت آتش هر قاعده از ضرب درجه‌های تطابق مقدم قواعد برای ورودی‌های مختلف به‌دست می‌آید. توابع شدت آتش نرمال‌شده‌ی قواعد به عنوان توابع پایه‌ی حالت در نظر گرفته می‌شوند. μ_i^j، شدت آتش نرمال‌شده‌ی قاعده‌ی iاُم برای حالت $s_j \in S$ و $N = |S|$ است. از آن‌جا که مراکز مجموعه‌های فازی متفاوت می‌باشند، ماتریس فضای حالت (شامل شدت آتش تمام قواعد فازی برای تمام نمونه‌داده‌های موجود ورودی) دارای مرتبه‌ی کامل است یعنی توابع پایه‌ی حالت، مستقل خطی می‌باشند. توجه شود در این‌جا فرض شده است توابع عضویت ورودی، محدب با مراکز یکتا هستند. خروجی سیستم و مقدار ارزش عمل تقریب‌زده شده $\widetilde{Q}(s,a)$ به صورت روابط (۷) و (۸)

محاسبه می‌شوند:

$$a_t(s_t) = \sum_{i=1}^{R} \mu_i(s_t) o^{ii^+} \quad (۷)$$

$$\widetilde{Q}_t(s_t, a_t) = \sum_{i=1}^{R} \mu_i(s_t) w_t^{ii^+} \quad (۸)$$

که o^{ii^+} و $w_t^{ii^+}$ به ترتیب عمل انتخاب‌شده و وزن مرتبط با آن در قاعده‌ی iاُم هستند. در الگوریتم‌های یادگیری کیوی فازی، برای انتخاب عمل در قواعد، از روش‌های مرسوم انتخاب عمل همچون حریصانه[۱۸]، شبه‌حریصانه[۱۹] و یا بیشینه‌نَرم[۲۰] استفاده شده است. هر چند در این‌جا نیز از این روش‌ها می‌توان استفاده کرد اما قضایا و لم‌های اثبات شده برای یادگیری سارسای فازی بر اساس یک رابطه‌ی جدید انتخاب عمل در هر قاعده به نام "بیشینه‌نَرم بهبود یافته" است که در رابطه‌ی (۹) معرفی شده است و در (۹)، T پارامتر دما[۲۱] است.

$$P(o_{ij}) = \frac{\exp(\frac{\mu_i w^{ij}}{T})}{\sum_{k=1}^{m} \exp(\frac{\mu_i w^{ik}}{T})} \quad (۹)$$

پس از محاسبه‌ی عمل نهایی a_t و اِعمال آن، عامل به حالت جدید s_{t+1} رفته و عمل جدید a_{t+1} با توجه به مقادیر وزن فعلی w_t انتخاب می‌شود. در ضمن سیگنال تقویتی r_{t+1} از محیط دریافت می‌گردد و مقادیر پارامترهای وزن هر قاعده به صورت رابطه‌ی (۱۰) به‌روز رسانی می‌شوند:

$$\Delta w_{t+1}^{ij} = \begin{cases} \alpha_t \times \Delta\widetilde{Q}_t(s_t, a_t) \times \mu_t(s_t) & \text{if} \quad j = i^+ \\ 0 & \text{otherwise} \end{cases} \quad (۱۰)$$

[18] Greedy
[19] ϵ-Greedy
[20] Softmax
[21] Temperature Parameter

که $\Delta\widetilde{Q}_t$، خطای تفاضل موقتی ارزش عمل بوده و به صورت رابطه‌ی (۱۱) محاسبه می‌گردد:

$$\Delta\widetilde{Q}_t(s_t, a_t) = r_{t+1} + \gamma Q_t(s_{t+1}, a_{t+1}) - Q_t(s_t, a_t) \qquad (11)$$

رویّه‌ی اجرایی یادگیری بر اساس الگوریتم FSL به صورت زیر است [۴]:

۱. مشاهده حالت s_{t+1} و دریافت سیگنال تقویتی r_{t+1}.

۲. توقف فرآیند یادگیری در صورتی‌که الگوریتم به تعداد دفعات لازم به موفقیت (رسیدن به هدف) رسیده باشد و یا تعداد مرحله‌ها به بیش‌ترین مقدار رسیده باشد.

۳. انتخاب عمل مناسب برای هر قاعده با استفاده از رابطه‌ی انتخاب عمل بیشینه‌نَرم بهبود یافته‌ی رابطه‌ی (۹).

۴. محاسبه‌ی خروجی نهایی a_{t+1} و مقدار ارزش عمل تقریب‌زده‌شده $\widetilde{Q}_t(s_{t+1}, a_{t+1})$ با استفاده از روابط (۷) و (۸).

۵. محاسبه‌ی $\triangle\widetilde{Q}_t$ و به‌روز رسانی w با روابط (۱۱) و (۱۰).

۶. محاسبه‌ی مقدار ارزش عمل تقریب‌زده‌شده جدید $\widetilde{Q}_{t+1}(s_{t+1}, a_{t+1})$ با استفاده از رابطه‌ی (۸).

۷. اِعمال خروجی نهایی به محیط.

۸. $t \leftarrow t+1$ و بازگشت به گام اول.

۲.۳ یادگیری کیو فازی (FQL)

یادگیری کیو فازی (FQL) از ترکیب الگوریتم یادگیری کیو و سیستم فازی حاصل می‌شود. روش یادگیری کیو فازی توسط لیونل جافی برای اولین بار در سال ۱۹۹۷ برای حل مساله‌ی آوردن گاری به مرکز ارائه شد. در واقع از سیستم فازی به عنوان تقریب‌زننده ارزش عمل در هر حالت استفاده

می‌شود. ساختار قواعد در FQL برای یک سیستم فازی سوگنوی مرتبه صفر با n ورودی و یک خروجی که دارای R قاعده است مشابه FSL است.

مهم‌ترین تفاوت روش‌های یادگیری کیوی فازی و یادگیری سارسای فازی در نحوه‌ی محاسبه‌ی خطای تفاضل موقتی ارزش عمل است؛ در الگوریتم‌های یادگیری کیوی فازی مقدار ارزش حالت-عمل به صورت رابطه‌ی (۱۲) محاسبه می‌شود که در آن دومین عبارت موجود در سمت راست رابطه‌ی متناظر با رابطه‌ی (۱۱)، حاصل‌ضرب فاکتور تخفیف در مقدار ارزش حالت بعدی یعنی مقدار $\gamma\widetilde{V}(s_{t+1})$ است.

$$\Delta\widetilde{Q}_t(s_t,a_t) = r_{t+1} + \gamma\widetilde{V}(s_{t+1}) - Q_t(s_t,a_t) \qquad (۱۲)$$

۴ گسترش‌هایی از روش یادگیری تقویتی فازی نقاد -تنها

در این بخش نمونه‌هایی کاربردی از پژوهش‌های انجام‌شده در راستای ارتقای روش‌های یادگیری تقویتی با معماری نقاد-تنها مرور می‌شوند.

۱.۴ مقداردهی اولیه مقادیر تالی قواعد فازی

یی[۲۲] و همکارانش در سال ۲۰۰۳ ایده‌ای برای فضای حالت و عمل پیوسته ارائه کردند که در آن از داده‌های آموزشی تولید شده توسط ناظر برای تنظیم اولیه‌ی پارامترهای بخش عملگر در معماری عملگر- نقاد استفاده شده است [۲۴]. در این روش ابتدا توسط یادگیری باناظر مقدار عمل برای هر حالت پیشنهاد می‌شود و سپس با استفاده از یادگیری تقویتی، مقدار نهایی پیرامون مقدار پیشنهادی تنظیم می‌گردد. یکی از ضعف‌های این روش وجود خطای زیاد در خروجی تنظیم شده بر اثر وجود ناسازگاری‌های موجود در داده‌های آموزشی ناظر است.

فتحی‌نژاد و همکارش در سال ۲۰۱۶ با کمک دانش ناظر به جای تعیین یک عمل برای هر حالت، ارزش اولیه‌ی عمل‌های نامزد را در یک کنترلگر با معماری نقاد-تنها تعیین نمود [۱۲]. این مقادیر به کمک یادگیری تقویتی، به صورت برخط در جهت بهبود کارایی کنترلگر تنظیم نهایی شده‌اند.

[22] Ye

در این روش برای تعیین مقدار ارزش هر عمل، به رفتار ناظر در زمان جمع‌آوری داده‌های آموزش توجه می‌شود. اگر ناظر هر بار در یک وضعیت خاص عمل‌های متفاوتی را انتخاب کند، ارزش هر عمل متناسب با تعداد انتخاب آن عمل در آن وضعیت خاص تعیین می‌شود.

از آنجا که خروجی نهایی سیستم فازی از ترکیب وزن‌دار تالی‌های منتخب در هر قاعده به‌دست می‌آید، برای هر خروجی یک ترکیب ممکن از عمل‌های نامزد هر قاعده محاسبه می‌گردد به گونه‌ای که ترکیب این عمل‌ها بتواند مقداری نزدیک به خروجی ناظر تولید کند. ارزش عمل‌ها در روند یادگیری بر همین اساس افزایش داده می‌شوند. ترکیب یادگیری باناظر و یادگیری تقویتی در این راهکار منجر به تسریع فرآیند یادگیری و بهبود کیفیت آموزش می‌شود. عملکرد کنترلگر حاصل از این روش در یک مساله‌ی ناوبری ربات با هدف پیگیری هدف و اجتناب از موانع با کاهش تعداد برخوردهای ربات به موانع و همگرایی سریع‌تر فرآیند یادگیری شده است.

۲.۴ تعیین مقدار تالی قواعد فازی از یک فضای پیوسته

در دو روش FQL و FSL، تالی هر قاعده در پایگاه قواعد فازی از میان اعضای مجموعه‌ای از عمل‌های گسسته‌ی ممکن، انتخاب و ارزش متناظر با هر عمل محلی به دلیل گسسته بودن عمل‌ها در یک جدول جستجو نگه‌داری می‌شد. تعریف مجموعه‌ای از عمل‌های محلیِ ممکن برای هر قاعده، نیاز به دانش طراحِ سیستم دارد. دقت عمل انتخاب شده‌ی نهایی برای هر تالی به تعداد اعضای مجموعه‌ی عمل‌ها در تالی هر قاعده بستگی دارد. برای رسیدن به دقت زیاد، تعداد اعضای مجموعه‌ی عمل‌ها به صورت تصاعدی افزایش می‌یابد که این امر منجر به کندی سرعت یادگیری می‌شود. برای حل این مساله روش گسسته‌سازی تطبیقی ارائه شده است [۳، ۱]. این روش در حل مساله‌ی حرکت ربات دو پا به‌کار گرفته شده است.

در این روش، فضای عمل طی چندین مرحله گسسته می‌شود. تعداد اعضای مجموعه‌ی عمل‌ها در تالی هر قاعده در تمام مراحل ثابت نگه داشته شده در عوض برای افزایش دقت، تعداد مراحل یادگیری افزایش داده شده است. در یک مساله با فضای عمل پیوسته، روش گسسته‌سازی ابتدا تعداد معدودی عمل به عنوان مجموعه‌ی عمل‌های نامزد برای تالی هر قاعده انتخاب می‌کند. سپس عامل شروع به یادگیری می‌کند تا دامنه‌ی برتر یعنی بازه‌ای خوش شانس تر در یافتن جواب بهینه را

به کمک مقادیر ارزش عمل در فضای عمل تشخیص دهد.آن‌گاه گسسته‌سازی روی این دامنه تکرار و یادگیری انجام می‌شود. این کار تا حصول نتیجه‌ای با دقت مناسب ادامه می‌یابد. نمونه‌ای از روش گسسته‌سازی تطبیقی، طی سه دور یادگیری برای یک قاعده به صورت زیر است:

- دور۱: {۰, ۱۰, ۲۰, ۳۰, ۴۰, ۵۰, ۶۰, ۷۰, ۸۰, ۹۰, ۱۰۰} با فرض اینکه عمل‌های ۴۰ و ۵۰ بالاترین مقدار ارزش در پایان دور اول را داشته باشند.
- دور۲: {۴۰, ۴۱, ۴۲, ۴۳, ۴۴, ۴۵, ۴۶, ۴۷, ۴۸, ۴۹, ۵۰} با فرض اینکه عمل‌های ۴۶ و ۴۷ بالاترین مقدار ارزش در پایان دور دوم را داشته باشند
- دور۳: {۴۶, ۴۶٫۱, ۴۶٫۲, ۴۶٫۳, ۴۶٫۴, ۴۶٫۵, ۴۶٫۶, ۴۶٫۷, ۴۶٫۸, ۴۶٫۹, ۴۷}

روش دیگری از نوع شکست بازه‌ها در سال ۲۰۱۷ ارائه شده است [۷]. این راهکار به جای افزایش عمل‌های نامزد به‌طور یکجا، پیشنهاد می‌کند عملیات بسط و پیشنهاد عمل‌های نامزد را حول مقادیری انجام دهد که بهترین نتیجه‌ی ارزیابی را داشته‌اند. روش مولد پیشنهادی در این پژوهش دو پارامتر را در بسط محدوده‌ی وزن‌های مولد اولیه دخالت می‌دهد: ۱) پارامتر تعیین‌کننده‌ی تعداد عمل‌های محلی نامزد که در عملیات بسط در راستای افزایش یا کاهش مقدار مولد ایجاد می‌شوند. به عبارت دیگر، این پارامتر تعداد عمل‌های نامزد پیشنهاد شده حول مقدار مولد را مشخص می‌نماید. ۲) پارامتر طول بسط که معرف اختلاف میان مقادیر پیشنهادی برای دو عمل محلیِ نامزد متوالی است. در این روش با پیشرفت فرآیند یادگیری و بهبود عملکرد عامل در محیطُ عملیات کاوش را به صورت محلی‌تری ادامه می‌دهد. به عبارت دیگر، پارامتر تعداد یک‌سویه‌ی عمل‌های نامزد در جریان یادگیری روندی افزایش و پارامتر طول بسط سیری کاهشی دارد. این روش گسسته‌سازی در یک مساله‌ی ناوبری ربات با هدف تعقیب دیوار و اجتناب از موانع آزموده شده است.

۳.۴ توسعه‌ی روابط در محاسبه‌ی خروجی نهایی

ناکاشیما[۲۳] و همکارانش در سال ۲۰۰۳ نحوه‌ی محاسبه‌ی ارزش عمل و خروجی را در روش FQL تغییر دادند [۱۷] و با استفاده از ترکیب وزن‌دار مقادیر ارزش عمل در تولید خروجی، از آن در یک

[23] Nakashima

مساله "زدن گل به دروازه‌بان" استفاده کردند. در این تغییر، ابتدا بردار ارزش W برای همه m عملِ نامزد در هر قاعده به صورت رابطه‌ی (۱۳) محاسبه می‌شود:

$$W_j(s_t) = \sum_{i=1}^{R} \mu_i(s_t) w^{i,j} \quad j = 1, \ldots, m \tag{۱۳}$$

که در $w^{i,j}$ پارامتر وزن یا ارزش مربوط به jاُمین عمل نامزد تالی قاعده iاُم است. خروجی نهایی نیز به صورت رابطه‌ی ۱۴ به‌دست می‌آید:

$$a(s_t) = \frac{\sum_{j=1}^{m} a_j \times W_j(s_t)}{\sum_{j=1}^{m} W_j(s_t)} \tag{۱۴}$$

کیم و همکارانش در سال ۱۹۹۹ برای محاسبه‌ی مقدار ارزش عمل خروجی نهایی در ساختار FQL از روش میانیابی استفاده کردند [۱۶]. این پژوهش‌گران با توجه به اینکه عمل نهایی به صورت حاصل‌جمع وزن‌دار عمل‌های انتخاب شده در هر قاعده است و مقدار آن با هیچ کدام از مقادیر عمل‌های نامزد برابر نیست، مقدار ارزشِ متناسب با عمل نهایی در هر قاعده را با استفاده از تکنیک‌های میانیابی محاسبه کرده‌اند. سپس با استفاده از رابطه (۱۵) ارزشِ عمل نهایی را تقریب زده‌اند:

$$\widetilde{Q}_t(s_t, a_t) = \sum_{i=1}^{R} \mu_i(s_t) w_t^{ia_t} \tag{۱۵}$$

در این رابطه $w^{i,a}$ مقدار ارزش عمل نهایی a در قاعده iاُم است که با استفاده از میانیابی و با توجه به مقادیر ارزشِ عمل‌های نامزد محاسبه می‌شود. این راهکار در نمونه‌ای از یک مساله تعادل گاری و اهرم ارزیابی شده است.

۴.۴ بهبود FQL با بهره گیری از معیارهای خبرگی

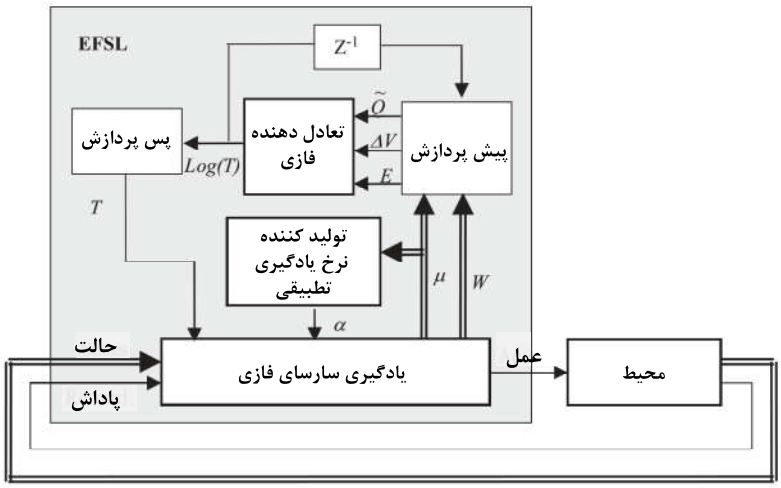
در روش FQL، $\Delta\widetilde{Q}$ در مقدار شدت آتش هر قاعده ضرب شده و نتیجه برای به‌روز رسانی مقدار ارزشِ عمل‌های نامزد آن قاعده استفاده می‌گردد. این نحوه تاثیرگذاری می‌تواند با اشتباه در انتخاب تعداد کمی از تالی‌های قواعد منجر به تولید عمل نادرست شوند و در نهایت جریمه‌ی عامل شوند. انتشار سیگنال تقویتی منفی (جریمه) بین همه قواعد باعث سردرگمی قواعدی می‌شوند که تالی مناسبی را انتخاب نموده‌اند. برای حل این مساله، سیستم استنتاج فازی به عنوان یک سیستم چند عامله در نظر گرفته شد. به این ترتیب برای هر قاعده‌ی فازی یک عامل مشخص شد. وظیفه‌ی هر عامل انتخاب مقدار مناسب عمل از میان عمل‌های ممکن برای تالی آن قاعده است [۱۱]. خروجی عامل‌ها در این روش با یکدیگر ترکیب شده و خروجی نهایی را می‌سازند.

مساله مهم در این راهکار نحوه‌ی تقسیم $\Delta\widetilde{Q}$ بین عامل‌ها (قواعد) است. مقدار مثبت $\Delta\widetilde{Q}$ به منزله‌ی پاداش و مقدار منفی آن به معنای جریمه است. برای تقسیم سیگنال تقویتی بین قواعد، عاملی که در وضعیت فعلی دارای عملی با مقدار ارزش بالاتر است، به عنوان عامل خبره‌تر شناخته می‌شود. هرگاه $\Delta\widetilde{Q}$ مثبت باشد، مقدار آن بین تمام عامل‌هایی با شدت آتشی بزرگ‌تر از یک حد آستانه منتشر می‌گردد و اگر مقدار $\Delta\widetilde{Q}$ منفی باشد عامل‌هایی با خبرگی پایین‌تر جریمه می‌شوند. این روش در مساله‌ی "آوردن گاری به مرکز" مورد ارزیابی قرار گرفته است.

۵.۴ یادگیری سارسای فازی بهبود یافته

الگوریتم یادگیری سارسای فازی بهبود یافته (EFSL)[۲۴] بهبود دیگری است که در سال ۲۰۱۰ با ارائه دو ایده‌ی "نرخ یادگیری تطبیقی" و "تعادل‌دهنده‌ی فازی" در روش FSL ایجاد شده است [۹]. شکل ۴ نمودار بلوکی این الگوریتم را نشان می‌دهد. از آنجا که در پیاده‌سازی این الگوریتم بهبودیافته، تنها از بردار شدت آتش و ماتریس پارامترهای وزنِ w بلوک یادگیری سارسای فازی استفاده شده است می‌توان از ساختار آن برای تمام الگوریتم‌های یادگیری تقویتی فازی با معماری نقاد-تنها استفاده کرد.

[24] Enhanced Fuzzy Sarsa Learning



شکل ۴: *نمودار بلوکی یادگیری سارسای فازی بهبود یافته [۹].*

تولید نرخ یادگیری تطبیقی روشی است که برای رفع معضل بیش برازش ارائه شده است. این روش در یادگیری برخط امکان کاوش یکنواخت و مناسب از محیط را فراهم می‌کند و از انجام کاوش‌های پی‌درپی تنها در نواحی خاصی از محیط که منجر به جریمه می‌شوند اجتناب می‌نماید. رابطه‌ی پیشنهادی این پژوهشگران به صورت رابطه‌ی (۱۶) و متناسب با عکس مقدار ملاقات فازیِ حالت جاری $FV(s_{t+1})$ است:

$$\alpha_t = \min(\frac{\beta_t \times R}{FV(s_{t+1})}, \alpha_{tmax}) \qquad (۱۶)$$

در این رابطه، R تعداد قواعد، α_{tmax} کران بالای α و β_t یک نرخ یادگیری مرسوم غیر افزایشی است. مقدار $FV(s_{t+1})$ به صورت رابطه‌ی (۱۷) تعریف می‌شود:

$$FV(s_t) = \frac{\mu(s_t)^T \times N_{t-1}}{sum(N_{t-1})} \qquad (۱۷)$$

که در آن N_t یک بردار N عضوی است و هر عنصر آن مجموع شدت آتش‌های قاعده‌ی iاُم است

به صورت رابطه‌ی بازگشتی ۱۸ تعریف می‌شود:

$$N_t = N_{t-1} + \mu(s_t) \tag{۱۸}$$

در این رابطه $\mu(s_t) = [\mu_1(s_t), \dots \mu_R(s_t)]^T$ بردار شدت آتش نرمال‌شده‌ی قواعد است. هر چه عامل "ناحیه‌ی قاعده‌ی فازی" iاُم (نواحی فازی تعریف شده توسط مقدم قاعده) را بیشتر ملاقات کند، عنصر iاُم در بردار N بزرگ‌تر می‌شود.

یکی از ویژگی‌های مناسب رابطه‌ی بیشینه‌نَرم به‌کارگیری ضریب دما در انتخاب تالی قواعد برای کنترل تعادل میان کاوش و بهره‌گیری است. در این راستا یک سیستم فازی سوگنوی مرتبه‌ی صفر برای تعیین مقدار ضریب دما پیشنهاد شده است [۹]. تفاضل بیشینه‌ی و کمینه‌ی ارزش عمل در حالت جاری، اختلاف ارزش حالت فعلی با ارزش حالت قبلی و درجه‌ی کاوش در حالت‌های قبلی، ورودی این سیستم فازی را تشکیل می‌دهند. مقدار ضریب دما، خروجی این سیستم تعادل‌دهنده‌ی فازی است.

۵ نتیجه‌گیری

در این مقاله روش‌های مختلف یادگیری تقویتی فازی با معماری نقاد تنها بررسی شد. دو روش پایه و معروف در این زمینه به نام‌های یادگیری سارسای فازی و یادگیری کیو فازی معرفی شدند. مهمترین تفاوت این دو روش در نحوه‌ی محاسبه‌ی مقدار تفاضل موقتی ارزش عمل است که در اولی مقدار ارزش عمل حالت بعدی ل و در دومی مقدار بیشینه‌ی ارزش‌های حالت بعدی لحاظ می‌شود. توضیح داده شد که این تفاوت باعث شده است که روش اول دارای تحلیل‌های ریاضی مثبتی در خصوص همگرایی باشد در حالی که برای روش دوم مثال‌های واگرایی وجود دارد.

پژوهش‌های دیگری که به بهبود این روش‌های پایه پرداخته‌اند بر نحوه‌ی تعیین مقدار اولیه‌ی ارزش عمل‌های تالی قواعد، تغییر مجموعه عمل نامزد در حین یادگیری، تعیین ضریب دما در فرمول انتخاب عمل، و فرمول جدید محاسبه‌ی خروجی نهایی سیستم فازی متمرکز بودند. با توجه به آنچه ارائه شد می‌توان به این نتیجه رسید که روش‌های ارائه شده برای تنظیم تالی قواعد یک سیستم سوگنو

مرتبه صفر مناسب هستند اما راهکاری برای تنظیم توابع عضویت ورودی و تولید قواعد به صورت خودکار ارائه نمی‌دهند.

مراجع

[۱] ف. آخوندی، ا. خانی، و. درهمی، به‌کارگیری آموزش تقویتی گسسته در فضای پیوسته با استفاده از ایده گسسته‌سازی تطبیقی، پانزدهمین کنفرانس ملی سالانه انجمن کامپیوتر ایران، تهران، (۱۳۸۸).

[۲] ف. اعلمی‌یان هرندی، و. درهمی، تنظیم پارامترهای مقدم و وزن قواعد فازی در یک طبقه‌بندی فازی، بیستمین کنفرانس ملی سالانه انجمن کامپیوتر ایران (CSICC 2015) – دانشگاه فردوسی مشهد، (۱۳۹۳).

[۳] ف. توکلی، و. درهمی، ع. کمالی‌نژاد، بکارگیری دومرحله‌ای یادگیری سارسای فازی در کنترل راه دوپا، چهارمین کنگره‌ی مشترک سیستم‌های فازی و هوشمند ایران، دانشگاه سیستان و بلوچستان، زاهدان، (۱۳۹۴).

[۴] و. درهمی، ف. اعلمی‌یان هرندی، م.ب. دولتشاهی، یادگیری تقویتی، انتشارات دانشگاه یزد، (۱۳۹۶).

[5] S. Abe and R. Thawonmas, A fuzzy classifier with ellipsoidal regions, *IEEE Transactions on Fuzzy Systems*, **5**(3)(1997), 358-368.

[6] F. Alamiyan Harandi and V. Derhami, A reinforcement learning algorithm for adjusting antecedent parameters and weights of fuzzy rules in a fuzzy classifier, *Journal of Intelligent and Fuzzy Systems*, **30**(4)(2016), 2339-2347.

[7] F. Alamiyan Harandi, V. Derhami and F. Jamshidi, A new framework for mobile robot trajectory tracking using depth data and learning algorithms, *Journal of Intelligent and Fuzzy Systems*, **34**(6)(2018), 3969-3982.

[8] V. Derhami, Similarity of learned helplessness in human beingandfuzzyreinforcementlearningalgorithms, *Journal of Intelligent and Fuzzy Systems*, **24**(2013), 347-354.

[9] V. Derhami, V. Johari Majd and M. Nili Ahmadabadi, Exploration and exploitation balance management in fuzzy reinforcement learning, *Fuzzy sets and systems*, **161**(4)(2010), 578-595.

[10] V. Derhami, V. Johari Majd and M. Nili Ahmadabadi, Fuzzy Sarsa learning and the proof of existence of its stationary points, *Asian Journal of Control*, **10**(5)(2008), 535-549.

[11] V. Derhami, V. Johari Majd and M. Nili Ahmadabadi, Improvement of fuzzy Q-learning using expertness criteria, *Proc. 10th annual Computer Society of Iran Computer Conference*, **1**(2005), 1002-1009.

[12] F. Fathinezhad, V. Derhami and M. Rezaeian, Supervised fuzzy reinforcement learning for robot navigation, *Applied Soft Computing*, **40**(2016), 33-41.

[13] H. Ishibuchi, K. Nozaki and H. Tanaka, Distributed representation of fuzzy rules and its application to pattern classification, *Fuzzy Sets and Systems*, **52**(1992), 21-32.

[14] J. Jang, C. Sun and E. Mizutani, *Neuro-Fuzzy and Soft Computing*. Prentice-Hall, upper Sanddle River, (1997).

[15] L. Jouffe, Fuzzy inference system learning by reinforcement methods, *IEEE Trans. Syst., Man, Cybern. C*, **28**(3)(1998), 338-355.

[16] M. S. Kim, G. G. Hong and J. J. Lee, Online fuzzy Q-learning with extended rule and interpolation technique, *Proc. IEEE Int. Conf. Intelligent Robots and Systems*, **2**(1999), 757-762.

[17] T. Nakashima, M. Udo and H. Ishibuchi, Implementation of fuzzy Q-learning for a soccer agent Proc. *IEEE Int. Conf. on Fuzzy systems*, **1**(2003), 533-536.

[18] K. Nozaki, H. Ishibuchi and H. Tanaka, Adaptive fuzzy rule-based classification systems, *IEEE Transactions on Fuzzy Systems*, **4**(3)(1996), 238-250.

[19] S. B. Roh, W. Pedrycz and T. C. Ahn, A design of granular fuzzy classifier, *Expert Systems with Applications*, **41**(16)(2014), 6786-6795.

[20] A. Sharifi, S.M. ALIYARI and M. Teshnehlab, Semi-polynomial Takagi-Sugeno-Kang Type Fuzzy System for System Identification and Pattern Classification, *Journal of Control*, **4**(3)(2010), 15-28.

[21] R. S. Sutton, Learning to predict by the methods of temporal differences, *Machine learning*, **3**(1)(1988), 9-44.

[22] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*, MIT Press Cambridge, (1998).

[23] L. X. Wang, *A course in fuzzy systems*, Prentice-Hall press, USA, (1999).

[24] C. Ye, N. H. C. Yung and D. Wang, A fuzzy controller with supervised learning assisted reinforcement learning algorithm for obstacle avoidance, *IEEE Transaction Systems, Man, Cybernetics*, **33**(1)(2003), 17-27.

[25] M. Zolghadri Jahromi and M. Taheri, A proposed method for learning rule weights in fuzzy rule-based classification systems, *Fuzzy Sets and Systems*, **159**(2008), 449-459.