
بسم الله الرحمن الرحیم

دانشکده مهندسی برق و کامپیوتر

گروه مهندسی کامپیوتر

پایان نامه

برای دریافت درجه کارشناسی ارشد

مهندسی کامپیوتر- هوش مصنوعی(رباتیک)

توسعه ایده‌های یادگیری تقویتی گسسته در یادگیری تقویتی پیوسته برای سیستم‌های چند عامله

استاد راهنما: دکتر ولی درهمی

استاد مشاور: دکتر علیمحمد لطیف

پژوهش و نگارش: بهناز حیدری دهکردی

پاییز ۱۳۹۲

این پایان‌نامه را تقدیم می‌کنم به:

همسر عزیزم که پشتوانه‌ی محکمی برای تمامی مراحل زندگی من است. همچنین به پدر مادر عزیزم که نگاه مهربانشان قوت قلب و امید زندگیست.

تشکر و قدردانی

اکنون که با لطف خدا و کمک‌های همیشگی او توانسته‌ام این پایان‌نامه را به انجام برسانم بر خود لازم می‌دانم از استاد بزرگوار جناب آقای دکتر ولی درهمی تشکر نمایم. ایشان همواره با نگاه و بررسی دقیق و هوشمندانه مرا یاری نمودند.

همچنین مراتب تشکر خود را از استاد مشاور خود آقای دکتر علیمحمد لطیف بیان کنم.

چکیده

در جهان پیچیده امروز برای انجام کارهای متفاوت گاهی توانایی یک فرد کافی نیست و مشارکت و همکاری افراد نیاز است. در دنیای کامپیوتر نیز سیستم‌های چند عامله متشکل از تعدادی عامل است که با یکدیگر در یک محیط در تعاملند. این سیستم‌ها ویژگی‌های خاصی دارند، از جلمه خودمختاری، عدم دسترسی به اطلاعات سراسری و به اشتراک گذاری دانش. در این سیستم‌ها تغییرات محیط وابسته به ترکیب عمل تولید شده از همه عامل‌ها می‌باشد. لذا، تولید عمل هر عامل نه تنها به وضعیت محیط بلکه به عملی که عامل‌های دیگر انتخاب می‌کنند وابسته است. بنابراین با توجه به پیچیدگی طراحی از روش-های یادگیری برای تنظیم پارامترهای انتخاب عمل عامل‌ها استفاده شده است. روش‌هایی که برای حل این گونه مسائل ارائه شده‌اند، اکثراً مبتنی بر اشتراک دانش عامل‌ها از طریق ایجاد توابع ارزش-عمل برای همه‌ی عمل‌های ممکن خود عامل و سایر عامل‌ها، در هر حالت است. با افزایش تعداد عامل‌ها ابعاد مسئله به صورت نمایی گسترش پیدا می‌کند. که باعث افزایش زمان یادگیری و افزایش حافظه مورد نیاز برای حل مسئله است. اکثر روش‌های ارائه شده با توجه به مطالعات انجام شده برای محیط‌ها با حالت و عمل گسسته تدوین شده‌اند؛ با توجه به اینکه مسائل دنیای واقعی مانند هدایت ربات‌ها ماهیت پیوسته دارند، نیاز به الگوریتم‌های پیوسته کارا داریم. در این پایان‌نامه دو ایده جدید برای حل مسئله همکاری در محیط‌های چند عامله با فضای حالت پیوسته ارائه شده است. مقادیر تابع ارزش حالت-عمل و ارزش حالت هر کدام تالی یک سیستم فازی سوگنو مرتبه صفر جداگانه هستند در مراحل آموزش مقادیر تالی قواعد تنظیم می‌شوند. تعداد ورودی‌های سیستم‌های فازی برابر با ابعاد فضای حالت است. ایده اصلی از روش یادگیری سارسا فازی ارائه شده است. نتایج تجربی بر روی مسئله قرار گرفتن متقارن عامل‌ها حول یک میز چندضلعی که نمونه ساده‌ای از فرم‌بندی است؛ حاکی از افزایش سرعت یادگیری و بهبود کارایی سیستم است.

کلید واژه: حالت پیوسته، سیستم چند عامله، سیستم فازی، فرم دهی، یادگیری تقویتی

فهرست مطالب

عنوان **صفحه**

فهرست علایم و نشانه‌ها

عنوان	علامت اختصاری
تابع ارزش عمل	Q
حالت	s
عمل	a
نرخ آموزش	α
پارامتر تخفیف	γ
تابع جایزه	r

فهرست جدول‌ها

عنوان **صفحه**

فهرست شکل‌ها

عنوان **صفحه**

فصل ۱- مقدمه

۱-۱- پیشگفتار

طبق تعریف هوش مصنوعی عامل هوشمند[1] یک موجودیت است که با توجه به دانشی که از محیط دارد، توسط حسگرها ادراکاتی را از محیط دریافت و عملی را در جهت بیشینه کردن معیار بازده‌اش انجام می‌دهد. لیکن حل بسیاری از مسائل پیچیده توسط یک عامل در زمان محدود امکانپذیر نیست. به عنوان مثال حمل یک میز بزرگ حداقل نیاز به دو عامل دارد. بدین منظور سیستم‌های چندعامله معرفی شده‌اند. این سیستم‌های چندعامله[2](MAS) محیط‌هایی هستند که در آن چند عامل حضور دارند که برای رسیدن به هدفی خاص در تعاملند. همکاری در سیستم‌های چندعامله عبارت است از انجام اعمالی به صورت جداگانه توسط عامل‌ها تا جایی که نتیجه نهایی سیستم را به یک هدف خاص برساند.

به صورت معمول سیستم‌های چند عامله ویژگی‌هایی دارد که همکاری بین عامل‌ها را پیچیده می‌سازد. اینکه عامل‌ها خودکارند و اطلاعات سراسری از محیط ندارند با این وجود باید به نوعی با یکدیگر به اشتراک دانش بپردازند.

در این پایان‌نامه به بررسی سیستم‌های چندعامله کاملاً تعاونی[3] می‌پردازیم. در این نوع سیستم‌ها هدف آموزش بیشینه کردن جایزه مشترک می‌باشد. با توجه به نوع روش همکاری عامل‌ها سیستم‌های کاملاً تعاونی به سه دسته تقسیم می‌شوند: روشهای مستقل از همکاری[4] که فقط در محیط‌های قطعی جوابگو هستند، روشهای همکاری مستقیم[5] که انتخاب‌های عمل تصادفی به نوعی حاصل همکاری یا مذاکره است، روشهای همکاری غیر مستقیم[6] که انتخاب عمل از بین مجموعه عمل‌ها که انتظار می‌رود منتج به مقادیر بهتری شوند، می‌باشند.

در بخش بعدی ابتدا به تشریح سیستم‌های چند عامله می‌پردازیم سپس روش‌های بیان یادگیری تقویتی چند عامله را بررسی می‌کنیم.

۱-۲- سیستم‌های چند عامله

سیستم‌های چند عامله در مبحث هوش مصنوعی توزیع شده قرار می‌گیرند. از دیدگاه هوش مصنوعی توزیع شده، یک سیستم چندعامله اجتماعی از عامل‌های مستقل برای حل مسئله است که در آن، هر عامل دارای یکسری خصوصیات خاص است. یک سیستم چندعامله، دربرگیرنده

[1] Intelligent Agents
[2] Multi agent system
[3] Fully cooperative
[4] Coordination-free methods
[5] Direct coordination methods
[6] Indirect coordination methods

جامعه‌ای از عامل‌های هوشمند و خود مختار[1] است که در یک محیط در کنار یکدیگر در حال کارند و سعی در انجام وظیفه‌ای خاص و رسیدن به هدفی مشخص دارند.
در حقیقیت در بسیاری از مسائل دنیای واقعی مانند: طراحی مهندسی، جست و جوی هوشمند، رباتیک سیستم‌های چند عامله مورد نیاز هستند[۱]. سیستم‌های چند عامله با سیستم‌های تک عامله متفاوتند. سیستم‌های چندعامله بهتر است یک سری ویژگی‌ها را داشته باشند از جمله: توسعه پذیری، تحمل خطا، همزمانی[2][۲].

۱-۲-۱- مشخصات سیستم‌های چند عامله

سیستم‌های چند عامله یک سری ویژگی دارند، آن‌ها را می‌توان به صورت زیر بیان کرد[۳]:

- دانش کافی و لازم برای حل مسئله در یک عامل وجود ندارد.
- کنترل سیستم توزیع شده است (یعنی یک سیستم کنترل مرکزی وجود ندارد).
- داده‌ها توزیع شده و غیرمتمرکز می‌باشند.
- یک سیستم چندعامله شامل تعدادی عامل است که:
  - از طریق برقراری ارتباط با یکدیگر تعامل می‌کنند.
  - قادر به عمل در یک محیط هستند.
  - حوزه‌های تأثیر متفاوتی دارند که می‌توانند با هم همپوشانی داشته باشند.

۱-۳- دلایل استفاده از سیستم‌های چند عامله

در اینجا دلایل استفاده از سیستم‌های چندعامله را بیان می‌کنیم.

- عامل به تنهایی نمی‌تواند همه کار انجام دهد. تقسیم کار بین عامل‌های مختلف مزایایی چون قابلیت انعطاف و توسعه پذیری را فراهم می‌کند.
- گاهی دانش یک عامل برای حل مسئله کافی نیست. دانشی که بین عامل‌ها پخش شده است یک دیدگاه کامل‌تر برای حل مسئله به عامل‌ها می‌دهد.
- استفاده از سیستم‌های چندعامله در پردازش توزیع باعث تسریع حل مسائل می‌شود.

۱-۴- تعاریف همکاری

تعاریف مختلفی برای همکاری عامل‌ها وجود دارد، از جمله همکاری چند عامل که برای انجام یک عمل توأم در یک محیط مشخص با یکدیگر به توافق می‌رسند[۴]و یا اینکه همکاری عبارت است

[1] Autonomous
[2] Parrallelism

از توانایی برای یافتن عمل بهینه. در [۵]تعریف ترکیبی ارائه شده عبارت است از: توانایی که عامل بتواند شرایط بهینه را برای رسیدن به هدف مشترک به وسیله پیدا کردن عمل توأم با سایر عامل-ها در محیط پویا، بسازد.

۱-۵- شیوه‌ی بیان مسائل چند عامله و ترکیب آن‌ها با یادگیری تقویتی

۱-۵-۱- بازی‌های اتفاقی[1]

یک بازی اتفاقی یک مجموعه به شکل $<n, S, A_{1...n}, T, R_{1...n}>$ است. که n تعداد عامل‌هاست. S مجموعه حالات است و A_i ها مجموعه عمل‌های عامل iام می‌باشند. T تابع احتمال گذر است و R_i هم تابع جایزه برای عامل iام است. در بازی‌های اتفاقی هر یک از عامل‌ها عمل خود را به طور همزمان انتخاب می‌کنند و جایزه را به طور شخصی بر اساس عمل کلی به دست می‌آورند. عمل توأم حالت بعدی و جایزه‌ها را تعیین می‌کند.

ما علاقه‌مند به بررسی فعالیت‌های تعاونی هستیم، بنابراین روی بازی‌های ماتریس تعاونی[2] متمرکز می‌شویم، که بازی‌های هماهنگی نام دارند. بازی‌های هماهنگی یک روش هستند که در انتخاب عمل توأم در CMRL[3] استفاده می‌شوند. گراف همکاری یک روش دیگر حل مسئله است. با استفاده از نظریه گراف‌ها عامل‌ها می‌توانند به وسیله گره‌ها بیان شوند، و ارتباط بین عامل‌ها با یال بیان شود[۶].

۱-۵-۲- بازی‌های هماهنگی[4]

از بازی همکارانه تک حالته برای مطالعه مسائل هماهنگی در CMRL استفاده کرده‌اند. در بازی‌های هماهنگی تک حالته همه‌ی عامل‌ها عمل‌ها را همزمان انتخاب می‌کنند و عامل‌ها پاداش یکسان بر اساس عمل کلی می‌گیرند.

تعادل نش[5] در[۷]یک ایده در حل بازی‌های اتفاقی است.

$$\begin{aligned} &(P_*^1, P_*^2) \\ &V^1(S, P_*^1, P_*^2) \geq V^1(S, P^1, P_*^2) \\ &V^2(S, P_*^1, P_*^2) \geq V^2(S, P_*^1, P^2) \end{aligned} \qquad (۱-۱)$$

[1] Stochastic Games
[2] Cooperative
[3] Cooperative Multi Agent Reinforcement Learning
[4] Coordination game
[5] Nash Equilibrium

در فرمول(۱-۱)، P نشان دهنده سیاست هر عامل است و V ارزش عملکرد را بیان می‌کند. P_*سیاست در حالت تعادل نش را بیان می‌کند. سیاست تعادل نش یک حالت پایدار است و این یعنی هیچ عاملی نمی‌تواند با تغییر سیاست خود ، تا زمانی که سایر عامل‌ها سیاست خود را تغییر دهند، سود ببرد.

بازی‌های هماهنگی یک روش بسیار ساده دارند، که برای هماهنگی دو عامله مورد استفاده قرار می‌گیرند، ولی مشکلاتی نیز در این روش وجود دارد. اول اینکه حالت‌ها و عمل‌ها به صورت محدود تعریف شده‌اند. هر چند با افزایش تعداد حالت‌ها و عمل‌ها محاسبات به طور نمایی افزایش پیدا می‌کنند و بنابراین پیدا کردن عمل توأم بهینه ساده نیست. لازم است که هر عامل عمل سایر عامل‌ها را بداند، تا بتواند عمل توأم بهینه را پیدا کند. اگر عامل‌ها همه‌ی اطلاعات سیستم را بدانند، به این معنی است که عامل اطلاعات کلی را در همه زمان‌ها می‌داند. و سیستم یک سیستم چند عامله نیست. سوم اینکه هیچ اثباتی برای همگرایی تعادل نش وجود ندارد.

۱-۵-۳- گراف‌های هماهنگی

یکی از مسائلی که در بیشتر الگوریتم‌های یادگیری تقویتی چندعامله دیده می‌شود، این است که عامل‌ها نیازمند آن هستند که عمل و تابع جایزه سایر عامل‌ها را بدانند تا بتوانند عمل توأم را پیدا کنند. بنابراین با افزایش تعداد عامل‌ها و عمل‌ها مسئله همکاری مشکل‌تر می‌شود. گراف همکاری[1] یکی از روش‌های خوب برای حل این مسائل است. گراف همکاری (CG) یک سری نیازهای همکاری‌های پویا را بیان می‌کند.

شکل هر گره در گراف همکاری نماینده یک عامل است و یال‌ها وابستگی بین عامل‌ها را نشان می‌دهند. فقط گره‌های مرتبط در هر مرحله زمانی باید عمل‌های خود را هماهنگ کنند.

گراف همکاری می‌تواند MAS را خیلی مختصر بیان کند، و فضای حالت را کاهش دهد، هر چند که باز هم مشکلاتی وجود دارد. در اینجا فرض می‌شود که بین همسایه‌ها ارتباط وجود دارد، اما تأخیرهای ناشی از ارتباطات، که در اثر تبادل اطلاعات مابین عامل‌ها به وجود می‌آید؛ در نظر گرفته نشده است. مسئله دوم این است که آیا همه عمل‌ها همزمان یا پشت سر هم انجام می‌شوند. گراف همکاری یک گراف وزن‌دار جهت‌دار نیست، اما یک گراف جهت‌دار است. بنابراین عامل از همسایه‌هایش تأثیر یکسانی می‌گیرد[۶].

[1] Coordination graph

۱-۶- دلایل استفاده از یادگیری تقویتی

یادگیری تقویتی یکی از انواع یادگیری است که در آن نیازی نیست عامل در ابتدا اطلاعی از هدف مورد نظر داشته باشد. بنابراین بدون دادن دانش خاصی در ابتدا و به وسیله یک سیگنال تقویتی عامل به سوی هدف هدایت می‌شود، به این دلیل نیاز به محاسبات ریاضی پیچیده و دانش خاصی برای یادگیری نیست؛ این ویژگی‌ها یادگیری تقویتی را به یک روش محبوب بدل کرده است.

۱-۷- مشکلات اصلی یادگیری تقویتی در مسائل چند عامله

در این بخش معایب اصلی روش یادگیری تقویتی را در سیستم‌های چند عامله بیان می‌کنیم:

۱-۷-۱- تنگنای ابعاد

هنگامی که تعداد عامل‌ها در یک سیستم افزایش یابد اکثر روش‌های کلاسیک ارائه شده یادگیری تقویتی با مشکل تنگنای ابعاد[1] روبرو می‌شوند؛ به این معنا که حجم اطلاعاتی که برای رسیدن به هدف نیاز است ذخیره شود به طور نمایی افزایش پیدا می‌کند. این امر حتی گاهی رسیدن به هدف را غیرممکن می‌کند.

۱-۷-۲- تقسیم جوایز

در سیستم‌های چند عامله، عمل هر عامل بر نتیجه حاصل تأثیر دارد با این وجود در اکثر روش‌های ارائه شده جوایز به طور یکسان بین عامل‌ها تقسیم می‌شود، این امر نوعی بی عدالتی است زیرا وقتی نتیجه نامطلوب است لزوماً عملکرد همه‌ی عامل‌ها نامطلوب نبوده است. البته با استفاده از یک سری تعاریف، جدید مانند خبرگی این مشکل تا حدودی حل شده است[۸].

۱-۷-۳- گسسته بودن بیشتر روش‌ها

اکثر روش‌های یادگیری تقویتی که برای سیستم‌های چند عامله ارائه شده است برای محیط-هایی با حالت و عمل گسسته‌اند(در فصل بعد بیشتر در این باره توضیح می‌دهیم). این امر گاهی باعث تشدید مشکل تنگنای ابعاد شده و گاهی با ماهیت مسئله‌ها که پیوسته هستند، در تناقض است.

[1] Curse of dimensionality

۱-۸- اهداف و نتایج حاصل از این پایان‌نامه

در این پایان‌نامه به دنبال ارائه روشی برای یادگیری تقویتی چند عامله هستیم که هم پیوسته باشد و هم بر تنگنای ابعاد غلبه کند. دو روش یادگیری تقویتی پیوسته برای سیستم‌های چند عامله ارائه شده است. مهم‌ترین مزیت این روش‌ها بر روش‌های قبلی ارائه شده، حل مشکل تنگنای ابعاد است. علاوه بر غلبه بر این مشکل؛ در پیاده‌سازی تجربی نیز به نتایج مناسبی در اثبات کارایی روش‌ها رسیده‌ایم.

در روش اول، روش کلاسیک یادگیری تقویتی تک عامله را توسط یک تقریب زننده فازی برای سیستم‌های چند عامله پیوسته سازی نموده‌ایم و یک فرمول جدید را برای محاسبه ارزش حالت و ارزش حالت-عمل به دست آورده‌ایم.

در روش دوم با ایده گرفتن از روش یادگیری سارسا فازی که یک روش تک عامله پیوسته است آن را برای سیستم‌های چند عامله گسترش داده‌ایم.

نتایج استفاده از دو روش فوق را بر روی مسئله هل دادن میز چندضلعی افزایش سرعت یادگیری و بهبود کارایی در مرحله تست است.

۱-۹- ساختار پایان‌نامه

در فصل دوم مفاهیم اولیه که در این پایان‌نامه استفاده‌شده؛ از جمله یادگیری تقویتی و یادگیری تقویتی در سیستم‌های چند عامله را شرح می‌دهیم مزایا، معایب و مثال‌هایی برای آنها می‌آوریم. در فصل سوم ایده‌های ارائه شده که روش‌هایی برای یادگیری تقویتی چند عامله پیوسته هستند را بیان می‌کنیم. در فصل چهارم نتایج شبیه‌سازی بر روی مسئله هل دادن میز چند ضلعی، را ذکر می‌کنیم و با یک روش گسسته مقایسه می‌کنیم. در فصل آخر به جمع بندی نهایی، نتیجه گیری و کارهای آتی می‌پردازیم.

فصل ۲- مفاهیم اولیه

۲-۱- مقدمه

سیستم‌های چند عامله در هوش مصنوعی جایگاه مهمی دارند. مقوله ارتباط عامل‌ها در این نوع سیستم‌ها مقوله‌ی پیچیده و با اهمیتی است. برای ایجاد ارتباط کارا بین عامل‌ها روش‌های متفاوتی وجود دارد که یکی از آن‌ها استفاده از روش‌های یادگیری است.

روش‌های متفاوتی برای یادگیری وجود دارد که با توجه به ماهیت سیستم‌های چند عامله یکی از روش‌های مناسب برای یادگیری در این سیستم‌ها یادگیری تقویتی است. یادگیری تقویتی در سیستم‌های چند عامله به صورت تکی و یا با ترکیب با سایر روش‌ها از جمله روش‌های ریاضیاتی، استفاده می‌شود.

در این بخش به بررسی مفاهیم اولیه استفاده‌شده در این پایان‌نامه را بیان می‌کنیم؛ همچنین روش‌های یادگیری تقویتی در سیستم‌های چند عامله و تک عامله می‌پردازیم و محاسن و معایب هر کدام از روش‌ها را بیان می‌کنیم.

۲-۲- یادگیری تقویتی

یادگیری تقویتی[۹] روشی برای حل مسائل با استفاده از یادگیری از سعی و خطا در یک محیط[1] پویا است. عامل یادگیری تقویتی، یاد می‌گیرد که در هر موقعیت چه کاری انجام دهد یعنی یک نگاشت از موقعیت‌ها به اعمال به دست آورد. در یادگیری تقویتی، به عامل یادگیر گفته نمی‌شود که عمل درست در هر موقعیت کدام است (یادگیری از نوع با ناظر[2] نیست)؛ بلکه بعد از انجام یک عمل در یک موقعیت، عمل انجام شده ارزیابی شده و به عامل گفته می‌شود عمل انجام شده چقدر خوب و یا بد بوده است (پاداش و یا جریمه داده می‌شود) و عامل یاد می‌گیرد که در موقعیت‌های مشابه این کار را انجام دهد و یا از آن احتراز کند. در بسیاری از کاربردهای واقعی، ممکن است عمل انتخاب شده نه تنها بر پاداش دریافت شده آنی بلکه بر پاداش‌های آینده نیز تأثیر بگذارد. از طرف دیگر، ممکن است عملی دارای سود آنی نباشد و در قدم‌های زمانی بعدی به پاداش‌های قابل توجهی منجر شود. یکی از نقاط قوت یادگیری تقویتی این است که راه حلی برای اندازه‌گیری کارآیی اعمالی ارائه می‌دهد که دارای سود آنی نیستند؛ اما در آینده منجر به سود و منفعت می‌شوند. عامل یادگیری تقویتی این پاداش با تأخیر را با یادگیری یک نگاشت از هر عمل ممکن به یک مقدار اسکالر انجام می‌دهد. این مقدار اسکالر، مجموع پاداش‌های آینده هر عمل با یک "ضریب فراموشی" برای هر پاداش در طول زمان است. ضریب فراموشی باعث می‌شود تا پاداش‌های دریافت شده بعدی (در آینده دورتر) نسبت به پاداش‌های فعلی دارای ارزش کمتری

[1] Environment
[2] Supervised

باشند. عامل یادگیری تقویتی سعی می‌کند، بهینه عمل کند؛ یعنی طوری عمل کند که در طول زمان بیشترین مقدار پاداش را دریافت کند. دو خصوصیت مهم یادگیری تقویتی که آن را برای مسائل مختلف مناسب می سازد، یادگیری از محیط به روش سعی و خطا و توانایی در نظر گرفتن پاداش‌های تأخیری می‌باشند. مثلاً عاملی که در یک محیط هزارتو قرار دارد تا زمانی که در مسیر قرار دارد و به هدف نرسیده جایزه قابل توجهی دریافت نمی‌کند.

یکی از چالش‌هایی که در یادگیری تقویتی می‌باشد و دیگر انواع یادگیری نیست تعادل بین کاوش[1] و بهره برداری از تجربیات[2]. عامل یادگیری تقویتی برای به دست آوردن پاداش بیشتر گاهی باید اعمال جدید و تجربه نشده را نسبت به اعمالی که در گذشته تجربه کرده و پاداش مناسبی گرفته است ترجیح دهد.

در کنار عامل و محیط، یادگیری تقویتی چهار جزء اصلی دارد: سیاست، تابع جایزه، تابع ارزش و یک مدل برای محیط(اختیاری) [۱۰].

سیاست[3] رفتار عامل یادگیرنده در زمانی خاص را بیان می‌کند. در واقع سیاست یک نگاشت از محیط به اعمال است. سیاست ممکن است یک تابع ساده یا یک پروسه‌ی پیچیده باشد.

تابع جایزه[4] در واقع هدف را معین می‌کند و در هر حالت یک سیگنال اسکالر است که میزان خوب یا بد بودن عملکرد را بیان می‌کند. هدف عامل یادگیری تقویتی این است که سیگنال جایزه را در طول رویداد ماکزیمم کند.

تابع ارزش[5] در واقع پس از حل مسئله یادگیری تقویتی تعیین شده و ارزش حالت‌ها را بیان می‌کند. به عبارت دیگر ارزش هر حالت میزان مجموع جایزه‌ی تخفیف یافته‌ای است که با شروع از آن حالت تا رسیدن به هدف به دست می‌آوریم. به عبارت دیگر تابع ارزش امید ریاضی جوایز کاهش یافته ای است که تا کنون دریافت شده است(۲-۱).

$$E(\sum_{t=0}^{\infty}\gamma^t r_t) \qquad (۲-۱)$$

مدل محیط[6] نیز یک تخمین از محیط است که بیان می‌کند در یک حالت با انجام یک عمل به چه حالتی می‌رویم.

فرم دیداری یادگیری تقویتی را می‌توان در شکل۲-۱: یادگیری تقویتی[۹] مشاهده نمود.

---

[1] Exploration
[2] Exploitation
[3] Policy
[4] Reward function
[5] Value function
[6] Model of the environment.